ARTICLE IN PRESS

INDIAN JOURNAL OF RHEUMATOLOGY XXX (2014) 1–5

Available online at www.indianjrheumatol.com and www.sciencedirect.com

Nuts and Bolts in Clinical Research

Demystifying statistics: How to choose a statistical test?

Subramanian Shankar*, Ramakant Singh

[a]Department of Internal Medicine, Armed Forces Medical College, Pune 411040, India

ARTICLE INFO

Article history:
Received 2 April 2014
Accepted 3 April 2014
Available online xxx

Keywords:
Chi square
t test
Regression analysis

ABSTRACT

The young researcher today is confronted with a choice of hundreds of statistical tests, both while reading research works or while planning own research. The principles guiding the choice of statistical tests are simple and this article, aimed at the young researcher, aims to demystify the same.

1. Introduction

With the advent of "evidence-based medicine", it became important for medical professionals to be well versed with the statistical methods not only to understand and critically analyse various research works but also to plan their own studies.[1] However, statistics is often regarded as a weak spot for many medical researchers, possibly because of the way it is taught in most medical schools with a lot of emphasis on technical details. Use of wrong or inappropriate statistical test is thus very commonly observed in various research works published in biomedical journals.[2,3]

Unlike in the past, the researchers of today neither have to perform statistical calculations manually nor do they have to be familiar with the mathematical principles of a statistical test. There are excellent softwares which can perform any statistical test in a very short time. The softwares however cannot select an appropriate statistical test to be used in a given situation. Therefore, it all boils down to understanding the basis of selecting a statistical test for analysis of a given data set, which remains a pure human effort. Although, there are well over 100 statistical tests in use, the great majority of research questions can be tackled by using only a handful of them. The test to be used is decided on the basis of the type of the research question, the type of data being analysed and the number of groups or data sets involved in the study.[4]

Choosing an appropriate test can be best compared to the process of learning how to drive a car. One can become an excellent driver without understanding much about the chassis, engine or fuel injection system of the car. In a similar way, the choice of statistical tests is independent of most of the theory governing various statistical tests.

This article is aimed at the young researcher and will focus on just one question; namely, which statistical test is to be used in which situation. It expects them to have no background understanding of statistical theory and practice. It is assumed

* Corresponding author.
E-mail address: shankarsid@gmail.com (S. Shankar).

http://dx.doi.org/10.1016/j.injr.2014.04.002

Please cite this article in press as: Shankar S, Singh R, Demystifying statistics: How to choose a statistical test?, Indian Journal of Rheumatology (2014), http://dx.doi.org/10.1016/j.injr.2014.04.002

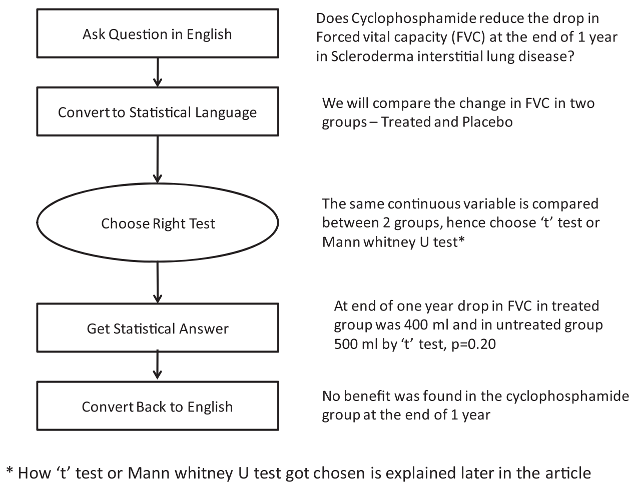

Fig. 1 – Five steps to choosing the correct statistical test with an example.

that the reader will eventually expand his/her knowledge over time about the various tests discussed in this article.

2. A step-by-step approach

To start with any research work, a research question is to be formulated in a language the researcher understands, namely, in plain English. Once the question is formulated, it has to be translated into statistical language which helps us in choosing the right sets of statistical tests. The answer that is obtained after applying the appropriate test is converted back to the original language to provide the final result (Fig 1).

The secret to choosing the correct statistical test therefore lies in articulating it correctly in plain English. It is often this first step that is not given adequate importance and which results in wrong choice of tests. The choice can best be visualized as a bilingual translation between English and statistical language.

This brings us to the next question. What is the statistical language? Statistics understands just one language: the language of "Variables".

3. What is a variable?

A variable is any feature or measurement in a given study that is to be analysed. Anything and everything we measure is a variable. For e.g. height, weight, smoking status, alcohol status, whether a person had a stroke or not; everything that can be measured is a variable. In all studies, we are usually doing one of the three things.

a. Assessing the characteristic of a variable
b. Comparing and contrasting 2 variables
c. Searching for associations between multiple variables

Variables are classified in 2 different ways. Based on their inherent nature/characteristic they are classified as Nominal (NV), Ordinal (OV) and Continuous (CV).[5] Based on the relationship they have with other variables they are classified as Dependant or Independent variable.

a. ***Dependant or response variable*** (DV): It is the variable of primary interest to the researcher. Each study has at least 1 dependent variable.
b. ***Independent or explanatory variable*** (IV): These are the variables that can influence/explain some fraction of DV. There can be many IV affecting a single DV.

Let us take a simple example. In a study that looks at multiple factors like age, smoking, diabetes, hypertension and hypercholesterolaemia affecting development of myocardial infarction, identify the DV and IV? Please pause and identify the DV and IV before going ahead.

The DV in this case is Myocardial infarction and the IVs are the other 5 factors. It is also apparent from the above example that both the DV and IV can be nominal, ordinal or continuous. After deciding the type of a variable, we will have to determine the type of analysis that can be done on a given set of observations (variables). Broadly, the analysis can be univariable, bivariable or multivariable. In some cases it gets influenced by time factor (survival analysis) and in others we modify it to compare with other tests (e.g., screening tests) (Fig 2).

The choice of tests also gets influenced by how the data is distributed in the population. For a normally distributed data,

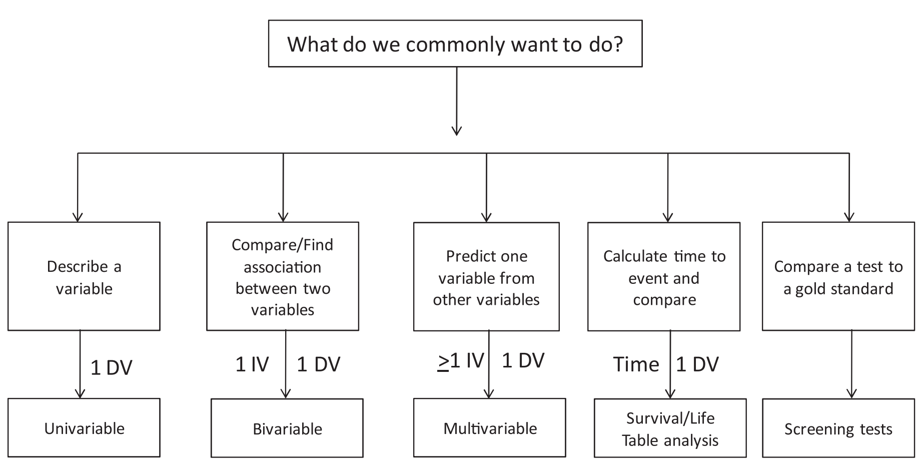

Fig. 2 – Various types of analysis. DV: dependent variable; IV: Independent variable.

Please cite this article in press as: Shankar S, Singh R, Demystifying statistics: How to choose a statistical test?, Indian Journal of Rheumatology (2014), http://dx.doi.org/10.1016/j.injr.2014.04.002

Data distribution curves

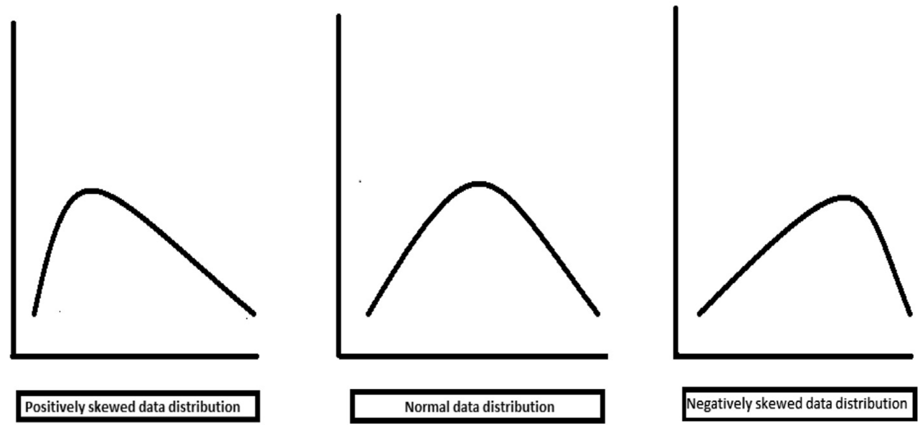

Fig. 3 – Types of distribution of data.

once chooses a parametric test while for a skewed data/small sample sizes one chooses a non-parametric test (Fig 3).

4. Statistical tests (parametric and non-parametric)

a. ***Parametric statistical test*** is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn. These tests assume the underlying source population(s) to be normally distributed. Examples of such tests are mean, paired t test, ANOVA etc.
b. ***Non-parametric tests*** are the ones that makes no such assumptions. In this strict sense, "non-parametric" is essentially a null category, since virtually all statistical tests assume one thing or another about the properties of the source population(s). Examples of non-parametric tests include the various forms of chi square tests, Fisher Exact Probability test, Mann–Whitney test, Wilcoxon Signed-Rank test, Kruskal–Wallis test and Friedman test.

That's all we need to know to choose the correct test. The rest of the article will take the reader through a brief description of statistical tests and various examples to clarify this concept.

4.1. *Univariate analysis*

Most studies will devote the first few paragraphs in the results section to describing the baseline characteristics of the patients. Here we are not making any comparisons and try to describe the characteristics using central tendency and dispersion. We only have to decide whether the given variable is Nominal (NV), Ordinal (OV) or Continuous (CV). In case, the variable is a CV, we further have to decide whether the given variables follow a normal distribution or not. There are various methods to check for normal distribution, e.g., plotting histograms or using one of the many "normality tests". If the CV is normally distributed, we use parametric tests else non-parametric tests are used to describe them. If you are not sure of whether the given variables are normally distributed or not, it is always better to use non-parametric tests.[6] For OV and NV only non-parametric tests are used.

Let us take a few examples to understand it better (Table 1).

Table 1 – Examples of univariable analysis.

S no	Clinical scenario	Type of variable	Distribution	Paired series	Test to choose
1	You recorded the weight of 300 children in a school. How can we present this data?	Continuous	Normal	No	Mean ± SD
2	Karan Johar has made 20 films. Each of his films are rated from * to *****. 16 were rated more than *** while 4 of them rated between * and **	Ordinal	Skewed	No	Median ± IQR
3	There are 30 Govt. Doctors in a room (pay range between Rs. 50,000 to Rs. 150,000 per month). Mr Mukesh Ambani walks in and joins them (monthly pay Rs. 100 crores). What is the average pay of the 31 people in room? (Can we use mean?)	Continuous	Skewed (grossly skewed)	No	Median ± IQR
4	You recorded the weight of 300 children in a school before and after drinking horlicks for 3 months. How do we find that the increase in weight (if at all) was significant?	Continuous	Normal	Yes	Paired t test
5	10 athletes from "Elite academy" ranked 15, 22, 34, 54 and 76 amongst 100 athletes in a decathlon event. They underwent a special training session at USA for 2 months and in the next event amongst the same set of 100 athletes held in September, they ranked 10, 18, 22, 31 and 48. Did the special training help them?	Ordinal	Skewed	Yes	Wilcoxon rank sum test

Please cite this article in press as: Shankar S, Singh R, Demystifying statistics: How to choose a statistical test?, Indian Journal of Rheumatology (2014), http://dx.doi.org/10.1016/j.injr.2014.04.002

Table 2 – Matrix for choosing appropriate tests in bivariable analysis.

Independent variable	Dependant variable		
	Continuous variable	Ordinal variable	Nominal variable
Continuous variable	Pearson's correlation coefficient	None	Chi square test for trend or Point biserial corr. coeff.
Ordinal variable	None	Spearman correlation coefficient	Chi square test for trend or Point biserial corr. coeff.
Nominal variable	Student's "t" test	Mann–Whitney "U" test	Chi square test/ Fischer's exact test

4.2. *Bivariate analysis*

These are the commonest types of test that we encounter in scientific literature. The key to understanding this is the fact that both DV and IV can be nominal, ordinal or continuous. Thus there are a possible 9 types of test that can exist. No tests have yet been devised to describe the relationship between an ordinal and a continuous variable effectively. Thus we just have 7 types of tests that govern this sphere as shown in Table 2.

Let us understand these with an example each (Table 3).

4.3. *Statistical tests for multivariate analysis*

In many clinical situations, study groups would be having various factors called risk factors (IV) to a single event or outcome (DV). Multivariable analysis is a statistical tool for determining the unique contributions of each risk factor to the outcome. For example, we cannot effectively test whether smoking increases the likelihood of coronary heart disease by randomly assigning persons to groups who smoke and groups who do not smoke because smokers may have other independent risk factor for CAD like preexisting diabetes and dyslipidemia, sedentary lifestyle and male sex. With multivariable analysis, we can demonstrate that even after adjusting for male sex, sedentary lifestyle, preexisting diabetes and dyslipidemia, smoking has an independent relationship with coronary artery disease. This is done mathematically using statistical softwares which controls other factors while isolating the effect of a single factor of interest.

The types of multivariable analysis are commonly used in clinical research are easily understood by examples given in Table 4. The proportional hazards test is part of a special subgroup of analysis called survival analysis whose primary characteristic is the fact that it is time dependent. Detailed discussion on regression analysis and survival analysis will be outside the scope of this article.

5. Conclusion

The choice of choosing a statistical test is primarily determined by defining the study question in plain language. Thereafter we translate the same in the language of variables. One then identifies the dependent and independent variable and also looks at normality of data. These few steps seen in conjunction with the tables above should be sufficient for the beginner to choose an appropriate statistical test.

Table 3 – Examples for choosing appropriate tests in bivariable analysis.

S No	Situation	DV	IV	Test to be used
Compare two variables (DV and IV) in one group				
1	Height of 100 children aged 5–10 years was measured and so was their weight. The relationship of height and weight was to be examined.	**Height** (CV)	**Weight** (CV)	Pearson's correlation coefficient
2	50 patients of RA were asked to mark their overall well being in a scale of 1–10 (patients general assessment or PGA). The Physicians then assessed them and made their own assessment from 1–10 (Physician global assessment or PhGA). How do we depict this relationship.	**PGA** (OV)	**PhGA** (OV)	Spearman correlation coefficient
Compare one variable in two or more groups				
3	In a medical college with 100 students (60 Boys, 40 Girls), 20 students scored distinction in Ophthalmology. These included 12 girls. Is there a significant difference in getting distinction between boys and girls?	**Distinction by boys** (NV) Can also be depicted as ratio (8/60)	**Distinction by girls** (NV) Can also be depicted as ratio (12/40)	Chi square test
4	A study reported that 3 sessions of cognitive behaviour therapy (CBT) might affect the pain scales in fibromyalgia. You measured the pain scales of 100 women, 50 of whom received CBT as per protocol. The results are given below. Is the result significant?	**Pain scale** (CV)	**CBT** yes/no (NV)	Mann–Whitney "U" test
5	A study reported that RA duration might be associated with osteoporosis (OP). You measured the presence of OP (defined as BMD < -2.5) of100 women and noted their disease duration. How would you depict the result?	**Osteoporosis** (NV)	**Duration of RA** (CV)	Chi square test for trend or Point biserial corr. coeff.
6	Is APGAR score related to cerebral palsy? You measured the APGAR score of 500 new borns and saw how many of them developed CP at the end of 2 year of follow-up. How would you depict the result?	**Cerebral Palsy** (NV)	**APGAR score** (CV)	Chi square test for trend or Point biserial corr. coeff.

DV: dependent variable, IV: independent variable, CV: continuous variable, OV: Ordinal variable, NV: nominal variable.

Please cite this article in press as: Shankar S, Singh R, Demystifying statistics: How to choose a statistical test?, Indian Journal of Rheumatology (2014), http://dx.doi.org/10.1016/j.injr.2014.04.002

Table 4 – Examples of multivariable and survival analysis.

Example	Dependant variable	Independent variable	Test used	Special feature
How do multiple factors affect bone mineral density (BMD) in RA?	BMD (CV)	Age (CV) Sex (NV) Duration (CV) Steroid (NV) Sharp score (CV)	Linear regression analysis	Variable coefficients have a linear relation with outcome
How do various factors affect development of myocardial infarction?	Myocardial infarction (NV)	Age (CV) Sex (NV) Diabetes (NV) Hypertension (NV) Smoking (NV) Smoking score (CV)[a]	Logistic regression analysis	Model constrains probability of outcome to 0 to 1
How does warfarin compare with dabigatran in preventing stroke when given for non-valvular atrial fibrillation?	Stroke (NV) (time dependant)	Warfarin Vs Dabigatran	Proportional hazards regression	Length of time to a discrete event Useful for longitudinal studies in which persons may be lost to follow-up

[a] Smoking can be quantified in any of the established ways and can be looked at as a continuous variable.
DV: dependent variable, IV: independent variable, CV: continuous variable, OV: Ordinal variable, NV: nominal variable.

Conflicts of interest

All authors have none to declare.

REFERENCES

1. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992;268:2420–2425.
2. Lang T. twenty statistical errors event you can find in biomedical research articles. *Croat Med J*. 2004;45:361–370.
3. Karan J, Goyal JP, Bhardwaj P, Yadav P. Statistical reporting in Indian pediatrics. *Indian Pediatr*. 2009;46:811–812.
4. Parikh MN, Hazra A, Mukherjee J, Gogtay N, eds. *Research Methodology Simplified: Every Clinician a Researcher*. New Delhi: Jaypee Brothers; 2010:121–128. Hypothesis testing and choice of statistical tests.
5. Dhir V, Arya V. Data types: the first step to planning your study. *Indian J Rheumatol*. 2014;9:27–31.
6. Nayak BK, Hazra A. How to choose the right statistical test? *Indian J Ophthalmol*. 2011, Mar-Apr;59:85–86.

Please cite this article in press as: Shankar S, Singh R, Demystifying statistics: How to choose a statistical test?, Indian Journal of Rheumatology (2014), http://dx.doi.org/10.1016/j.injr.2014.04.002