Scholarly Research Journal for Interdisciplinary Studies,
Online ISSN 2278-8808, SJIF 2019 = 6.38, www.srjis.com
PEER REVIEWED & REFERRED JOURNAL, NOV-DEC, 2019, VOL- 7/55

STATISTICAL TOOLS IN RESEARCH AND DATA ANALYSIS

Ashwani Kumar, Ph. D.

Asst. Prof. Dronacharya P. G. College of Education Rait ,Kangra

Abstract

Statistical methods involved in carrying out a study include planning, designing, collecting data, analysing, drawing meaningful interpretation and reporting of the research findings. The statistical analysis gives meaning to the meaningless numbers, thereby breathing life into a lifeless data. The results and inferences are precise only if proper statistical tests are used. This paper will try to acquaint the reader with the basic research tools that are utilised while conducting various studies. The article covers a brief outline of the variables, an understanding of quantitative and qualitative variables and the measures of central tendency. An idea of the sample size estimation, power analysis and the statistical errors is given. Finally, there is a summary of parametric and non-parametric tests used for data analysis.

***Keywords**: statistical tools, degree of dispersion, measures of central tendency, parametric tests and non-parametric tests, variables, variance.*

Introduction

Statistics is a branch of science that deals with the collection, organisation, analysis of data and drawing of inferences from the samples to the whole population. This requires a proper design of the study, an appropriate selection of the study sample and choice of a suitable statistical test. An adequate knowledge of statistics is necessary for proper designing of an epidemiological study or a clinical trial. Improper statistical methods may result in erroneous conclusions which may lead to unethical practice.

VARIABLES

Variable is a characteristic that varies from one individual member of population to another individual. Variables such as height and weight are measured by some type of scale, convey quantitative information and are called as quantitative variables. Sex and eye colour give qualitative information and are called as qualitative variables. **Quantitative variables**

Quantitative or numerical data are subdivided into discrete and continuous measurements. Discrete numerical data are recorded as a whole number such as 0, 1, 2, 3,… (integer), whereas continuous data can assume any value. Observations that can be counted constitute the discrete data and observations that can be measured constitute the continuous data. Examples of discrete data are number of episodes of respiratory arrests or the number of re-intubations in an intensive care unit. Similarly, examples of continuous data are the serial

serum glucose levels, partial pressure of oxygen in arterial blood and the oesophageal temperature.

A hierarchical scale of increasing precision can be used for observing and recording the data which is based on categorical, ordinal, interval and ratio scales Categorical or nominal variables are unordered. The data are merely classified into categories and cannot be arranged in any particular order. If only two categories exist (as in gender male and female), it is called as a dichotomous (or binary) data. The various causes of re-intubation in an intensive care unit due to upper airway obstruction, impaired clearance of secretions, hypoxemia, hypercapnia, pulmonary oedema and neurological impairment are examples of categorical variables. Ordinal variables have a clear ordering between the variables. However, the ordered data may not have equal intervals. Examples are the American Society of Anesthesiologists status or Richmond agitation-sedation scale. Interval variables are similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced. A good example of an interval scale is the Fahrenheit degree scale used to measure temperature. With the Fahrenheit scale, the difference between 70° and 75° is equal to the difference between 80° and 85°: The units of measurement are equal throughout the full range of the scale. Ratio scales are similar to interval scales, in that equal differences between scale values have equal quantitative meaning. However, ratio scales also have a true zero point, which gives them an additional property. For example, the system of centimeter is an example of a ratio scale. There is a true zero point and the value of 0 cm means a complete absence of length. The thyromental distance of 6 cm in an adult may be twice that of a child in whom it may be 3 cm.

Classification of variables

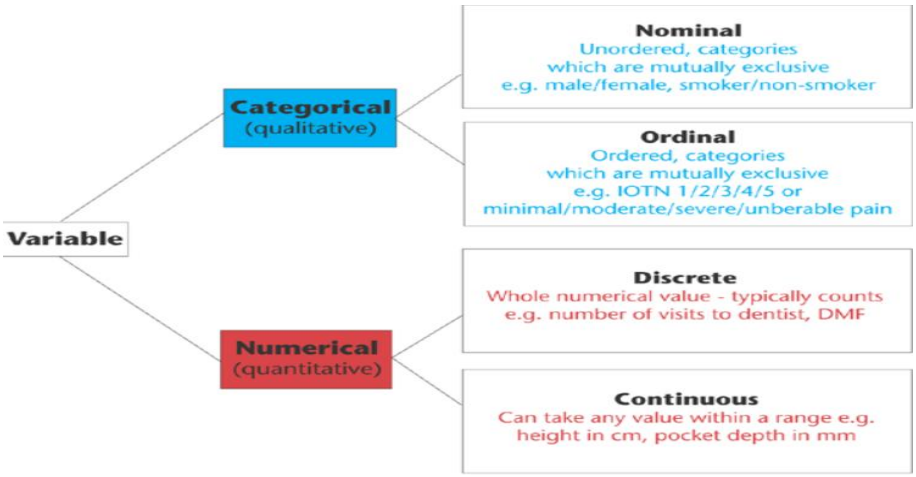


Figure No. 1

STATISTICS: DESCRIPTIVE AND INFERENTIAL STATISTICS

Descriptive statistics try to describe the relationship between variables in a sample or population. Descriptive statistics provide a summary of data in the form of mean, median and mode. Inferential statistic use a random sample of data taken from a population to describe and make inferences about the whole population. It is valuable when it is not possible to examine each member of an entire population. The examples if descriptive and inferential statistics are illustrated in Table 1. Descriptive statistics The extent to which the observations cluster around a central location is described by the central tendency and the spread towards the extremes is described by the degree of dispersion.

Measures of central tendency

The measures of central tendency are mean, median and mode Mean (or the arithmetic average) is the sum of all the scores divided by the number of scores. Mean may be influenced profoundly by the extreme variables. For example, the average stay of organo phosphorus poisoning patients in ICU may be influenced by a single patient who stays in ICU for around 5 months because of septicaemia. The extreme values are called outliers. The formula for the mean is

Mean, $\mathbf{x} = \sum \mathbf{x_i} / \mathbf{n}$ where x_i = each observation and n = number of observations.

Median is defined as the middle of a distribution in a ranked data (with half of the variables in the sample above and half below the median value) while mode is the most frequently occurring variable in a distribution. Range defines the spread, or variability, of a sample. It is described by the minimum and maximum values of the variables. If we rank the data and after ranking, group the observations into percentiles, we can get better information of the pattern of spread of the variables. In percentiles, we rank the observations into 100 equal parts. We can then describe 25%, 50%, 75% or any other percentile amount. The median is the 50th percentile. The inter quartile range will be the observations in the middle 50% of the observations about the median (25th–75th percentile). Variance is a measure of how spread out is the distribution. It gives an indication of how close an individual observation clusters about the mean value. The variance of a population is defined by the following formula:

$$\sigma 2 = \frac{\sum (x - \overline{x})^2}{N}$$

where σ2 is the population variance, X is the population mean, Xi is the i th element from the population and N is the number of elements in the population. The variance of a sample is defined by slightly different formula:

$$s2 = \frac{\sum (x - \overline{x})^2}{n-1}$$

where s2 is the sample variance, x is the sample mean, xi is the i th element from the sample and n is the number of elements in the sample. The formula for the variance of a population has the value 'n' as the denominator. The expression 'n−1' is known as the degrees of freedom and is one less than the number of parameters. Each observation is free to vary, except the last one which must be a defined value. The variance is measured in squared units. To make the interpretation of the data simple and to retain the basic unit of observation, the square root of variance is used. The square root of the variance is the standard deviation (SD). The SD of a population is defined by the following formula:

$$\sigma = \sqrt{(x - \overline{x})^2 / N}$$

where σ is the population SD, X is the population mean, Xi is the i th element from the population and N is the number of elements in the population. The SD of a sample is defined by slightly different formula

$$s = \sqrt{(\sum (x - \overline{x})^2 / n-1)}$$

where s is the sample SD, x is the sample mean, xi is the i th element from the sample and n is the number of elements in the sample. An example for calculation of variation and SD is illustrated in Table 2

Normal distribution or Gaussian distribution

Most of the biological variables usually cluster around a central value, with symmetrical positive and negative deviations about this point.[1] The standard normal distribution curve is a symmetrical bell-shaped. In a normal distribution curve, about 68% of the scores are within 1 SD of the mean. Around 95% of the scores are within 2 SDs of the mean and 99% within 3 SDs of the mean [Figure 2].

Skewed distribution

It is a distribution with an asymmetry of the variables about its mean. In a negatively skewed distribution [Figure 3], the mass of the distribution is concentrated on the right of Figure 1. In a positively skewed distribution [Figure 3], the mass of the distribution is concentrated on the left of the figure leading to a longer right tail.

Inferential statistics

In inferential statistics, data are analysed from a sample to make inferences in the larger collection of Table 2:

Example of mean, variance, standard deviation Example: The weight of five patients undergoing laparoscopic cholecystectomy was 80, 70, 80, 80, 90

Mean weight = (80+ 70 +80+ 80 +90) / 4 = 100

Variance = (80 -90) ^{2} (70 -90)^2+ (80 -90)$^{2+}$ (90- 90)^2 / 5 -1

= 100 + 400 +100+ 0 /5- 1

= 600/4

=150

SD = 150

=12.24

SD – Standard deviation

the population. The purpose is to answer or test the hypotheses. A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon.

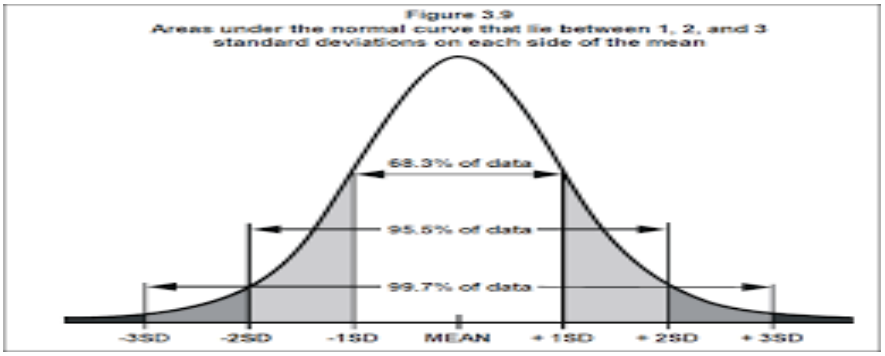
Hypothesis tests are thus procedures for making rational decisions about the reality of observed effects. Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).

In inferential statistics, the term ‘null hypothesis’ (H0 ‘H-naught,’ ‘H-null’) denotes that there is no relationship (difference) between the population variables in question.

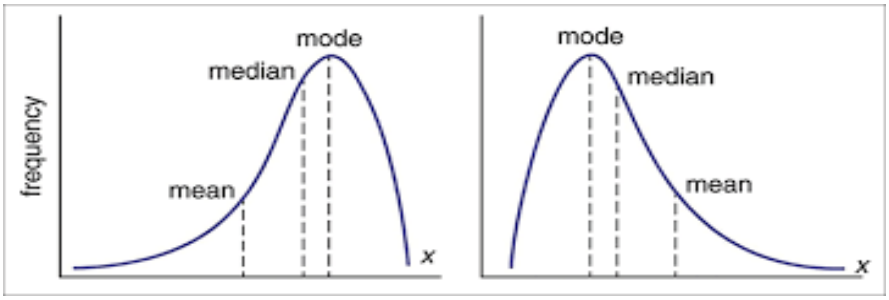
Alternative hypothesis (H1 and Ha) denotes that a statement between the variables is expected to be true.

The P value (or the calculated probability) is the probability of the event occurring by chance if the null hypothesis is true. The P value is a numerical between 0 and 1 and is interpreted by researchers in deciding whether to reject or retain the null hypothesis

If P value is less than the arbitrarily chosen value (known as α or the significance level), the null hypothesis (H0) is rejected . However, if null hypotheses (H0) is incorrectly rejected, this is known as a Type I error. Further details regarding alpha error, beta error and sample size calculation and factors influencing them are dealt with in another section of this issue by Das S



Graph No.1



Graph No. 2 Curves showing negatively skewed and positively skewed distribution

PARAMETRIC AND NON-PARAMETRIC TESTS

Numerical data (quantitative variables) that are normally distributed are analysed with parametric tests. Two most basic prerequisites for parametric statistical analysis are

: • The assumption of normality which specifies that the means of the sample group are normally distributed

• The assumption of equal variance which specifies that the variances of the samples and of their corresponding population are equal. However, if the distribution of the sample is skewed towards one side or the distribution is unknown due to the small sample size, non-parametric statistical techniques are used. Non-parametric tests are used to analyse ordinal and categorical data. Parametric tests The parametric tests assume that the data are on a quantitative (numerical) scale, with a normal distribution of the underlying population. The samples have the same variance (homogeneity of variances). The samples are randomly drawn from the population, and the observations within a group are independent of each other.

The commonly used parametric tests are the Student's t-test, analysis of variance (ANOVA) and repeated measures ANOVA. Student's t-test Student's t-test is used to test the null hypothesis that there is no difference between the means of the two groups. It is used in three circumstances:

1. To test if a sample mean (as an estimate of a population mean) differs significantly from a given population mean (this is a one-sample t-test) The formula for one sample t-test is t X u = − SE where X = sample mean, u = population mean and SE = standard error of mean

2. To test if the population means estimated by two independent samples differ significantly (the unpaired t-test). The formula for unpaired t-test is: t X X x x = 1 2 1 2 − SE − where X1 − X2 is the difference between the means of the two groups and SE denotes the standard error of the difference. To test if the population means estimated by two dependent samples differ significantly (the paired t-test). A usual setting for paired t-test is when measurements are made on the same subjects before and after a treatment. The formula for paired t-test is: t d d = SE where d is the mean difference and SE denotes the standard error of this difference. The group variances can be compared using the F-test. The F-test is the ratio of variances (var l/var 2). If F differs significantly from 1.0, then it is concluded that the group variances differ significantly.

3. Analysis of variance The Student's t-test cannot be used for comparison of three or more groups. The purpose of ANOVA is to test if there is any significant difference between the means of two or more groups.

4. In ANOVA, we study two variances –

(a) between-group variability and

(b) within-group variability.

The within-group variability (error variance) is the variation that cannot be accounted for in the study design. It is based on random differences present in our samples. However, the between-group (or effect variance) is the result of our treatment. These two estimates of variances are compared using the F-test. A simplified formula for the F statistic is: F MS MS b w = where MSb is the mean squares between the groups and MSw is the mean squares within groups. Repeated measures analysis of variance As with ANOVA, repeated measures ANOVA analyses the equality of means of three or more groups. However, a repeated measure ANOVA is used when all variables of a sample are measured under different conditions or at different points in time. As the variables are measured from a sample at different points of time, the measurement of the dependent variable is repeated. Using a

standard ANOVA in this case is not appropriate because it fails to model the correlation between the repeated measures: The data violate the ANOVA assumption of independence. Hence, in the measurement of repeated dependent variables, repeated measures ANOVA should be used. 

Non-parametric tests

When the assumptions of normality are not met, and the sample means are not normally, distributed parametric tests can lead to erroneous results. Non-parametric tests (distribution-free test) are used in such situation as they do not require the normality assumption.[15] Non-parametric tests may fail to detect a significant difference when compared with a parametric test. That is, they usually have less power. As is done for the parametric tests, the test statistic is compared with known values for the sampling distribution of that statistic and the null hypothesis is accepted or rejected. The types of non-parametric analysis techniques and the corresponding parametric analysis techniques are delineated in Table 5. Median test for one sample: The sign test and Wilcoxon's signed rank test The sign test and Wilcoxon's signed rank test are used for median tests of one sample. These tests examine whether one instance of sample data is greater or smaller than the median reference value.

Sign test

This test examines the hypothesis about the median θ0 of a population. It tests the null hypothesis H0 = θ0. When the observed value (Xi) is greater than the reference value (θ0), it is marked as+. If the observed value is smaller than the reference value, it

Situation	To use Parametric	To use Non parametric
1.Data type	Ratio or Interval	Ordinal or Nominal
2. Usual central measure	Mean	Median
3.Correlation test	Pearson	spearman
4.Independent measure two groups	Independent measure t-test	Mann whitny-test
5.Independent measure more than two groups	One way independent measure ANOVA	Kruskal –wallis test
6.Repeated Measure two conditions	Matched Pair t-test	Wilcoxon test
7.Repeated measure more than two conditions	One way repeated measure ANOVA	Friedman's test

Table No. 1

is marked as − sign. If the observed value is equal to the reference value (θ0), it is eliminated from the sample. If the null hypothesis is true, there will be an equal number of + signs and − signs. The sign test ignores the actual values of the data and only uses + or − signs. Therefore, it is useful when it is difficult to measure the values. **Wilcoxon's signed rank test**

There is a major limitation of sign test as we lose the quantitative information of the given data and merely use the + or – signs. Wilcoxon's signed rank test not only examines the observed values in comparison with θ0 but also takes into consideration the relative sizes, adding more statistical power to the test. As in the sign test, if there is an observed value that is equal to the reference value θ0, this observed value is eliminated from the sample. Wilcoxon's rank sum test ranks all data points in order, calculates the rank sum of each sample and compares the difference in the rank sums.

Mann–Whitney test

It is used to test the null hypothesis that two samples have the same median or, alternatively, whether observations in one sample tend to be larger than observations in the other. Mann–Whitney test compares all data (xi) belonging to the X group and all data (yi) belonging to the Y group and calculates the probability of xi being greater than yi: P (xi > yi). The null hypothesis states that P (xi > yi) = P (xi < yi) =1/2 while the alternative hypothesis states that P (xi > yi) ≠1/2.

Kolmogorov-Smirnov test

The two-sample Kolmogorov-Smirnov (KS) test was designed as a generic method to test whether two random samples are drawn from the same distribution. The null hypothesis of the KS test is that both distributions are identical. The statistic of the KS test is a distance between the two empirical distributions, computed as the maximum absolute difference between their cumulative curves.

Kruskal–Wallis test

The Kruskal–Wallis test is a non-parametric test to analyse the variance. It analyses if there is any difference in the median values of three or more independent samples. The data values are ranked in an increasing order, and the rank sums calculated followed by calculation of the test statistic.

Jonckheere test

In contrast to Kruskal–Wallis test, in Jonckheere test, there is an a priori ordering that gives it a more statistical power than the Kruskal–Wallis test.

Friedman test

The Friedman test is a non-parametric test for testing the difference between several related samples. The Friedman test is an alternative for repeated measures ANOVAs which is used when the same parameter has been measured under different conditions on the same subjects.

Tests to analyse the categorical data Chi-square test,

Fischer's exact test and Mc Nemar's test are used to analyse the categorical or nominal variables. The Chi-square test compares the frequencies and tests whether the observed data differ significantly from that of the expected data if there were no differences between groups (i.e., the null hypothesis). It is calculated by the sum of the squared difference between observed (O) and the expected (E) data (or the deviation, d) divided by the expected data by the following formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

A Yates correction factor is used when the sample size is small. Fischer's exact test is used to determine if there are non-random associations between two categorical variables. It does not assume random sampling, and instead of referring a calculated statistic to a sampling distribution, it calculates an exact probability. Mc Nemar's test is used for paired nominal data. It is applied to 2 × 2 table with paired-dependent samples. It is used to determine whether the row and column frequencies are equal (that is, whether there is 'marginal homogeneity'). The null hypothesis is that the paired proportions are equal. The Mantel-Haenszel Chi-square test is a multivariate test as it analyses multiple grouping variables. It stratifies according to the nominated confounding variables and identifies any that affects the primary outcome variable. If the outcome variable is dichotomous, then logistic regression is used.

SOFTWARES AVAILABLE FOR STATISTICS, SAMPLE SIZE CALCULATION AND POWER ANALYSIS

Numerous statistical software systems are available currently. The commonly used software systems are Statistical Package for the Social Sciences (SPSS - manufactured by IBM corporation), Statistical Analysis System ((SAS - developed by SAS Institute North Carolina, United States of America), R (designed by Ross Ihaka and Robert Gentleman from R core team), Minitab (developed by Minitab Inc), Stata (developed by Stata Corp) and the

MS Excel (developed by Microsoft). There are a number of web resources which are related to statistical power analyses.

A few are:

• **Stat Pages.net** - provides links to a number of online power calculators

• G-Power - provides a download able power analysis program that runs under DOS

• Power analysis for ANOVA designs an interactive site that calculates power or sample size needed to attain a given power for one effect in a factorial ANOVA design

• SPSS makes a program called Sample Power. It gives an output of a complete report on the computer screen which can be cut and paste into another document.

SUMMARY

It is important that a researcher knows the concepts of the basic statistical methods used for conduct of a research study. This will help to conduct an appropriately well-designed study leading to valid and reliable results. Inappropriate use of statistical techniques may lead to faulty conclusions, inducing errors and undermining the significance of the article. Bad statistics may lead to bad research, and bad research may lead to unethical practice. Hence, an adequate knowledge of statistics and the appropriate use of statistical tests are important. An appropriate knowledge about the basic statistical methods will go a long way in improving the research designs and producing quality medical research which can be utilised for formulating the evidence-based guidelines.

REFERENCES

Winters R, Winters A, Amedee RG. Statistics: A brief overview. Ochsner J 2010;10:213-6.

Sprent P. Statistics in medical research. Swiss Med Wkly 2003;133:522-9.

Kaur SP. Variables in research. Indian J Res Rep Med Sci 2013;4:36-8.

Satake EB. Statistical Methods and Reasoning for the Clinical Sciences Evidence-Based Practice. Ist ed. San Diego: Plural Publishing, Inc.; 2015. p. 1-19.

Wilder RT, Flick RP, Sprung J, Katusic SK, Barbaresi WJ, Mickelson C, et al. Early exposure to anesthesia and learning disabilities in a population-based birth cohort. Anesthesiology 2009;110:796-804.

Manikandan S. Measures of central tendency: Median and mode. J Pharmacol Pharmacother 2011;2:214-5.

Myles PS, GinT. Statistical Methods for Anaesthesia and Intensive Care. Ist ed. Oxford: Butterworth Heinemann; 2000. p. 8-10.

Binu VS, Mayya SS, Dhar M. Some basic aspects of statistical methods and sample size determination in health science research. Ayu 2014;35:119-23.

. Nickerson RS. Null hypothesis significance testing: A review of an old and continuing controversy. Psychol Methods 2000;5:241-301.

Ntrolled study to compare the effectiveness of intravenous dexmedetomidine with placebo to attenuate the hemodynamic and neuroendocrine responses to fixation of skull pin head holder for craniotomy. North J ISA 2016;1:16-23.