Clinical Research Methods

Primer of Epidemiology VI: Statistical analysis of research data

VIDHYA VENUGOPAL, ARUN PULIKKOTIIL JOSE, DIMPLE KONDAL

INTRODUCTION

Statistical principles are the cornerstone of research. This article aims to help the reader apply the right type of statistical analysis and improve communication between researchers and statisticians as well as help understand published studies. For more details, the readers should refer to standard textbooks such as that by Norman and Streiner.[1]

This article uses the Cardiometabolic Risk Reduction in South Asia (CARRS) study as an example to explain various statistical principles.[2]

POPULATION AND SAMPLE

To understand the scope of statistical analysis, it is important to know the distinction between population and sample (Fig. 1). The population (also commonly called the target population and sampling frame for an observational study) refers to the entire universe of measurements that the researcher is interested in and is therefore unobserved. However, the researcher, has the task of making inferences regarding the unknown population values (hereafter referred as parameter) based on the data he/she collects as part of the research study. The subset of data that is observed is referred to as the study sample. In an observational study, the study sample is drawn at random and is expected to be representative of the underlying population. The study sample provides us an estimate (hereafter referred as the sample statistic) of the unknown population parameter.

For example, we need to estimate the mean systolic blood pressure (BP) of individuals aged 18–60 years in the population (parameter). We draw a representative sample of individuals aged 18–60 years from the population and based on the sample, we estimate the mean systolic BP (statistics).

The size of the sample is often an important consideration for a researcher before starting a study. A large sample size would be ideal; however, it is often harder to obtain due to constraints of resources and time. Calculation of the optimal sample size for a study based on a primary study question/hypothesis, acceptable margin of precision, is an important first step in any study protocol to make meaningful statistical inference.

For example, the target population for the CARRS study was

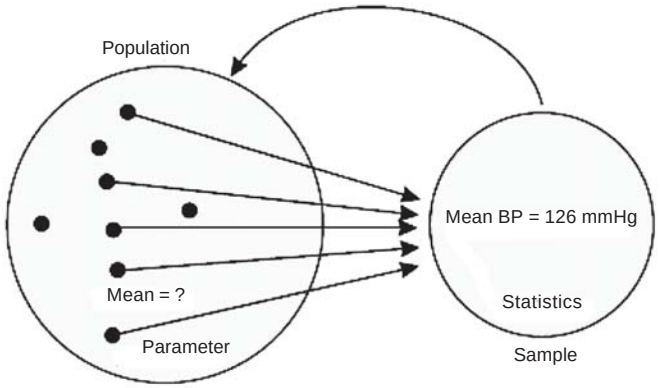

FIG 1. Population and sample BP blood pressure

individuals living in an urban location of South Asia. It was impossible to study and enumerate the entire population; therefore, a sample of individuals from Delhi and Chennai were chosen. Data from the sample were used to estimate the underlying cardiovascular disease and risk factor burden in the above-mentioned target population.

How was the sample size chosen? The sample chosen was calculated in such a way that the numbers would be adequate to represent the entire city of Delhi. It should give us the mean BP or prevalence of hypertension with 95% certainty. For higher certainty such as 99%, the sample size will be bigger. In addition, the sample size chosen should have an adequate power (statistical power is the likelihood that a study will detect a real effect consistently when there is a real effect). If the sample size is small, the power to detect a true difference would be small; this is called a type II error. Conventionally, for the estimation of sample size, power is set at 80%. The second element in the estimation of sample size is α (alpha) error, which is linked to the 'p' value, and denotes the probability of finding a difference when none exists. This is conventionally set at 5%. In addition, if we were measuring mean BP in a population (say, for Delhi in the earlier example), we should have an indication of what the mean would be from previous studies (if none are available, we may do a pilot study with a small sample and get an indication as to what that would be). In addition, we will need the standard deviation (SD) (discussed later) and using a standard formula, we will obtain the sample size. Sample size also depends on the design and purpose of the study; however, these considerations are beyond the scope of this article and the readers are referred to Norman and Streiner.[1]

In the CARRS example, the WHO had estimated the minimum sample size of 250 that would be needed to represent a true mean of BP in a narrow age band. Given that BP varies between men

Centre for Chronic Conditions and Injuries, Public Health Foundation of India, Plot 47, Sector 44, Gurgaon 122002, Haryana, India
VIDHYA VENUGOPAL, ARUN PULIKKOTIIL JOSE

Centre for Chronic Disease Control, New Delhi, India
DIMPLE KONDAL

Correspondence to DIMPLE KONDAL; *dimple.kondal@phfi.org, dimple@ccdcindia.org*

[**To cite:** Venugopal V, Jose AP, Kondal D. Primer of Epidemiology VI: Statistical analysis of research data. *Natl Med J India* 2021;**34:**352–8.]

and women and age deciles and also accounting for other issues such as the design, we increased the number so as to reach a 'reasonable number'. For example, if the sample size needed for men in the age group of 20–30 years is 250 and we also need to include women assuming the same number is needed for women, the sample size will be 500. If we add another decile of age 31–40 years, then the sample size will be 1000 (two age groups and two gender groups).

DATA AND VARIABLES

Several types of data can be collected from the study sample. These measurements are called variables as they vary from individual to individual. The research question and the resources available to the study typically dictate the procedures for data collection. The choice of the statistical method for analysis relies on the study design, nature of the outcome, exposure and other factors in the study. The distribution of the outcome variable is an important consideration while choosing the appropriate method for statistical inference as well as interpreting results. The different types of distributions will be discussed later.

Broadly, there are two types of variables: (i) quantitative (numeric) and (ii) qualitative (categorical). An example of the quantitative variable is BP or body weight and qualitative variable is gender or smoking status. Quantitative variables can be further stratified as continuous and discrete. Continuous variables refer to those that can take any value along a continuum, whereas discrete variables are ones that can take only a pre-specified set of values. Qualitative variables can be classified as nominal or ordinal. Nominal categorical variables take on distinct values with no inherent ordering, whereas the values of an ordinal categorical variable follow a set order (Fig. 2).

Table I gives a summary of the types of variables along with examples from our study.

DESCRIPTIVE STATISTICS

The first step in any statistical analysis is a description of the data collected. This provides insights into the quality of the data as well as an overview of the primary variables that pertain to the research question. Data can be described or summarized using graphs and/or tables. Table II gives an overview of descriptive statistics and graphs by different types of variables.

DISPLAY OF DATA

Graphs

The choice of graph depends on the nature or type of the variable. Categorical variables are often displayed using bar charts or pie charts. They provide information at a glance on the proportion of measurements that fall into the categories of the variable. They can also be used to explore the frequency of missing values in the variable (by including a category for missing on the X-axis of the graph).

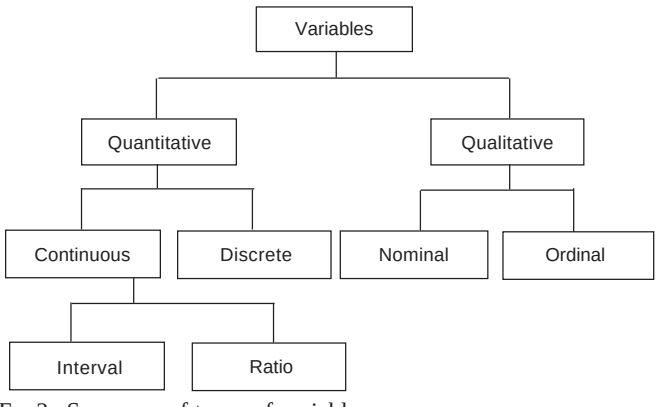

FIG 2. Summary of types of variables

TABLE I. Summary of type of variables with examples

Type of variable	Description	Examples
Quantitative		
Discrete	Variable that can take only pre-specified whole number values	Number of hospital admissions, number of fractures and members in the family
Continuous		
Interval	A variable that can be measured along a continuum; has a meaningful difference between two values; has arbitrary zero point (e.g. BMI cannot have a zero but a BMI of 18 could be a good starting point to give a range of values)	BMI, IQ, temperature in centigrade
Ratio	A variable that has a meaningful zero point; where there are equal intervals between values	Weight, pulse rate, respiratory rate, temperature (kelvin)
Categorical		
Nominal	A categorical variable without an intrinsic order	Gender (male/female), smoking status (current/former)
Ordinal	A categorical variable with some intrinsic order or numeric value	NYHA or CCS class, cancer staging

BMI body mass index IQ intelligent quotient NYHA New York Heart Association CCS Canadian Cardiovascular Society

TABLE II. Overview of different descriptive statistics and graphs by type of variable

Variable	Statistic	Graph
Continuous (ratio or interval)	Measure of location: Mean Measure of dispersion: Variance or SD	Histogram, box plot
Categorical (ordinal or nominal)	Measure of location: Mode, median Measure of dispersion: IQR	Bar chart, pie chart

SD standard deviation IQR interquartile range

Continuous variables can be summarized using histograms and box plots. Continuous variables can also be summarized using bar charts and pie charts by categorizing the variable. These graphical techniques are explained in the following examples:

- *Bar chart*: The categorical variables are commonly illustrated through a bar chart, and it shows the frequencies/percentages or the relative frequencies of the characteristics in different categories as bars. Figure 3 shows the prevalence, awareness, treatment and control of hypertension in adults over 20 years of age in three South Asian cities. The categories of hypertension (prevalence, awareness, treatment and control)

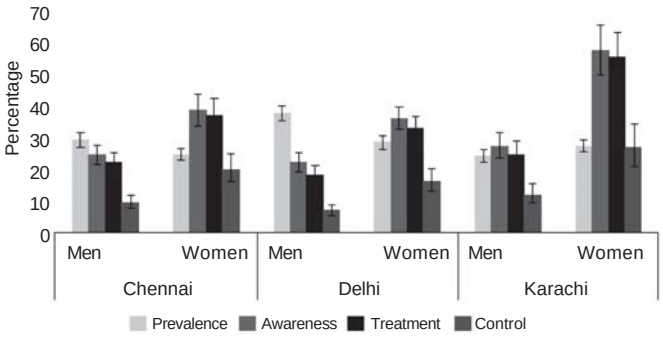

FIG 3. Bar chart showing the prevalence, awareness, treatment and control of hypertension among participants of CARRS (Cardiometabolic Risk Reduction in South Asia Surveillance Cohort[3]) study

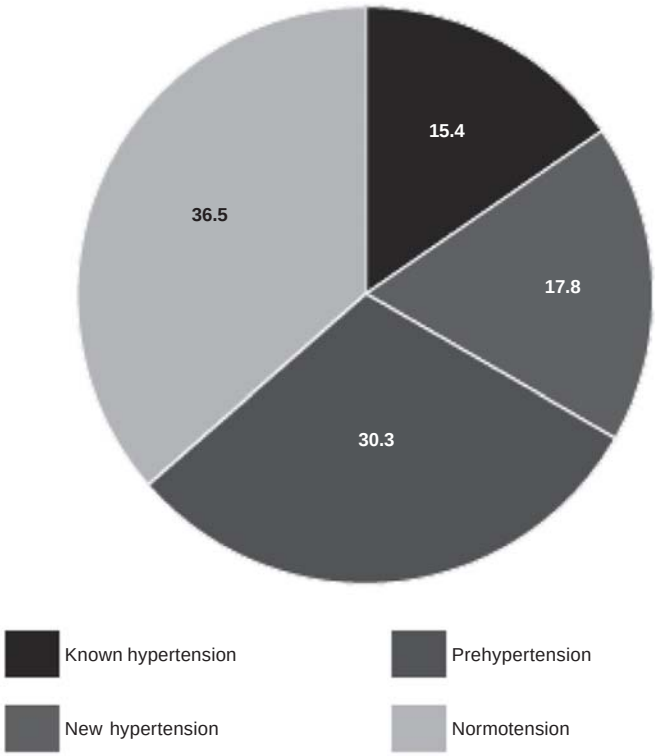

FIG 4. Pie chart showing the distribution of hypertension from the CARRS (Cardiometabolic Risk Reduction in South Asia Surveillance Cohort) study

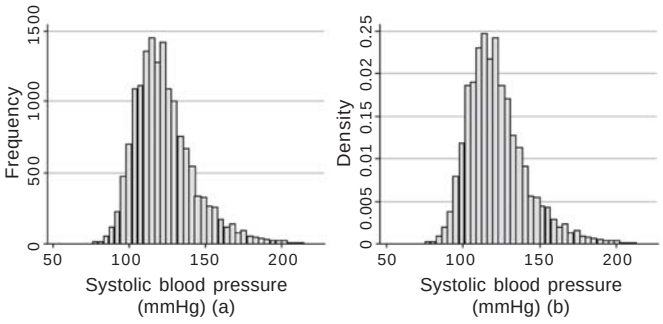

FIG 5. (a) Histogram showing the distribution of systolic blood pressure – frequency presented; (b) histogram showing the distribution of systolic blood pressure – density presented

by men and women are placed on the X-axis and the percentage of adults are placed on the Y-axis.

- *Pie chart*: It is most commonly used to show percentage breakdowns. Figure 4 shows the distribution of known hypertension, prehypertension, new hypertension and normotension.
- *Histogram*: Frequency distributions are usually presented by histograms as shown in Fig. 5 for the systolic BP data. In a histogram, we can use either frequencies or percentages; the shape of the histogram will be the same. Histogram is a set of vertical bars whose area is proportional to the frequencies presented.

Tables

Descriptive statistics are usually provided in tables. There are two ways of reporting descriptive statistics for continuous variables: measures of location and measures of dispersion.

Measures of location typically include mean, median and mode. The mean represents the arithmetic mean of all the values of the variables. In simple terms, the mean is calculated as the sum of all the observations divided by the number of observations.

The median represents the middle value of the variable such that half of the values fall above it and half of the values fall below it. To calculate the median, arrange the parameter values in ascending order. In case the number of observations (n) is odd, the median is the ($[n + 1]/2$)th observation in the list. In case n is even, the median value is the average of the ($n/2$)th and ($[n + 2]/2$)th observations in the list.

The mode is the value of the variable that occurs most frequently.

The median is generally preferred to the mean in case the variable has outliers because the latter is more sensitive to extreme values.

Outliers are values in the data that are distant or extreme in value compared with the rest of the observations. Researchers should identify such observations and deem whether they represent actual data or are erroneous measurements.

The mean age of the CARRS participants was 42.4 years. The median age was 41.0 years. The mean, median and mode will be close to each other in large samples, and the distribution will be narrow.

- *Box plot* (basic anatomy): The box plot, also known as box-and-whisker plot, shows a certain location in the distributions, i.e. first quartile, third quartile, median and 25th and 75th percentile values. The box plot of the systolic BP is shown in Fig. 6. A box is drawn with the bottom as the first quartile (Q1) and the top as the third quartile (Q3). The length of the box provides the interquartile range (Q3–Q1) and represents the middle 50% of the data. The horizontal line in the box represents the median value. The whiskers or straight lines mark the full extent of the data and are drawn on either end of the box to the maximum and minimum values.

The box plot presents a great deal of information. Figure 6 shows that the median value of systolic BP is 123 mmHg. The values 75–113 mmHg are in the first quartile, 114–123 in the second quartile, 124–136 in the third quartile, and 137–225 in the fourth quartile. The data also have outlying values (>170).

Measures of dispersion correspond to the measures of location such that they provide an idea of the spread of values

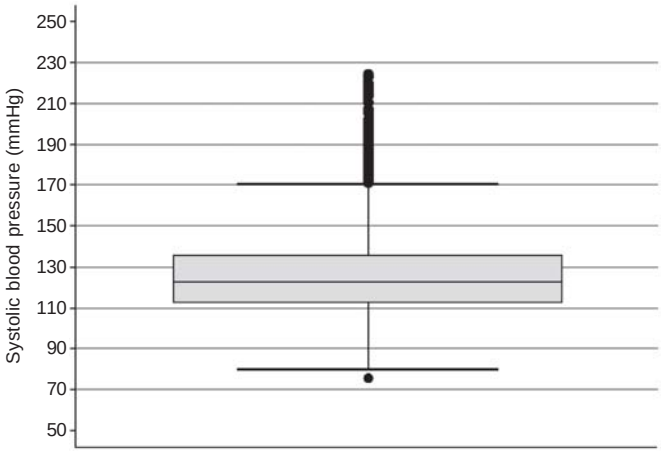

FIG 6. Box-and-whisker plot of the distribution of systolic blood pressure of 5285 individuals

in the study sample. Commonly used measures of dispersion include the variance and the range.

The range is the difference between the largest and the smallest values of the variable, whereas the variance provides an estimate of deviations of individual data points from the mean. It is calculated by summing the squared deviations and dividing that by *n*.

The SD is just the square root of the variance. It is useful in terms of describing the data as it has the same units as the variable itself. These measures of dispersion along with the measures of location provide the researcher an idea of the characteristics of the data collected as part of the study.

In an ordered set of data, quartiles represent the points that divide the data into four equal groups. The first, second and third quartiles are denoted by Q1, Q2 and Q3, respectively. The second quartile also represents the median value of the data. The difference between Q3 and Q1 is known as the interquartile range.

Coefficient of variation (CV)

It is a measure commonly used in laboratory sciences that utilizes the mean and the SD. It is the ratio of the SD by the mean expressed as a percentage. CV=(SD/mean)×100. A smaller value of the CV is better as it indicates less variability with respect to the mean (i.e. the values are more consistent in magnitude). The CV is unitless and therefore can be compared across samples.

Measures of shape

The pattern of variation of a variable is depicted by its distribution. The distribution of a continuous variable could be either symmetric or skewed. Symmetric distributions have the same size and shape on both sides of the centre or midpoint. Positively skewed distributions have a longer tail to the right of the centre, whereas negatively skewed distributions have a longer tail to the left of the centre. These different types of distributions are illustrated in Fig. 7.

Random variables and probability distribution

The previous section deals with the description of the types of data that can be collected as part of a study and the statistics that can be computed using them. These variables observed in the sample as well as the statistics such as sample mean are called random variables. As we observe only a sample of data points from the population, it is likely that each time we choose a sample, there may be differences in the outcomes. The set of all such possible outcomes along with its relative frequency of occurrence are known as the probability distribution.

Normal distribution

Most continuous variables and sample statistics are said to have a bell-shaped distribution. This implies that most of the values of the variable are centred around a mean value and that values that are further away from the mean have lower relative frequency than those around the mean. An example of this type of distribution is shown in Fig. 8.

The normal distribution provides the basis for several parametric statistical tests owing to its symmetry. As shown in the figure, mean ±1 SD will include about 68% of the sample values, the mean ±2 SD will include about 95% of the sample values and the mean ±3 SD will include about 99% of the sample values. This feature of the normal distribution is useful while constructing confidence intervals, which will be discussed later.

Standard normal distribution

A standard normal distribution is a special type of normal distribution that has a mean=0 and variance/SD=1. This means that this type of distribution is centred around 0 and is perfectly symmetrical where the mean, median and mode are similar and equal to 0. This came from the theory of numbers where the numbers are distributed from "– to +" and given the orderly distribution, the mean and median are 0.

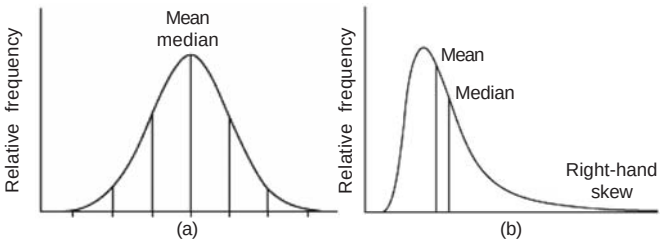

FIG 7. Different shapes of distributions: (a) Characteristics of a normal distribution; (b) characteristics of a skewed distribution

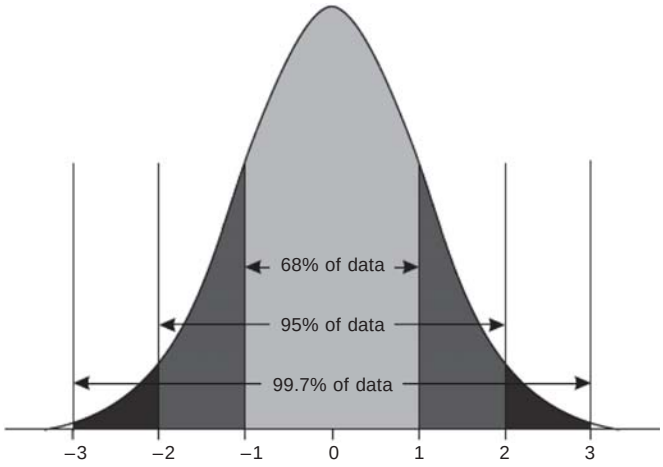

FIG 8. Standard normal distribution curve

Statistical inference

Statistical inference deals with the use of the data from the study sample to arrive at conclusions regarding some unknown population parameter. There are two main steps to statistical inference: (i) estimation and (ii) hypothesis testing.

Estimation deals with the prediction of an unknown population parameter based on sample data with some level of precision.

Hypothesis testing deals with analysing whether a population parameter is equal to some prespecified value. It allows us to determine whether enough statistical evidence exists to conclude that a belief (or hypothesis) about the population parameter is supported by the data. Techniques for hypothesis testing can also be used to answer questions comparing two population parameters.

TYPES OF HYPOTHESIS

The concept of hypothesis testing is linked closely to the formulation of the research question. The hypotheses include two conflicting statements regarding the true value of the unknown population parameter. These are known as the null and alternative hypotheses.

For example, if the research question is to test whether a population mean is equal to a particular value (say, heart rate mean is 80), the null and alternate hypotheses will be as follows:

- Null hypothesis (H_0): The population heart rate mean, $\mu=80$
- Alternate hypothesis (H_A): The population heart rate mean, $\mu\neq80$

In this case, we could also frame the H_A as $\mu>80$ or $\mu<80$. These represent one-sided alternate hypotheses, whereas the original example represents a two-sided hypothesis.

Often, the research question involves comparing means between two groups/populations. In such a case, the null and alternate hypotheses will be as follows:

- H_0: There is no (statistically) significant difference between the two groups with respect to the outcome, i.e. H_0: $\mu1=\mu2$.
- Alternative hypothesis: There is a (statistically) significant difference between the two groups with respect to the outcome.
- One-sided hypothesis: H1: $\mu1<\mu2$; H1: $\mu1>\mu2$, or two-sided hypothesis H1: $\mu1\neq\mu2$.

We generally assume the H_0 to be true and attempt to show that the H_A is true. The process of hypothesis testing starts with the computation of a test statistic based on sample data. The test statistic is constructed keeping the research question in mind.

Given the test statistic computed from the sample data, the researcher has to then compute the chance of getting a difference as big as the one observed, given that the H_0 is true.

TYPES OF ERRORS

Elements of hypothesis testing

These include size of the test (α) and power of the test ($1-\beta$).

As described earlier, α is referred to as the significance level of the test and is important in the context of p values and confidence intervals. $1-\beta$, i.e. 1–(type II error) gives us the power of the test. The findings from a study with a higher power are considered robust and credible (Table III).

There is always a trade-off between these two types of errors. The common practice is to fix the value of α and try to minimize α (or maximize $1-\beta$, i.e. power) by increasing sample size or choosing different outcome variables.

Confidence intervals and levels

The confidence interval is an interval estimate, computed from the statistics of the observed data that might contain this true value of an unknown population parameter. The interval has an associated confidence level that, loosely speaking, quantifies the level of confidence that the parameter lies in the interval. For example, a 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population. The confidence level is designated prior to examining the data. Most commonly, the 95% confidence level is used.

p value

p value refers to the probability of observing a test statistic as extreme or more extreme than the one calculated from the study sample, under the assumption that the H_0 is true.

The p value of a test is used to make a rejection decision. If $p>\alpha$ (usually >0.05), do not reject H_0, and if $p<\alpha$, reject H_0.

One of the most common mistakes researchers make is to overemphasize the importance of the p value. It is important to note that statistical significance does not guarantee clinical significance. Furthermore, a significant p value suggesting statistical association can by no means be taken as evidence of causation. The significance of the p value is tied closely to the sample size of the study, and clinically important associations may often be missed due to insufficient data. Therefore, combining statistical and clinical evidence is crucial for sound inference from research studies. p value only tells us as to how often an observed finding or more extreme than the observed one could occur due to chance alone.

CHOOSING THE APPROPRIATE STATISTICAL TEST

Once the data from the sample have been successfully collected and cleaned by removing outliers and erroneous observations, they are ready to be analysed. The choice of the method of analysis relies largely on the nature of the outcome variable. It also relies on the distribution of the outcome, specifically in the case of a continuous variable, whether it is normally distributed or not. Some of the questions to ask before embarking on the statistical analysis are:

- What is the nature of the outcome variable? What is the nature of the predictor of interest?
  - Are both continuous?
  - Are both categorical?
- Is one continuous and the other categorical?
- If the outcome is a continuous variable, is it normally distributed?
- If the outcome is a categorical variable, are there two groups or more than two groups?

TABLE III. Type I and type II error

Actual	Decision	
	H_0 true	H_0 false
H_0 true	Correct decision	Type I error
H_0 false	Type II error	Correct decision

H_0 null hypothesis Type I error: difference between groups is inferred to be true when in actuality there is no difference Type II error: failing to find a difference between groups when in actuality there is a difference

- Are the groups being compared related, i.e. are the observations between the groups being compared paired versus independent? Examples of paired observations could be BP measurements taken on the same individual before and after an intervention, heart rate on the same individual before and after exercise, etc. The idea is to identify beforehand any inherent relatedness between the groups being compared as part of the statistical analysis so as to appropriately account for it.

A univariate analysis considers a single variable at a time and is typically a descriptive analysis of the outcome and predictor variables in the study. A bivariate analysis considers two variables at a time, typically the outcome along with each predictor of interest. Finally, a multivariable approach to analysis refers to analysing the data as a whole, with the outcome and all the relevant predictor variables as part of the statistical model.

Parametric and non-parametric tests

In cases when the outcome variable is non-normally distributed (e.g. skewed data) or the data are on the ordinal scale, a non-parametric test is applied. This type of test is also used if the sample size is small and the assumptions of normality and of parametric testing are at risk of being violated. When the outcome variable is normally distributed, the parametric tests are applied. The distribution of the variable is checked by plotting histogram, normal probability plot and quantile–quantile plot. The normality assumption can also be checked using different statistical tests such as the Kolmogorov–Simonov test or Shapiro–Wilk test.

Table IV shows different types of statistical comparisons and the appropriate parametric as well as non-parametric tests for each. Table V lists the tests to be used for comparing proportions in categorical outcome data.

Studying relationship or association

Apart from measuring difference between the study sample and study population or between two study samples, another important aspect of research is to study the relationship between the variables being studied. Relationship is measured using either correlation or regression analysis. The correlation provides the strength of the linear relationship between two variables. It does not provide cause-and-effect relationship. In contrast, regression analysis provides the cause-and-effect relationship (e.g. body mass index [BMI] and BP) and is used when the goal is to predict the value of one characteristic from the knowledge of the other.

In correlational analysis, the correlation coefficient is used to calculate the strength of the relationship between two variables. The value of correlation coefficient ranges between –1.0 and 1.0. A value of exactly 1.0 means that there is a perfect positive relationship between the two variables, i.e. a positive increase in one variable is related to a positive increase in the second variable. Similarly, a value of –1.0 means that there is a perfect negative relationship between the two variables, which means that the variables move in opposite directions—for a positive increase in one variable, there is a corresponding decrease in the second variable. If the correlation is 0, there is

TABLE V. For categorical outcome data

Type of comparison	Example	Tests
Comparing proportions: Proportion in two groups	CHD status by gender	Chi-square test
Special case: Proportion in two groups, when any cell has expected value <5	Comparison of a rare adverse event between two treatment groups	Yates correction or Fisher exact test
Comparison of proportions in matched (paired) samples	Elevated biomarker (yes/no) before and after treatment	McNemar Chi-square test
Multiple groups	CHD status by gender adjusting for age groups	Mantel–Haenszel Chi-square test

CHD coronary heart disease

TABLE IV. Commonly used statistical tests for categorical versus quantitative variable

Type of comparison	Parametric tests	Non-parametric tests
Comparing means pre-post differences from a single sample	Paired *t* test Examples: Average. SBP pre- and post-intervention Two groups that are matched on a confounder	Wilcoxon signed rank test Similar situation as paired *t* test; however, the quantitative variable is non-normal For example, comparison of C-reactive protein levels before and after statin treatment in a group of post-MI patients
Comparing means (two samples)	Student *t* test Comparison of average. SBP between two independent groups, e.g. males and females	Wilcoxon rank sum test/Mann–Whitney U test Similar situation as Student *t* test; however, the quantitative variable is non-normal For example, comparison of C-reactive protein levels between two independent groups, for example, healthy controls and post-MI patients
Comparing multiple samples	One-way ANOVA followed by *post hoc* ANOVA, if required Extension of Student *t* test situation For example, comparison of average SBP across three age categories	Kruskal–Wallis test followed by *post hoc* ANOVA, if required Extension of Wilcoxon rank sum test situation For example, comparison of C-reactive protein between three independent groups, for example, healthy controls, post-MI patients and post-MI + patients with diabetes

SBP systolic blood pressure MI myocardial infarction ANOVA analysis of variance

no relationship between the two variables. The strength of the relationship varies in degree based on the value of the correlation coefficient.

Linear regression

In statistics, linear regression is a linear approach to modelling the relationship between a dependent variable (BP and pulse rate) and one or more independent variables (age and BMI). The two variables being studied are either two continuous variables or one continuous and one categorical variable. In simple linear regression, only one independent variable is used to measure the association with an outcome. The multiple regression analysis is used when there are more than one independent variables.

Logistic regression

Logistic regression is an advanced non-parametric test that is used to study the relationship between a dichotomous dependent variable (dead or alive) and, in most cases, a continuous or categorical independent variable (age and weight). A scatterplot would not be able to depict the true relationship between such variables due to the dichotomous nature of the dependent variable. A solution to this problem is to transform the scores so that the curve fits a cumulative probability curve for the logistic distribution. However, the resultant curve usually resembles an S-shaped curve, unlike the straight line seen in linear regression.

SURVIVAL ANALYSIS

Survival analysis is a branch of statistics in which the time to an event is of interest. It is used for analysing the expected duration of time until one or more events happen. The event can be death, disease occurrence, disease recurrence, recovery or other experience of interest depending on what is being studied. Survival analysis in studies related to cardiology is usually done to understand or compare efficacy of two or more modalities of management, for example, to compare the time-to-death/ cardiovascular event with use of two different classes of drugs for the management of hypertension. Techniques of survival analysis include the calculation of survival rates, life-table analysis and Kaplan–Meier method. The Kaplan–Meier survival curve is the most commonly used method to estimate survival probabilities and graphically display the survival rates. A detailed explanation of these methods is beyond the scope of this article and readers are encouraged to read more about these techniques from standard books on biostatistics.

ACKNOWLEDGEMENTS

We are grateful to Dr D. Prabhakaran, Vice President for Research, Public Health Foundation of India (PHFI), for critical inputs in developing this manuscript. We also thank Sanjana Bhaskar, Research Assistant, Centre for Environment Health, PHFI, for editorial assistance and referencing.

Conflicts of interest. None declared

REFERENCES

1 Norman G, Streiner D. *Biostatistics: The bare essentials*, 3rd ed. Hamilton, Canada:BC Decker; 2008.
2 Nair M, Ali MK, Ajay VS, Shivashankar R, Mohan V, Pradeepa R, *et al.* CARRS surveillance study: Design and methods to assess burdens from multiple perspectives. *BMC Public Health* 2012;**12:**701.
3 Prabhakaran D, Jeemon P, Ghosh S, Shivashankar R, Ajay VS, Kondal D, *et al.* Prevalence and incidence of hypertension: Results from a representative cohort of over 16 000 adults in three cities of South Asia. *Indian Heart J* 2017;**69:**434–41.