Basics of statistics for postgraduates

R. P. Nerurkar
Department of Pharmacology, T. N. Medical College, Mumbai, India

Address for correspondence: Dr. R. P. Nerurkar, Department of Pharmacology, T. N. Medical College, Mumbai, India.
E-mail: rpnerurkar@rediffmail.com

Statistics is frequently considered by postgraduates to be a tool for analyzing data after a dissertation work is complete. However, it is mandatory to use statistics right from the planning stage of a study. This article covers the area of basic statistical concepts with the aim of guiding postgraduates toward choosing correct statistical methods for a particular research question and dataset.

STATISTICAL CONCEPTS

Definition of statistics

Statistics is a way of thinking about the variable events. The relative frequency with which an event occurs is called its probability (*P* value). By convention, events with a probability of 5% or less ($P < 0.05$) are considered rare or significant.

The objective of research is to find out about the population at large. However, it is generally not possible to study the whole population and research questions are addressed in an appropriate sample. The information obtained from a sample of individuals is used to make statements about a wider population of similar individuals. The procedure of drawing conclusions about the population based on study data is known as inferential statistics. The findings from a study give us the best estimate of what is true for the relevant population.

Types of data

In research, it is necessary to study certain characteristics in a group of subjects, such as age, sex, socioeconomic group, etc. Each of these characteristics may vary from person to person and is referred to as a variable. The values taken by these variables are referred to as data.

Data collected during a study may fall into one of the following three types of data:

1. Nominal (categories, attributes) e.g., sex (male or female), religion (Hindu, Muslim, Christian, Others), blood groups (O, A, B and AB), yes/no type (patient responded or not, cured or not cured, hypertensive or normal, smoker or nonsmoker).
2. Ordinal (graded) e.g, severity of pain, itching or erythema may be graded as absent = 0, mild = 1, moderate = 2, severe = 3, socioeconomic status, degree of cigarette smoking (nonsmoker, ex-smoker, light smoker, heavy smoker).
3. Interval/ratio type (measurements) e.g., age, height, area of the lesions, blood glucose levels, etc.

Descriptive statistics

Descriptive statistics includes measures of central tendency and variability. This type of statistics is commonly used to summarize data about sociodemographic and clinical features.

Measures of central tendency include mean, median and mode.

1. The mean is the arithmetic average of data from interval or ratio scales. Mean values are affected by extreme values (outliers).
2. The median reflects the 50th percentile score or the middle value when the data is arranged in an ascending order. Median value is not affected by extreme values.
3. The mode is the most frequently occurring value of the distribution.

Measures of variability include range, interquartile range,

How to cite this article: Nerurkar RP. Basics of statistics for postgraduates. Indian J Dermatol Venereol Leprol 2008;74:691-5.
Received: August, 2008. **Accepted:** November, 2008. **Source of Support:** Nil. **Conflict of Interest:** None Declared.

standard deviation (SD) and standard error of mean (SEM).

The range denotes the spread between extreme values (minimum and maximum). Interquartile range is the data included between the 25th and the 75th percentile of a distribution.

SD describes the variability of data about the sample mean and it therefore describes the variability within a given sample. The SEM helps describe the distribution of means of several samples about a true population mean. It describes the variability of mean between samples. SEM is given by the formula SD/√N.

Finally, confidence interval (CI), which is derived from the SEM, defines the interval likely to include a true population value, based on statistical values and probability characteristics of data distribution. Mean $\pm$ 1.96 SEM gives the 95% CI. It gives the range of values that will contain the true population mean with a probability 0.95. A wider CI implies wider variance. Studies with a large sample size will provide a narrow CI.

Categorical (nominal) data can be summarized as frequencies or percentages.

Distribution of data: Normal distribution

The most important distribution in statistics is called normal distribution. It is often called Gaussian distribution. The term "normal" does not mean that the distribution is common or typical. Normal distribution curve is a frequency distribution curve and is unimodal, symmetrical and bell-shaped [Figure 1].

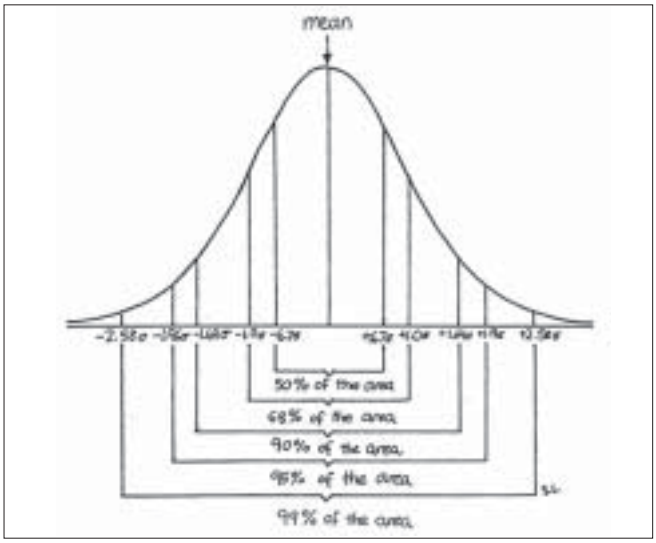

Figure 1: Normal or Gaussian distribution curve

Many statistical tests (*t*-test, analysis of variance, correlation and regressions) are based on the assumption that the collected data passes the normality test (Gaussian distribution). When data do not have a normal distribution, we can either transform the data (e.g., by taking logarithms) or use a method that does not require the data to be normally distributed. However, this process requires a little more expertise in biostatistics.

Presentation of data

After the data is summarized, it needs to be presented in the tables, graphs or diagrams.

Line diagrams are generally used to show an event in relation to time. Bar diagrams are generally used to provide visual comparison of figures. Pie diagrams are used to show the relative frequency (or percentage) of many parts of a whole. Histograms are used for frequency distribution. Scatter diagrams are useful to show an association between the variables.

USES OF STATISTICS

Statistical methods are helpful for:

1. Summarizing the data.
2. Making the estimates for the population.
3. Defining the "normal" range.
4. Testing the association between the attributes.
5. Measuring the correlation between the variables.
6. Computing one variable in terms of the others (regression).
7. Testing the significance of difference within a group or between the groups.

HYPOTHESIS TESTING

To answer questions such as "Is there a difference between two means?" or "Is there an association between various observations?," we use various tests of significance. These tests are aimed at assessing whether the null hypothesis is likely to be correct. The null hypothesis states that there is no difference between the groups with respect to the measurements made.

The significance test chosen is dependent on the type of data we are dealing with, whether it has a normal distribution, and the type of question being asked. Once the distribution is known, you can tell if the null hypothesis should be tested using parametric or nonparametric methods. Analysis using parametric tests relies on the data being normally distributed.

Nonparametric tests are used when the data is not normally distributed. The various tests of significance give us the test statistic value (*t*, *F*, *U*, *r*, etc.) and the *P*-value is then calculated using standard tables provided. If $P < 0.05$, the null hypothesis is rejected and we conclude that there is a significant difference or association between the groups.

Importance of sample size in planning and interpretation of medical research

The role of statistics in medical research starts at the planning stage of a clinical trial or laboratory experiment. This is carried out to establish the design and the size of the experiment that will ensure a good prospect of detecting the effects of clinical or scientific interest. The increasing volume of research by the medical community often leads to an increase in the number of contradictory findings and conclusions.

If the sample size is too small, the study may fail to detect a true difference that actually exists (Type II or beta error, False -ve conclusion). By convention, the beta error should be 0.2 or 20% or less i.e., sample size should be large enough to have a maximum 20% chance of the result being false negative (i.e. *P*-value being more than 0.05 even when there is significant difference between the two groups). In other words, we say that the study did not have enough power to detect a true difference. Then, the power of a study is termed as 1-beta and should be 0.8 or 80% or more.

If the sample size is too large, it may be concluded that even a very small difference is statistically significant but, in actuality, this difference is not clinically significant. We conclude that a difference exists when actually it does not exist. This is known as Type I error or alpha error (False +ve conclusion). By convention, this error (level of significance) should be 0.05 (5%) or less. A very large sample size also increases the cost and causes delay in completion of a research project.

For details of calculation of the sample size before starting a study, please refer to the article by Zodpey (Sample size and power analysis Indian J. Dermatol Venereol Leprol Mar-Apr 2004; 70:p 123-128).

CHOICE OF STATISTICAL TEST (WHY A PARTICULAR TEST IS CHOSEN)

It depends on the objective or the goal of the study [Table 1].

1. Is the goal to compare the means, compare the percentages or is it detecting an association or a establishing a correlation?
2. How many groups are to be compared?
3. What is the type of data collected? (Numerical, ordinal or nominal data).
4. Is it a paired data (in the same patient before and after treatment) or unpaired data (independent groups).
5. Is the data normally distributed (Gaussian distribution)?

TIPS AND GUIDELINES FOR DESCRIPTIVE AND INFERENTIAL STATISTICS

1. Numerical data is summarized in the form of mean and its variability is expressed as SD or standard error. Ordinal data can also be summarized as mean and SD, but median and interquartile range are preferred.
2. Nominal data is summarized in percentages.
3. All the above tests shown are inferential tests. After the

Table 1: Choice of a statistical test in common research setting

Type of data (parameter under study)	No. of groups	Unpaired data between independent groups	Paired data within the same group or dependant e.g., in the same patient at a different time interval
Numerical (interval or continuous data)	2	Unpaired t	Paired t
e.g., age, body weight, vapometer,	More than 2	One way ANOVA and post hoc	Repeated measures ANOVA and
spectrometer reading s such as L, a, b		tests (Tukey, Dunnet, etc.)	post hoc tests (Tukey, Dunnet, etc.)
Ordinal (graded) e.g., pain intensity graded	2	Mann Whitney or Wilcoxon's	Wilcoxon's matched pair signed
as no pain = 0, mild = 1, moderate = 2,		rank sum test	rank test
severe = 3, erythema graded, wrinkles	More than 2	Kruskal Wallis and post hoc tests	Friedmann's and post hoc
graded, clinical evaluation parameters			tests (Dunn's)
are graded as 0, 1, 2 and 3			
Nominal e.g., sex (M or F), cured (yes or no),		Fisher's exact test, χ^2 test,	McNemar, Cochran Q
adverse effect (seen or not seen),		χ^2 test for trend	
religion (Hindu, Muslim or Others)			

test is applied, it gives the *P*-value.

3.1 If *P*-value < 0.05, it is considered significant. It means that the probability of getting the observed difference between two or more groups by mere chance variation is less than five in 100 or 5%, and we conclude that the observed difference is not due to a chance variation.

3.2 *P*-value is also known as Type I or alpha error (false positive) i.e., it indicates the probability of the observed difference being due to pure chance.

4. In analysis of variance (ANOVA), if $P < 0.05$, it means that a significant difference exists among the various group means. It does not indicate which two means have a significant difference.
5. *Post hoc* tests are used whenever ANOVA is significant, which indicates between which two means there is a significant difference. (*Post hoc* tests are used for multiple comparisons.)
6. Tests applied to numerical data are known as parametric tests (*t* test, one way or repeated measures ANOVA). Tests applied on the ordinal or nominal data are also called as nonparametric tests (Wilcoxon, Mann Whitney, Friedmann, Kruskal Wallis, etc.).
7. If the numerical data is not normally distributed (non-Gaussian distribution), nonparametric tests are used.

Statistical methods used depend on the research question or the goal of the study

1. Is there a difference of means between groups? - We use *t* tests or ANOVA.
2. Is there any association between the variables (comparison of proportions) - We use χ^2 or Fisher's exact test.
3. When the research study involves studying association between the risk factor (exposure) and the outcome - We use odds ratio (for case control studies), relative risk or risk ratio (for cohort studies, usually prospective studies), absolute risk reduction and number needed to treat.
4. When looking for a correlation between variables - We use correlation analysis (Karl Pearson, Spearman rank correlation coefficients, etc.).
5. Is there a difference in the occurrence of an event over a time period - We use survival analysis, Kaplan Meir's curves, log rank tests or Mantel Hanzel test, Cox proportional hazard, etc.

Common errors in statistics (examples of the misuse of statistics)

Incorrect analysis of data is probably the best known misuse of statistical methods. Mishandling of statistical analysis can lead to incorrect answers and conclusions.

1. The *t* tests are widely used to compare two groups of measurements. However, it is inappropriate to use *t* tests for the data that is not normally distributed (non-Gaussian distribution) (Reference Altman DG 180 p 1473).
2. Use of unpaired *t* test in case of paired data or paired *t* test for unpaired data.
3. Use of multiple *t* test for the comparison the means of more than two groups.
4. The χ^2 test is a test of association and is used for the comparison of proportions. However, Fishers exact test is used for a smaller sample size of 50 or less and if any cell value in a 2 x 2 table is equal to or less than 5.
5. Using Karl Pearson correlation for the data that is not normally distributed or is on an ordinal scale, the Spearman rank is a more appropriate method.
6. Using one-sided *P* value instead of two-sided *P* values (one-sided *P* values are to be used only when we assume that the data in the other group is always on the same side as compared with the first group, e.g., if we assume that after treatment the blood pressure will always decrease in ALL patients).

Which statistical software should I use?

There are many statistical softwares available for carrying out the analysis, such as SPSS, SAS, NCSS, Epi-info, etc. The Graphpad Instat software is very popular as its demo version can be freely downloaded from the website www.graphpad.com. However, one may use any of the commercially available softwares as well and may get equally good results.

This article has summarized commonly used statistical methods in clinical research. An attempt has been made to introduce the subject of statistics to postgraduates ready to undertake research, although with an attendant risk of oversimplification at times. Readers are requested to refer to the bibliography for further details and consult an expert biostatistician for the detailed analysis of their data.

FURTHER READING

1. Swinscow TD. In: Campbell MJ, editor. Statistics at square one. 9th ed. London: BMJ Publishing group; 1997.
2. Dawson B, Trapp RG. Basic and Clinical Biostatistics. 4th ed. Boston: McGraw Hill; 2004.
3. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 8. Miscellaneous topics. J Assoc Physicians India 1991;39: 621-4.
4. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 7. Interval data (III). J Assoc Physicians India 1991;39:549-53.
5. Nanivadekar AS, Kannappan AR.Statistics for clinicians. 6. Interval data (II). J Assoc Physicians India 1991;39:477-82.
6. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 5.

Interval data (I). J Assoc Physicians India 1991;39:403-7.
7. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 4. Ordinal data. J Assoc Physicians India 1991;39:273-7, 281.
8. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 3. Nominal data (II). J Assoc Physicians India 1991;39:194-8, 222.
9. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 2. Nominal data (I). J Assoc Physicians India 1990;38:931-5, 974.
10. Nanivadekar AS, Kannappan AR. Statistics for clinicians. 1. Introduction. J Assoc Physicians India 1990;38:853-6.
11. Bowalekar SK. Statistics in medical research--V. Some non-parametric tests. J Postgrad Med 1994;40:96-101.
12. Bowalekar SK. Statistics in medical research--IV. Sampling distribution, statistical testing of hypothesis and student's t-test. J Postgrad Med 1994;40:46-51.
13. Bowalekar SK. Statistics in medical research--III. Correlation and regression analysis. J Postgrad Med 1993;39:235-43.
14. Bowalekar SK. Statistics in medical research--II. Measures of central tendency. J Postgrad Med 1993;39:166-73.
15. Bowalekar SK. Statistics in medical research--I. J Postgrad Med 1993;39:105-10.
16. Zodpey SP. Sample size and power analysis in medical research. Indian J Dermatol Venereol Leprol 2004;70: 123-8.