# Interpretation of Data and Statistical Fallacies

Prof. Usha Jogi
Lecturer,
Department of Business Administration,
Shri G. K. and C. K. Bosamiya Arts and Commerce College,
Jetpur, Gujarat (India)
Email: usha.gkck@gmail.com

*Abstract: Although statistics is essential to almost all sciences-social, physical, and natural and is very widely used in almost all spheres of huminity. It is not without limitations. The most significant limitation of statistics is the misuse of it. The use of statistic by the experts who are well experienced and skilled can analyze and interpreted the statistical data in correct way. Wrong interpretation of the data results in statistical fallacies which may arise in the collection, presentation, analysis, and interpretation of data. This paper throws light on the factors which give rise to misinterpretation of data and statistical fallacies. The study of these factors leads to minimize the statistical fallacies. Utmost care and precaution should be taken for the interpretation of statistical data. "Statistics should not be used as a blind man uses a lamp-post for support instead of illumination." Statistical interpretation depends not only on statistical ideas but also on ordinary clear thinking. Clear thinking is not only essential in interpreting statistics but is sufficient in the absence of specific statistical knowledge. Statistics neither proves nor disproves anything. It is merely a tool which, if rightly used may prove extremely useful but if misused by inexperienced, unskilled, and dishonest statisticians might lead to very fallacious conclusions "Statistics are like clay of which you can make a god or a devil as you please." The use of statistics by the experts who are well experienced and skilled in the analysis and interpretation of statistical data reduce the chances of mass popularity of this important science. "Science of statistic is the useful servant but only of great value to those who understand its proper use."*

## I. MEANING AND NEED

Interpretation of statistical data involves:

I. Summarization and analysis of data.
II. Prediction or inference.

Interpretation of data means to draw inference on the basis of colleted data after proper statistical analysis. The interpretation can be true if the collection, presentation analysis is properly done. Sometime wrong interpretation of the data may arise which leads to statistical fallacies. It may happen due to presentation, collection, analysis of the data wrongly.

## II. IMPORTANT FACTORS FOR MIS-INTERPRETATION OF DATA OR STATISTICAL FALLACIES

The factors giving rise to misinterpretation of the data or statistical fallacies are;

I. Bias
II. Inconsistency in definition
III. False generalization
IV. Comparison between unlike components
V. Wrong interpretation of statistical measures
VI. Wrong interpretation of index components of time series
VII. Technical errors
VIII. Errors in selection of data distrust of statistic

### A. BIAS

Bias is very common in statistical investigation and it leads to wrong conclusions. In any statistical investigation biased errors may be due to,
- Bias on the part of investigator whose personal prejudices are affecting the results of the enquiry.
- Bias in the measuring instrument in recording the observations.
- Bias due to wrong collection of the data and wrong formula used for the analysis of the data.

- Bias in the technique of approximation: while rounding off each individual value is either approximated to next highest or lowest number so that all the errors move in the same direction. e.g. if the figures are 405 and 495 they are recorded as(rounded off) to 400 and 500 respectively.

### B. INCONSISTENCY IN DEFINITION

In order the comparisons of phenomenon to be valid if it is necessary that the definitions of the object should remain constant throughout. If there is need of change in definition a footnote should be given so as to make valid comparisons.

**For Example,**

To construct cost of living index number the proper unit for enumeration is household. It should be clear whether household consist of a family blood relations or people taking food in common kitchen or all the persons living in the house or the persons listed in the ration card only. The concept of household for enquiry is to be decided in advance.

So, in order to have meaningful comparison of the results the same pattern of classification should adopt throughout the analysis and also for further enquiries in the same subject.

### C. FALSE GENERALIZATION

Many times we commit errors in drawing conclusions or making generalizations on the basis of data which is not representative of the population.

**For Instance,**

a. "The number of car accidents committed in a city in a particular year by women drivers is 10 while those committed by man drivers is 40. Hence, women are safe drivers." This statement is obviously wrong because nothing is given about total number of men and women drivers in the city in the given year.

b. "It has found that 25% of the surgical operations by surgeon X are successful." This conclusion is not true. Out of 4 surgical operations if 3 operations on particular day are unsuccessful the 4th must be success is not true. It may be happen that 4th operation is also unsuccessful or it may happen that all 4 operations are successful. The above statement means that if the numbers of operations are larger we should expect 25% of the operations to be successful.

c. Incomplete data causes to fallacious of conclusion. Let us consider the score of these two students A and B.

|           | Test-1 | Test-2 | Test-3 | Average Score |
|-----------|--------|--------|--------|---------------|
|           |        |        |        |               |
| A's Score | 50     | 60%    | 70%    | **60%**       |
| B's Score | 70%    | 60%    | 50%    | **60%**       |

If we consider average score which is 60% in each case we will conclude that the level of intelligence of the two students is same at the end of the year. But this is wrong conclusion. If we observe carefully A has improved constantly while B has deteriorated consistently.

### D. IMPROPER COMPARISONS

The comparisons will be meaningful only if the things are really a likeotherwise the conclusions drawn will be wrong.

For instance to make valid comparisons of index number it is essential that the commodities are of same quality or grade. Both the index numbers should be calculated by the same formula. If one index number is calculated by Laspeyre's formula then both index numbers are not comparable.

### E. WRONG INTERPRETATION OF STATISTICAL MEASURES

The different type of measures like averages, dispersion, skewness etc. calculated from the collected data are many times misused to present the information for personal motives which misguide the public.

#### a. *Wrong interpretation of averages*

1. Let us consider the industry the average salary of the workers is say Rs.5, 500 per week. If the salaries of few bosses are added the average wage per worker comes out to Rs. 7,500. Suppose if we say that average salary of workers is Rs.7, 500

per week. It shows very good impression and it shows that the living standard of workers is good but the real picture is entirely different. So, for extreme observations the arithmetic mean gives wrong interpretation which leads to misleading conclusions.

2. Vishal gets pocket money allowance of Rs.12 per day. He felt that this was less amount and he enquired about the pocket money of his friends and he get the following data including Vishal's allowance in Rs. 12, 18, 10, 5, 25, 20, 22, 15, 10, 10, 15, 13, 20, 18, 10, 15, 10, 18, 15, 12, 15, 10, 15, 10, 12, 18, 20, 5, 8, 20.

He presented this data to his father and requested for increase in his allowance. His father was statistician replied that Vishal's amount of allowance was more than average amount. Comment on father's statement.

### SOLUTION

The frequently distribution of the daily allowance is as follows.

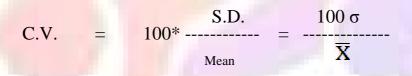| Daily allowance (in Rs.) | Tally Marks | Frequency | fx |
|---|---|---|---|
| 5 | II | 2 | 10 |
| 8 | I | 1 | 8 |
| 10 | IIIIIII | 7 | 70 |
| 12 | III | 3 | 36 |
| 13 | I | 1 | 13 |
| 15 | IIIIII | 6 | 90 |
| 18 | IIII | 4 | 72 |
| 20 | IIII | 4 | 80 |
| 22 | I | 1 | 22 |
| 25 | I | 1 | 25 |
| **Total** | | **N=30** | **∑fx=426** |

The average daily allowances are given by:

$$\overline{X} = \frac{\sum Fx}{N} = \frac{426}{30} = 14.20$$

Since maximum frequency is Rs.7 for daily allowances Rs.10 which gives mode. Hence, modal value is Rs.10 Vishal's allowances are Rs.12 which is less than arithmetic mean X. But his father pointed out that his amount was more than average amount by referring to mode at the average which is Rs.10.

### b.   *Wrong interpretation of measures of dispersion and skewness*

Let us have two series of height in meters and weight in kgs for 10 persons the range for series of height is less than that series of weight. It may be concluded that the series of weight is more variable than the series of height. Without calculating S.D. of both the series however this is wrong. Calculate coefficient of variation (C.V.) for both the series.

$$C.V. = 100 * \frac{S.D.}{Mean} = \frac{100\,\sigma}{\overline{X}}$$

Which is pure number independent of units of measurement, distribution with smaller C.V. is said to be more homogenous or less variable than the other and the series with greater C.V. is said to be more variable than the others.

After training the average daily wage in a factory had increased from Rs.8 to Rs.12 and S.D. had increased from Rs.2 to Rs.2.50. After training the wage has become higher and more uniform. Comment on the statement.

**Solution:**

It is given that after training the average hourly wages of workers have increased from Rs.8 to Rs.12. this implies that the total wages received per hour by all the workers together have increased but we cannot conclude that the wage of each individual has increased. Let us calculate coefficient of variation (C.V.) of the wages of workers to know about uniformity.

**C.V. of wages before training -**

$$C.V. = \frac{100*S.D.}{Mean} = \frac{100*2}{8} = 25$$

**C.V. of wages after training –**

$$C.V. = \frac{100*S.D.}{Mean} = \frac{100*2.5}{8} = 20.93$$

Since C.V. wages after training is less than C.V. of wages before training. It may be concluded that the wages before more uniform after the training.

### c. Wrong interpretation of correlation, co-efficient, and regression coefficients

Sometimes correlation coefficient may be wrongly used to study the relationship between two series which are really independent of each other.

E.g. correlation may be studied between.
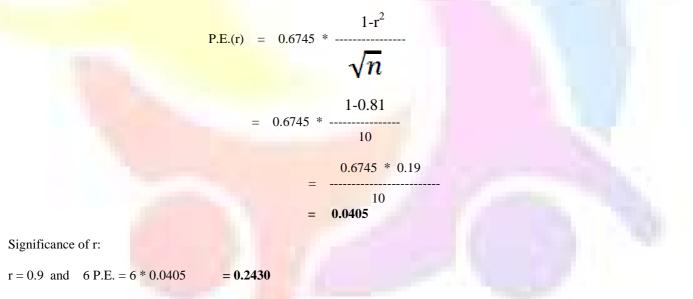1. The series of height and income of individuals over a period of time.
2. The series of intelligence and size of shoe.

If we observe higher degree of correlation coefficient between the size of the shoe and the intelligence which means that larger the shoe size more intelligent the person is. The conclusion is absurd or wrong because two series are really independent such a correlation is called chance correlation or nonsense correlation.

**Example-1**
Suppose that the correlation coefficient between the two series of marks of 10 students in mathematics and statistics is

A preferable error of correlation coefficient is:

$$P.E.(r) = 0.6745 * \frac{1-r^2}{\sqrt{n}}$$

$$= 0.6745 * \frac{1-0.81}{10}$$

$$= \frac{0.6745 * 0.19}{10}$$

$$= \mathbf{0.0405}$$

Significance of r:

r = 0.9 and    6 P.E. = 6 * 0.0405        **= 0.2430**

Since r is much greater than 6 P.E. (r) if hence r is highly significant.

**Remarks:** r id significant we can say that higher the mark of student in mathematics higher is the score in the statistics and vice-versa. However it does not mean that all the students who are good in mathematics are also good in statistics and vice-versa.

From the above example it is clear that "The coefficient of correlation expresses the relationship between two series and not between the individual items of the series.

**Example- 2**
"If the correlation coefficient between two variables is zero then the variables are independent "comment.

**Solution**

The statement is wrong $r_{xy} = 0$ does not mean that x and y are independent. It simply means that there is absence of linear relationship between x and y. the variables may be related in same other form quadratic, logarithmic, exponential or trignometric form.

### Example-3
A correlation between the variables has a value r = 0.6 and correlation between other two variables is 0.3. Does it mean that the first correlation is twice stronger than the second? Discuss the above closeness of relationship.

#### Solution
It is wrong to say that correlation in the first case is twice as strong as correlation in the second case.

$r_1 = 0.6$ coefficient of determination $(r_1)^2 = 0.36$
$r_2 = 0.3$ coefficient of determination $(r_2)^2 = 0.09$

Thus the correlation in the first case ($r_1 = 0.6$) is 4 time a strong as the correlation in second case $r_2 = 0.3$ the closeness of relationship between two variables is not proportional.

### Example-4
Give the comments on the following statement.
$r_{xy} = 0.9$, $b_{yx} = 2.04$, $b_{xy} = -3.2$

#### Solution
Regression coefficient cannot have different signs. Here, $b_{yx}$ is positive and $b_{xy}$ is negative which is not positive. Again $r_{xy}$, $b_{yx}$, $b_{xy}$ must have the same sign.

## F. WRONG INTERPRETATION OF INDEX COMPONENTS OF TIME SERIES

Sometime index numbers are wrongly interpreted.

**For example,**
If the consumer price index is higher for Delhi than that of Mumbai does it mean that Delhi is more expensive than Mumbai.

#### Solution
If the consumer price index is higher for Delhi than that of Mumbai generally it can be said that Delhi is more expensive than Mumbai but this is not absolutely correct because…

1. Due to the variations in the conclusion pattern even with the same prices the living at one place may be expensive than the living at the other place.
2. The sources of obtaining the quotations of prices and quantities consumed of variance things may be different. This may make all differences.

The various components of time series may be analyzed and interpreted incorrectly.

**For Example,**
We may fit linear trend equation to the given set of data by the principle of least squares and use it to and corresponding trend values and future estimates these value and estimates will be valid if the data really exhibits a linear trend.

The linear trend is commonly used but it is rarely observed in economic and business data.

## G. TECHNICAL ERRORS

Some of the technical errors in statistical work which result in wrong conclusions are :
1. The errors due to improper choice of sampling technique for the selection of sample e.g. we may use random sampling where specified sampling may be appropriate.
2. Size of the sample is inadequate.
3. Error due to improper choice of the statistical formula. e.g. we may use media where made is more appropriate or the use of arithmetic mean where geometric mean is appropriate.
4. The errors due to bias in the estimation method. This is due to improper choice of estimation technique.
5. The error committed in the use of ratio and percentage. e.g. if the price of commodity increases from Rs. 25 to Rs.75 during a particular period the one generally concludes that there is 300 % Increase in price during the given period. This is wrong the actual increase is 200%

Increase in Rs is (Rs.75 – Rs.25) = Rs.50

Hence increase in Rs.100 $= \dfrac{50}{25} * 100 = 200$

### H. EFFECT OF WRONG INTERPRETATION OF DATA - DISTRUST OF STATISTIC

The wrong interpretation of data promotes distrust of statistics. The improper use of statistics tools by irresponsible, dishonest people with an improper statistical bend of mind has led to the public distrust in statistics due to which public loses it belief faith and confidence in the science of statistics and starts blaming it. The reason behind blaming may be;

a. Figures are innocent and believable but they do not have the lable of quality on their face.

b. Arguments are put forward to find certain results which are not true by making use of inaccurate figures or by using incomplete data and hence, distorting the truth.

### III. CONCLUSION

Hence, if statistics and its tools are misused. The fault does not lie with the science of statistics but it is the people who misuse it, are to be blamed. Utmost care and precaution should be taken for the interpretation of statistical data. "Statistical should not be used as a blind man uses a lamp-post for support instead of illumination."