Great Plains Sociologist

Volume 10 | Issue 2 Article 2

1997

Data Analysis Made Easy: An Undergraduate Student's Guide to Choosing Appropriate Statistical Tests for Social Research

A. Olu Oyinlade
Nebraska Wesleyan University

The Great Plains Sociologist Vol. 10, No. 2 Spring 1998

Data Analysis Made Easy: An Undergraduate Student's Guide to Choosing Appropriate Statistical Tests for Social Research.

A. Olu Oyinlade
Department of Sociology, Anthropology, and Social Work
Nebraska Wesleyan University

Abstract

This article is written as a guide for undergraduate students in using statistics in the social sciences. Some general guidelines are provided for deciding which statistic to use with different types of data (nominal, ordinal, interval, and ratio). Four sections are presented: identifying variables, choosing appropriate statistics, computation, and understanding results. This article is not written as a "nuts and bolts" guide to teaching all of statistics but instead is a guide to help sudents. Instructors of this material may also benefit from these discussions.

Many students in sociology, psychology, and other areas of social sciences and education often regard a course in quantitative research methods as their toughest class. This is usually not because they lack adequate knowledge of research procedures, but because they are generally not well prepared for quantitative work. They lack adequate statistical preparation; hence, research methods cause fear. In my experience teaching quantitative research methods and assisting students in research, I have discovered that the problems most students have are basic, and when they are guided through these problems, a magical bulb lights in their heads, making data analysis easy and fun to conduct.

This article is written for undergraduate students in research methods to provide some guidance on how to resolve some of their common problems. It is not intended to teach the nuts and bolts of statistical data analysis. Rather, it is intended as an aid to help

students choose appropriate statistical tests, given the nature of their data.

Common Problems

The common problems can be divided into four basic areas: (1) identifying variables, (2) choosing the appropriate statistical test, (3) computing the statistic, and (4) understanding the results. Each of these four problem areas is discussed below.

Area 1: Variables Identification.

Before students can successfully conduct data analysis, they must be able to properly label their variables. Undergraduate students typically deal with only two types of variables--independent and dependent. These variables must be properly labeled "X" (independent) and "Y" (dependent) in the student's statistical spread sheet. An independent variable is the variable that effects a change in another variable, the dependent variable. For example, in a hypothesis that states that income buys happiness, the independent variable is income, and the dependent variable is happiness. Happiness is dependent on income. If these variables are not properly identified, students risk seeking a relationship different from their intentions. If happiness is erroneously treated as the independent variable in the above hypothesis, students will be testing to see if happiness buys income, not the stated hypothesis.

Area 2: Choosing The Appropriate Statistical Test.

This is the most critical of all problems that impede students' mastery of the quantitative research process. After properly labeling the research variables, many students have a knowledge deficiency in choosing the right statistical test for data analysis. Here is a simple two-step guide to help students deal effectively with this problem.

Step 1: Nature of Data. To do a successful data analysis, students must first understand the nature of their data. Data are generally classified in one of four major categories: nominal, ordinal,

interval, or ratio (Levin and Fox, 1991; Judd, Smith and Kidder, 1991; Babbie 1992).

Nominal data are categorical or discrete in nature. They represent data classification by name only, and no arithmetic properties such as multiplication, subtraction, addition and division can be carried out upon them. Even when numerals are assigned to nominal categories, they serve only as labels for the researcher's convenience of organizing data, not for mathematical calculations. For example, party affiliation can be Democrat or Republican, and gender can be male or female. These categories are only different, hence, it makes no sense to multiply, divide, add, or subtract them (Bordens and Abbot, 1991). Because nominal data have no arithmetic properties, they are the lowest form of measurement.

Ordinal data are the next highest level of data. These data combine the nominal quality of categorization with ranking or ordering. That is, ordinal data can be ranked or ordered from low to high and vice versa. For example, we can say that twenty is greater than ten, and thirty is greater than twenty, but we cannot "infer from the numbering any absolute quantity, nor could we infer that the intervals between numbers are equal" (Singleton, Straits and Straits, 1993:112). This means that we cannot say that twenty is twice as big as ten but only that twenty is larger than ten. Like nominal data, one "cannot perform most mathematical (statistical) operations in analyzing the data. We cannot add, subtract, multiply, or divide; we can only rank things: $1<2$, $2<3$, $1<3$" (Singleton, Straits and Straits, 1993:112). Data obtained through attitude, opinion, social class standing, preference ratings, religiosity, and happiness scales are examples of ordinal data.

Interval data have all the properties of ordinal plus one major additional quality: the values on the variable being measured are equidistant from one another. That is, the distance between ten and twenty is exactly the same between thirty and forty; however, interval data do not possess an absolute zero. Zero does not mean a complete absence of something, so, like ordinal data, you cannot claim that one value is twice as large or small as the other. The only time you can claim that one value is twice as large or small as the other is when the data are ratio, the highest level of data.

Ratio data have an absolute zero or a true zero point, which makes it possible to perform all mathematical functions. Examples of interval data include IQ scores, class examination scores, and temperature, while ratio data include number of births, age, years of education completed, number of divorces, and unemployment rate. In most social science research, little distinction is made between interval and ratio data. They are treated the same way and analyzed with the statistical procedures applicable to ratio data. As a rule, students must strive to measure their variables at the highest level possible. "Do not settle for a lower level of measurement when you can be more precise by measuring a variable at a higher level" (Singleton, Straits and Straits, 1993:114).

Any student who does not have a good understanding of what these data categories mean must seriously study them. They are simple and basic, yet they are the key to a student's ability to conduct data analysis.

Step 2: Descriptive and Inferential Decisions. Students must decide whether they need descriptive and/or inferential statistics.

Descriptive Statistics

Descriptive statistics are procedures for organizing, summarizing, and interpreting data (Monette, Sullivan and DeJong, 1990). They can be classified under three categories: univariate, bivariate, and multivariate. Univariate statistics describe only one variable, labeled X, in the student's statistical spread sheet. They include frequency distribution, measures of central tendency (mean, median, mode), and measures of dispersion (range, quartile, variance, standard deviation).

Bivariate descriptive statistics describe the relationship between two variables, independent and dependent, while a multivariate analysis describes the relationship among three or more variables of which at least one variable is independent and another is dependent. Most undergraduate research is bivariate so the rest of this paper will focus on bivariate analysis.

Common statistics for bivariate descriptions, especially at the undergraduate level, include contingency tables containing raw

observed frequencies (see Table 1) and percentage tables containing percentages derived from frequencies (see Table 2). The conventional rules for constructing both contingency and percentage tables are as follows (Monette, Sullivan and DeJong, 1990).

1. The tables are constructed to describe frequencies or percentages of categorical data only. Noncategorical data must be transformed into categories before they can be expressed in tables.

2. When data, such as ordinal, are categorized in ranks (high to low), the categories should be ordered as demonstrated in Tables 1 and 2, with columns ranging from the lowest on the left to the highest on the right. Rows are ordered from the highest on top to the lowest at the bottom.

3. The independent variable is placed at the top of the table, representing the columns, while the dependent variable is located on the side of the table displayed in rows.

4. The squares, called *cells*, contain values called *cell frequencies*. The totals at the end of each row or column are referred to as *marginals*.

5. A table is conventionally identified by the number of rows and columns it contains. A table with two rows and two columns becomes a 2 X 2 (read 2 by 2) table while a table with three rows and three columns is a 3 X 3 table. Also, because the number of rows is always designated first, a 2 X 3 table is not the same as a 3 X 2 table. Tables 1 and 2 are examples of 2 X 3 tables.

Other bivariate descriptive statistics measure the strength of relationship or association between independent and dependent variables. Depending on the type of association being sought, strength of relationship can vary from a perfect negative association

Table 1: Example of a Contingency Table Showing Raw Frequencies.

Dependent Variable	Independent Variable: Light	Medium	Heavy	
Heavy	30	40	50	120
Light	80	20	40	140
Marginal	110	60	90	260

Table 2: Example of a Percentage Table.

Dependent Variable	Independent Variable: Light	Medium	Heavy
Heavy	37.3%	66.6%	55.5%
Light	72.7%	33.3%	44.4%
	100%	100%	100%
	(110)	(60)	(90)

(-1.00) to a perfect positive association (+1.00). It can also vary from no relationship (0.00) to +1.00. A +1.00 relationship means that both independent and dependent variables vary together in the same direction 100% of the time, while a -1.00 indicates an inverse (opposite directions) relationship between the two variables 100% of the time. A 0.00 strength shows that the two variables have no patterned relationship. When the strength of a relationship is less than perfect, the higher the value of the strength (positively or negatively),

the stronger the relationship (Monette, Sullivan and DeJong, 1990). Also, the "+" and "-" signs are applicable to ordinal and higher level data only. They do not apply to nominal data, because they carry no meaning for categorical data.

Specific bivariate descriptive statistics that measure the strength of association between two nominal variables include the phi coefficient (ø) and lambda (λ). The phi coefficient ranges from -1 to +1 and it is applicable only when the research variables can be dichotomized in a 2 X 2 table only. When the table is larger than 2 X 2, lambda, with a value range from 0 to +1 should be used. When the phi coefficient is squared ($ø^2$), it produces a proportional reduction in error value (PRE value) which tells the extent to which the independent variable helps to reduce error in predicting the values of the dependent variable. Lambda value is already presquared, so it automatically has the PRE effect. This is why lambda is usually low and also never negative (Monette, Sullivan and DeJong, 1990).

The strength of the relationship between two ordinal variables can be determined using Spearman's rho (r_s) or Kendall's tau (τ). Both measures range from -1 to +1, but rho is better when dealing with fully ordered data with no tied scores while tau is better for data with ties. In addition, tau is better when the student is not sure about the expected pattern of relationship between two variables and, therefore, has difficulty determining which variable is independent and which is dependent. In this case, tau allows the student to use either variable as independent and still obtain a valid strength of relationship between the two variables. Also, tau automatically carries the PRE effect, whereas rho has to be squared (r^2_s) to have PRE effect (Monette, Sullivan and DeJong, 1990).

The most common measure of association with interval or ratio data is Pearson's Correlation Coefficient, popularly known as Pearson's r. Pearson's r, like rho and tau, ranges from -1 to +1. When squared (r^2) it is termed the coefficient of determination. The value of r^2 tells the extent to which a change in the dependent variable is a direct result of a change in the independent variable (Bowen and Weisberg, 1980). One note of caution for the student to remember about Pearson's r is that it estimates linear relationships only and,

therefore, may underestimate the strength of a relationship that is not linear, such as those that are curvilinear.

Inferential Statistics

Inferential statistics, like descriptive ones, can be bivariate. They are computed to allow the researcher to generalize from a sample to the population from which the sample was drawn (Monette, Sullivan and DeJong, 1990). These types of analyses are based on the probability that an event will occur. More precisely, they specify the probability of occurrence of the strength of the relationship obtained through descriptive statistics.

To establish the probability that a given relationship between two variables will occur at a sufficient frequency to allow generalization over a larger population, a test of significance is required. This test helps to establish that the probability of a given relationship is not due to accident but due to an actual pattern of existing relationship between your variables (Cuzzort and Vrettos, 1996). In most sociological and social sciences research, the significance test is most commonly determined at an alpha (α) or probability (p) level of .05. This means that the strength of the relationship obtained from descriptive statistics is allowed to occur no more than 5% of the time by chance or error. Put another way, a given pattern of relationship between the variables is expected to occur at least 95% of the time as a direct result of true interaction between your variables to claim the relationship as significant.

To test for significance, a null hypothesis must be developed that states that there is no significant association between the independent and the dependent variables. An acceptable p level must be determined. If the conventional $p < .05$ (read probability less than .05) is used in social science, and the p value obtained from the test of significance is $< .05$ (e.g., .03), reject the null hypothesis. This means that a statistically significant association exists between the two variables. If the obtained p value is greater than .05 (e.g., .06), fail to reject the null hypothesis. The null is correct, and there is no statistical relationship.

Since the desirable level of p is at the discretion of the researcher, students must guard against setting it too high or too low. For example, if the p is set high at .30 and the obtained p level is .288, the student will claim that a significant relationship exists between two variables, but this is not a reliable claim, and any generalization from it will be equally unreliable as well as erroneous. Why? Simply because "a statistically significant finding is one that has a low probability (usually < .05) of occurring as a result of error variance alone" (Leary, 1991:162). So, at p = .288, you are allowing chance or error to influence the relationship between your variables approximately 29 percent of the time. This means you would claim a true relationship exists and reject the null hypothesis, when, in fact, the relationship is due to a high probability of chance or error, and the null should not have been rejected. This type of error of rejecting the null hypothesis (when it should not have been rejected) is called a Type I (alpha) error.

A Type II (beta) error is also possible. This usually happens when the p is set too low, such as $p < .001$, and one fails to reject the null hypothesis that should have been rejected. At this level, you are being very strict regarding the extent of chance you could tolerate when you make a decision regarding your null hypothesis. The down side of this is that you would fail to reject the null if you obtained a p value such as .01, .02 or .002. By failing to reject the null, you indicate that the null is correct, there is no statistically significant relationship between your variables, while in fact, the relationship is significant. To be on the safe side, it is best to use the conventional $p < .05$ for most social science research.

When the phi and lambda are calculated for strength of association, the most common inferential statistics used to test for significance is the Chi Square (χ^2). The t- test is used with Spearman's rho, Kendall's tau, and Pearson's r. Also, with Pearson's r, a regression analysis can be conducted to help predict the value of the dependent variable, given certain values of the independent variable. See Table 3 for more details on how to select appropriate statistics.

Table 3: Oyinlade's Guide To Choosing Appropriate Common Statistical Tests.

RESEARCH VARIABLES		
INDEPENDENT	**DEPENDENT**	**STATISTICAL TEST***
1. NOMINAL	NOMINAL	**A.** Phi (with Chi Square) **B.** Lambda (with Chi Square) **C.**Table (with Chi Square) **D.** Percentage (with Chi Square)
2. ORDINAL	ORDINAL	**A.** Spearman's Rho (with T-test) **B.** Kendall's Tau (with T-test)
3. INTERVAL or RATIO	INTERVAL or RATIO	**A.** Pearson's r (with T-test) **B.** Regression Analysis (to predict value of Y)
4. NOMINAL or RATIO	INTERVAL	**A.** Lambda (with T- test for independ ent variable) **B.** Lambda (with ANOVA (F test) for independent variables with more than two independent variables) **C.** Convert nominal into dummy and use #3
5. INTERVAL or RATIO	NOMINAL	**A.** convert interval into nominal and use #1. **B.** Convert nominal into dummy and use #3
6. NOMINAL<---->ORDINAL		**A.** Lambda (with Mann-Whitney U test) **B.** Convert ordinal into nominal and use #1

	C. Convert nominal into dummy and use #3
7. ORDINAL<----->INTERVAL	**A.** Treat ordinal as interval & use #3 **B.** Treat interval as ordinal & use # **C.** Convert ordinal to nominal & use #4. **D.** Convert both data into nominal & use#1

NOTE: Only higher order data can be converted into lower order data. Lower order data cannot be transformed into higher order--except in situations where ordinal data are treated as interval data or when nominal is converted into dummy variables.
*Tests of significance are in the parentheses after tests of strength of association.

Area 3: Computation.

It is impossible to instruct how to compute data analysis in one article since there are numerous statistical software programs that students use. My only advice is that students learn a computer program that will perform the analyses they intend to do. A program with a pull-down menu, such as found in Macintosh computers and PC windows programs, is generally easier for students than DOS programs. When using a program with a pull-down menu, students need only locate appropriate tests from command menus, click the mouse, and their analysis is automatically computed. Students must remember to designate the X and Y variables in the data before running the program, otherwise, no statistics will be computed.

Area 4: Understanding Results.

No statistical computation is useful unless the student is able to interpret the results from a lengthy computer printout. What is the easiest way for students to understand and interpret results? How are students able to select necessary information from the output of the

computer? Here is a little guide to help students interpret results with a minimum of headaches.

Results of Tests of Significant Difference: When you perform a Chi Square, ANOVA, Mann-Whitney U test or the t-test (see Table 3), you have done a test of significant difference. That is, you have tested to see if the elements or categories within the independent variable are significantly different in their effects on the dependent variable. To make this interpretation, some items of information on the printout are essential for you to report. They are observed frequencies for each element in the variables, value of the test performed for strength of association (e.g., phi or lambda), value of the test performed for significance (e.g., Chi Square value, F value for ANOVA, U for Mann-Whitney or t value for the t-test), degree of freedom (df), and probability (p) value. These items are usually sufficient in making a general statement of significant difference, and they are best presented in a table.

Results of Tests of Significant Relationship: These tests establish the existence (or lack of existence) of a specified pattern of relationship between variables. They are designed to specify the degree and pattern of co-variation between variables, indicating how two variables occur or vary together. To establish this relationship, the student must first test for strength of co-variation or association between two variables with Pearson's r, Spearman's rho or Kendall's tau, among others. The student will then conduct a t-test for the significance of the strength of association at a specified p level. Also, a simple linear regression analysis can be performed to predict the value of the dependent variable given a value of the independent variable (see Table 3).

To interpret the results of any of the tests of significant relationship and make a general statement of relationship between two variables, students only needs to focus on interpreting a few items from their printout. They are degree of freedom (df), the value of test performed for strength of association (rho, tau, r), the t value for test of significance, and the probability (p) value. When a simple linear regression is performed, it is often sufficient for students to report values for df, t, F, r, and p, especially at the undergraduate

level, in establishing that a specific pattern of co-existence or co-variation exists (or fails to exist) between two variables.

Summary and Conclusion.

Data analysis does not have to be a deterrent to effective quantitative research. Statistical difficulties that often scare students away from quantitative research can be classified in four areas: variables identification, choosing appropriate statistical tests, computation, and understanding results.

By following guidelines presented in this paper, students will find it easier to resolve most data analysis problems and have fun doing research, however, since this paper is not designed to teach statistics per se, students interested in quantitative research must develop their statistical skills sufficiently to understand the nature of the data, the purpose of a statistical procedure, and the meaning of the findings of such procedures. With a little statistical background, the guidelines provided in this paper could prove invaluable to an undergraduate student doing quantitative research.

References

Babbie, Earl. 1992. *The Practice of Social Research.* Belmont, California: Wadsworth Publishing Company.

Bowen, Bruce D. and H. F. Weisberg. 1980. *An Introduction to Data Analysis.* San Francisco: W. H. Freeman and Company.

Cuzzort, R. P. and James S. Vrettos. 1996. *Statistical Reason.* New York: St. Marin's press, Inc.

Judd, Charles M., Eliot R. Smith and Louise H. Kidder. 1991. *Research Methods in Social Relations.* Fort Worth: Holt, Rinehart & Winston, Inc.

Leary, Mark R. 1991. *Introduction to Behavioral Research Methods.* Belmont, CA: Wadsworth Publishing Co.

Levin, Jack and James A Fox. 1991. *Elementary Statistics.* New York: Harper Collins Publishers.

Monette, Duane R., Thomas J. Sullivan and Cornell R. DeJong. 1990. *Applied Social Research.* Fort Worth: Holt, Rinehart and Winston, Inc.

Singleton, Royce A., Bruce C. Straits and Margaret M. Straits. 1993. *Approaches to Social Research.* New York: Oxford University Press.