Levels of Measurement and Choosing the Correct Statistical Test

Most textbooks distinguish among nominal, ordinal, interval, and ratio scales based on a classification system developed by Stevens (1946). Choice of the statistical analyses in the social sciences typically rests on a more general or cruder classification of measures into what I will call "continuous" and "discrete." Continuous refers to a variable with many possible values. By "discrete" I mean few categories. I, as well as others, often use the terms "dichotomous," "binary," "categorical," or "qualitative" synonymously with "discrete." [1] This general characterization of a dependent (response) variable as discrete or continuous relates to two general classes of commonly employed statistical tests—those based on the normal distribution and those based on the binomial distribution (or its relatives, the multinomial and Poisson distributions). Normal theory plays an important role in statistical tests with continuous dependent variables, such as t-tests, ANOVA, correlation, and regression, and binomial theory plays an important role in statistical tests with discrete dependent variables, such as chi-square and logistic regression.[2]

This general characterization of a dependent (response) variable as discrete or continuous relates to two general classes of commonly employed statistical tests—those based on the normal distribution and those based on the binomial distribution (or its relatives, the multinomial and Poisson distributions). The normal distribution plays an important role in statistical tests with continuous dependent (response) variables, such as t-tests, ANOVA, correlation, and regression, and the binomial distribution plays an important role in statistical tests with discrete dependent variables, such as logistic regression.[3] It is the dependent variable rather than the independent variable that is critical because the error or residual distribution, which based on y_i [remember $e = \left(y_i - \hat{y}_i\right)$], is used in the inference process.

Ordinal scales with few categories (2,3, or possibly 4) and nominal measures are often classified as discrete and are analyzed using binomial class of statistical tests, whereas ordinal scales with many categories (5 or more), interval, and ratio, are usually analyzed with the normal theory class of statistical tests. Although the distinction is a somewhat fuzzy one, it is often a very useful distinction for choosing the preferred statistical test, especially when you are starting out.

Type of Dependent (Response) Variable (or Scale)	**Level of Measurement**	**General Class of Statistic**	**Examples of Statistical Procedures**
Discrete	nominal, ordinal with 2, 3, or 4 levels	Binomial, multinomial, Poisson	chi-square, logistic regression
Continuous	ordinal with more than 4 categories, interval, ratio	normal	t-test, ANOVA, regression, correlation

Classifying the independent and the dependent variable as continuous or discrete will determine the type of analyses that are likely to be appropriate in a given situation.

		Dependent Variable	
		Discrete	**Continuous**
Independent variable	**Discrete** (binary and categorical)	Chi-square Logistic Regression Phi Cramer's V	t-test ANOVA Regression Point-biserial Correlation
	Continuous	Logistic Regression Point-biserial Correlation	Regression Correlation

[1] Mathematicians will define discrete variables more generally in a way that will include many if not most of the variables that social scientists view as "continuous" in common practice. For example, Hays (1994) gives "If a random variable can assume only a particular finite or a countably infinite set of values, it is said to be a discrete random variable." (p. 98)
[2] As we will discover later, the Pearson chi-square test really uses a normal distribution as an approximation, but the binomial (or multinomial) distribution is central to most statistics used with categorical dependent variables. I have placed chi-square with the binomial theory class of statistics, therefore, because the normal distribution is really just used as an efficient substitute for the binomial distribution.
[3] As we will discover soon, the *z*-proportions and Pearson chi-square tests really use a normal distribution as an approximation, but the binomial (or multinomial) distribution is central to many statistics used with categorical dependent variables. I have placed chi-square with the binomial theory class of statistics, therefore, because the normal distribution is really just used as an efficient substitute for the binomial distribution.

Ordinal Analyses

The above contrast between discrete and continuous variables is an oversimplification, and there are good reasons to treat dependent measures with three or four ordinal categories differently. Although in practice, many researchers may use binomial and normal theory statistics when there are few ordinal values on the dependent variable, analyses designed specifically for ordinal variables, such as log-linear models, and ordinal logit and probit models, will provide optimal statistical tests (Agresti, 1984, 2002; Long, 1997; Cliff, 1996; Wickens, 1989). Analyzing ordinal dependent variables with few categories as if they were nominal is problematic (e.g, Pearson's chi-square), because important information about ordering is lost; whereas analyzing these (few-category) ordinal variables as if they were continuous risks inaccurate statistical tests because of violations of error distributional assumptions.

Controversies and Common Practice

There is a longstanding debate about how to classify measurements and whether levels of measurement can be a successful guide to the choice of data analysis type (e.g., Borgatta & Bohrnstedt, 1980; Michell, 1986; Townsend & Ashby, 1984; see Hayes & Embretson, 2012; Scholten & Borsboom, 2009 for recent discussions of the controversy). [4] In my view, there are two relevant issues that should be distinguished from one another. The first issue is a more philosophical concern about whether psychological (or other social) phenomena can be reliably and validly represented by numeric ordinal data. I happen to believe that there is a wealth of psychometric research that has already established this to be the case (e.g., Bendig, 1954; Symonds, 1924; Matell & Jacoby, 1971), but I will leave this controversy to those more qualified to consider deep epistemological dilemmas (which are, frankly, often over my head).

The second issue is more of an empirical or statistical question about whether scales, such as Likert-type scales, with some sufficient number of several ordinal response options will provide accurate results when normal distribution statistical tests (e.g., t-tests, ANOVA, OLS regression) are used. There seems to be fairly good evidence from simulation studies that suggests that if there are 5 or more ordered categories there is relatively little harm in treating these ordinal variables as continuous (e.g., Johnson & Creech, 1983; Kromrey, & Rendina-Gobioff, 2002, Muthén & Kaplan, 1985; Zumbo & Zimmerman, 1993; Taylor, West, & Aiken, 2006). There appears to be added benefit to additional ordinal values up to some point (at least to 7- or 9-point scales). Note that this distinction applies to the dependent variable used in the analysis to the response categories used in a survey whenever multiple items are combined (e.g., by computing the mean or sum)—a composite measure that will have many values and will usually be considered continuous. There are other concerns that are important, such as the distribution of the variable. Normal distribution statistics, such as OLS regression and ANOVA, are remarkably robust to small or moderate departures from normality if sample sizes are even moderate (e.g., N = 20-40; Myers, Well, & Lorch, 2010; Stonehouse & Forrester, 1998), more substantial departures can usually be addressed with robust approaches (e.g., robust standard errors in regression or structural modeling, bootstrap estimates). When sample sizes are small and distributions are highly nonnormal, nonparametric tests (e.g., Mann-Whitney U test, Kruskal-Wallis ANOVA) may be optimal (see Sheskin, 2011, for an extensive list).

Problems with Crude Categorization and Artificial Dichotomization

One needs to be careful about converting continuous variables into dichotomous or categorical variables. One example is the practice of doing a "median split," which puts those with scores above and below the median into two categories, but other methods of artificial categorization can be just as problematic. Generally, a great deal of useful information is discarded, but other statistical issues arise. Although many papers have been published as far back as the 1940s on this topic, the practice of dichotomizing continuous variables is still quite prevalent. A paper by MacCallum, Zhang, Preacher, and Rucker (2002)

[4] My intention is not to try to resolve the debate, but to offer a general simple heuristic as a starting place for deciding which type of analysis is used in common practice in the social sciences for general types of dependent variables. In reality, there are a number of other factors that must be considered in deciding on the most appropriate and statistically accurate analysis, including the distribution of the dependent variable, whether it is count data, and sample size among others. Think about the system I propose here as a kind of analysis triage or grand organizational scheme and trust that I will cover some of the caveats and other special considerations as we go along.

is a superb overview of the problems and potentially serious consequences of this practice. The same principal generalizes to reducing the number of ordinal outcome values or dichotomizing ordinal variables because of low frequency (Murad, Fleischman, Sadetzki, Geyer, & Freedman, 2003).[5] Although these concerns are usually raised about analysis strategies, the same concern could be raised, of course, about our choices of measures when we design a study if we use too few ordinal response options to assess an underlying continuous variable.

References and Further Readings

Agresti, A. (1984). *Analysis of ordinal categorical data.* NY: Wiley.
Agresti, A. (2002.) *Categorical Data Analysis, second edition*. NY: Wiley.
Bendig, A. W. (1954). Reliability and the number of rating scale categories. *The Journal of Applied Psychology, 38*, 38±40.
Borgatta, E.F., and Bohrnstedt, G.W. (1980). Level of measurement - Once over again. *Sociological Methods and Research, 9*, 147-160.
Cliff, N. (1996). Ordinal methods for behavioral data analysis. Mahwah, NJ: Erbaum.
DeCoster, J., Iselin, A. M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological methods, 14*(4), 349-366.
Hayes, H., & Embretson, S.E. Psychological measurement: Scaling and analysis. In H. Cooper; P.M. Camic, D.L. Long, A.T. Panter, and D. Rindskopf (Eds). *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 163-179) Washington, DC: American Psychological Association.
Hays, W.L. (1994). *Statistics, Fifth Edition*. Belmont, CA: Wadsworth.
Johnson, D.R., & Creech, J.C. (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review, 48*, 398-407.
Kromrey, J. D., & Rendina-Gobioff, G. (2002). An empirical comparison of regression analysis strategies with discrete ordinal variables. Multiple Linear Regression Viewpoints, 28(2), 30-43.
Long, J.S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.
MacCallum, R.X., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 1-40.
Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study. *Educational and psychological measurement, 31*, 657-674.
Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398_407.
Murad, H., Fleischman, A., Sadetzki, S., Geyer, O., & Freedman, L. S. (2003). Small samples and ordered logistic regression: Does it help to collapse categories of outcome?. *The American Statistician, 57*(3), 155-160.
Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189.
Myers, J. L., Well, A., & Lorch, R. F. (2010). *Research design and statistical analysis*. New York: Routledge.
Scholten, A. Z., & Borsboom, D. (2009). A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology*, *53*(2), 69-75.
Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures, fifth edition*. Raleigh, NC: Chapman & Hall/CRC Press.
Stevens, S.S. (1946). On the theory of scales of measurement. *Science,* 103, 677-680.
Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. Journal of Experimental Psychology, 7, 456-461.
Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and psychological measurement*, *66*(2), 228-239.
Townsend, J. T. and Ashby, F. G. (1984), Measurement Scales and Statistics: The Misconception Misconceived, *Psychological Bulletin, 96*, pp. 394-401.
Wickens, T.D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.
Zumbo, B.D., & Zimmerman, D.W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34,* 390-400.

.

[5] DeCoster and colleagues (2009) discuss a few of the exceptions in which fewer crude categories may be useful.