PERSPECTIVE

Vol. 4, 2018

ISSN 2394–3750 EISSN 2394–3769

Science & Technology

Establishing chi square as a non-parametric test in descriptive statistics

Adeniran Adetayo Olaniyi

Department of Transport Management Technology, Federal University of Technology, Akure (FUTA), Ondo State, Nigeria
Phone: 07036196773
Email: 4tynil@gmail.com

Article History
Received: 24 May 2018
Accepted: 05 July2018
Published: July 2018

Citation
Adeniran Adetayo Olaniyi. Establishing chi square as a non-parametric test in descriptive statistics. *Science & Technology,* 2018, 4, 106-113

ABSTRACT

This study establishes Chi-Square as a non-parametric test in descriptive statistics and differentiates between descriptive statistic and inferential statistics. The essence of statistics is highly connected to the test of raw facts which is also referred to as a test of significance, and also connected to the importance of getting a representative sample from the population (estimation). Statistics involves two classifications which are descriptive statistics and inferential statistics. Generally, statistical data are described in graphics and pictures; they are presented using measures of central tendency and dispersion. The fact that statistical data is being described in pictures, graphs, central tendency, and dispersion does not mean it is a descriptive statistics. Descriptive statistics have to do with the measures of association which indicate whether the variables are related or not; and to determine if the sample is well represented from the population. The measure of association is commonly carried out by Spearman rank correlation and Chi Square. The study justifies the difference between descriptive and inferential statistics; it separate descriptive statistics from a mere data description; and reveals that the population has to do with sampling which is highly connected with parameters. Furthermore, probabilistic sampling is connected with parametric test, and non-probability sampling is connected with non-parametric test. The study further establishes that Chi Square is a non-parametric test used in descriptive statistics if and only if the data type is nominal;

also, it is a parametric test used in inferential statistics if the data type is non-nominal and the population is small such that there is no need for sampling technique.

Keywords: Chi Square, Non-parametric test, Descriptive statistics

Acknowledgement and dedication
I give kudos to Professor Samson O. FADARE, whom directly and indirectly impacted me with knowledge of research during my Master programme. Under his supervision and tutelage, I was able to critically differentiate between descriptive statistics and inferential statistics. He is a man of high mental magnitude and has risen above the tyranny of the flesh. He is dependable, reliable, and I therefore dedicate this article to him.

The study is structured as detailed below:

- Background to the study
- Problem statement
- Statistics
- Descriptive statistics and inferential statistics
- Descriptive detailing of descriptive statistics and inferential statistics
- Sampling
- Parametric and non parametric tests
- Hypothesis testing
- Chi Square test

1. INTRODUCTION

1.1. Background to the Study

Generally, statistical data are described in graphics and pictures; they are presented by using measures of central tendency and dispersion. The fact that statistical data is being described in pictures, graphs, central tendency, and dispersion does not mean it is a descriptive statistics. Descriptive statistics have to do with the measures of association which indicate whether the variables are related or not; and to determine if the sample is well represented from the population.

The essence of statistics is highly connected to the test of raw facts which is also referred to as a test of significance, and also connected to the importance of getting a representative sample from the population. There are different sampling strategies, including random sampling. A true random sample (probability sampling) means that everyone in the target population has an equal chance of being selected for the sample. Another key component of proper sampling is the size of the sample. Obviously, the larger the sample size, the better, but there are tradeoffs in time and money when it comes to obtaining a sample which comes in the form of non-probability sampling. Sampling is highly connected with hypothesis.

A hypothesis is a predictive statement, capable of being tested by scientific methods that relates an independent variable to some dependent variable. It is a medium of comparing means of two or more samples and to determine if they are significantly different from each other. Such comparisons can only be conducted when the researcher has interval level data. The use of interval level data is preferred by most researchers because it provides a more precise measurement of the phenomena under consideration. It is often impossible to conduct such under nominal and ordinal data type, hence researchers must then turn to another set of statistical tools that allow the testing of hypotheses using nominal and ordinal data. These tools are referred to in the field of statistics as non-parametric tests.

1.2. Problem Statement

Globally, researchers in the field of Management, Social sciences, Medicine, Physical sciences, Agricultural sciences, Engineering, and other disciplines have different view and perspective about Chi Square test during analysis. Some have the perspective that Chi-Square is a non-parametric test in inferential statistics; Chi-Square is a parametric test in inferential statistics; Chi-Square is a non parametric test for descriptive statistics; and data representation in pictures, graphs, measures of central tendency and dispersion are descriptive statistics. Their view seems conflicting and confusing the present day scholars, most especially young and upcoming scholars. This study is therefore carried out to clarify the confusion and establish justifiable facts regarding Chi-Square, and also differentiates between descriptive statistics, inferential statistics.

2. LITERATURE REVIEW

2.1. Statistics

Statistics is a scientific method of analyzing masses of numerical data so as to summarize the essential feature and relationships of the data in order to generalize from the analysis to determine patterns of behavior, particular outcomes or future tendencies. It is upon which statistical inference is based (Terry, 2007).

Generally, statistics are grouped into descriptive statistic and inferential statistic based on the types of data involved. The types of statistical data are nominal, ordinal, interval, and ration scale data types. They are also referred to as levels of measurement. Any data type that is nominal and ordinal measure belongs to descriptive statistics while interval and ratio scale belong to inferential statistics.

2.2. Descriptive Statistics and Inferential Statistics

Statistical analysis or statistics is the test of significance and the estimation of the extent to which a population is represented by the sample. Statistical analysis can either be descriptive statistics or inferential statistics. Descriptive statistics and inferential statistics are critical in determining the level of relationship between dependent and independent variables. The difference between descriptive statistics and inferential statistics is that descriptive statistics have to do with the measure of association which highlights potential relationships between variables. This is applicable if and only if the data types are nominal or ordinal in nature. Inferential statistics also measure the relationship, direction, and extent between dependent and independent variables. This is applicable if and only if the data types are interval or ratio scale in nature.

Measures of association indicate whether the variables are related. The two common methods used to determine the measures of association are Spearman rank correlation and Chi Square (Stephaine, 2018). This study, however establishes Chi Square as non-parametric technique in a descriptive statistics. In order to establish parametric and non-parametric test, the understanding of sampling is very crucial.

2.3. Descriptive Detailing of Descriptive Statistics and Inferential Statistics

The usefulness of descriptive detailing of data is to provide basic information about variables in a dataset. It should be noted that whether descriptive or inferential statistics, the analysis of data using descriptive tools cannot be overemphasized. Hence, descriptive detailing is quite different from descriptive statistics. Descriptive detail of data consists of the following measures:

1. Graphical and pictorial representations;
2. Measures of central tendency; and
3. Measures of dispersion.

Graphical and Pictorial Representations

Graphical and pictorial representations of statistical analysis enhance researchers and readers to have proper understanding of individual variables. Graphs and pictures provide visual representation of the data. Some of the representations include, Histogram, Scatter plot, Geographical Information Systems (G.I.S), Socio-grams which are meant for complex variables.

Measures of Central Tendency

Measures of central tendency are the most basic and, often the most informative description of a population's characteristics that describes the average member of the population of interest. There are three measures of central tendency which are Mean, Median, and Mode.

Measures of Dispersion

Measure of dispersion provides information about the spread of a variable's value. There are four key measures of dispersion, Range, Variance, Standard, Deviation and Skew.

2.4. Sampling

Sampling is the process of examining a representative set of items (people or things) out of the whole population or universe. The purpose is to gain understanding about some feature or attribute of the whole population based on the characteristics of the sample, upon this parametric test and non-parametric test will be categorized. It is important to note that when the hypothesis is wrong, it might be as a result of slight misrepresentation of sample (Terry, 2007). The types of sampling are in two categories:

1. Probability sampling also known as random sampling;

2. Non probability sampling.

Probability sampling is the type of sampling where each item in the population has an equal probability of being chosen. From probability sampling, parametric test comes in. Hence parametric test is associated with probability sampling. On the other side, non-probability sampling is associated with non-parametric test.

Non probability sampling: This is a sampling technique where samples are gathered in a process that does not give all individual in the population equal chances of being selected. This might be as a result of time limitation, money and workforce inadequacies. Since it is almost impossible to randomly sample the entire large population and it is often necessary to employ another sampling technique. Subject in a non probability sample are usually selected on the basis of their accessibility or by the purposive personal judgment of the researcher (Oskar, 2008).

The side-effect of the non- probability sampling method is that an unknown proportion of the entire population cannot be sampled. This entails that the sample may or may not represent the entire population accurately. Therefore, the results of the research cannot be used in generalizations pertaining to the entire population. The types of non-probability sampling are:

a. Convenience sampling;
b. Consecutive sampling;
c. Quota sampling;
d. Judgmental sampling (purposive sampling);
e. Snowball sampling.

There are conditions arising for adopting non probability sampling which also applies to non-parametric test, they are:

a. When demonstrating that a particular trait exists in the population;
b. When the researcher aim to do a qualitative, pilot or exploratory study;
c. When randomization is impossible like when the population is almost limitless;
d. When the research does not aim to generate results that will be used to create generalizations pertaining to the entire population;
e. When the researcher has limited budget, time and workforce (Explorable, 2009).

The major advantage of non-probability sampling when compared to probability sampling is that it is cost and time effective. It is also easy to use and can be used when it's impossible to conduct probability sampling (Stephaine, 2018). It has been earlier discussed that statistical analysis involves two classifications which are descriptive statistics and inferential statistics.

2.5. Parametric and Non-Parametric Tests

The statistical analysis or significance tests covered depend to a greater or lesser extent on the assumption or presence of the normal distribution. They are concerned with the parameters of the distribution, e.g. mean, proportions, e.t.c. Hence, they are referred to as parametric test. Altman and Bland (2009) revealed that numerical data (quantitative variables) that are normally distributed are analyzed with parametric test. There are two basic prerequisites for parametric test, they are:

1. The assumption of normally which specifies that the means of the sample group are normally distributed; and
2. The assumption of equal variance which specifies that the variances of the samples and of their corresponding population are equal (Zulfiqar and Bala, 2016).

Non-parametric test

In the case whereby non-normal data or data with extreme values or data that are not enough is known to be able to make any assumption about the type of distribution. In such circumstances, non-parametric or distribution-free tests may be used. Also, non-parametric tests can be used on data ranked in some order (decreasing or increasing order). However, if the distribution of the sample is skewed towards one side or the distribution is unknown due to the small sample size, non-parametric statistical techniques are used (Belwick, *et al.*, 2004).

Non-parametric tests are used to analyze ordinal and nominal (categorical) data. It should be noted that the assumptions of normality, and distributed parametric tests results to erroneous results therefore the non-parametric tests are used in such situation as they do not require the normality assumption (Nahm, 2016). Hence, they are referred to as distribution-free test. Although non-parametric test fail to detect a significant difference when compared with a parametric test, such that they usually have less power.

According to Abraham (2018), non-parametric test for nominal data can be analyzed with;

a. The Sign test;
b. The Binomial test;

c. The Fischer's Exact Test;
d. The McNemar's test;
e. The Cochran's test; and
f. The Chi Square test.

Also, non-parametric test for ordinal data can be analyzed with;

a. The Mann-Whitney U test;
b. Wilcoxon signed-ranks test;
c. The Kruskal-Walis one-way analysis of ranks;
d. The Friedman's ANOVA; and
e. Spearman's rho.

Advantages of non-parametric test

1. No assumptions need be made about the underlying distributions;
2. They can be used for data ranked in some order; and
3. The mathematical concepts are simpler than that of parametric tests.

Disadvantages of non-parametric test

1. They are less discriminating than parametric test; hence they are more prone to error and less powerful; and
2. Although simple, the arithmetic may take a long time.

A very good example of non-parametric test is the Chi Square test. This is an important extension of hypothesis testing and is used when it is wished to compare an actual, observed distribution with a hypothesized or expected distribution. It is often referred to as "Goodness of fit" test in contingency tables. It can be used to determine how well empirical distributions, i.e. those obtained from sample data, fit theoretical distributions such as the Normal, Poison and Binomial.

For more emphasis, non-parametric test is best applicable for qualitative study, based on descriptive statistics, meanwhile the sampling must be non-probabilistic in nature and the data type or data scale must be nominal and ordinal scale of measurement. If hypothesis testing is not properly guided by these rules, the end result of the research analysis will give spurious, nonsense, and problematic results which will be misleading.

Table 1 Summary of statistical guidelines

Scale of Measurement	Sampling or Data Collection	Statistics	Technique	Variance	Test	Analysis
Interval	Probability	Inferential	Quantitative	Homogeneity	Parametric	Regression analysis, Karl Pearson's coefficient of correlation
Ratio	Probability	Inferential	Quantitative	Homogeneity	Parametric	Regression analysis, Karl Pearson's coefficient of correlation
Nominal	Non-probability	Descriptive	Qualitative	Heterogeneity	Non-parametric	Chi Square
Ordinal	Non-probability	Descriptive	Qualitative	Heterogeneity	Non-parametric	Charles Spearman's coefficient of correlation

Source: Marie (2008); Author, 2017

2.6. Hypothesis Testing

Hypothesis is usually considered as the principal task in research. Its main function is to suggest new experiments and observations. In fact, many experiments are carried out with the deliberate object of testing hypotheses. Decision-makers often face situations wherein they are interested in testing hypotheses on the basis of available information and then take decisions on the basis of such testing. In social science, where direct knowledge of population parameter(s) is/are rare, hypothesis testing is often used for deciding whether a sample data offer such support for a hypothesis that generalization can be made (Adeniran, *et al*., 2017).

Steps involved in hypothesis testing

Generally, there are five steps involved in hypothesis testing, they are;

1. State the assumptions. The assumptions are:

   Null Hypothesis H_0 (there is no significant relationship between the dependent variable(s) and the independent variable(s).

   Also, Alternate Hypothesis H_1 or H_a (there is a significant relationship between the Dependent variable(s) and the independent variable(s).
2. Determine the table value from the Degree of Freedom and from the already set critical region or significance level (0.05, or 0.01, or 0.025). There are different Degrees of Freedom for each test; it depends on the type of test being carried out (parametric or non-parametric test).
3. Calculate or compute the test statistics using either the parametric test or the non-parametric test.
4. State the decision rule. The decision rule for the calculated test states that if the calculated value of the test statistics is more or greater than the table value, the Null Hypothesis will be rejected, but if the calculated value of the test statistics is less than the table value, the Null Hypothesis cannot be rejected. The decision rule for the computed test states that if the significance level of the computed test statistics (P.value) is less than the chosen critical region commonly 0.05, the Null Hypothesis will be rejected, but if the significance level of the computed test statistics (P.value) is more than the chosen critical region commonly 0.05, the Null Hypothesis cannot be rejected.
5. The calculated value will be compared with the table value, or the significance level of the computed test (P.value) will be compared with the chosen critical region and the decision rule either to reject or not reject the Null Hypothesis will be concluded (Adeniran, *et al*., 2017).

It is important to be meticulous about the term "*reject*" and "*cannot reject*". Even if the Null Hypothesis was decided not to be rejected that does not mean that it will be accepted. A Null Hypothesis is a hypothetical statement that is expected to be disproved therefore it can never be accepted but it might not be rejected. Instead of using the statement *"accept the Null Hypothesis"* it is best to adopt the statement that "*Null Hypothesis cannot be rejected*".

This study explains the use of Chi Square to test hypothesis involving nominal data measures. It should be pointed out by way of a cautionary note that statistics designed to test hypotheses for nominal data are no better than the data which they are designed to analyze. The lower level of precision possible using nominal or ordinal measures makes the non-parametric statistics somewhat less accurate for hypothesis testing.

2.7. Chi-Square Test

The Chi-Square distribution is a theoretical or mathematical distribution which has wide applicability in statistical analysis. The term 'Chi Square' (pronounced with a hard 'ch') is used because the Greek letter χ is used to define this distribution. It can be seen that the elements on which this distribution is based are squared, so that the symbol χ^2 is used to denote the distribution (Stephanie, 2018).

The Chi Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi Square test is that no relationship exists with the categorical variables in the population; they are independent. Also, it is commonly used to evaluate tests of independence when using a cross tabulation (also known as a bivariate table). Cross tabulation presents the distributions of two categorical variables simultaneously, with the intersections of the categories of the variables appearing in the cells of the table. The Test of independence assesses whether an association exists between the two variables by comparing the observed pattern of responses in the cells to the pattern that would be expected if variables were truly independent of each other (Stephanie, 2018).

In the same vein, χ^2 statistic appears quite different from the other statistics because it can be used for achieving the goodness of fit test and the test of independence. For both of these tests, the data obtained from the sample are referred to as the observed numbers of cases. These are the frequencies of occurrence for each category into which the data have been grouped. In the Chi

Square tests, the null hypothesis makes a statement concerning how many cases are to be expected in each category if this hypothesis is correct. The Chi Square test is based on the difference between the observed and the expected values for each category.

The Chi Square statistic is defined as

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

Where O_i is the observed number of cases in category i, and E_i is the expected number of cases in category i. The Chi Square statistic is obtained by calculating the difference between the observed number of cases and the expected number of cases in each category. This difference is squared and divided by the expected number of cases in that category. These values are then added for all the categories, and the total is referred to as the Chi Squared value.

Issues regarding Chi Square result

There are a number of important considerations when using the Chi Square statistic to evaluate a cross tabulation.

1. Chi Square value is extremely sensitive to sample size, when the sample size is too large (approximately 500), almost any small difference will appear statistically significant.
2. It is sensitive to the distribution within the cells and the statistical software tool gives warning message if cells have fewer than five (5) cases. This can be addressed by using categorical variables with a limited number of categories (e.g., by combining categories if necessary to produce a smaller table) (Stephanie, 2018).

In order to test the association between variables, Chi Square tests are used. Chi Square are suitable for nominal type of data (i.e., data that are put into classes: e.g. gender (male, female).type of job (skilled, unskilled, semi skilled) to determine whether they are associated. Nominal data as the word connotes are data that the researcher nominate values to when coding. A Chi Square is said to be significant if there is an association between two variables, and non significant if there is not an association.

Also, Chi Square is said to be useful in a parametric test if the population is small. This is so because a small population requires no statistical estimation for sample size, hence it is needed for significance testing. But it cannot be suitable for parametric test if the population is large. This is so because in order to have an accurate sample size estimated from the population such that the population will be adequately represented in the sample; there is need for random sampling technique which is probability sampling. Hence, probability sampling cannot work for Chi Square but non-probability sampling.

3. CONCLUSION AND RECOMMENDATION

The two major types of statistics are descriptive statistics and inferential statistics. It should be noted that whether descriptive or inferential statistics, the analysis of data using descriptive tools cannot be overemphasized. Hence, a research design can either adopt an inferential statistical analysis or descriptive statistical analysis, but cannot refer to descriptive detailing as descriptive statistics.

Moreover, if the data types in the research are the combination of nominal, ordinal, interval, and ratio scale data types, or the combination of one descriptive statistical data type and one inferential statistical data type, then the use of both descriptive and inferential statistical analysis can be said to be adopted in the research design. The two measures commonly used for descriptive statistical analysis are Spearman rank correlation and Chi Square. The study has clearly justified Chi Square as non-parametric technique in a descriptive statistics.

Finally, the study has justified that Chi Square is a non-parametric test used in descriptive statistics if and only if the data type is nominal and sampling is non-probabilistic. Also, Chi Square is a parametric test used in inferential statistics if and only if the data type is non-nominal in nature, and the population is small such that it requires no need for random sampling technique. Whether descriptive or inferential statistics, descriptive detailing is quite important and the fact that descriptive detailing is carried out does not mean it is a descriptive statistics. Hence there is difference between descriptive detailing and descriptive statistics.

REFERENCE

1. Abraham S. F. (2018). Module 9: Non-parametric Tests. NOVA Southeastern University.
2. Adeniran, A. O., Adekunle, E. A., and Oyedele, O. J. (2017). Establishing the Concept of Research Hypothesis through the Relationship between Demand in Nigeria International Air Passenger Traffic and Economic Variables. International Journal of Economic Behavior and Organization, 5(5):105-113. doi: 10.11648/j.ijebo.20170505.12

3. Altman, D. E., Bland, J. M. (2009). Parametric Versus Non-parametric Methods for Data Analysis. British Medical Journals, 338: 31-67.
4. Belwick, V., Cheek, L., and Ball, J. (2004). Statistics Review 10: Further Non-parametric Methods. Crit Care, 8: 196-199.
5. Explorable (2009). Quantitative and Qualitative Research. Retrieved on 19th January, 2018, from: http://explorable.com/quantitative-and-qualitative-research/
6. Marie Diener-West (2008). Use of the Chi Square Statistic in a Test of Association between a Risk Factor and a Disease. Johns Hopkins University, Pp. 1-42.
7. Nahm, F. S. (2016). Non-parametric Statistical Tests for the Continuous Data. The Basic Concept and the Practical Use. Korean J Anesthesiol, 69: 8-14.
8. Oskar, B. (2008). Research Designs. Retrieved on 19th January, 2018, from: http://explorable.com/researchdesigns/
9. Stephanie (2018). Chi Square Statistics in Research. Retrieved on 17th January, 2018 from: http://www.statisticssolutions.com/using-ChiSquare-statistics-in- research/
10. Terry, L. (2007). Quantitative Techniques. Sixth Edition, 2007. Book Power/ELST
11. Zulfiqar, A. and Bala, B. (2016). Basic Statistical Tools in Research and Data Analysis. Indian Journal of Anaesthesia, 60(10): 790. Retrieved on 18th January, 2018 from: http://www.ncbi.nlm.nih.gov/pmc/articles/