THE INTERNATIONAL JOURNAL OF
SCIENCE & TECHNOLEDGE

Possible Errors in Quantitative Data Analyses

Michael Dabi
Lecturer, Evangelical Presbyterian University College, Ghana

Abstract:
The undertaking of research in the sciences or business normally requires an understanding and correct application of carefully selected statistical methods and tests. Most often than not, statistical analyses used for certain types of data collected in research are faulty and misleading. There is often the temptation to overuse statistics to produce fanciful outputs by researchers at one extreme, while others at the other extreme just apply simple descriptive statistics to the data without inferring the sample results to the population as should be. This paper attempts to examine various statistical analyses and the most likely appropriate tests that should go with them with a focus on the Likert scale controversy. The most commonly and frequently used tests were gathered through a review of scientific and business publications and distinguished into either parametric or non-parametric tests. The resulting outcome presents a quick table with a handy explanation for choosing statistical tests based on the kind of data a researcher may be dealing with.

1. Introduction

1.1. Why Data Analysis?
Data analysis seeks to identify relationships, associations, trends and etc that may exist within or among a dataset collected from a population; and later testing how true or real these findings (relationships, associations and trends) are.
Statistical analysis is the collection, examination, summarization, manipulation, and interpretation of quantitative data to discover its underlying causes, patterns, relationships, and trends. A vast amount of raw data can be summarized into meaningful outcomes that allow for the identification of interesting and vital trends. These can be made possible via descriptive and inferential statistics.

1.2. What Is a Statistical Test?
A statistical test is an assessment of a null hypothesis to test the probability that the null hypothesis is valid. In quantitative research, a parameter of interest in the population to the researcher is investigated by using probability to select a sample from the population. Descriptive statistics are then used to present the parameter of interest, then using inferential statistics, we infer back to the population. The sample data may observe a difference, correlation or other trends. In inferring to the population, the researcher normally does so by formulating a null hypothesis which in principle must be rejected with the help of a suitable test statistic to show that the results of the study are significant and hence safe to be inferred to the population. This is the whole idea about a statistical test.
The results and discussions on empirical studies usually adopt the use of statistical tests in the analyses and interpretation of data. It is in the best interest of a researcher to know what these tests are, what they are used for, and when they can be used. This will help the researcher in two ways: to help him interpret the results of other publications accurately and to help him convey his thoughts in his data to the scientific world more expertly. Studies have confirmed the readers who were familiar with tests such as Pearson's chi-square, descriptive statistics, Fisher's exact test and t-tests should be able to correctly interpret results from about 70% of medical articles (Reed *et al*., 2003 in du Prel *et al*., 2010).
Though valuable and apparently indispensable, statistical tests may however not be necessary for every study.

1.2.1. Types of Statistical Tests
Generally there are two categories of statistical tests based primarily on the type of data and underlying assumptions. These are parametric and non-parametric tests.
Parametric tests always make assumptions about population parameters such as the mean and the standard deviation, hence the name parametric. For example it assumes that the population from which the sample data was drawn has a normal distribution with respect to the variable of interest. A second requirement is that the variable of interest must be measured on at least the interval or ratio scales. Examples of parametric tests are the t-tests, ANOVAs, Pearson's correlation coefficient.
Nonparametric tests on the other hand make no assumptions about the parameters in the population. They do not meet the requirements of parametric tests and hence applying parametric tests would be grossly inappropriate. Examples include the chi-square, Spearman's correlation etc. They are well suited for data that contain nominal and ordinal variables.

It is worth noting that non-parametric tests can be used on parametric data by removing the requirement of a normal distribution assumption. The disadvantage however is that some vital information about quantitative data (interval/ ratio) may be lost in the process and hence reduces the power to detect a statistically significant difference.
Since these two broad tests rely strongly on the type of data set available, we would look at the various types of data a researcher may be confronted with.

1.3. Types of Data
The choice of test to analyze the data should normally be based on the type of data and some other properties of the data. Data can be described in a number of ways. The most popular descriptions have been based on the scales or levels of measurement on which the variables in the data were measured. This study will follow the works of Wilson (2005) to use the scale of measurement to differentiate data (or variables, to be used interchangeably) into two broad types: qualitative (nominal and ordinal) and quantitative (interval and ratio) data.

1.3.1. Nominal Data
Qualitative data (called non-parametric data in some quarters) are categorical in nature – that is, they are separated into categories without mathematical properties.
The first of these, nominal data, measured on the nominal scale, simply gives names without any meaningful rank to things. For example, in a study where the variable gender is requested from respondents, the responses (male/female) are nominal since male and female are just names of two categories with no order or rank whatsoever. Another example is where respondents are made to respond to a question with a dichotomous response scale, yes/no. Finally, eye colour (brown/blue/green) and religion (Christianity/Islam/Hinduism/Traditionalist) will yield nominal data. As can be seen, nominal data are not numerical so cannot assume mathematical properties. They can assume dummy figures during analyses, but one should not be tempted to use these numbers to perform mathematical operations. Measures of dispersion such as the mean and the standard deviation would be misleading. Most convenient descriptive graphing for this type of data is the bar graph or chart and the pie chart. Drawing histograms for nominal data will be inappropriate since the jointed bars will assume a continuous scale.

1.3.2. Ordinal Data
Ordinal data (measured on the ordinal scale) has all the properties of the nominal: having a name and being in categories. It however has another property of having the categories arranged in a meaning order. For example rating a performance into one of their three categories: good, better or best; or 1^{st}, 2^{nd}, or 3^{rd}. A more colourful example will be Likert scale ratings: strongly agree/ agree/ neutral/ disagree/ strongly disagree. Each category is ranked higher or better than the preceding one or otherwise. Since they are qualitative data, they do not allow mathematical operations. They have similar properties as nominal data when it comes to graphing them.

1.3.3. Interval Data
Quantitative data (interval and ratio) have all the properties of qualitative data (name, categories, ranked) with some added properties. The general property of quantitative data is that they are numerical and not dummy numbers.
Interval data has the added property of having equal intervals between categories or levels. Equal intervals should allow some mathematical operations such as sums and differences on the numbers. The commonest example of the interval scale is temperature. The intervals across a thermometer are equal. So that we can subtract 36 degrees celcius from 50 degrees celcius. Interval data do not however possess a true starting point or zero, hence mathematical operations such as divisions cannot be done on them. Also, since the data is numerical, means and standard deviations can be calculated.

1.3.4. Ratio Data
Ratio data has the advantage of having an absolute zero: absence of the variable measured. Money is measured on a ratio scale since the absence of money is represented by an absolute zero, and $300 is thrice as much as $100. Other variables like interest rates, height, weight, blood cholesterol levels will yield ratio data.
Descriptive graphs such as the histogram and the line graph are most appropriate for quantitative data.

2. Possible Errors in Application of Statistical Tests

2.1. Type 1 and Type 2 Trrors
Though a Type III error has been rumoured in the statistics community, we currently still stick to only two types of statistical errors – type 1 and type 2. A type I error is the probability of rejecting a true null hypothesis, denoted by the symbol alpha (α). In simple terms, it is detecting an effect in a study which was actually non-existent; likened in legal terms to finding an innocent man guilty of a crime. On the other hand, a type II error is the probability of accepting (failing to reject) a false null hypothesis. Again, simply put, it is failing to detect an effect in a study when there was really an effect. This also is synonymous in the legal system to a judge finding a guilty man innocent. Type 1 errors are generally considered to be more grievous than type II errors (but with some exceptions especially in manufacturing where a bad product might be shipped to a consumer – type 1 error, compared to a good product being taken as bad and destroyed – type II error).

In data analysis, type I errors correspond to using parametric tests such as the t-tests to analyze qualitative (nonparametric) data (data measured on the nominal and ordinal scales). Thus, effects would be detected that are not real but just due to the manipulation. Type II errors similarly occur when non-parametric tests such as the chi-squared are used to analyze quantitative (parametric) data (data measured on the interval and ratio scales). Here, real effects in the data may go unnoticed or undetected. Both types of errors may give misleading results and lead to incorrect conclusions and inappropriate recommendations. The study will investigate one alleged type I error that has been prevalent in the analyses of survey data – the Likert scale data analysis.

2.2 The Likert Scale Controversy

Losby and Wetmore (2012) aptly explained what a Likert scale is:

"A Likert scale is an ordered scale from which respondents choose one option that best aligns with their view. It is often used to measure respondents' attitudes by asking the extent to which they agree or disagree with a particular question or statement. A typical scale might be "Strongly disagree, Disagree, Neutral, Agree, Strongly agree.""

In order to take advantage of more powerful statistical techniques, researchers often assume that the intervals on a Likert scale are equal. This argument has been raging behind the scenes for years now with mostly psychologists supporting the idea and researchers from the pure mathematical mainstream taking the opposite stance. We will take a look at these arguments.

The debate as to whether the Likert scale produces ordinal or interval data, and as to whether parametric tests should be applied to ordinal data has been around for long (e.g. Stevens, 1955; Baker, Hardyk and Petrinovich, 1966; Labovitz, 1967; and Thomas, 1982). These researchers have been distributed across the opposite sides of the divide. Fairly recent contributions have equally featured from Jamieson (2004) and Lubke and Muthen (2004).

Michell (1986) sets the ball rolling with the claim that much of the impasse relating to data or scale type and its associated statistics can be attributed to differences in perspectives on measurement from various disciplines. This explains why the argument is mainly between psychologists and mathematicians who differ slightly in opinion as relating measurement, though other disciplines have come up supporting either side.

Those advocating for Likert scales to be treated as interval scales have posited two main reasons both based on assumptions. The first is the assumption of equal intervals on the Likert scale or standardization of these intervals to suit an interval scale (Baker, Hardyk, and Petrinovich, 1966; Labovitz, 1967). These researchers have shown empirically that there is little or no difference if an ordinal scale as we know it is treated as an interval scale after these assumptions and standardization.

The other assumption is the one of normality for the shape of the distribution (Gaita, 1980 and Borgatta & Bournstedt, 1980). These believe that if an ordinal data passes the normality test or assumption, then it can be treated as an interval scale.

These arguments have been countered by other researchers (e.g. Stevens, 1955; Marcus-Roberts and Roberts, 1987) who have practically demonstrated that using parametric analysis, such as means, standard deviations and Pearson's correlation on ordinal data will produce very strange results.

The 21st century has seen similar arguments. For example Lubke and Muthen (2004) found evidence to support their claim of the possibility to observe true parametric values in factor analysis with likert scale data provided assumptions about normality etc are met; the same notion shared by Glass *et al*. (1972) when he found that F tests in ANOVA could return accurate p-values on likert items under the same assumptions of normality and equidistant intervals.

On the other hand Jamieson (2004) has argued that since the Likert scale presents responses as mere ordered categories, the intervals between the scale values should not be taken to be equal. Hence, any numerical operations applied as a result of the assumption of equidistance would be invalid. Only none parametric statistics should be applied.

3. Attempts at Calming the Storm

Allen and Seaman (2007), in an investigation of Likert scales and available data analyses had contended that the underlying reason for treating ordinal data as interval data may be the more power that the parametric tests have over their non-parametric alternatives. They also suggest the ease with interpretations of parametric tests to be another reason. But they warned against the temptation, especially without consideration of the values of the dataset and objectives of the analysis, which may end up to be misleading and misrepresenting of the findings of the study. To achieve the equidistant property required of interval scales, Allen and Seaman had suggested a combination of Likert scales into indexes so as to add values and variability to the data; then meeting the assumptions of normality, parametric analyses can be duly followed. This they claim could be achieved by the use of a continuous line or track bar to yield a continuous line measure.

Other writers have suggested a misunderstanding of the Likert scale as the underlying problem (e.g. Uebersax, 2006; Boone and Boone, 2012). These writers have therefore suggested a plausible solution to the gridlock. For example Uebersax (2006) has found no reason to join in the debate but rather advised researchers to make the distinction between a Likert item (a single question or statement on the Likert scale) and the Likert or Likert-type scale itself (a multi-item scale where people's attitudes or opinions on a construct are measured by adding or averaging their responses across all items. The Likert-type scale is so-called because it is just a variation of the traditional Likert scale with responses other than the traditional agree-disagree continuum). He argued that when responses are added or averaged across several items, an overall score measurement is produced that can safely be treated with parametric methods.

This stance was later supported by Boone and Boone (2012). They also believed that once a researcher understood the distinction between Likert-type data (actually should be Likert item per our definitions) and Likert scale data, the decision with the appropriate analyses method will follow easily. Following up closely with Uebersax, they made reference to the original Likert scale created by Likert where he combined responses from a series of questions to create a measurement scale; with the data analyses being based on the composite score from the series of questions that represented the new scale and not the individual questions. In this combined form, a quantitative measure is provided which can be used in analyses involving interval and ratio scales. They concluded that Likert data (from Likert items) will only yield ordinal data whereas Likert scale data (created by calculating a composite score (sum or mean) from four or more likert items) should be measured on the interval scale.

4. Conclusion

Different disciplines have different views on measurements on the Likert scale. The writer will not like to take a biased stance but offer suggestions that may help to avoid open pitfalls in analyses. Though it does not make sense using arithmetic on Likert data (because the assigned numbers are merely numerals and not numbers, so that the mean or average of fair and good cannot be fair-and-a-half (Kuzon Jr. et al., 1996) just by mere reason of numerals attached to them), it is interesting to note that some parametric tests such as ANOVA applied on Likert data may yield similar results to its more appropriate non-parametric alternative, the Kruskal-Wallis test. This makes the argument still interesting. The writer can only advise researchers to exercise caution and make some recommendations to be taken into consideration when planning analyses for data produced from Likert scales.

Having seen both sides of the argument, we can draw conclusions by considering the following suggestions from Knapp (1990) and Grace-Martin (2010)) when choosing statistical tests.

i. Take a stance – choose a measurement perspective. As much as each party will try to dissuade the other from their view point, the fact remains that this impasse is here to stay for a long time to come. It is only appropriate to take a stance carefully depending on the type and objectives of the research at hand.
ii. If you plan to use a parametric test for categorical data, other conditions such as normality and equal variances should be met.
iii. Ultimately, find out the non-parametric equivalent of the parametric test you intend to use. Table 1 shows a set of commonly used parametric tests and their non-parametric equivalents.

Parametric test	Corresponding non-parametric test	Comments or Purpose
t-test for independent samples	i. Mann-Whitney U test ii. Wilcoxon rank sum test	Compares the means between two unrelated groups. Tests whether there is a significant difference between the two means
Paired t-test or t-test for related samples or t-test for dependent samples	i. Wilcoxon matched pairs signed-rank test ii. McNemar's test for symmetry	Compares the means between two related groups
Pearson's correlation coefficient	Spearman rank correlation coefficient	Assesses the linear association between two variables
One-way ANOVA	Kruskal-Wallis analysis of variance by ranks	Determines whether there are any significant differences between the means of two or more independent (unrelated) groups.
Two-way ANOVA	Friedman two-way analysis of variance	Compares groups classified by two different factors
Linear regression	Spearman rank correlation coefficient	Predicts an outcome from an independent variable

Table 1: Summary of common parametric tests and their equivalent non-parametric counterparts. From Smith (n.d).

iv. Finally, some researchers have advised the use of more stringent significance levels such as 0.01 and 0.005 when one intends to apply parametric tests to non-parametric data to ensure the results are very strong and beyond doubt (e.g. Grace-Martin, 2010).

5. Recommendations

The study will like to recommend a collection of analyses/tests at a glance for researchers to fall on when in doubt of the type of analyses or test to use for a certain type of data. The collection was arrived at after reviews of several articles and books including Goldin et al. (1996), Neideen and Brasel (2007), *When to use a statistical test* (n.d), Wilson (2005), Stevens (1992), . The focus is more on the outcome (dependent) variable (s). For detailed analyses, a researcher may have to look out for the characteristics of the independent variable (s) as well. These are shown in Table 2.

	Situation / Purpose	Measurement scale (of dependent variable)	Suggested Appropriate Test/ Analysis
1	To test whether a sample of data came from a population with a specific distribution.	Nominal	Chi-square goodness of fit
2	Relationship between variables	Nominal	i. Chi-square test for independence ii. Fisher's exact test iii. McNemar's test of symmetry
3	To predict a single outcome variable from a single independent variable	Ratio	Linear regression
4	Test for normality	Nominal or ordinal	Kolmogorov–Smirnov test
5	Test for normality	Any	Shapiro-Wilk test
5	Correlation/Relationship between two variables	Interval or ratio	Pearson's correlation coefficient
6	Correlation/Relationship between two variables	Ordinal	Spearman's rank correlation coefficient
7	Comparing two or more samples that are independent	Ordinal	Kruskal-Wallis test
8	Comparing two related or matched samples, or repeated measurements on a single sample	Ordinal	Wilcoxon signed-rank test
9	Comparison of more than two paired samples	Ordinal/ interval	Friedman test
10	To determine whether there are any significant differences between or among the means of two or more independent (unrelated) groups	Interval / ratio	One-way ANOVA
11	To predict a single outcome variable from a single independent variable	Ordinal / interval	Ordered logistic regression
12	To predict an outcome from two or more independent (categorical) variables	Nominal	Factorial logistic regression
13	To compare two or more regression lines to each other	Interval/ ratio	Analysis of covariance (ANCOVA)
14	Predicting a single outcome from several independent variables	Nominal	i. Discriminant Analysis ii. Logistic regression
15	Predicting several dependent variables from a set of categorical variables	Interval/ ratio	Multivariate analysis of variance (MANOVA)
16	To analyze the structure of the interrelationships among a large number of variables to determine a set of common underlying dimensions (factors).	Interval/ ratio	Factor analysis
17	To identify and measure the associations among two sets of variables	Interval/ ratio	Canonical correlation
18	To emphasize variation and bring out strong patterns in a dataset	Ratio	Principal Component Analysis (PCA)
19	To further investigate the relationship between two sets of variables	Ratio/ interval	Canonical Component analysis (CCA)
20	to provide a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects.	Interval ratio	Multidimensional Scaling (MDS)
21	To provide a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects.	Ordinal	Non-metric MDS
22	Predict a single outcome from several independent variables	Interval/ ratio	Multiple regression
23	Predict several outcome variables from several independent (numerical) variables	Interval/ ratio	Canonical Analysis
24	To understand consumer preferences for products or services	Nominal/ ordinal	Conjoint analysis
25	grouping objects of similar kind into respective categories	Nominal/ordinal/ interval	Cluster Analysis

Table 2: Statistical tests/analyses at a quick glance

6. References

i. Allen, I.E., & Seaman, A.S. (2007). http://asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html
ii. Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens's proscriptions on statistics. Educational and Psychological Measurement, 26, 291- 309.
iii. Boone, H.N. & Boone, D.A. (2012). Analyzing Likert data. Journal of Extension [On-line], 50 (2), Article 2TOT2. Available at: http://www.joe.org/joe/2012april/tt2p.shtml
iv. Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. Sociological Methods and Research, 9, 147-160.
v. Choosing the correct statistical test (n.d). Retrieved June 15, 15 at: http://bama.ua.edu/~jleeper/627/choosestat.html
vi. Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. Psychological Bulletin,87, 564-567.
vii. Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the analyses of variance and covariance. Review of Educational Research, 42 (3), 237-288.
viii. Goldin, J., Zhu, W., Sayre, J.W. (1996). A review of the statistical analysis used in papers published in Clinical Radiology and British Journal of Radiology. Clinical Radiolology, 51(1): 47–50.
ix. Grace-Martin, K. (2010). Can Likert Scale Data ever be Continuous? [Web log post]. Retrieved from: http://www.theanalysisfactor.com/can-likert-scale-data-ever-be-continuous/
x. Jamieson, S. (2004). Likert scales: how to (ab)use them. Medical Education, 38 (12), 1217-1218.
xi. Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing Research, 39 (2), 121–123.
xii. Kuzon, W.M. Jr, Urbanchek, M.G., McCabe, S. (1996). The seven deadly sins of statistical analysis. Ann Plastic Surg, 37, 265–272.
xiii. Labovitz, S. (1967). Some observations on measurement and statistics. Social Forces, 46, 151-160.
xiv. Landau, S. & Everitt, B.S. (2004). A handbook of statistical analyses using SPSS. Florida: Chapman and Hall/ CRC Press LLC.
xv. Losby, J. & Wetmore, A. (2012). CDC coffee break: using Likert scale in evaluation survey work. Retrieved June 15, 2015 at: http://www.cdc.gov/dhdsp/pubs/docs/cb_february_14_2012.pdf
xvi. Lubke, G. H., & Muthen, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. Structural Equation Modeling, 11, 514-534.
xvii. Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. Journal of Educational Statistics, 12, 383-394.
xviii. Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. Psychological Bulletin, 100, 398-107.
xix. Neideen, T. & Brasel, K. (2007). Understanding statistical tests. Journal of Surgical Education, 64 (2), 93-96.
xx. Reed Iii, J. F., Salen, P., & Bagher, P. (2003). Methodological and statistical techniques: what do residents really need to know about statistics? Journal of medical systems, 27(3), 233- 238. In: du Prel, J.B., Hommel, G., Röhrig, B., & Blettner, M. (2010): Choosing a statistical test. Part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int; 107 (19): 343–348. DOI: 10.3238/arztebl.2010.0343.
xxi. Smith, P.F
xxii. Stevens, J. P. (2009). Applied multivariate statistics for the Social Sciences (2nd ed.). USA: Routledge.
xxiii. Stevens, S. S. (1955). On the averaging of data. Science, 121, 113-116.
xxiv. Thomas, H. (1982). IQ interval scales, and normal distributions. Psychological Bulletin. 91, 198-202.
xxv. Uebersax, J.S. (2006). Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement website. Available at: http://john-uebersax.com/stat/likert.htm. Accessed: June 15, 2015.
xxvi. What statistical analysis should I use? Statistical analyses using Stata (n.d). UCLA: Statistical Consulting Group. Retrieved June 15, 2015 from: http://www.ats.ucla.edu/stat/mult_pkg/whatstat/
xxvii. When to use a particular statistical test. (n.d). Retrieved June 15, 2015 from: http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdf
xxviii. Wilson, J.H. (2004). Essential statistics (1st ed.). USA: Pearson.