South Afr J Anaesth Analg. 2022;28(5 Suppl 1):S80-84
https://doi.org/10.36303/SAJAA.2022.28.5.2916

South Afr J Anaesth Analg
ISSN 2220-1181 EISSN 2220-1173

FCA REFRESHER COURSE

Commonly used statistical tests and their application

MM Kebalepile, P Motshabi Chakane

Department of Anaesthesia, School of Clinical Medicine, Faculty of Health Sciences, Charlotte Maxeke Johannesburg Academic Hospital, University of the Witwatersrand, South Africa
Corresponding author, email: *moses.kebalepile@wits.ac.za*

The use of statistics to derive insights from the data has provided researchers with empirical tools to describe, diagnose, predict, and prescribe actions, interventions and generate knowledge from varied data sets. Depending on the types of data (structured, semi-structured, and unstructured), different statistical models and tests can be performed to generate the said insights and new knowledge. Correctly identifying the data types (whether the data is discrete or continuous) aids researchers in identifying the applicable statistical tests. The distribution of the data also has an implication for determining the relevant statistical test. When continuous data are said to be normally distributed, the types of tests that can be used to compare groups differ to when the data are found to be not normally distributed. The same is true for discrete data, in that there are certain assumptions that need to be satisfied to identify the correct statistical tests to use in comparing the data groups. In this paper, the discussion will be focused on nine commonly used statistical tests. These common tests include the t-test, paired t-test, analysis of variance (ANOVA) test, Wilcoxon rank-sum test (WRST), Wilcoxon signed-rank test (WSRT), nonparametric ANOVA, chi-square (X^2) test, Fisher's exact test, and McNemar's test. The conditions for the use of these tests will be described, and the applications thereof will be discussed. Therefore, the purpose of this paper is not to describe the mathematical theories behind different statistical tests but to introduce the tests and their applications.

Keywords: t-test, paired t-test, analysis of variance test, ANOVA, Wilcoxon rank-sum test, Wilcoxon signed-rank test, nonparametric ANOVA, chi-square test, Fisher's exact test, McNemar's test

Background

While research is described as a systematic inquiry into nature and society,[1] statistics is defined as a combination of multiple approaches and processes that enable researchers to describe, summarise, interpret and analyse research data.[2,3] These approaches and processes are all used to create knowledge and derive insights from the research data or data sets in general. Research has always been accepted to be either basic or applied. Knowledge generated from the former often has no obvious practical use, while the latter often is the research type of choice when societal questions require data-driven solutions.[1]

There are multiple paradigms (such as verification and falsification) associated with knowledge generation, but in the current paper, only the falsification schools of thought will be described. Falsification, generally associated with the philosopher Karl Popper, is premised on the idea that a hypothesis can be tested against available data, and from these tests, truths or knowledge can be created.[1] On the contrary, verificationists learn and derive new knowledge through observations of nature and societies. Only when repeated observations are verified does the knowledge get accepted as new knowledge. Generation of knowledge through statistically tested hypotheses and falsified evidence gives scientific research tools for theories that can be validated. Therefore, statistical methods provide a conceptual and computational framework for answering research questions that can lead to the generation of new knowledge.

The use of statistics in medical and clinical research has grown.[4,5] With this increase in the use of statistics in clinical research, the quality and correct use of statistical techniques and tests has been evaluated.[4] It has often been found that some researchers in the clinical setting have used statistical methods incorrectly.[4] The reason for the reported incorrect use of statistical methods and techniques might be related to the complexity of some statistical methods.[5] This finding, that researchers may sometimes use statistical methods incorrectly, suggests that there is a need to develop basic competence in commonly used statistical tests and methods. The correct use of statistical tests and methods will enhance the practice of evidence-based medicine.

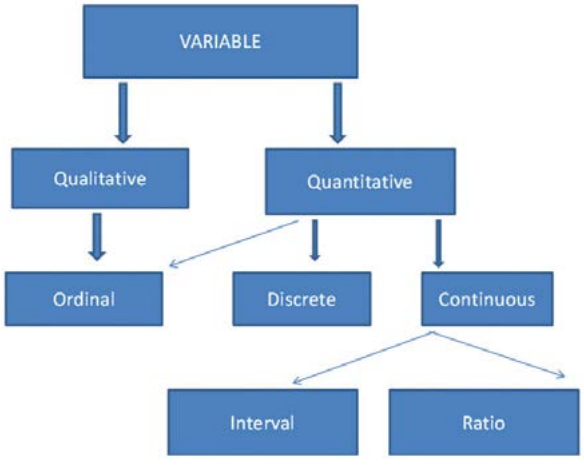

Figure 1: Classification of variables[3]

Correct use of statistical tests and methods to derive knowledge or make inferences depends on knowing the type of data that requires analysis. This starts with knowing the attributes and characteristics of the data that is to be analysed. These attributes can be called variables and they represent the studied subject's unique features, i.e. the age of patients in a sample that is being studied.[3] Ali and Bhaskar[3] described the variables as either quantitative or qualitative. Figure 1 by Ali and Bhaskar summarises the types of variables that are used in statistics.[3]

Knowing the type of data and its structure allows for correctly choosing the test to use for hypothesis testing and knowledge creation when using the falsification paradigm. Hypothesis testing is the statistical practice of checking for significant difference between two hypotheses (the null hypothesis and the alternative thereof).[6] The tested statistical difference is about a studied population parameter and its value estimated from a sample from that population. In principle, the null hypothesis, denoted H_0, is a postulation that suggests that there is no significant difference between population parameter and the value estimated from a sample.

Therefore the alternative hypothesis would premise that there is a statistically significant difference between a tested value of a population parameter and its estimated value.[6] There is always a risk of making an erroneous decision in deciding on the hypothesis to accept (the null or alternative hypothesis).[7] If a study rejects a null hypothesis that is true, that error is called a Type 1 error, and the risk associated with committing a Type I error is called alpha (α).

When the opposite error occurs, accepting the null hypothesis that should otherwise have been rejected, a Type II error would have occurred and the risk of a Type II error is denoted beta (β).[7,8] The reliability of a null hypothesis, or the strength of its evidence as a test, is tested against a probability statistic called the *p*-value.[8] The *p*-value is described as 'the probability of obtaining an effect equal to or more extreme than the one observed considering the null hypothesis is true.'[8] In other words, it is the probability that an observed value would have occurred by chance. Different research disciplines may use different probability thresholds. In medical research, this probability is set at a threshold of 5% ($p = 0.05$). Therefore, in the case of the *p*-value in a hypothesis test, being found to be less than 5%, the evidence is considered strong to support rejecting the null hypothesis. As in all probability, the *p*-value also ranges between 0 and 1. The closer the value is to zero, the stronger the evidence to reject a null hypothesis. Figure 2 is a demonstration of the probability density curve developed in R statistical programming language: version 4.2.1. and annotated in Microsoft paint.

Assumptions

Different statistical tests can be used for hypothesis testing. There are certain statistical assumptions that must be satisfied for the said tests to be used. Foremost in the assumptions required is the nature of the probability distribution of the data on which the hypothesis is assessed. Most tests require that data be normally distributed. Normally distributed variables in the data are symmetrical around the mean and show no kurtosis or skewness.[9] Figure 3 shows the different probability distributions, and depending on the data distributions, either metric or nonmetric data, parametric or nonparametric tests will be used. The critical assumption of normality may not be critically important if the sample is large. Figure 3 was developed in R statistical programming language: version 4.2.1.

Both descriptive statistics and inferential statistics require that the data distribution be known so that the correct test or measure (measures of dispersion and measures of central tendency) may be used. On describing normally distributed continuous data, the correct measure of central tendency would be the mean, but if the data follow a distribution that is not normal, the median might be a more correct measure. This alternative measure can be applicable when the data is positively skewed. Another assumption that may influence the type of tests used is the sampling. Randomisation in sampling would lead to a preference of some tests over others.[9] The assumption is that the data have been acquired/collected through a random sample. Another assumption that is important when using statistical tests is the presence of outliers in the data. Outliers in data sets can imply that the assumption of normality is not met. Homogeneity of variances is another assumption that is particularly important for multiple samples or groups where the samples are assumed to

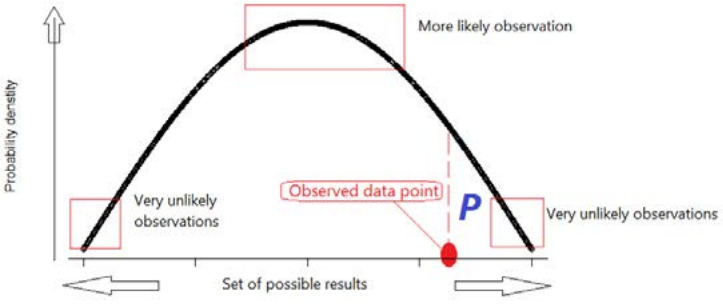

Figure 2: Probability density curve demonstrating the p-value

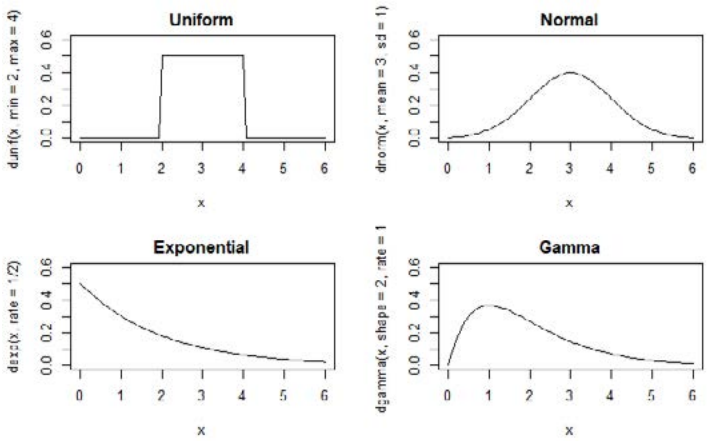

Figure 3: Different types of probability distributions

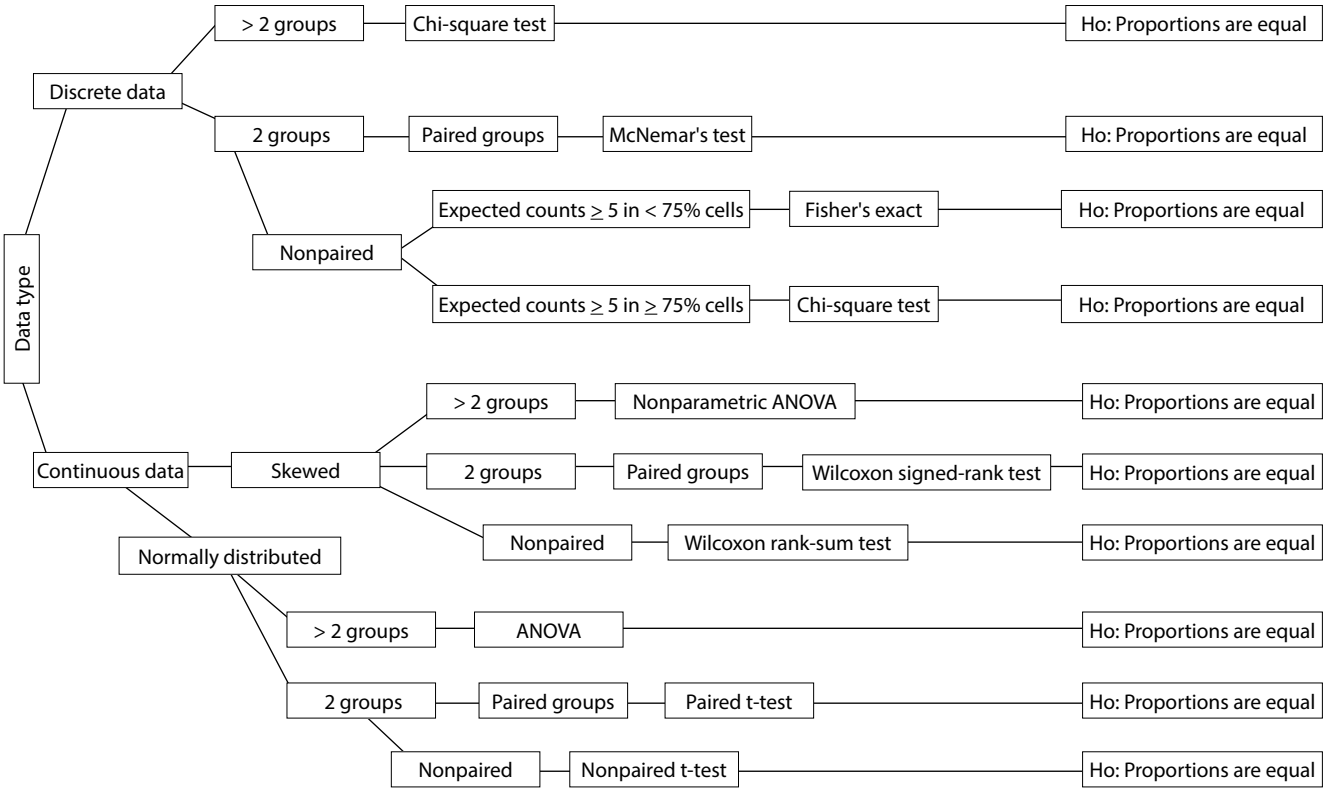

Figure 4: Common statistical tests

be derived from populations that have equal variance.[9] Although there are other assumptions that can be discussed,

Descriptive statistics

Descriptive statistics can be frequencies, graphs of proportions, and other measures describing the data set. The most used descriptive statistics are the measures of central tendency, measures of position, measures of frequency, and the measures of dispersion or variation.[10] Measures of central tendency include mean, median and mode. The aim and use of measures of central tendency such as the mean, is to describe the data by ways of the central position in the data.[10] The mean, also known as an average, can either be arithmetic, geometric or harmonic.

The median is the centre-most value when the data points in data sets are arranged in ascending order or ordered from smallest to largest value in the data set. The mode is the most frequently occurring value in the data set.[10] Although the measures of central tendency are adequate to describe the data set by way of the centre position, they are said to be limited in terms of describing the variation or variance in the data set. It is for this reason that the measures of dispersion, such as standard deviation and variance are used to describe variations in the data set. These measures only apply to ordinal, interval, and ratio data that can be ranked.[10] Descriptive statistics are generally the first step in analysing the data. They allow the research to readily explore the data and make informed decisions about the inferential tests and techniques that can be applicable for their type of data. It is on this background that this review will present some common statistical tests, and their applications. In this paper, general statistical assumptions were discussed, and under the discussion of common statistical tests, the test-specific assumptions are described.

Common statistical tests

Depending on the type of data being analysed, i.e. whether continuous or discrete, and the properties of that data, such as the distribution, the researcher can decide on the correct type of test to perform. Put differently, one always needs to establish if the data is metric or nonmetric. Metric data is scaled data often represented as interval or discrete data and continuous data.[11] Nonmetric data lacks a meter through which the distance between scales can be quantified. They include nominal, ordinal and binary data.[11]

If continuous data is normally distributed, and the intended comparative analysis is between two groups, such as testing for difference in the means of two groups, which are paired, then the correct test would be the paired t-test. If the two groups are not paired, then the test is a normal nonpaired t-test. If the comparison is between more than two groups, the differences in the means of the groups can be established using an analysis of variance (ANOVA) test. The null hypothesis in these tests would be that the means are equal (for t-test and ANOVA) or that the mean differences are equal (for the paired t-test).[12]

If the continuous data are skewed, and comparing paired two groups, the relevant test would be the Wilcoxon signed-rank test (WSRT). If the two groups are not paired, then the test recommended is the Wilcoxon rank-sum test (WRST). If the

skewed continuous data analysed is derived from more than two groups, then the correct test would be a nonparametric ANOVA.[12]

However, when analysing discrete data, and testing equality of proportions between two nonpaired groups, the first consideration to make would be to establish if the expected counts in the groups are at least greater than or equal to five in more than 75% of the cells. If this assumption is true, a chi-square (X^2) test would be adequate to test if the proportions are equal. However, if the expected counts are found to be greater than five in less than 75% of the cells, then Fisher's exact test would be more appropriate. On the other hand, if the two groups are paired, McNemar's test would be correct.[12]

When analysed proportions are of more than two groups in the discrete data, a chi-square test would also be applicable.[12]

T-test

There are three variations of the t-test, which include the one sample t-test, independent samples t-test, and paired samples t-test.[13] The first is used to test if the sample mean is statistically different from the population (population from which the sample was derived) mean. The unpaired, called independent t-test, evaluates the means difference between two unpaired samples or groups.[13] An example of the use of unpaired t-tests can be when the researcher wishes to establish if the mean blood pressure of males differed significantly from that of females after administering a certain pharmaceutical agent known to affect blood pressure. In paired t-test, as the name suggests, the means compared are of the same subject and recorded at different time points (t_0 and t_1).

Analysis of variance test

As with the t-test, there are variations (such as one-way ANOVA and one-way repeated measures ANOVA) of the ANOVA, and the ANOVA in general is useful when the means compared are of more than two groups. The test statistic and *p*-value then indicates the significant difference of at least one group to the rest. Therefore, that significant *p*-value, would be associated with a specific group indicating between group difference.[13]

Wilcoxon rank-sum test

The WRST is the applicable test when the assumption of normality has not been fulfilled and the t-test is not applicable for evaluating the equality of means for nonpaired two groups. Therefore, in the nonparametric test, the test is for equality of medians. Through a series of computational steps, the WRST uses a test statistic called W to falsify the null hypothesis that states that the proportions are equal, or the population is the same. If W is much smaller than the first median (μ_0), then the null hypothesis is not supported, and it can be concluded that the medians in this case are not equal and the populations are different.[14]

Wilcoxon signed-rank test

The WSRT is the alternative to the paired t-test, which is applicable when the distribution of the data is not normal. As an alternative, the WSRT is a test for location and seeks to evaluate that a distribution is symmetric about the hypothesised value.[15] However, the WSRT assumes that there is symmetry between the distribution of the differences between the two studied samples' parameter such as the median.[15] Through a series of steps, the WSRT converts the nonparametric problem to a parametric problem that could be computationally tested in a similar mechanical approach as with the t-test.[15]

Chi-square test

The X^2 is an association test and seeks to establish the relationship between two categorical variables. It can also be used as a goodness of fit test.[16] Using a published equation, the chi statistic is compared to a critical value, taking into consideration the degrees of freedom.[16] A *p*-value less than 0.05 or a set probability threshold, provides evidence to rejecting the null hypothesis and concludes that there is association between studied variables.[16]

Recommendations and conclusions

It is established in this paper that empirically using data and falsification approaches to derive insight can lead to knowledge creation. The knowledge created in this approach can be assessed independently and consequently validated. It is therefore crucial in research and knowledge creation to be conversant with the various statistical tests that can be used to test hypotheses and create new knowledge.

Insufficient knowledge of common statistical tests and their associated assumptions may lead to erroneous or incorrect use of the various tests. It is therefore recommended that a basic knowledge of data types, descriptive statistical tests, and common hypothesis testing tests be acquired for the effective use of research, particularly in clinical research.

In conclusion, hypothesis testing is a powerful statistical tool to explore data and create knowledge. Using correct statistical tests to test hypotheses provides evidence to use in support of tested theoretical positions and knowledge systems. This support for tested theories can provide useful knowledge for better clinical practices and approaches.

ORCID

MM Kebalepile https://orcid.org/0000-0002-5346-5798

P Motshabi Chakane https://orcid.org/0000-0001-9990-6336

References

1. Katzenellenbogen JM, Joubert G, Abdool Karim SS. Epidemiology: A research manual for South Africa. 2nd ed. Cape Town: Oxford University Press Southern Africa; 1997.
2. Heumann C, Schomaker M, Shalabh S. Introduction to statistics and data analysis: With exercises, solutions and applications in r. Springer International Publishing; 2017. https://doi.org/10.1007/978-3-319-46162-5.
3. Ali Z, Bhaskar SB. Basic statistical tools in research and data analysis. Indian J Anaesth. 2016;60(9):662-9. https://doi.org/10.4103/0019-5049.190623.

4. Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. The use of statistics in medical research. Am Stat. 2007;61(1):47-55. https://doi.org/10.1198/000313007X170242.
5. Evans SR. Common statistical concerns in clinical trials. J Exp Stroke Transl Med. 2010;3(1):1-7. https://doi.org/10.6030/1939-067x-3.1.1.
6. Yarandi HN. Hypothesis testing. Clin Nurse Spec. 1996; 10(4):186-8. https://doi.org/10.1097/00002800-199607000-00009.
7. Riou B, Landais P. Principles of tests of hypotheses in statistics: Alpha, beta and p. Ann Fr Anesth Reanim. 1998;17(9):1168-80. https://doi.org/10.1016/S0750-7658(00)80015-5.
8. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: An explanation for new researchers. Clin Orthop Relat Res. 2010;468(3):885-92. https://doi.org/10.1007/s11999-009-1164-4.
9. Verma PJ, Abdel-Salam A-S. Testing statistical assumptions in research. Wiley; 2019. https://doi.org/10.1002/9781119528388.
10. Yellapu V. Descriptive statistics. Int J Acad Med. 2018;4(1):60-63. https://doi.org/10.4103/IJAM.IJAM_7_18.
11. Kent RA. Analysing quantitative data: Variable-based and case-based approaches to non-experimental datasets. Sage; 2015. https://doi.org/10.4135/9781473917941.
12. Du Prel JB, Röhrig B, Hommel G, Blettner M. Choosing statistical tests: Part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010;107(19):343-8. https://doi.org/10.3238/arztebl.2010.0343.
13. Mishra P, Singh U, Pandey CM, Mishra P, Pandey G. Application of student's t-test, analysis of variance, and covariance. Ann Card Anaesth. 2019;22(4):407-11. https://doi.org/10.4103/aca.ACA_94_19.
14. Perolat J, Couso I, Loquin K, Strauss O. Generalizing the Wilcoxon rank-sum test for interval data. Int J Approx Reason. 2015;56(Part A):108-21. https://doi.org/10.1016/j.ijar.2014.08.001.
15. Ramachandran KM, Tsokos CP. Chapter 12 - nonparametric statistics. In: Ramachandran KM, Tsokos CP, editors. Mathematical statistics with applications in r. 3rd ed. Academic Press; 2021. p. 491-530. https://doi.org/10.1016/B978-0-12-817815-7.00012-9.
16. Franke TM, Ho T, Christie CA. The chi-square test: often used and more often misinterpreted. Am J Evaluation. 2011;33(3):448-58. https://doi.org/10.1177/1098214011426594.