Choosing the correct statistical test in research

Nalika Gunawardena[1]

(Key words: Statistical test; research)

Gone are the days where researchers had to perform statistical calculations manually. Nowadays, many researchers have access to computer software which will perform whatever statistical testing the researcher needs to perform. However, statistical software does not have the ability make decisions on the statistical test to be used in a given situation. In other words, they do not have the capability to match the correct statistical test for the correct situation and data sets. If the researcher blindly orders the software to perform all possible statistical tests, the software will present him or her with a whole array of tests, a mix of relevant and irrelevant. Therefore, knowledge on choosing the correct test is essential for the researcher. The aim of this article is to present an easy and user-friendly method to choose the correct statistical test. This method is applicable to both descriptive and experimental study designs.

When should one choose the statistical test to be used? Is it at the stage of analysis?

Ideally presentation of results and statistical tests to be performed to achieve each of the specific objectives should be decided upon at the stage of planning of the research project.

The method that is presented in this article requires the researcher to respond to a checklist of four questions[1] and to follow a selected flow chart[2] until the person reaches the test that is relevant to the situation in which he or she needs to apply statistics.

Following is the checklist of four questions.

Q1. What scales of measurements have been used?

Q2. Which hypothesis has been tested?

Q3. If a hypothesis of difference has been tested, are the samples independent or dependent?

Q4. How many sets of measurements are involved?

[1]*Senior Lecturer in Community Medicine, Department of Community Medicine, Faculty of Medicine, University of Colombo*

For example, for weight observations or measurements in research, the researcher may assign the study units a number using relevant units such as kilogrammes and this number would reflect the actual magnitude of the weight of the study unit. When observing or inquiring into the ethnicity or socio-economic status the researcher may assign a code number to the individual. The basis on which the numbers are assigned to observations determines the scale of measurement being used. Traditional classification of scales of measurement describes three types namely, Nominal scale, Ordinal scale and Interval scale[1].

Nominal scale

If the researcher simply uses numbers to label categories which do not represent any order, then the scale of measurement used is the Nominal scale. In a descriptive study the researcher may measure sex of the study units, and assign numbers '1' for males and '2' for females. By doing this, the researcher is not indicating that females are 'more' or 'less' in relation to sex indicating that categories do not represent any order. In an experimental study, the researcher may use nominal scale to categorize study units in the experimental group and control group based on type of the side effect that they may experience following administering of the drug/placebo. Though the researcher may also assign consecutive numbers for different things like side effects, it does not indicate that one side effect is higher or lower than the other. Nominal scale is considered the lowest form of scale of measurement as it does not provide any information on the relationship between the categories.

Results of research data measured in nominal scale would be presented using frequency distributions.

Ordinal scale

If the researcher uses numbers to label categories in which an order can be identified, then the scale of measurement used is the Ordinal scale. In other words, here the relationship between categories in terms of 'greater than' or 'lesser than' status can be identified. In descriptive studies the researcher may categorize people based on their level of

satisfaction to a health service provided using categories of 'highly satisfied', 'somewhat satisfied' and 'dissatisfied'. This categorization also provides information on the relationship between them. In experimental studies we may use ordinal scale to categorize study units in both experimental and control groups, based on the type of response that they will experience following administering a drug using the categories such as 'very good response', 'good response' and 'poor response'. Again, this categorization provides information on the relationship between the categories. It indicates that 'very good response' would be or is better than 'good response' and 'good response' better than 'poor response'. However, it should be noted that ordinal scale does not numerically quantify how much greater 'very good response' is, when compared to a 'good response'.

Results of research data measured in ordinal scale would be presented using frequency distributions.

Interval scale

When the researcher assigns the numbers to observations, if the difference (interval) of two such numbers that have been assigned are meaningful (numbers are assigned to weight observations or measurements and the difference between two such weight observations or measurements will denote how much 'greater than' or 'lesser than' one measure is from the other), this scale of measurement is called interval scale. This happens, as in interval scale the numbers that are assigned represents the actual magnitude of it. This

happens as theoretically the distances between successive points in the interval scale are equal.

In descriptive studies we may measure heights of individuals and assign them the values in the interval scale using the unit centimetre (e.g. centimetre or cm). A person who is assigned the number 150 cm, we know is taller than the person assigned the number 140cm and also we know that the difference of height between these two persons is 10cm.

In experimental studies we may be able to use interval scale to measure the response to a drug among study units in both experimental and control groups, in terms of improvement of haemoglobin (Hb) level. We would then know that a person who has an improvement of 3g/dl has responded to the drug better than a person who had an improvement of 1g/dl and that the difference of improvement is 2g/dl. We also know that, the difference between two persons of 2g/dl improvement and 3g/dl improvement is equal to the difference between two persons of 4g/dl improvement and 5g/dl improvement.

Results of research data measured in interval scale would be presented using measures of central tendency (mean, mode, median) and dispersions (standard deviation, standard error).

At this stage, when we have determined the type of scale that has been used to measure the outcome that has to be statistically tested, we can make some decisions regarding the collective group of statistical tests that we may have to use[1,2] (Table 1).

Table 1: Groups of statistical tests to be used for data measured using different scales of measurement

Nominal	chi square or one of its variations	
Ordinal	ordinal tests or non-parametric tests (Mann-Whitney U test, Kruskal-Wallis H test, Wilcoxon T test, Spearmen r)	
Interval	Measure is normally distributed in the population	**YES**-parametric tests (t test, ANOVA F test, Pearson r)
		NO- consider as ordinal scale Ordinal tests or non-parametric tests (Mann-Whitney U test, Kruskal-Wallis H test, Wilcoxon T test, Spearmen r)

As shown above, when the scales of measurements are either nominal or ordinal, the groups of statistical tests to be used can be decided without answering any further questions. In case of interval scale, we need to answer a further question which inquires whether or not the measure is normally

distributed in the population. The normally distributed measures will conform to a normal curve, if we do this measurement on the whole population and plot a graph with values of the measurement presented in the X axis and frequency of occurrence presented in the Y axis. The normal

curve as Gaussian described it is actually a theoretical distribution[3]. The good news for the researchers is that many person-related measurements we perform come quite close to this ideal distribution[1]. As shown in Table 1, if search of the body of literature indicates that the measure is normally distributed in the population, the researcher should make the decision to use one of the parametric statistical tests to test for significance. If it is found that the measure that the researcher is dealing with is NOT normally distributed in the population, the researcher has to treat these measurements as measures done using ordinal scale.

Upon answering the first question in the checklist, the researcher is able to pick up the relevant flow chart to be used (Figures 1-3).
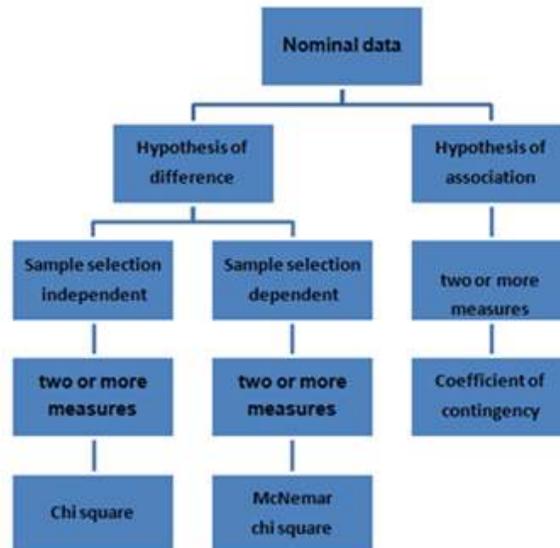

Figure 1: *Flow chart on statistical tests to be used for data measured in Nominal scale*

(Adapted from Hennekens CH, Buring JE, Mayrent SL. Epidemiology in medicine. 1st ed. USA: Lippincott Wiliams and Wilkins; 1987)
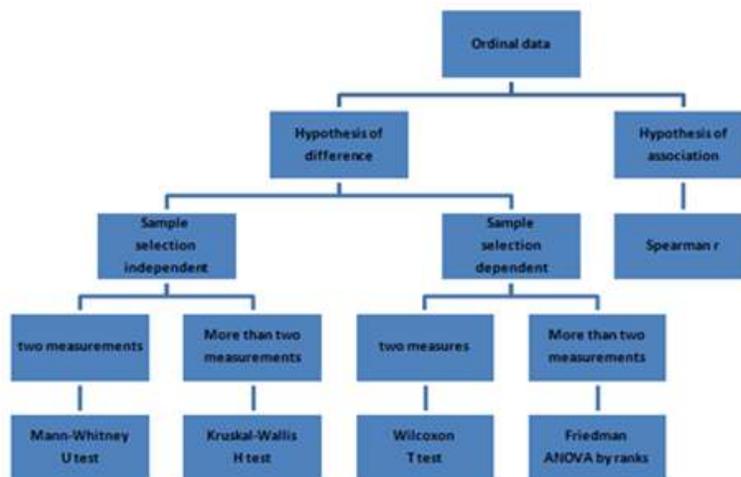

Figure 2: *Flow chart on statistical tests to be used for data measured in Ordinal scale*

(Adapted from Hennekens CH, Buring JE, Mayrent SL. Epidemiology in medicine. 1st ed. USA: Lippincott Wiliams and Wilkins; 1987)
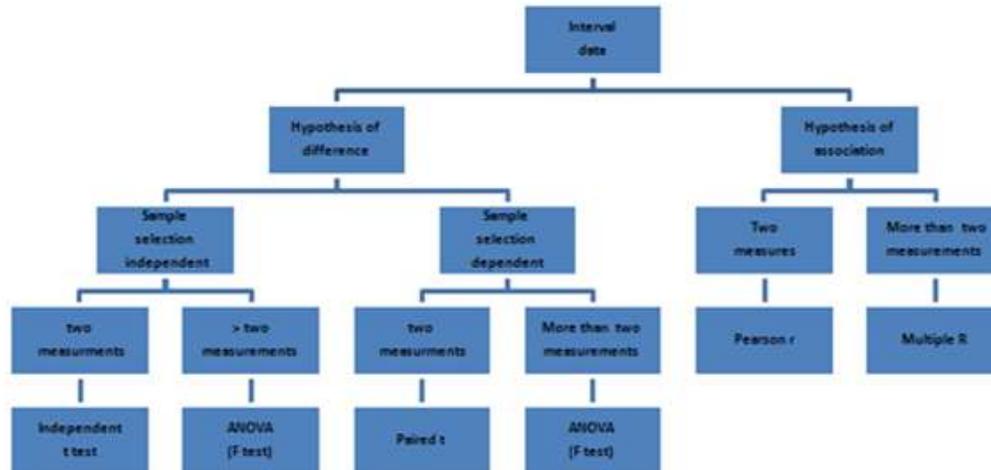
Figure 3: *Flow chart on statistical tests to be used for data measured in Interval scale*

(Adapted from Hennekens CH, Buring JE, Mayrent SL. Epidemiology in medicine. 1st ed. USA: Lippincott Wiliams and Wilkins; 1987)

Q2. Which hypothesis has been tested?

The second question to answer is about the hypotheses to be tested.

There are only two types of hypotheses that can be statistically tested in research[1]. They are either a hypothesis of difference or a hypothesis of association.

Hypothesis of difference states that the difference that is shown in the results obtained from the samples are also different in the larger populations from which the samples came.

In descriptive studies the researcher may test either the hypothesis of difference or the hypothesis of association. For example, if in a descriptive study, the research question to be answered by statistical testing is 'whether there is a difference in the prevalence of childhood malnutrition in urban and rural sectors' or 'whether the mean heights of the three groups of basketball players are different' the hypothesis to be tested is a hypothesis of difference. In contrast to this, if the research question to be answered by statistical testing in a descriptive study is 'whether there is a relationship between stunting and wasting of pre-school children in rural sector' it indicates that the hypothesis of association is being tested. Hypothesis of association states that the relationship of the two (or more) sets of outcome that we see in the results obtained from the sample is also present in the larger populations from which the sample came. Testing of hypothesis of association involves measurements of two or more sets of outcome within a single sample whereas the testing of hypothesis of difference will always involve a measurement of a single outcome made on two or more samples.

Experimental research should always test only the hypothesis of difference. For example, the research question to be tested using statistical testing in an experimental study will be 'whether the outcomes are different in the study and control groups'.

Once the researcher answers this second question in the checklist, he/she is now able to pick up the path to follow in the chosen flow chart (Figures 1-3).

Q3. If a hypothesis of difference has been tested, are the samples independent or dependent?

Third question is applicable only if the hypothesis of difference is being tested. As indicated earlier, testing for hypothesis of association involves measurements of two or more sets of outcomes within a single sample; hence this checklist question becomes irrelevant.

In the instances where the hypothesis of difference is being tested, if the selection of one of the samples is in any way influenced by the selection of the other samples we call them dependently selected. An example of dependent sample selection will be when 'matching' criteria have been used in selecting the groups to be tested in either descriptive or experimental studies. An experimental study in which the same subject acts as his or her own control (within group design) is another example of an instance in which samples are dependent on each other. Following checking

on selection of sample, if the researcher is convinced that selection of one sample has not, in any way, influenced the selection of the other sample, he or she should consider them as independent samples. In descriptive studies or in experimental studies, if the samples are selected using random sampling techniques the samples are independent.

Upon answering this third question in the checklist, the researcher is only one step away from the pathway to choosing the correct statistical test (Figures 1-3).

Q4. How many sets of measurements are involved?

The last question to answer is the easiest. It is about the number of sets of groups or outcome measures that are involved in the analysis. The question inquires whether the hypothesis of difference is testing only two groups or whether it is more than two. For example, if in a descriptive study, the research question to be answered by statistical testing is 'whether there is a difference in the prevalence of childhood malnutrition in urban and rural sectors' it indicates that two groups are being tested. If an experimental study involves three groups (one experimental group and two control groups or two experimental groups and one control group) and, if the research question was 'whether the response to the drug is different among the three groups', it indicates that more than two groups are being tested. Similarly, if in a descriptive study, the hypotheses of association is being tested and the question to be answered by statistical testing is 'whether there is a relationship between stunting and wasting of pre-school children in rural sector' it indicates that two sets of outcomes are being tested. If the research question is to test 'whether there is a relationship between stunting, wasting and head circumference of pre-school children in rural sector', it indicates that more than two outcomes are being tested.

Answering this fourth question in the checklist and following the flow chart appropriately, it will now allow the researcher to choose the correct statistical test to be used (Figures 1-3).

References

1. Sprinthall RC. Basic statistical analysis. 8th ed. USA: Allyn & Bacon; 2006.

2. Hennekens CH, Buring JE, Mayrent SL. Epidemiology in medicine. 1st ed. USA: Lippincott Wiliams and Wilkins; 1987.

3. Pipkin BF. Medical statistics made easy. Edinburgh: Churchill Livingstone; 1984.