International Journal of Innovative Research in Science and Engineering
Volume 7, Issue 12, 2021
www.ijirse.com
ISSN: 2454-9665

Fundamental Statistical theory interspersed with thoughts and Suggestions

Dr. Prabhakar Singh

Associate Professor, Department of Statistics, Harish Chandra P G College, Varanasi, U.P, India

ABSTRACT

It is always difficult to write papers and give talks that contain statistical information because the presentation methods used for analytic results have an impact on the reader's or audience's comprehension. Furthermore, rather than being clearly explained, fundamental statistics are frequently misunderstood or obscured by the author. Because one can never be certain of a reader's or audiences statistical sophistication, it is easy for the author to write material that muddles rather than clarifies the results. The reader or audience is then left wondering whether the author's work consists of "lies, damn lies, or statistics." This paper contains a brief discussion of fundamental statistical theory interspersed with some thoughts and suggestions on the use and presentation of statistics statistical findings After all, if the intended audience misunderstands or misinterprets the information presented, the paper or presentation is rendered useless.

KEYWORDS: Statistics, graphs, tables, presentations

Corresponding Author;- prabhakarsingh06725@gmail.com

INTRODUCTION

Why might the use of statistics in a paper or presentation be problematic?

1. An author may wish to examine a sample from a specific population.

Some conclude that the results characterise the population of the sampling frame.

• Caution is advised because the sampling design has a significant impact on how generalizable the results are.

2. An author may be attempting to demonstrate the nature and strength of a relationship between a dependent variable and a set of independent variables.

• Caution should be used when communicating this because the analysis does not prove causality, as many papers claim.

3. An author may need to reduce a very large set of questions from a survey to a few simple, easy-to-understand factors;. • Caution is advised because the technique used in the reduction process can significantly affect the interpretation of the factors.

4. An author could be demonstrating an accurate and sophisticated analysis technique.

• Caution is advised because the impact is greatly diminished if the audience is unable to follow the statistical

ISSN: 2454-9665

discussion.

With all of this in mind, there are several annoying statistical errors that frequently appear in manuscripts and presentations:

Misunderstanding the concept of significance:

In essence, there is no distinction between "significant" and "not significant." All of the findings are statistically significant, with p-values ranging from 0 to 1. To give a p-value meaning, one must consider the costs of being incorrect. Consider the following straightforward medical example:

"The likelihood of surviving for a year after the surgery will be about 75%," a surgeon tells a patient considering surgery. What should the patient make a decision on? If the patient had late-stage terminal cancer, he might welcome and accept the operation. If the patient does not have a life-threatening condition, he has a reasonable reason to reject the operation as too risky.

Before making a decision, one must consider the costs and benefits of the decision appropriate level of importance far too often; people mistakenly believe that a p-value of 0.05 or 0.01 indicates that a statistical result is significant. Although these are traditional values, they are not magical! An author must always be prepared to answer the question, "What are the costs of being wrong and the benefits of being correct?" before deciding whether a value is significant or not. Exploratory research can accept a much higher p-value than large-scale research. When the null hypothesis is rejected, the p-value is simply the probability of making an error.

Confusing strong tendency with proof:

In general, statistics is inductive reasoning for the science of “good enough.” It is about reasonable conclusions being drawn from the observation of strong tendencies. But this is not proof. One looks at the p-value to determine if the likelihood of an error occurring is sufficiently small to reject the null hypothesis, and asks, “Is that good enough?” Statistics simply provides the odds that something may or may not be true,when the analytic processes meet all of the underlying assumptions. Consider the case of a highly visible US political figure involved in beef futures trading a few years ago:

This person, a novice futures trader, claimed to have increased the value of her beef futures portfolio from an initial sum of a $1,000 to over $100,000 dollars in less than a year. The obvious questions was “Did she cheat?” If one samples the historical records of novice futures traders, the likelihood of this happening is on the order of one in severalhundred million.

This does not prove she cheated; it only provides the probability that random variability alone could account for this increase assuming that the historical information is still relevant. The reader must interpret the information and come to his or her own conclusion. In fact, it might beexpected that two readers might come to different conclusions after reading the same journal article. It all depends upon the readers’ willingness to accept the benefits and costs associated with the possible conclusions and their outcomes.

Using incorrect statistical measures:

A commonly seen problem in papers and presentations isthe use of arithmetic means (averages) with ordinal

ISSN: 2454-9665

variables such as dislike-like scales.

Consider a simple three point scale (1 = dislike, 2 = neutral, and 3 = like). Although the sum of two "dislikes" equals 2, two "dislikes" do not equal a "neutral". But the previous must be true touse an arithmetic mean, which is based on addition. The median is the appropriate statistical measure in this case and thus nonparametric methods should be used to analyze this data. The basic types of variables and appropriate methods for measuring their centers are summarized as follows:

- Nominal scaled variables are categories without inherent order. A simple example here is"gender" as changing the order from "male to female" to "female to male" does not affectanything. For variables like this, only modes are appropriate.
- Ordinal scaled variables are categories with inherent order. For this type of variable, bothmodes and medians are appropriate. One of these variables commonly seen on surveys is "highest level of education attained" In this case, reshuffling the order of the categories tobe listed as: Masters; High school diploma; Doctorate; Some college; Bachelors; etc. does not make sense and now only modes can be used.
- Interval scaled variables are quantities without an inherent zero. The most common of these is "temperature." Zero in Celsius is a lot warmer than zero in Fahrenheit, which is much warmer than zero in Kelvin. There is no universally agreed upon zero. Modes, medians, andarithmetic means are appropriate in this situation.
- Ratio scaled variables are quantities with an inherent zero. For example, it is generally clear to everyone that when "the number of people in a room" is zero, the room is empty of people.These variables can support modes, medians, arithmetic means, and geometric means.

Misunderstanding hypothesis test errors:

There are two types of error possible for hypothesistests which are not interchangeable and which do not sum to a value of one:

- Type I error occurs when a true null hypothesis is rejected. The maximum acceptable probability of this error occurring is called the significance level, denoted by α. Once a test statistic has been calculated, the appropriate probability distribution is used to convert the test statistic into the observed probability of a Type I error occurring. This is the p-value that should be included in research papers and presentations.
- Type II error occurs when a false null hypothesis is accepted. The probability of such an error is denoted by β. The "power" of a test is the probability of such an error not occurring$(1-\beta)$. β depends upon three interacting factors: α, the sample size, and the size of the difference that you wish to detect relative to the population variability. The smaller this difference is, the harder it is to detect and thus the larger β becomes.

It is important to note that an inverse relationship exists between Type I and Type II error and that the only way to reduce both simultaneously is to increase the sample size. If a researcher rejects his null hypothesis, a Type I error may have occurred and the p-value applies. If the null hypothesis is accepted, a Type II error may have occurred and then β applies. As the calculationof β can be extremely difficult, it is not often taught in a basic statistics course. This may account for why it is very frequently ignored in research papers and presentations.

ISSN: 2454-9665

Unreported βgreater than 0.50 are not uncommon in research studies! (Murphy & Myors, 1998)

Using inappropriate ranking:

Often, researchers have respondents rank order their preferences for a rather long list of items. Then, the researcher compares these rankings as if they are in fixed orders. Yet, subjects rarely have the capacity to rank order more than a few items with any degree of accuracy. In psychology, the upper bound on a subject's absolute judgment about a stimulus is known as "channel capacity." In general, this is considered to be about 6.5 categorieson average (Miller, 1956). With this in mind, the number of items to be rank ordered should be limited to no more than 4 or 5. When the number of items to be ranked is large, the lower rankeditems become stochastic in nature. That is, if one were to do a test-retest of the rankings, the firstfew usually remain rather stable but the latter become mixed up. Treating these lower ordered ranks as meaningful is quite improper.

Inadequately describing data:

When an author describes data from a sample or infers statistical properties for the population from which the sample was drawn, he most oftenindicates the sample size, the mean and the standard deviation for the variable under consideration. While this is typical for many published papers, it is quite inadequate. Considerthe four samples summarized in this manner in Table 1. As these are quite large samples, it is not uncommon for authors to automatically assume that the observations are normally distributed. Furthermore, the summary measures suggest that the samples are similar; however,the distributions of these samples are most definitely not. Visually examine the same samples inFigure 1. In the first distribution, the mean is the most likely observation. In the second distribution, the mean is as equally likely to be observed as any other value. In the third distribution, the mean is the least likely observation. In the fourth distribution, the mean is substantially to the right of the mode and median. An author needs to know how his data is distributed and should never assume that distributions have similar shapes!

INSERT FIGURE 1 HERE

Two statistical measures can help describe these anomalies: skewness and kurtosis. Skewness measures symmetry, with negative values indicating a long left tail and positive values indicatinglong right tail. Sample 4 should have a positive value while the other three samples should have a skewness measure of zero as they are symmetric. Because of the skewness of sample 4, reporting both the mean and the median can help the reader understand the degree of skew of thedistribution if he is unfamiliar with the skewness measure. Kurtosis, the measure of the height of a distribution relative to its range (i.e. "peakedness"), is especially useful in distinguishing the first three samples as they are all symmetric. A kurtosis of zero matches a normal distribution, while a negative value indicates a flatter and wider shape than the normal distribution and a positive value indicates a more pointed and narrow shape than the normal distribution. Table 2 provides a more adequate summary description of the four samples than is usually provided in papers and presentations.

Equating "Unlikely" with "Impossible": Authors often report a p-value equal to 0.000 as thisis what was provided by their software. A probability of zero means it never happens, which is commonly phrased as "impossible." Statistical routines such as SPSS or SASS are programmed to report numbers less than 0.0005 as

ISSN: 2454-9665

0.000. But this is a computer phenomenon, not a statistical property. The computer is indicating that making a mistake by rejecting the null hypothesis is "somewhat unlikely." The difference between "impossible" and "somewhat unlikely" is not trivial. This value should always be reported as "less than 0.0005."

Confusing standard deviation with standard error:

When describing the possible means of apopulation distribution, the standard error should be used instead of the standard deviation. The Central Limit Theorem provides that distribution of all possible sample means is normally distributed when the sample size is large enough. The mean and its standard error are then adequate to describe this distribution. The standard error is easily calculated as the standard deviation of the sample divided by the square root of the sample size. Thus, larger sample sizes result in smaller standard errors, not smaller standard deviations. For the four samples in Figure 1, the estimated population means and corresponding standard errors are all 5.000 and 0.003 respectively It is important to note that the sampling distribution of a mean does not representthe population distribution, but rather the distribution of all possible sample means.

Poorly annotating tabular material:

Tables are often simply plunked down in the text or shown on a power point slide with little explanation. The author seems to assume that the audience is be able to knowledgably interpret the information without any guidelines as to whatthe values mean. An early cell in the table should be footnoted to explain how the cell is to be read. A very simplified example is presented in Table 3. This minor addition allows the reader to quickly verify his understanding of the table. Explaining how to read a data table is essential as the complexity of the table increases. The more complex the table, the greater the need is to provide information which clarifies how the table is to be read.

Using perplexing variable names:

Authors often use variable names that were constructed for computer analysis routines. Variable names, especially in older software packages, are often restricted to 7 or 8 alphanumeric characters. As such, these names may have little intrinsic meaning as they are simple codes used by researchers to identify the variables in their own minds. In manuscripts or presentations, variables must be identified in a way that is understood by the reader or audience. These can be quite different than the variable names reported in the initial computer outputs. It is really a simple matter to edit these names prior to submission by using a word processor.

Inappropriately using pie charts:

This particular graph is hardly ever informative and often abused. Pie charts are rarely interpreted correctly as the values are encoded as relative angles, yet our visual system is designed for decoding relative lengths. Using a bar chart or even a data table is easier understood than a pie chart. The fact that the values total to 100% can be readily stated with an annotation. After all, a graph is about the communication of complex information. Done right, it is a powerful tool. Done wrong, it confuses the reader and dilutes the message.

CONCLUSION

Authors often simply claim variables are significant or not significant without providing the relative p-values.

ISSN: 2454-9665

NS (non significant) is often substituted for the p-value when it is above 0.05. For example, in an article published recently in a much respected journal, variables were omitted if they had a p-value of 0.051 or higher, but the variables with p-values of 0.050 were included. No explanation was given as to why this cutoffwas used. The explanation Vis a footnote that read "variables with a p-value above 0.05 were outside the desired significance level," did not clarify the issue. This gives the appearance that difference of one-thousandth in the probability of a Type I error is the distinction between a variable being important and a variable being totally useless. But, this is not necessarily true.What this actually shows is a lack of understanding by the author about what statistics reallydemonstrate!

REFERENCES

1. Kaplan, A. (1964). The Conduct of Inquiry San Francisco, California: Chandler Publishing Co.
2. Miller, G.A. (1956). The Magical Number Seven, plus or minus two: Some limits on ourcapacity for processing information. The Psychological Review, 63(2), 81-97.
3. Murphy, K.R. & Myors, B. (1998) Statistical Power Analysis: A simple and general model for traditional and modern hypothesis tests. Mahwah, New Jersey: Lawrence Erlbaum Associates.
4. Krishna Moorthy. R, Investment and Tax Planning for the NRIs - (I), Fortune India (Monthly), December, 1989, Vol.111, No.4, p.24 8 25.
5. Bishwajit Bhattacharyya, Direct Tax Reforms and NRIs, Financial Express, October 5, 1989, p.4 8 12.
6. Ministry of external affairs (2017). Population of Overseas Indians (PDF). Ministry of External Affairs (India). 31 December 2017. Retrieved 28 Jan 2018.
7. NRI Investment in India: Changes since December 1985, (1988), IIC Publications, New Delhi.
8. Piparaiya. R.K, Expatriate Funds: Rhetoric and Realities-I, Indian Express,
9. February 14, 1983.
10. Rao. K.V, Why Incentives fails to lure NRIs, Financial Express, January 9, 1989, p.3.
11. Piparaiya. R.K, Expatriate Funds: Rhetoric and Realities-I, Indian Express, February 14, 1983.
12. Dhawan. O.P, Non-resident Indian Investment Facilities,(1985), 4th Ed., Standard Booksellers, Dariba Kalan, Delhi.
13. Gupta, Poonam (2005-12-01). Macroeconomic Determinants of Remittances: Evidence from India. International Monetary Fund. ISBN 9781451862430. Retrieved 2009-03-14.
14. Clear Tax (2018). NRI Status and NRI Taxation, RNOR - Resident but NotOrdinarily Resident. Clear tax.in. Retrieved 2018-02-15.
15. UN (2015), International migrant stock 2015: graphs: Twenty countries or areas of origin with the largest Diasporas populations (millions). United Nations Population Division.
16. Banerji, Kalvan (1988). Indian Banks Overseas - An Untapped potential, The Journal of the Indian institute of Bankers, Diamond Jubilee Special Issue, 1928- 1988,59 (1), 57-62.