A Brief Outline of Bio-Statistics in Medical Research

Dar Nazir Ahmad [1], Sofi Mushtaq Ahmad [2] , Nayak Bilal Gul [3], Khan NA[4], Lone MM[5],KS Raina[6]

[1]. Statistician, HBCR, Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)
[2]. Senior Resident , Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)
[3]. Senior Resident , Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)
[4]. Professor, Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)
[5]. Professor and Head Department of Radiation Oncology,SKIMS, Srinagar Kashmir(India)
[6]. DEO,HBCR, Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)

Abstract- Bio-statistics is the important branch of statistics and is related to medical field. Also it plays an important role in the field of research. The main role of statistics in research is to designing a research, analyzing data and draw meaningful conclusions. The meaningful conclusion can be drawn by using proper statistical tests. Statistics also helps to reduce large volume of raw data which must be suitably reduced so that the same can be read easily and can be used for further analysis. This article covers the brief outline of data, qualitative data, and quantitative data. Also give a brief outline of measures of central tendency (location), measures of variability (dispersion) , statistical inference(parametric and non-parametric tests) and give brief outline of sample size calculations.[4]

Index Terms- Biostatistics, data, Parametric and non-parametric, measures of location, measures of dispersion (variation)

I. INTRODUCTION

The process of converting data into meaningful information requires a special approach called statistics. Statistics is the branch of methods for making wise decisions in the face of uncertainty. In other words, it can be defined as collection, summarization, organization, analysis and interpretation of numerical data. Biostatistics is the science that helps in managing medical uncertainties. It is mainly consists of various steps like generation of hypothesis, collection of data and application of statistical analysis. An ample knowledge of bio-statistics is important for research scholars, medical students, and nursing students so that they can design epidemiological study accurately and draw meaningful conclusions and inadequate knowledge of biostatistics leads to biased results and conclusion which may lead to erroneous conclusions and also may lead to unethical practice.

II. DATA

Data means raw facts and figures, from which a meaningful conclusion may be drawn. If you want to understand a phenomenon of any disease you must have a data so that you will be able to know the pattern and incidence of disease. Data will be further divided into two categories qualitative (categorical) and quantitative (numerical) data. Data which is non-numerical called qualitative data e.g, gender, habitat, health status, etc. Data which is numerical called quantitative data e.g, number of patients , pulse rate , etc . Quantitative data may be further divided into categories discrete and continuous .[2]

Discrete data can take on only integer values whereas continuous data can take on any values. For instance the number of cancer patients treated by a hospital each year is discrete but height of the cancer patient is continuous. Some data are continuous but measured in a discrete way e.g, your age , it is common to report your age as say 28.Classification of data shown in [1] [Figure 1].

A: Statistics: Descriptive and inferential statistics

When we analyze our data it is necessary to use both descriptive and inferential statistics. Descriptive statistics are used to describe our data, summarize our data and organize our data so that we can draw meaningful results. The important descriptive statistical measures that are used to describe, summarize and find out the central value and location of the data are Measures of location and Measures of variability.[11]

1. Measures of Central Tendency:

Measures of location(central tendency) gives the central value of the data that is used to represent all the values of series. The three important measures of location are Mean , Median ,and Mode . Geometric and harmonic mean are not used most often.[3]The most important are summarized below:

1.1 Mean

Mean is the first and simplest measure of location. It is the most frequently used measure of location. It can be defined as sum of observation divided by the number of observation. The most important drawback is, mean of a particular group is affected by very small and very large number.

Mean=Sum of observation /number of observation

Eg; Birth weights of new born babies are
3.3 , 6.1 , 5.8 , 3.8 , 2.7 , 4.1 , 3.4 , 3.9 , 5.1 , 3
Mean=$\sum xi/n$
Mean =3.3+6.1+5.8+3.8+2.7+4.1+3.4+3.9+5.1+3/10
Mean= 4.12kg
i.e Mean birth weight of new born babies are 4.12kg

1.2 Median

Median is defined as middle of observation. It divides the whole data into two equal parts one part comprising all the values less than median and second part comprising all the values greater than median. Median is not affected by extreme values.

Median is the only average used for dealing with the qualitative of data. In median we have two cases odd and even case.

In odd case we arrange the distribution into ascending (descending) order and distribute the series into two parts and middle one is median. In even case we arrange the distribution into ascending (descending) order and calculate average between the two middle values and the middle value is median.

Eg, in odd cases:- Birth weights of new born babies are
(3.3 , 6.1 , 5.8 , 3.8 , 2.7 , 4.1 , 3.4 , 3.9 , 5.1)

Arrange the data in ascending order
(2.7 , 3.3 ,3.4 , 3.8 , 3.9 , 4.1 , 5.1 , 5.8 , 6.1)

3.9 is the median birth weight of new born babies.
Eg , In even cases:- Birth weights of new born babies are
(3.3, 6.1 , 5.8 , 3.8 , 2.7 , 4.1 , 3.4 , 3.9 , 5.1,3)

Arrange the data in ascending order
(2.7 , 3 , 3.3 ,3.4 , 3.8 , 3.9 , 4.1 , 5.1 , 5.8 , 6.1)

Arithmetic mean between (3.8 + 3.9)/2= 3.85
3.85 is the median birth weight of new born babies.

1.3 Mode:

Mode is the most frequently occurring value in a set of data. Mode is particularly useful in the study of popular sizes. Mode is the average to be used to find the ideal size in a series.
Eg
No mode
Raw data : 10.3 4.9 8.9 11.7 6.3 7.7
One Mode
Raw data: 6.3 4.9 8.9 6.3 4.9 4.9
More than 1 mode
Raw data : 21 28 28 41 43 43

Table 1: Summary of Central tendency

Measure	Descriptive
Mean	Balance point
Median	Middle value when ordered
Mode	Most frequent

2. Measures of Dispersion (variability):

Measures of dispersion convey information regarding the amount of variability present in the data. If all the values are same there will be no dispersion and if all the values are different there will be dispersion. There will be two possibility of dispersion one is if all the values are close to each other there will be less amount of variability present in the data and second is if all the values are too much scattered there will be large amount of dispersion present in the data. For among the measures of dispersion range and standard deviation (SD) is most often used measures of dispersion. For the comparison point of view we use mostly the co-efficient of variation (co-efficient of standard deviation).

2.1 Range

Range is the simplest measure of dispersion .It can be defined as difference between the two extreme items of series. The utility of range is that it gives us an idea of variability very quickly.
Range= (highest value of series- lowest value of series)

2.2 Standard Deviation (SD):

The standard deviation is mostly used in research studies and is regarded as the very satisfactory measures of dispersion. Standard deviation can also be defined as the positive square root of the mean of the squared deviation of the values of mean.The standard deviation describes how much individual measurement differs on the average from the mean.
SD=

A large standard deviation shows that there is a wide scatter of measured values around mean and small standard deviation shows that individual values are concentrated around the mean with little variation among them.

2.3 Variance:

Variance quantifies the amount of variability or spread about the mean of the sample.

3 . Normal Distribution or Gaussian distribution

Mean and SD are summary measures of the silent features of the data. But in many cases, such summary doesn't adequately describe the full characteristics of the data set. To study scatter in detail, the frequency of subjects with values in specified short intervals are plotted against the intervals. This plot is in terms of bars drawn adjacent to one another with area representing the frequency . This plot is called histogram. When midpoints of the top of the bars are joined by straight line we get frequency polygon. Smoothened shape of this polygon is called frequency curve, and the curve formed in the bell shaped curve. This curve signifies that the frequency is highest for middle of the variables and decline both sides are similar. Mean, Median and Mode are same .Curve is symmetrical. These are the broad features of what is known as Gaussian curve or Normal curve. This frequency distribution is known as Normal or Gaussian distribution.

4 Statistical Inference:

Statistical inference means to draw inference about the characteristics of the population. It deals with the estimation of population parameter and statistical tests of significance is drawn on the basis of sample statistics and the findings are expected to be applicable for the entire target population. Some of the below mentioned terms used in statistical inference :

(a) Hypothesis : Hypothesis means assumption regarding the population.
(b) Null Hypothesis (Ho): A null hypothesis is usually statement that there is no difference between groups or

that one factor is not dependent on another and corresponds to the no answer.

(c) Alternative Hypothesis(H_A) or (H1) : Alternative hypothesis which is complimentary to the null hypothesis.

(d) P-value: The p-value (probability value) is the probability of event occurring by chance if the null hypothesis is true. The p-value always lies between (0 and 1), also it is interpreted by the researchers in deciding whether to reject or retain the null hypothesis . If p<0.05,it means that data is statistically significant and if P>0.05 data is not Statistically Significant. [12][Table 2]

(e) Type I-error:- p[reject Ho/when it is true]=α

(f) Type II-error:- p[accept Ho/ when H1 is true]=β

(g) Power = p[reject Ho/H1 is true]=1- β

Table 2 : P values interpretation

P	Result	Null Hypothesis
<0.01	Result is highly significant	Reject(null hypothesis) Ho
≥0.01 but <0.05	Result is significant	Reject (null hypothesis) Ho
Value ≥0.05	Result is not significant	Do not reject (null hypothesis)Ho

4.1 Parametric and non-parametric test

Parametric test means which deals with the parameters of the population and if the distribution follows normal distribution we use parametric tests. Hypothesis tests which are based on knowledge of the probability distribution that the data follow are known as parametric test, often data do not conform to the normality of the distribution in these situation we can use non-parametric test of the distribution(sometimes referred to as distribution free test or rank method).Non-parametric tests are particularly useful when the sample size is small and when the data are measured in categorical scale[5][6].The commonly used parametric tests are used in research methodology i.e Students t-test and Analysis of variance (ANOVA).

4.1.(a) Student's t-test

Students t-test is a parametric test , it is applicable to find a significant difference between two means . It is used in three situations:-

(a) To test if the sample mean differs significantly from the population mean (one sample test)

(b) To test if the population means estimated by the two independent samples differ significantly (unpaired t-test).

(c) To test if the population means estimated by the two dependent samples differs significantly (paired t-test).

4.1.(b) Analysis of variance (ANOVA):

When data are normally distributed students t-test can be used to assess the significance of the means of the sample. To compare the difference between three or more independent groups simultaneously, an analysis of variance, which is parametric test can be used and when there is only one qualitative variable which defines the groups, a one –way ANOVA is performed. In ANOVA we will study two types of variation Between groups and within groups. The ANOVA separates the total variability in the data into which can be attributed to difference between the individuals from the different groups (between group variation) and the variation between the individual within each group (within group variation), sometimes called unexplained or residual variation.

Non-Parametric tests

4.1.2 (a) Chi-square test:

Chi-square test is non-parametric test and is a measure of significance. This test will be used to find out the association between two events in binomial or multinomial samples. It helps you to decide if the relationship exists but not how strong it is .Chi-square test will be used for categorical data. In medical research when a researcher wants to study the association or relation between two or more variables ,this type of relationship will be studied by correlation or regression . Some times we are interested in studying the association between a continuous variables grouped into categories [Anemia:mild , moderate , and severe] and discontinuous variables grouped into categories [Economic status: lower , middle and upper] and association between two continuous variables grouped into categories [height: short , medium and tall] [weight: light , moderate and heavy] . [8]

The important assumption of chi-square test is if any of the expected frequency is less than 5 (i.e E<5) in any one of cell, we use Fisher's exact test. Fisher's exact test is used to determine if there are non-random associations between two categorical variables.[10]

4.1.2 (b) Sign test:

Sign test is simplest of all non-parametric test and is based on the median of the distribution. The test will be used to comparing the single sample with some hypothetical value ( ) for the median in the population. If our sample comes from this population, then approximately half of the values in our sample should be greater than ( )and half of the sample should be less than ( ).In sign test we use signs positive (+) and negative (-) to every observation . When reference value is less than observed value plus sign will be used and when reference value is greater than greater than observed value negative sign will be used. And when reference value is equal to observed value it will be eliminated.

4.1.2 (c) Wilcoxon Signed rank Test:

The wilcoxon signed rank test takes account not only of the signs of the difference but also their magnitude and therefore is more powerful test. Also individual difference is calculated for each pair of results. Ignoring zero difference, there are then classified as being either positive and negative. In addition, the difference are placed in order of size, ignoring their signs and are ranked accordingly. The smallest difference thus get the value (1), the second smallest gets the value (2), etc up to the largest difference.

4.1.2 (d) Mann- Whitney U-test:

Both sign test and wilcoxon signed rank test are helpful non-parametric test and alternative for Mann-Whitney U test is one sample t test and paired test . A non-parametric alternative to the unpaired t-test is provided by wilcoxon rank sum test which is known as Mann-Whitney U-test. This is generally considered when comparison is done between two independent groups.

The proper description of various tests is described in Table 3.

5 Software's used for statistical Analysis are mentioned below:

Software's which are used for data analysis are called statistical software. There was many software's available for analysis point of view, some are free and downloadable from the site.

Common statistical software's:

1. *SPSS- Statistical Software for social Sciences (V23)*
2. *SAS – Statistical Analysis System(v9.3).*
3. *Systat - (13)*
4. *BMDP – Biomedical Data Processing package*
5. *S plus – (6.2)*
6. *Epiinfo (WHO- 3.5.4) version 7 is latest -Free*
7. *Stata (13)*
8. *Gauss or Pass*
9. *Glim (14.2)*
10. *R software 3.1.2*
11. *statpages.org*
12. *G-power*
13. *statstodo.com*

III. Conclusion

Apart from the knowledge of medicine, knowledge of biostatistics plays a vital role in the field of research. The main purpose of this article is to aware medical students what is the applicability and usage of biostatistics in medical field. This brief overview of biostatistics gives a quick summary methods for medical students in their short project and thesis work.

Financial support and sponsorship:

Nil

Conflict of interest:

There are no conflicts of interest.

References

[1] Perrie A,sabin C (Eds).Describing data.Medical statistics at galance.UK: Blackwell Science Ltd;2000 pp 16-9.

[2] Kuzma JW ,Bohnenblust SE (Eds). Summarizing data : Basic statistics for the science.London:Mayfield publishing company;2001 pp 44-54.

[3] Manikandan S. Measures of central tendency: median and mode. J Pharmacother 2011;2:214-5

[4] Sprent p. statistics in medical research.Swiss Med wkly 2003;133:522-9

[5] Bewick V , check L,Ball J.statistics review 10:Further non-parametric methods.crit care 2004;8:196-9

[6] Altman DG, Bland JM.parametric v non-parametric methods for data analysis.BMJ 2009;338:a3167

[7] Kaur SP .variables in research. Indian J Res Rep Med Sci 2013;4:36-8

[8] Ali Z,Bhaskar SB.Basic statistical tools in research and data analysis.indian J Anaesth 2016;60:662-9

[9] Magnello ME karl person and the origin of modern statistics : An elastician becomes statistician , Rutherford J , Vol.1 2005-2006 avaliable online at :http://Rutherford.org

[10] Rana R,Singhal R, Chi-square test and its application in hypothesis testing.J pract cradiovasc Sci 2015;1:69-71

[11] Indrayan A ,satyanarayan L.Simple biostatistics for MBBS,PG entrance and USMLE 4th edition Delhi: Academa publisher 2013

[12] Perrie A,sabin C (Eds).Describing data.Medical statistics at galance.UK: Blackwell Science Ltd;2000 pp 42-3

Authors

First Author – Mr. Nazir Ahmad Dar, Statistician (HBCR) Department of Radiation Oncology, Sheri-i-kashmir institute of Medical sciences (SKIMS) Soura Srinagar 190011, e-mail:johnnazirahmad@gmail.com, M.No:09906697883,07006421669

Second Author – Dar Nazir Ahmad, Statistician, HBCR, Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)

Third Author – Sofi Mushtaq Ahmad, Senior Resident , Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)

Fourth Author – Nayak Bilal Gul, Senior Resident , Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)

Fifth Author – Khan NA, Professor, Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)

Sixth Author – Lone MM, Professor and Head Department of Radiation Oncology,SKIMS, Srinagar Kashmir(India)

SeventhAuthor – KS Raina, DEO,HBCR, Department of Radiation Oncology, SKIMS, Soura Srinagar Kashmir (India)

Figure 1.Classification of data analysis

Table 3: Various Tests

Type of data	Tests
Categorical vs. categorical	Chi-square test and f-test
Categorical vs. continuous	T-test, Mann-whitney Wilcoxon test, ANOVA, Kruskal Wallis , Repeated Measure Friedman test
Continuous vs. continuous	Correlation -r Spearman -ρ
	Intra Class Correlation