Chapter-10

Basic statistical concepts in epidemiological studies

Amna Rehana Siddiqui

Introduction

Designing, executing, and analyzing a research study demands skills based on principles of epidemiology and biostatistics.[1] Epidemiological designs are based on robust sampling techniques and sample size calculation. The analytical research question has a certain degree of uncertainty which must be quantified to solve it for getting an unbiased answer, (Does passive smoking exposure during pregnancy reduces birth weight of a new born) hence such a process goes through following six stages. (1) formulation of a hypothesis (Mothers exposed to passive smoking during pregnancy will deliver neonates with a birth weight of 200 gms less than that of neonates born to mothers not exposed to passive smoking) (2) defining the outcome variable precisely under the current hypothesis, with the statistical paradigm in the background (Birth weight measured in gms with the accuracy of two decimal points using a digital scale without clothes measured on day zero of birth), (3) designing the study for a specific population ascertaining specified experiences that are hypothesized leading to the outcome (An observational prospective cohort study will be designed to enroll women in first trimester as the two groups exposed and not exposed to passive smoking to follow up until giving birth). (4) Collection of data (ascertaining exposure to passive smoking exposure at defined intervals during follow up), to be followed by data management, ensuring the quality of data collection, storage, and analysis. Data analysis comprises of transforming raw data into usable information for reaching conclusions and decision making.

First step of analysis will involve organizing and summarizing data (mean birth weight of neonates with determining the spread of birth weight in two groups) followed by estimation and testing of hypothesis using a test of significance, (5) testing the hypothesis (comparing mean birth weight of neonates delivered to exposed and unexposed groups using a statistical test of significance e.g. Student's t test) whether the outcome is dependent on the hypothesized experience. The null statistical model would assume that specified experience (exposure to passive smoking) will have no change in the outcome (birth weight of neonate will not differ whether exposed to passive smoking or not), with an alternative hypothesis that will refute the null hypothesis, assuming that the specified experience will have a change in the outcome variable (birth weight of newborn exposed to passive smoking will differ by 200gms). Study is implemented to collect data on the experience, the defined outcome, searched characteristics related to experience and outcome. (6) Finally interpreting the data analysis from statistical summaries and reaching the appropriate conclusions (e.g. considering that the pregnant women in the two groups may have differed by nutritional status, parity, poverty, if more of them in passive smoking exposed group, thereby lower birth weight resulted because of poor nutrition, low and high parity, and poor socioeconomic status; this will require multivariable analysis to exclude the extraneous effects other than that of passive smoking).

This chapter introduces basic concepts in biostatistics required to understand published scientific papers. The major topics include, scales of measurement, common statistics used to summarize data, assessment of probability of possible outcomes, and making inferences from the given data considering confidence intervals and testing of hypothesis. Good quality data provides basis for inferences and help to reach important conclusions.[2] The above-mentioned outline will be used for describing basic statistical requirements for each stage using examples.

Descriptive Studies

Descriptive studies are designed to identify distribution of risk factors and disease related outcomes in any population or sub-group of population, e.g. proportion of current smokers in university students, number of unimmunized children in a village, correlation between two variables (age and systolic blood pressure), mean birth weight of neonates born to adolescent mothers. Researchers collect information (data) using various tools designed to measure specific characteristics of study population (humans, animals, or objects). Such information required by researchers is collected through measuring tools (e.g. questionnaires) either designed by investigators or preferably taken from validated tools reported in literature.

Scales of Measurements

The scales of measurement need to be identified thoughtfully keeping the study research question, objectives, hypothesis, especially in terms of whether the desired tool /questionnaire will be able to answer the health problem related question, and can measure the desired characteristics. Measurement of desired characteristics like, age, sex, marital status, parity, weight, height, morbidity, and many others is done using scales of measurement (Table-I).[3,4]

Summary Statistics and data display

Qualitative observations using nominal or ordinal scale are described in terms of percentages and proportions, and displayed in tables and bar charts. Ordinal scales with some order are classified as nominal scales as the difference between two adjacent categories is not the same. Bar charts can show complex data in a simple way for nominal and ordinal type of variables. Pie chart is another choice for showing qualitative data variables.

Continuous scales for which the difference between numbers (like 1 and 2) have a numerical scale (e.g. weight, height) are called quantitative observations. Numerical data can be displayed in a large variety of frequency tables (using

Table-I: Measurement Scales for variables

#	*Type of measurement*	*Charac-teristics of variable*	*Example*	*Statistical measures for descriptive reporting*
I.	Nominal (Qualitative)	Unordered categories Binary, dichotomous, categorical (Yes or No)	Race, Sex, Blood group, type of anemia, educated or not, having pain	Absolute and relative frequencies
II.	Ordinal (Qualitative)	Ordered categories with intervals that are not quantifiable	Degree of pain, grades of tumor, level of comfort, APGAR score, level of education, severity of pain	Mode, Median, 1st and 3rd quartile, interquartile difference, minimum & maximum
III.	Interval Continuous (Quantitative)	spectrum with quantifiable intervals	Age, Body weight, height, waist circumference, body mass index, years of schooling, duration of pain	Mean (Standard deviation(SD)), median, quartiles, minimum, maximum, percentiles
	Discrete (ordered) (Quantitative)	Values have a finite scale	# of fractures, # of cigarettes, parity, # of patients visiting a clinic	

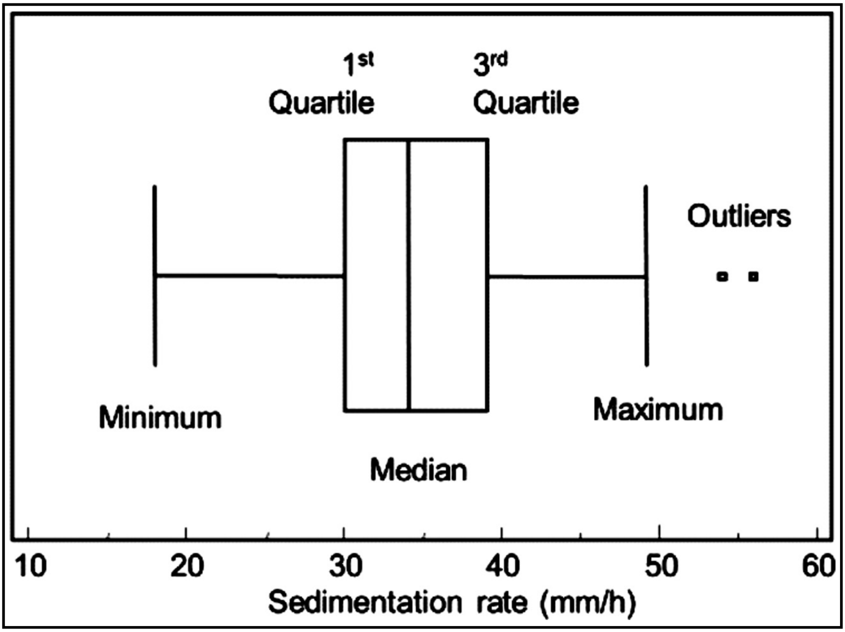

Fig.1: Box and whiskers plot Reproduced with permission from: Pupovac V & Petrovečki M. Summarizing and presenting numeral data. Lessons in Biostatistics. Biochemia Medica. 2011. Vol 21 (2):106-10

class limits), histograms, line graphs and box and whiskers plots (Fig.1).[5] Data in Fig.1 (n=312) shows a sedimentation rate as median of 34 mm/h, 1st quartile 30 mm//h, 3rd quartile 39 mm/h, minimum percentile (18 mm/h) and the maximum percentile (49 mm/h), with two outliers. (Fig.1)

Descriptive statistics

It is used for quantitative data and done by assessing the measures of central tendency and dispersion (Fig.2).[5] Summarization of data as descriptive statistics is done by assessing the central tendency of data, a value around which data are centered. The three most common measures of central tendency are mean, median, and mode. Arithmetic mean or simply mean is the most commonly used measure as it uses all the information in a data, however, it is sensitive to extreme values, Median is resistant to extreme values, and is the central value when data are arranged in an order. Mode is the most frequently occurring value. A symmetrical distribution cannot be assumed when mean and median differ from each other substantially.[1] A histogram is useful to check the distribution of numeral variables.

When the distribution of a variable is symmetric around mean, median, and mode are identical. In right- skewed data, median is less than mean, whereas in the left- skewed data, mean is less than median.[6] A symmetric distribution is 2/3rds of the values fall under one SD. When data for a continuous scale variable is to be shown in a table, mean is used as a cutoff for making class intervals. If the distribution is asymmetric then the cut off at median can be used to categorize a variable for showing data in a table.

Variability of a dataset is described by range, standard deviation, and variance. Range is the difference between minimum and maximum value of data. Standard deviation (SD) is a measure of spread of its mean value and needs to be reported along with mean value when symmetrically distributed.[7] A distribution with a greater spread (SD=20) has more variability than the one with lesser spread (SD =10).[6] The SD informs about the extent the data points cluster around the mean value. Measures of dispersion complement measures of central tendency; if the SD is small it means that mean is describing the scores in dataset, whereas when the SD is high it still states that mean is the best representative of all the scores in a dataset, but many values lie far away from mean (e.g. birth weight of a neonate). A Z score is a standard score and is placed on a symmetrical curve. Z-score of a measurement X indicates how many standard deviations is the measurement away from the mean? A positive Z-score indicates that the measurement is above the mean and a negative Z-score indicates that the measurement is below the mean. Sixty eight percent of the data will fall within 1 standard deviations of the mean, 95% of the data will fall within 2 standard deviations of the mean, and almost all 99.7% of the data will fall within 3 standard deviation of the mean (Fig.2).

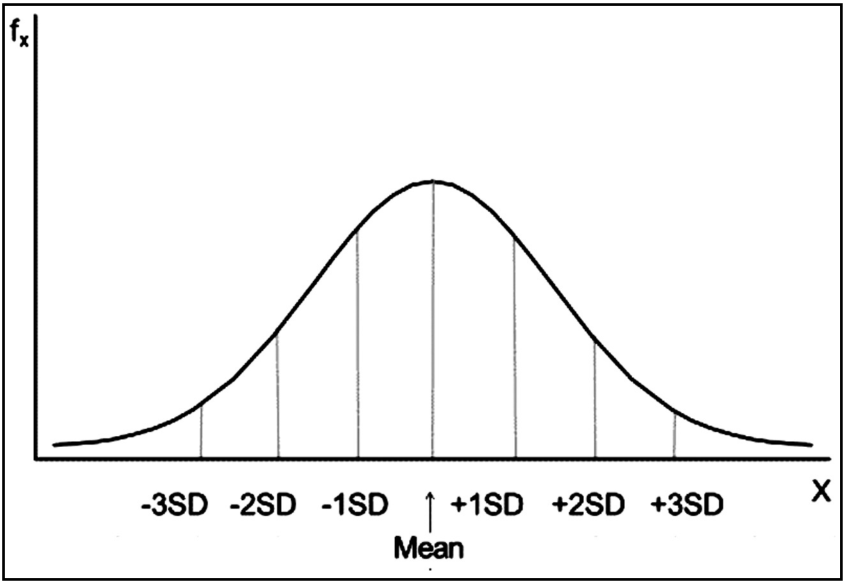

Fig.2: Mean and standard deviation (SD) as measure of central tendency and spread in a symmetric distribution [Reproduced with permission from: Pupovac V & Petrovečki M. Summarizing and presenting numeral data. Lessons in Biostatistics. Biochemia Medica. 2011.Vol 21 (2):106-10]

Descriptive characteristics when summarized for a population are called parameters, e.g. National Census data from Pakistan reported average household size in any year, whereas when average household size is determined by using a sample of Pakistani households' form the city of Karachi it is called a statistic.[8]

Inferential statistics

For research purpose estimation of population parameters is done through the information contained in the sample. Therefore, the sample selection needs to be done carefully (randomly or any other similar method) as statistical assumptions and tests of significance are based on random samples selected from population of interest with symmetrical distribution.[8] If X bar is the statistic (sample mean) and several samples from the population are taken, each will yield a different X bar, hence such a sampling distribution of means will provide a population mean (μ). The standard deviation of the sampling distribution of mean is called standard error of mean. This is used in estimating the confidence interval around the point estimate, that indicates the variability of the point estimate.[9] According to the central limit theorem (CLT), a statistical theory, states that a sufficiently large sample size from the population with a definitive variance level, the means of all samples obtained from such a population will be approximately equal to the mean of the population.[9] The interval estimates consist of two numerical values defining a range of values that, with a specified degree of confidence (e.g. 95%) includes the parameter being estimated. The confidence interval estimation captures the variability of the randomly drawn sample.

Hypothesis testing

A good hypothesis is based on a good research question, that is simple, specific and stated in advance. A specific hypothesis clarifies study population, variables, and test of significance. E.g. it is hypothesized that mean birth weight of neonates born to smoker mothers will be 200 gms less than the mean birth weight of neonates born to non-smoker mothers, when numerical outcome data (birth weight is a continuous variable) are hypothesized to be compared Student's t test as test of significance will be used. Here null hypothesis will be that mean birth weights do not differ by mother's smoking status, with an alternative that smoking status will show a mean difference of xxx gms. Likewise, such a hypothesis could be stated for low birth weight (LBW) using it as a binary/dichotomous/categorical outcome variable; hypothesizing that twice the number of LBW newborns will be borne to smoker mothers than non-smokers, when adjusted for maternal nutrition, reproductive and demographic variables. Here if the study design is prospective cohort study then twice more LBW of the magnitude of 2.0 can be measured as Relative Risk (RR), and Chi square will be used as test of significance as now the LBW is categorical variable (as Yes and No). In statistical hypothesis testing, there are two types of errors that could occur sometimes. (a) Type I error occurs when the test shows significant difference but in truth no such effect exists. (b) Type II error occurs when a true difference or effect is present between the two groups but due to small number of participants, or poor conduct of measurements the statistical analysis is unable to show a significant p value. The concern for false positives and false negatives need to be considered when designing, analyzing and interpreting the research study results.[8]

P value

Probability is the proportion of times that event (mean birth weight < 200 gms) occur over a long series of repeated studies. Probability is quantified as a future event as the number between zero to one; zero indicating impossibility and one indicates certainty[9]. Let us assume that at statistic of 0.14 was obtained when comparing mean birth weights with a corresponding p value of 0.90. This p value means that there are 90 chances in 100 that a t statistic of 0.14 (or larger) will be obtained if the null hypothesis were true. Because this probability is so high, there is very little evidence to reject the null hypothesis. Therefore, when we fail to reject the null hypothesis, it is concluded that there is not sufficient evidence to support a statistically significant difference between the two groups. However, if the p value was 0.001 then the conclusion would be that there is one in thousand chances that such a result will be obtained if null hypothesis was true; therefore, it is a very extreme and low probability (1 in 1000) therefore we reject the null hypothesis.

Summary

Understanding of statistics is crucial for health care workers including doctors, nurses, and those involved in decision making process for patients, system, and research.[10] Understanding published literature is necessary to update rapidly changing advances in health care and management. If the application of mean and standard deviation is applied to not-normally distributed data, then a different and error prone interpretation could lead to misinterpretation of data leading to limited understanding of a specific problem. Improvements in health care are based on measuring and comparing outcomes in two sets of conditions (medical or surgical, short or long follow up, day care or admitted care). Statistical tests applied to compare two sets of procedures are based on assumptions of data distributions. Application of correct tests of significance is necessary, if the data are not normally distributed, non-parametric tests of significance are applied to determine the meaningful difference in the two procedures. Interpretation of p values setting of alpha, and understanding of type I and type II errors would make a lot of difference when interpreting data for evidence and decision making.

In this chapter we have only discussed comparing one outcome and one variable, or procedure, however, there are multivariable techniques to compare two or more procedures. Besides these, there are sometimes problems with missing data, low response, loss to follow up and several statistical procedures can handle such missing information; but the responsibility to develop full understanding that which methods would be appropriate in specific situation rests with the reader.

REFERENCES

1. Williams OD. Chapter 4: Basic biostatistics concepts and tools. Basic Epidemiology 2nd Edition. World Health Organization 2006.
2. Hulley SB, Newman TB, Cummings SR. Chapter 4: Planning the measurements, precision, Accuracy and Validity. Pages 32-34. Designing Clinical Research Stephen B Hulley. 4th Edition 2013. Wolters Kluwer, Lippincott Williams & Wilkins.
3. Basic & Clinical Biostatistics. Chapter 3: Exploring and Presenting Data: pages 20-25. Beth Dawson- Saunders & Robert G Trapp. 2nd Edition, 1993. Appleton & Lange.
4. Divisi D, Di Leonardo G, Zaccagna G, Crisci R. Basic statistics with Microsoft Excel: a review. J Thorac Dis 2017;9(6):1734-1740. doi: 10.21037/jtd.2017.05.81
5. Pupovac V and Petrovečki M. Summarizing and presenting numeral data. Lessons in Biostatistics. Biochemia Medica. 2011;21(2):106-110.

6. Chan YH. Biostats 101: Data presentation Singapore Medical Journal 2003. Vol 44 (6): 280-285.
7. Spriestersbach A, Röhrig B, du Prel JB, Gerhold-Ay A, Blettner M. Descriptive Statistics: The Specification of Statistical Measures and Their Presentation in Tables and Graphs. International Dtsch Arztebl Int 2009;106(36):578–831
8. Kocher MS & Zurakowski D. Clinical Epidemiology and Biostatistics: A primer for Orthopedic Surgeons. Journal of Bone and Joint Surgery 2004. Vol 86-A No 3: 607-620
9. Daniel WW: Biostatistics A foundation for Analysis in the health Sciences 6th Edition 1994. Wiley Series in Probability and Mathematical Statistics – Applied.
10. Lim E. Basic statistics (the fundamental concepts). J Thorac Dis 2014;6(12):1875-1878

[1] Dr. Amna Rehana Siddiqui, MBBS, FCPS, MSPH, PhD.
Associate Professor,
Community Health Sciences,
Aga Khan University,
Karachi - Pakistan.
E-mail: amnarehana@gmail.com