THE DATA-ITS ADEQUACY AND ACCURACY IN PRESENTATION; PROBABILITY, DISPERSION AND THE CORRELATION

(Statistical Applications in Medical Research)

*Kumar Ashok

ABSTRACT

The data forms the main base at the very grass root level and is the BACK BONE for application of all the statistical methods, be it in the field of medical sciences, social sciences or any such allied field. The present paper, in this direction, attempts to HIGHLIGHT the main considerations and points to be taken care of in DATA ANALYSIS, to wards statistical Inference(s).

KEY WORDS : Data, Statistics, Primary and Secondary Data, Sampling, Probability, Null Hypothesis, Statistical Significance, Data Variability, Correlation, Multiple Correlation.

* Dr. Ashok Kumar, Associate Professor (Biostatistics), Department of Community Medicine, Universal College of Medical Sciences & Teaching Hosptial, Bhairahawa, Nepal

1. INTRODUCTION

The data forms the main base at the very grass root level and is the back-bone, for application of all the statistical methods, whileanalyzing various facts quantitatively and deriving inferences thereof, through appropriate tests. This recommends the need for special care to be taken thereof, while dealing with data, in any field of application. The present text in this direction attempts to highlight the main considerations and points to be taken care of in data analysis; along with introductory concepts of the Probability, Dispersion and the Correlation, in various applications.

For Correspondance
Dr. Ashok Kumar, Associate Professor (Biostatistics)
Department of Community Medicine
Universal College of Medical Sciences & Teaching Hosptial
Bhairahawa, Nepal
E-mail: aku@rediftmail.com

2. THE DATA

The Data,** plural of the word datum, is generally referred to, as a set of numerical observations corresponding to different values of a variable which are collected for a predetermined purpose. Apart, the various characteristic features it must possess, are that

- ❖ These are numerical statements of facts, collected systematically and with reasonable degree of accuracy, for a pre determined purpose.
- ❖ These are comparable and subject to analysis and interpretation.
- ❖ These always correspond to a group of units and not the individuals and are affected to a great extent by a multiplicity of causes.
- ❖ These are subject to computations and estimates and that the interpretations derived upon thereof, are true only on average and generalized basis.
- ❖ These are always liable to be misused and result to biased conclusion.
- ❖ These result to, fallacious conclusions if mis-interpreted i.e., not referred to, in right perspectives with full reference to context.

It may also be pointed out that when the word 'STATISTICS' is USED in plural sense, then it is also referred to as data, like Statistics(Data) of Height, Weight, Income, Export, Body Temperature, Pulse Rate, RBC/WBC counts.

**** STATISTICAL DATA-Their Care and Maintenance**

D.J.FINNEY, ISAS Bulletin No.1 A publication of Indian Society of Agricultural Statistics, New Delhi-110012.

2.1 The Data Presentation:

The next foremost problem after proper collection of data; whether primary data(data which are collected for the first time by the researcher himself) 'or' the secondary data (data which have already been collected by some other agency and the researcher has simply to find the source from which he can procure it); is its presentation, followedby analysis and interpretation; as the collected data in its original form, commonly called the RAW data, is generally so huge and complex that no meaningful information can be drawn from it, unless it is properly presented and summarized,systematically.

The main objective of data presentation thus, lies in that, the collected data must be so rearranged and presented in the form of CLASSIFIED data, such that its main characteristic features emerge away in respect of its similarities and dissimilarities and that there is no ambiguity of any kind in its presentation.

The broad categories of data presentation are through TABLES, DIAGRAMS, GRAPHS, and CHARTSand through AVERAGES as a main device of data summarization, since "an average value" being a SINGLE VALUE; is representative of the whole group or data, it corresponds to.

2.2. Data Analysis and Statistical Inference:

As is well known. that of the two common methods of studying any population or distribution i.e. the method of complete enumeration (census method) and that of SAMPLING; most of studies in empirical research are based on sampling and as such great care is to be taken, since deriving a wrong interpretation will give a totally wrong picture not for only the sample but for the whole POPULATION or UNIVERSE.

While the data analysis part concerns mainly with the selection of suitable and proper statistical methods as per need; the inference part relates with drawing correct results and interpretations thereof and is inductive in nature; for the simple reason that we adopt the approach of INDUCTION i.e. from PARTICULAR to GENERAL in studying a population through SAMPLING.

The Statistical inference thus is mainly concerned firstly to SPECIFY, the population, the sample in selected from; and then to estimate the population values (PARAMETERS) on the basis of sample values (STATISTICS) with reasonable degree of accuracy at specified probability level and this is the stage which necessitates the use of 'probability' in statistical applications and we may rightly say 'Statistics' as the "Science of Estimates and Probabilities" which otherwise is also known as the "Science of Averages" as most of the statistical inferences hold good only on average basis".

3. The Probability:

Probability is the base of all statistical inferences as all the results in the theory of statistics are derived upon, only at specified probability levels. The main characteristic features of probability are:

- ❖ Numerically, probability always lies between O and I (both inclusive).
- ❖ Probability is always related either with "Random Experiment" i.e. the experiment whose outcome cannot be specified before hand,(as is usually in deterministic experiment) 'or' "an" "event" (i.e. the ultimate outcome of a TRIAL like tossing an unbiased coin, throwing a dice etc.)

- ❖ An event with probability zero is certain to not happen and called an "IMPOSSIBLE event" while that with probability one is certain to happen and is called a "SURE event".
- ❖ Mathematically the probability of happening of an event E is defined as the ratio of favourable cases of that event to the total number of cases related with that event, which are all exhaustive, equally likely and mutually exclusive.

In symbols, $p(A) = m/n$
m = favourable cases
n = total number of cases

To quote with :
Probability of head in tossing an unbiased coin $= p(H) = 1/2$

Probability of an Ace in selecting a card from the pack of cards$=p(A)=4/52=1/13$

If, p and q denote respective probabilities of success and failure; then

$$\boxed{p+q=1}$$

3.1 Probability in Statistical Inference:

As already outlined 'statistical inference' may be rightly defined as the process of drawing inference(s) about PARAMETERS (Population Values) on the basis of corresponding STATISTICS (sample Values) ;through testing the Acceptance or Rejection of the Null Hypothesis(Ho),also called the Hypothesis of ZERO Difference, through suitable Test of Significance, at specified probability level.

In fact, while testing Ho a Test can commit Two TYPES of Error as under:

Ho	Test Result	
	Accept Ho	Reject Ho
True	✓	Type I Error
False	Type II Error	✓

Thus, Rejecting a True Hypothesis is the the Type I error, while Accepting a False Hypothesis is the type II error. It is also evident that both of these errors cannot be MINIMIZED at the same time. As such after FIXING the PROBABILITY LEVEL of type I error; type II error is MINIMIZED to all possible extent. This probability level at which Type I error is fixed; is called the LEVEL OF SIGNIFICANCE.

The probability levels commonly used are respectively 0.05 and 0.01 which in terms of percentage are respectively 5 percent and 1 percent levels of significance. A 5 % significance level indicates that maximum chance of committing type I error "or" in general, Test Result being wrong is 5 out of 100.It is also pertinent at this stage toknow about STATISTICAL SIGNIFICANCE. If, in testing significance of OBSERVED DIFFERNCE between specified values like two MEANS, two VARIANCES, two PROPORTIONS; by appropriate TEST of SIGNIFICANCE at a given probability level- (1) The Null Hypothesis Hocan be accepted then it refers to that the said difference is, NOT significant or INSIGNIFICANT and that it is only due to chance or error factor and NOT a real one (2) Hois rejected then the said difference is said to be SIGNIFICANT and that it is not only due to chance or error factor, but a REAL one.The significance of "observed difference" can also be tested on the basis of corresponding 'p' value of the Test Statistic.As per popular notions (i) If, $p>0.05$ the said difference is NOT significant (ii) If. $p<0.01$, The said difference is HIGHLY significant.

4. The Dispersion:

While, among various characteristics of a distribution 'or' a population, central tendency (a characteristics of first order) provides us that value which is representative of the whole distribution and where majority of its items have got a tendency to concentrate or cluster upon; the DISPERSION of the distribution (a characteristic of second order) provides us an idea about data variability i.e. the manner in which different items of the distribution are concentrated 'or' scattered 'or' distributed around the central value.

Among various measures of dispersion like Range, Quartile Deviation, Mean Deviation, Variance, Standard Deviation Coefficient of Variation; the most simple is the RANGE as the difference between the highest and lowest values; while the most popular and commonly used is the VARIANCE 'or' its positive square root, known as the STANDARD DEVIATION To quote with, given n values **x**1 ,**x**2………………**x**n of a variable x:

Range = **X**H-**X**L

XH= Highest Value
XL = Lowest Value

Used only in particular cases like Record of Daily Temperature, Annual Rainfall, Number of defective Items in Quality Control Charts.

Standard Deviation = $\sigma = \sqrt{\text{variance}}$

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

Treated as the best measure of dispersion since it possesses all the characteristics of a satisfactory measure of dispersion.

Coefficient of variation (C.V.)

$$= \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

Used (i) as a relative measure of dispersion as compared to mean (ii) to test the consistency of the given data (iii) to know the relative variation of the data related with different "units of measurements".

It is to be mentioned that all the above measures correspond to uni-variate distribution i.e. distributions with a single variable. However, if we have to study the dispersion in case of Bi-Variate distribution i.e. simultaneous variation in the values of two variable say X and Y at the sametime, the measure which is used is the covariance and defined as :

Covariance between X and Y = Cov (X,Y)

$$= \frac{\sum(X\text{-}X)\ .(Y\text{-}Y)}{n}$$

5. The Correlation:

The term correlation is used to study the relationship between two such variables in which by changing the values of one variable the values of other variable also change. To quote with, if X and Y are two variables such that by changing the values of X,Y also changes 'or' by changing the values of Y,X also changes; then there is correlation between them and they are said to be correlated; eg.(Area, Production);(Height, Weight);(Age of husband, Age of Wife);(Rain-fall, Agril Production); (Price, Supply);(Price Demand);(Pressure, Volume); (Body Temperature, Pulse Rate); (Body weight, B.M.I.).

Further, the correlation can be (a) Positive 'or' Negative (b) Perfect or Incomplete (c) Simple 'or' Multiple.

5. (a) (i) Positive Correlation

When the two variables X and Y move in the same direction i.e. when X is increasing Y is also increasing (x↑ y↑)' or' When X is decreasing Y is also decreasing (x↓y↓) the correlation between X and Y is said to be positive correlation e.g.(Area, Production);(Height, Weight):(Body Temperature, Pulse Rate);(Body weight,B.M.I.).

a) (ii) Negative Correlation :

When two variables X and Y move in opposite directions i.e. When X increases Y decreases (x↑ y↓)'or' when X decreases Y increases (x↓ y↑) the correlation between X and Y is said to be Negative Correlation, e.g. (Pressure, Volume); (Price, Demand) ; (Height ,B.M.I.).

5. (b) (i) Perfect Correlation:

If two variables X and Y are correlated in such a way that by changing X the Values of Y change by a fixed quantity 'or' in a fixed proportion, there exists a Perfect Correlation between X an Y.

A perfect correlation can either be a perfect positive correlation 'or' a perfect negative correlation, depending upon the type of correlation between X and Y. To quote with: Table 1

Positive Correlation (i)	Perfect Positive Correlation (ii)	Perfect Positive Correlation (iii)
X Y	X Y	X Y
0 8	0 10	0 10
1 10	1 12	1 12
2 17	2 14	3 16
3 20	3 16	6 22
4 25	4 18	10 30

(ii) If the correlation is NOT a perfect, it is said to be an Incomplete Correlation.

5.(c) (i) Simple Correlation :

If we study Correlation between Y and X it is said to be a simple Correlation.

(ii) Multiple Correlation:

If correlation is studied not only between Y and X, but between Y and a number of variables (**X**1 ,**X**2 ,…………**X**n) It is said to be a multiple Correlation. Such a correlation is studied when changes in the values of variable Y is not affected by changes in the values of only one variable X but by a number of variables **X**1,**X**2,………………**X**n.

Further, when in the case of a variable, the changes in which is governed not only by a single variable but by a number of variables, the correlation is studied between that variable and any one variable, nullifying the effects of all the remaining variables and is said to be a PARTIAL Correlation.

ZERO Correlation:

If corresponding to X and Y, there is no change in the values of Y by changing X 'or' in the values of X by changing Y, then there is zero correlation 'or' NO correlation between X and Y.

The type of CORRELATION can also be studied graphically by SCATTER DIAGRAMS by potting the given values of (X, Y) or graph. If the trend of points is from (i) downwards to upwards it records a positive correlation, (ii) upwards to downwards it records a negative correlation. while (iii) if points are Scattered haphazardly then it shows that there is no correlation between X and Y.

5 (d) Extent or Degree of Correlation:

The degree of correlation between two variables X and Y shows the extent of relationship between them, i.e. by changing any one of them to what extent the other is changing.

This is measured by Karl Person's Coefficient of Correlation between X and Y; which is defined as:

$$r_{XY} = r_{YX} = r = \frac{\text{Cov}(X,Y)}{\sigma X\ \sigma Y}$$

$$= \frac{\text{COV}\ (X,Y)}{\sqrt{\text{Var}\ X\text{Var}Y}}$$

$$= \frac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{\sqrt{\Sigma(x-\bar{x})^2,\ \Sigma(Y-\bar{Y})^2}}$$

Where, ($\bar{X}$, $\bar{Y}$) = Means of (X,Y) (σX, σY) = Standard Deviation of (X,Y)

n = number of pairs of observation of (X,Y)

(e) Properties of Correlation Coefficients:

i. This coefficient always is between -1 and +1 (both inclusive).
ii. This is independent of units of measurement of X and Y.
iii. This is independent of change of origin and scale.

f) INTERPRETATION of the given value of 'r':

i. r= -1 This indicates a perfect negative correlation between X and Y; indicating that X and Y move in opposite directions and that corresponding to a change in X,Y changes by a fixed quantity, or in a fixed proportion.

ii. r lies between -1 and 0 (both exclusive):

Say, r = -0.32

This indicates a negative Correlation between X and Y, meaning thereby when X increases Y decreases 'or' when X decreases Y increases.

iii. r = 0,this indicates a zero correlation between X and Y.

iv. r lies between 0 and +1. (both exclusive):

Say, r = 0.56

This indicates a positive correlation between X and Y, meaning thereby that when X increases Y also increases 'or' when X decreases Y also decreases.

v. r=+1, This indicates a perfect positive Correlation between X and Y; meaning thereby that X and Y both move in the same direction and that corresponding to change in X,Y changes by a fixed quantity 'or' in a fixed proportion.

It may be pointed out at this stage that above Karl Person's correlation Coefficient 'r'; is termed as the correlation coefficient of zero order 'or' also as the LINEAR correlation Coefficient between X and Y.

(a) In case of data dealing with qualitative characteristics like Intelligence, Beauty, Obesity, Deafness, Blindness, etc. The correlation co-efficient is calculated by SPEARMAN'S rank correlation Coefficient, $\mathbf{r}_s$ defined as:

$$\mathbf{r}_s = \frac{1-6\Sigma d^2}{n(n^2-1)}$$,in case of 'untied' ranks

And,

$$\mathbf{r}_s = \frac{1-6\ [\Sigma d^2 + \frac{\Sigma(p^3-p)}{12}\ ]}{n(n^2-1)}$$ in case of 'TIED' ranks

Where, n= number of pairs of ranks
d = difference of two ranks in a pair
p = number of tied or equal ranks.

6. Multiple Correlation:

When the changes in the values of Y are not affected by changes in the values of only one variable X; but by a number of variables X_1,X_2………………….X_n, the correlation between Y and (X_1,X_2…………X_n) is called the Multiple Correlation., and the Coefficient used to measure this is called the Coefficient of Multiple Correlation denoted by **R**Y.X1,X2,X3,………….Xn.

REFERENCES

- *CROXTON,Fredrik E. and COWDEN, Dudley J.;Applied General Statistics:Prentice Hall of India, Pvt.Ltd.*
- *Grewal P.C. Numerical Methods of Statistical Analysis; Sterling Publications,Pvt.Ltd.*
- *GuptaS.P.Practical Statistics: S. Chand and Co. New Delhi.*
- *Kapur J.N. and SaxenaH.C, Mathematical Statistics ;S. Chand and Compay New Delhi.*
- *Kumar Ashok An Introduction to Sampling; Student's Friends Publications, Allahabad.*
- *ROTHMAN, Kenneth J. and Greenland, sander: Modern Epidemiology: Lippincott- Raven Publishers.*
- *SinhaH.C. and PillaiS.K.,Statistical Methods for Biological Workers; Ram Prasad and Sons Publications, Agra.*

Madhu Pokharel
Cell.: 9847056868

S.K. Offset Press
Kaligandaki Offset Press
Milan Chowk, Butwal-11, Rupandehi, Nepal
Remember us for all kind of printing jobs.