II.A. SHOULD I MEASURE MY VARIABLE AS CONTINUOUS OR CATEGORICAL?

Which scenario is worse?: (a) a dependent variable (DV; e.g., consumption) measured as a continuous variable, even though you know this is potentially an unreliable and invalid measure of your focal construct, but a continuous scale enables the use of more powerful parametric modeling techniques; or (b) the construct measured more cautiously, resulting in the DV only available as categorical data, and thus nonparametric modeling approaches are required.

My view on this is that ultimately the reliability and validity of the data, and the analysis conducted, are the most important issues. I used a very simple measure of my construct in my research, because the pilot of a more complex tool indicated that certain social classes were overestimating their consumption (from the literature and from a validation measure included in the original tool). Hence, the data were categorical. To model the variable, I used log-linear analysis (to define underlying relations between categorical variables) and then logistic regression and discriminant analysis to develop and confirm the models.

Someone argued that perhaps overestimation was not such a great problem, because the constant term in the multiple regression equation takes care of this. (Is that right?) If overestimation was by a uniform amount, then multiple regression took account of this, and, therefore, multiple regression was the appropriate statistical tool. My argument was that the overestimation on the DV was not uniform across all social class groups, and, therefore, this was not a feasible route. Although my rationale is fairly sound (I think), I raise this question, as I imagine this must be an issue for many who need a measure of consumption, which theoretically lends itself to being measured as a continuous variable, but in reality a less thorough measure has to be used.

Professor Ulf Böckenholt
University of Illinois

This question is interesting because it touches on a variety of issues which, however, are not primarily of a statistical nature. In my opinion, the main challenge here is to specify how people arrive at their judgments. For example, one important question is whether the judgment is based on a recall or an inference process. An extensive discussion of these issues can be found in the book series on survey questions by Schwarz and Sudman (1996). The choice of a statistical tool (i.e., regression vs. log-linear models) is of minor importance in this context.

Asking for categorical judgments may simplify the response task but does not necessarily lead to more accurate results. For example, Schwarz (1996) demonstrated that the reported number of hours people spend on certain activities (e.g., watching TV during 1 week) depends heavily on the category labels of the response scale. His explanation for this finding is that people anchor their judgments on the response categories when they cannot or do not feel like recalling the actual amount of time spent.

REFERENCES


Professors Eric Bradlow and Wes Hutchinson
University of Pennsylvania

Well, in general, the “garbage in, garbage out” truism applies. So, if the continuous DV is truly unreliable and invalid, do not use it. Presumably, this question is directed at more legitimate tradeoffs when the continuous variable has some degree of reliability and validity, but appreciably less than that of the dichotomous variable. The strength of dichotomous data based on a high validity probability model is that the main source of error is sampling error, whose characteristics are well known and whose statistical power is often higher than one may expect. The relative value of the “richer” continuous data can be assessed using standard analyses of statistical power if its only problem is reliability (i.e., an unbiased source of error). However, if measurement problems create biases, then the best way to use such data is to explicitly model the biases, the
sources of error, and the underlying model of interest. Such modeling, however, can be nonstandard and complicated. A good example of the value of this approach can be found in recent work on the calibration of subjective probabilities that models both sampling and response error (e.g., Juslin, Olsson, & Bjorkman, 1997).

REFERENCE

Professor Ramya Neelamegham
University of Colorado

As you correctly point out, it is much better to use more reliable and valid data. This is especially true for the problem you highlight in which there are readily available models and software for analysis of categorical data. We consider two scenarios.

Scenario 1: The continuous DV is measured with error. However, there is no systematic pattern in the size of this error across different social groups (i.e., measurement error is uncorrelated with the independent variables).

Let $Y^*$ denote the true DV. We wish to model the relation between different independent variables (e.g., social groups) and $Y^*$. Assume there are two social groups; we use a dummy variable ($D_1$) to denote the first group, and the second group will be treated as the base. (e.g., see Greene, 1997, pp. 379–389, or Hardy, 1993, for a description of dummy variables).

We wish to estimate:

$$ Y^* = \alpha + \beta_1 D_1 + e $$

(1)

The observed data, however, are error prone. Let us denote the observed measure by $y$ and the error in the observed data by $\delta$. Thus, the relation between the true measure and observed measure is given by (as per “classical test theory” described in any book on measurement)

$$ y = Y^* + \delta $$

(2a)

The model we estimate using the observed data is

$$ y = \alpha' + \beta_1' D_1 + e. $$

(2b)

The question you raise is fundamentally this: What is the relation between the coefficient estimates in Equations 1 and 2b? Substituting 2a in 2b we get:

$$ Y^* + \delta = \alpha' + \beta_1' D_1 + e, $$

so

$$ Y^* = (\alpha' - \delta) + \beta_1' D_1 + e $$

Thus, the constant term in Equation 2b is adjusted compared to the constant term in Equation 1 by $\delta$. If this error is positive (e.g., respondents’ overestimation of consumption), we would expect the estimated intercept term to be greater than the true intercept.

The intercept usually interests us far less than the slope parameter. The least squares estimator of $\beta_1$ in Equation 1 is given by

$$ \beta_1 = \frac{\text{Cov}(Y^*, D_1)}{\text{Var}(D_1)} $$

(3)

Similarly, the least squares estimator of $\beta_1'$ in Equation 2b is given by

$$ \beta_1' = \frac{\text{Cov}(y, D_1)}{\text{Var}(D_1)} $$

(4a)

where $\sigma^2_0$ denotes the variance of the measurement error $\delta$.

The standard error of $\beta_1'$ is the square root of the variance of $\beta_1'. Clearly, the standard error of $\beta_1'$ is greater than the standard error of $\beta_1$. This result implies that the larger the variance of the measurement error ($\sigma^2_0$), the harder it will be to detect significant effects for $D_1$ given error-prone data. Thus, measurement that is not precise only hurts you in that the likelihood of obtaining significant results is lessened—the tests become conservative.

These results suggest that even in the case of similar levels of overestimation by different social groups, if you do not know the variance of the measurement error ($\sigma^2_0$), you may be better off using the categorical measure that is more valid and reliable.

Scenario 2: Now imagine that the continuous DV is measured with error, and in addition, there is a systematic pattern in
the size of this error across different social groups (i.e., measurement error is correlated with the independent variables).

As in the first scenario, we denote the measurement error by \( \delta \). In this case, however, \( \delta \) varies by social group. We state this analytically as

\[
y = Y^* + D_1 \delta.
\]  
(5)

As stated for the first scenario, when fitting a model like Equation 1, \( \beta_1 = \text{Cov}(Y^*, D_1)/\text{Var}(D_1) \).

Using observed data \( y \) (given in Equation 5) and fitting a model like Equation 2b, we obtain the coefficient estimate for the slope to be

\[
\beta_1' = \text{Cov}(y, D_1)/\text{Var}(D_1)
\]

\[
= \text{Cov}(Y^*, D_1 + D_1 \delta, D_1)/\text{Var}(D_1)
\]

\[
= \text{Cov}(Y^*, D_1)/\text{Var}(D_1) + \text{Cov}(D_1 \delta, D_1)/\text{Var}(D_1)
\]

\[
= \text{Cov}(Y^*, D_1)/\text{Var}(D_1) + \delta \text{Var}(D_1)/\text{Var}(D_1)
\]

\[
= \beta_1 + \delta
\]

Hence, in this case the estimated coefficient varies from the required coefficient estimate by \( \delta \); \( \delta \) represents the difference in the overestimation error by different groups. If an estimate of such error is available, it is still possible to use ordinary least squares (OLS). For example, if you know from previous studies or other measures that Group 2 tends to overestimate their consumption by 10% more than Group 1, we can use this information to estimate \( \delta \) and thereby infer \( \beta_1 \) from the estimated \( \beta_1' \). In the absence of such information, you are, of course, better off using a less error-prone measure of consumption, and if it is categorical, using logistic regression.

Some references that discuss such errors in measurement in the linear model context are Maddala (1989, chap. 11) and Bollen (1989, chap. 5).

**REFERENCES**


**II.B. SHOULD I ANALYZE FREQUENCIES OR PROPORTIONS?**

Suppose a researcher hypothesizes (H1) that participants will engage in more thinking during advertising processing in Condition A versus Condition B. The researcher further predicts (H2) that beyond this difference in the amount of thinking, participants in Condition A will experience more thoughts of a particular nature (called Type 1). Consider the following based on various cognitive response indexes:

<table>
<thead>
<tr>
<th>Thought Index</th>
<th>Condition A</th>
<th>Condition B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1 thoughts</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>All other thoughts</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Total thoughts</td>
<td>8.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Although a test of H1 would focus on whether total thoughts differed between conditions, less straightforward is how H2 should be tested. One approach could involve testing whether the number of Type 1 thoughts differed between conditions. Alternatively, one could make this comparison using the proportion of Type 1 thoughts (Type 1 thoughts divided by total thoughts). In this example, the two approaches would lead to different conclusions about whether the data support H2. What are the relative merits of these two approaches? More specifically, is it true that proportions should be used whenever significant differences exist between conditions in the total number of thoughts?

Professor Prashant Malaviya
University of Illinois at Chicago

The general answer is, no, it is not imperative that proportions be used. The correct method presumably depends on the specific hypothesis that is being tested. It would appear that there are two general hypotheses that one could consider: A researcher could make predictions about the relative amounts of different kinds of thoughts, or the predictions could be about the absolute amounts of each kind of thought.

To accurately test predictions about the relative amounts of various thoughts in different experimental conditions, proportions with respect to the total amount of thoughts generated would be required. Thus, such an analysis would be appropriate for hypotheses that say that irrespective of the amount of thinking that a respondent has engaged in, what is the relative level of valence among thoughts (i.e., relatively more positive than negative thoughts are listed; Greenwald, 1968), the relative depth of processing revealed by thoughts (i.e., relatively more detailed than superficial thoughts; Malaviya & Sterngal, 1997), or the relative types of elaboration that are prompted (i.e., relatively more item specific than relational thoughts; Malaviya, Kisielius, & Sterngal, 1996)?

In contrast, hypotheses that deal with absolute differences between various kinds of thoughts would not require normalization with respect to the total number of thoughts generated. Thus, such an analysis would be indicated when the responses are used to assess the accuracy of stimulus learning (i.e., use thought data to determine the level of accurate recall of a message's contents) or when valenced thoughts are used to determine overall level of evaluations (i.e., thoughts are used as...
another measure of evaluation; Cacioppo & Petty, 1979; Mackie & Worth, 1989).

REFERENCES


Professor Deborah Roedder John University of Minnesota

Whether absolute or relative measures of thoughts should be used depends on the particular theoretical prediction being forwarded. In this case, the prediction appears to be that Type 1 thoughts will be proportionately greater in Condition A than in Condition B. Although access to the specific prediction and theory would be useful here, it seems as though the issue is whether the thoughts of participants in Condition A are more focused toward Type 1 thoughts than those of participants in Condition B. This is a proportional hypothesis.

Is it true that proportions should be used whenever significant differences exist between conditions in the total number of thoughts? This question is analogous to asking whether the popularity of a particular brand in one area of the country versus another should be measured by absolute sales figures in both areas or by the market share in both areas. What is the answer? It depends on what you are exactly looking for and why. If you are asking me where we have the greatest sales volume, than I would base my comparison on the absolute sales figures. If you are asking me where we have the greatest market penetration, than I would base my comparison on the market share figures. In sum, one needs to look at the predictions, experimental context, and purpose of the measures to make the call.

Professor Kent Grayson London Business School

There are two ways to address this question. The first approaches the problem from a theory development perspective. The second approaches it from a more practical implementation perspective.

First, from a theory development perspective, one of the problems often faced in the evaluation of experimental data is the potential influence of alternative explanations. Researchers must do their best to eliminate alternative influences in their experiments, and good reviewers must do their best to envision the role of alternative influences. Therefore, before analyzing these data (and probably before collecting the data in the first place), the researcher must consider the potential influence of a correlation between total thoughts and Type 1 thoughts. In this experiment, there is a high correlation between total thoughts and Type 1 thoughts. Does this reflect what one might expect in the real world?

If there is a strong theoretical argument for a universally high correlation, then the rise in Type 1 thoughts can be explained by the rise in total thoughts, and vice versa. Thus, H1 and H2 are confounded. In this case, the researcher must reformulate the hypotheses by eliminating one (or perhaps combining them into one), and it does not matter whether the raw numbers or percentages are used.

However, if there is a strong theoretical argument that the correlation is universally low (or more specifically, that the correlation will change depending on the condition), then H1 and H2 are not confounded. In this case, the researcher must develop a convincing theoretical argument that the correlation is low or variable, perhaps also with support from pre-tests. Second, the researcher’s conditions must clearly create situations in which the correlations are expected to be low, different, or both.

However, the researcher’s argument for a variable correlation would be assisted by the creation of additional experimental manipulations that interact with Conditions A and B to create different data patterns. Given only two cells in which the ratio of Type 1 thoughts to total thoughts is 50%, it is difficult to argue that the ratio may vary. The argument would be strengthened if, in one or two additional cells, the ratio was significantly different from 50%.

Another way of saying this is that the data in the example reflect only a main effect. Compared with interactive effects, main effects are more vulnerable to alternative explanations. The researcher may argue that there are more Type 1 thoughts in Condition A versus Condition B, but someone may argue that it is not so much that the conditions produced different levels of Type 1 thoughts, but rather that the conditions produced different levels of overall thinking. Depending on the nature of the research, this alternative explanation may undermine the experiment.

Second, with regard to practical implementation, another question about the analysis of these data relates to whether the researcher is interested in raw numbers or in percentages. This may seem like a circular answer to the question of whether to use raw numbers or percentages, but the answer to many data analysis questions should be based on the researcher’s goals and hypotheses.
To illustrate what I mean, I think it is easier to consider an example using more tangible entities that thoughts. Suppose a researcher is measuring the emissions from two oil refineries. This researcher is interested in two things: (a) total emissions and (b) amount of a smog-producing emission called VOC. The data are as follows:

<table>
<thead>
<tr>
<th>Refinery X</th>
<th>Refinery Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of VOC (tons)</td>
<td>20</td>
</tr>
<tr>
<td>Total emissions (tons)</td>
<td>100</td>
</tr>
</tbody>
</table>

As in the thought data, the researcher has the option of comparing the raw numbers (20 vs. 40) or the percentages (20% vs. 20%). This question depends on why the hypotheses are being proposed. A researcher assessing environmental damage may care more about the raw numbers: Refinery Y pollutes the environment more than Refinery X. In this situation, the data about total emissions is not central to the research question, although it can help to put the researcher's conclusions about amount of pollution into context.

On the other hand, a researcher assessing the process efficiencies of each refinery may care a great deal about the amount of VOC in relation to the total because this indicates the relative efficiencies of the emissions controls in each. Because each refinery produces the same percentage of VOC, the researcher may have evidence that the two refineries have similar efficiencies. Although in both cases there is still a potential confound, the practical nature of the hypotheses reduces the negative effects of this confound.

## II.C. ANALYSIS OF CONSTANT SUM SCORES

We often see data collected using constant sum scales. That is, people are asked to allocate, say, 100 points to some number of possible response variables. If each variable was a Likert-type scale, we might use a simple one-way analysis of variance (ANOVA) to examine differences between mean responses. But, these responses are dependent on one another. Does this pose a problem to analysis? What is the appropriate way to analyze differences in the mean number of points allocated using constant sum scales?

---

**Professor Jan-Benedict Steenkamp**

*Catholic University of Leuven, Belgium*

This is a real problem and one of the reasons why I hardly use constant sum scales. You cannot use them in a regression either, unless you omit one category. I would analyze such data using ANOVA or t tests, with one category left out. Of course, the data for k - 1 categories are still (somewhat) dependent on each other (if you have five attributes, and Attribute A gets 40 points, you know that the importance of B is 60 or less), but the strict mathematical dependence is still more limited than for the last attribute. Moreover, I would tend to use a more conservative p value.

By the way, in my extensive research experience with real consumers, I have typically found that consumers find it really hard to complete that task. Their numbers hardly ever add up to 100 (unless you use a computer-interactive program with the restriction built in that the responses have to add up to 100; however, in that case, respondents get irritated very quickly and start to reduce points for attributes at random, so I do not feel that is a good way to go). It is obviously easy to rescale the answers to 100, but it does mean that the dependence is further reduced.

---

**Professor Greg Allenby**

*Ohio State University*

A good way to think about data from constant sum scales is as outcomes from a repeated choice process. For example, a survey of physicians may ask them to allocate 100 points across 10 commonly prescribed drugs, with the points reflecting the frequency that they prescribed the drugs to the last 100 patients with a particular disease. This kind of data leads to the use of discrete choice models for the analysis of the data (e.g., a logit model), with the choices weighted by the points. One can then analyze the choices themselves (or the probabilities underlying the resulting multinomial outcomes) or attempt to model the probabilities in a conjoint-like (or regression-like) fashion. In the former analysis, the observed outcomes are point estimates of the multinomial probabilities (with variance $p(1 - p)/n$) and in the later, more sophisticated analysis, for example, maximum likelihood methods must be used.

---

**Professor Sachin Gupta**

*Northwestern University*

The two key characteristics of the data that need to be taken into account in the test are (a) each response is a $k \times 1$ vector, and (b) the $k$ responses of each participant are nonindependent because they sum to 100. Consider the $k$ responses as probabilities $(p_1, p_2, \ldots, p_k)$. The general approach to the test is to transform this vector of probabilities into the unbounded real space and then apply a standard test such as Hotelling $T^2$. At least two alternative transformations are possible:

1. Compute the $(k - 1)$ vector of cumulative probabilities; that is, $(p_1, (p_1 + p_2), \ldots, (p_1 + p_2 + \ldots + p_{k-1}))$. Transform this vector of cumulative probabilities using the inverse of the cumulative distribution function of the standard normal; that is, obtain, $M = (\Phi^{-1}(p_1), \Phi^{-1}(p_1 + p_2), \ldots, \Phi^{-1}(p_1 + p_2 + \ldots + p_{k-1}))$. 

---
2. Use the logit transformation to obtain the \((k-1)\) vector:

\[ L = (\ln(p_1/p_k), \ln(p_2/p_k), \ldots, \ln(p_{k-1}/p_k)). \]

Now, use the Hotelling \(T^2\) to test for differences in mean values of \(L\) or \(M\). Both transformations recognize the two characteristics of the data noted previously. The Hotelling \(T^2\) test assumes normality of the data. The test is discussed in standard multivariate texts such as Morrison (1976). The logit transformation may be derived from the log probability model.

REFERENCES


of variances, and more important, covariances, which will take into account how the categories are related. The results will reflect the relative popularities of the categories, incorporating their special interrelations (due to the constraint of the summed score) as best as possible into the model.

REFERENCE


A related question follows.

II. D. ANALYSIS OF FREQUENCIES DISTRIBUTED ACROSS CATEGORIES

Suppose I have observations each of which can fall into only one of four categories. How do I test for differences in proportions of membership in each of the categories? I have seen a chi-square that provides a test of the null hypothesis that all categories are equal. However, I have seen no test for differences between specific categories.

Here is an example for 100 observations. I can classify the observations as follows:

<table>
<thead>
<tr>
<th>Category</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category A</td>
<td>11</td>
</tr>
<tr>
<td>Category B</td>
<td>13</td>
</tr>
<tr>
<td>Category C</td>
<td>35</td>
</tr>
<tr>
<td>Category D</td>
<td>41</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

A chi-square test may tell me that the proportions in all categories are not equal. What I want to know is if the proportion of the population in Category D is larger than that in Category C. Also, because I could do six possible pairwise tests with this data, is there an appropriate control for family-wide error similar to a Bonferroni adjustment in ANOVA?

The chi-square on number of categories minus one (in this case, three) degrees of freedom is computed:

$$X^2 = \sum_{i=1}^{\frac{1}{4}} \frac{(o_i - e_i)^2}{e_i}$$

no matter the actual values of the expected frequencies (equal or not). The degrees of freedom capture the element expressed in the previous question—that these counts are not independent.

As this questioner implies, the chi-square is a sort of global test of fit—addressing the question of whether each of the observed cell counts match our a priori guesses as represented by the expected cell counts, overall, that is, across all four cells. The question asks how to do the more microtests of one cell frequency versus another, analogous to contrasts between subsets of means once one has discovered a significant main effect on some factor in ANOVA.

Your best bet is probably to do a follow-up chi-square (with 1 df) on the pair of cells of particular interest, probably with a conservative alpha level. The question hypothetically poses a particular interest in comparing Categories C and D—if this comparison is indeed the main focus, there is no need to adjust alpha further for multiple comparisons, but indeed, if the analyst is going to be comparing many of the pairs of proportions, particularly in an exploratory manner, it would be easy enough, and recommended, to adjust alpha (e.g., .01/3 = .0033; you would not really want to do all six, they would be redundant). Good thinking. A conservative alpha level does not compensate for the negative correlations among the category counts, but it is a beginning.

The follow-up index simplifies to

$$X^2_f = \frac{(n_C - n_D)^2}{(n_C + n_D)}$$

the math supporting this simplification is found in the Question Appendix following the references for this question.

Alternatively, you may create a confidence interval (say 99% rather than 95%) for the proportions in Categories C and D and check whether they overlap (and so are not statistically different) or not (thus, are significantly different). Each confidence interval would be computed:

$$p + \left( \frac{1}{2n} \pm 2.58 \sqrt{\frac{p(1-p)}{n}} \right)$$

(cf. Minium, 1978, p. 446) for $p_p = p_C$ and $p_d = p_D$, each in turn, the proportions of the 100 points allocated to Categories C and D, and $N$ is the total number of points allocated. In this example, $N = 100, p_C = .35$ and $p_D = .41$, and their respective 99% confidence intervals would range [.232, .478] and [.288,
or, to be conservative (minimizing slightly the difference in frequencies between the two categories), use the equation with a "correction for continuity" as per Snedecor and Cochran (1980, p. 122):

$$X^2 = \frac{(n_c - n_D)^2}{(n_c + n_D)}$$

(This statistic is found discussed in Snedecor & Cochran, 1980, pp. 120–125, in the context of comparing proportions obtained in related samples, analogous to a matched $t$ test of means on dependent samples—in our application, two data points from the same sample. Their example is predicated on a two-way structure, essentially a question of homogeneity of proportions; that is, Categories A through D crossed with Sample 1 or 2, but their argument is fundamentally one that rests on the binomial, a condition that holds here, within the context of our polytomous, multinomial categories.)

II.E. IS A LOGIT ALWAYS BETTER THAN THE ANALYSIS OF PROPORTIONS?

Is a logit analysis, in which zero through one observations are treated as the unit of observational analysis, always superior to cumulating such zero through one observations into appropriate sets, each with its proportion (and then analyzing the proportions as the units of observation)? That is, when are the benefits of the increased sample size of a logit model outweighed by the failure of its assumptions, so that the analysis of proportions, even with the reduced sample size (because there are fewer proportions than individual observations), becomes more appropriate.

Professor Laura Koehly
University of Iowa

When fitting a logit model, or any log-linear model, we are modeling the cell frequencies, and in effect, the cell probabilities. Reducing the data to proportions would definitely be misguided. The assumptions underlying log-linear models include (a) poisson or multinomial sampling models and (b) independent random sampling. The assumption most often violated in the log-linear modeling framework is that of independent observations. Commonly, complex sampling designs, such as cluster sampling and stratification, are employed. There has been considerable work on modifying the inferential procedure to account for dependencies in your responses. One possibility is to correct the degrees of freedom of the test statistic by adjusting the sample size with a design effect factor (Kish & Frankel, 1974). This approach is adopted by Altham (1976), Brier...
and Reitz and Dow (1989). Koch, Freeman, and Freeman
(1975) discussed a weighted analysis approach that can
also be applied. If the analysis involves a logistic regres-
sion model, then the generalized estimating equations ap-
proach, developed by Liang and Zeger (1986), can be used
to adjust for dependencies due to cluster sampling.

REFERENCES

Altham, Patricia. (1976). Discrete variable analysis for in

Brier, Stephen. (1980). Analysis of contingency tables under cluster sam-

Cohen, J. E. (1976). The distribution of the chi-squared statistic under clus-
tered sampling from contingency tables. *Journal of the American Statis-


Kish, Leslie, & Frankel, Martin R. (1974). Inference from complex samples
(with discussion). *Journal of the Royal Statistical Society, Series B*, 36,
1–37.

Koch, Gary G., Freeman, Daniel H., Jr., & Freeman, J. L. (1975). Strategies in
the multivariate analysis of data from complex surveys. *International


Rao, J. N., & Scott, Allen J. (1984). On chi-squared tests for multiway con-
tingency tables with cell proportions estimated from survey data. *Annals of
Statistics*, 12, 46–60.

sample units in contingency tables. *Journal of Mathematical Sociology;
14*, 85–96.

Editor: Depending on your research purposes, there may be a
couple of options. There are simple z tests to compare two
proportions (or one empirical proportion against a hypothe-
sized value). If you had proportions obtained from multiple
groups, you may be tempted to do multiple pairs of such z
tests, but you would need to correct your alpha rate for
computing many nonindependent tests, and furthermore,
you would likely be working within a larger structure (e.g.,
a factorial), and the data would be analyzed with greater il-
 lumination if the design were captured in the character of
the model.

Even if the proportions you were trying to model showed
greater variability than the simple binary values on which your
questions seems focused, statisticians have long counseled that
transformations be taken, so as to minimize issues like the de-
pendence of means ($p$, the proportion) and variances ($p \times (1-p)$)
in these applications (e.g., in preparation for ANOVA, arc sines
of square roots of proportions, or square roots of counts; cf.

Log-linear and logit models would be the easiest to defend
for your data circumstances. User friendly introductions to
the material include Demaris (1992), Feinberg (1981),
and Knoke and Burke (1980). You may find valuable the discus-
sions regarding Questions II.F and II.G in this special issue.

REFERENCES


(2nd ed.). Cambridge, MA: MIT Press.

Knoke, David, & Burke, Peter J. (1980). *Log-linear models*. Beverly Hills,
CA: Sage.

(7th ed.). Ames: Iowa State University Press.

II.F. CAN I DO AN ANOVA ON A
BINARY DV?

I have a 2 x 2 factorial design with a dichotomous DV. A sam-
ple of 200 individuals will be assigned to each cell. I would
like to run an ANOVA on the data, but ANOVA requires the
DV to be measured on a metric scale (continuous). Under
what conditions, if any, may I legitimately run an ANOVA to
analyze a dichotomous DV?

____________________

Professor Ramya Neelamegham
University of Colorado

As you correctly state in your question, when the DV is metric
it is conventional to use ANOVA. In the case of a nonmetric
dV, such as a dichotomous measure, a discrete choice model
such as the logit model is appropriate.

ANOVA is a special case of the general linear model (cf.
regression) that assumes that the DV is continuous. There are
several problems with using such linear models when the DV
is categorical. To illustrate one such conceptual problem, let $y$
denote the DV that takes on either value of zero or one. We
collect data from N participants and for each participant we
observe whether $y$ equals 0 or $y$ equals 1. We wish to explain
the participants' responses using a set of independent vari-
bles, denoted by $X(N \times p) = \{x_i\}, x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$; $X$ can
consistent of a set of variables identifying experimental condi-
tions, individual difference measures, interactions, and so on.
If we wish to use a linear model framework to explain $y$ with
the $X_i$ variables, we use standard notation and obtain (e.g., on
standardized variables)

$$
y_i = \beta_0 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon_i = X_i \beta + \epsilon_i \quad \text{for } i = 1 \ldots N.
$$

Furthermore, in accordance with this linear model, we assume
that $E(y_i) = X_i \beta$ (i.e., the errors are random and typically as-
sumed to be normally distributed and so cancel out). Given that
yi can be either 0 or 1, the expectation of yi is the probability of observing yi = 1; that is, \( E(y_i) = Pr(y_i = 1) = X\beta \). However, \( X\beta \) is not bounded to lie between 0 and 1; hence, with this model we can obtain probability values that are negative or greater than one. Clearly, this is a problem.

Other problems with using a linear model in the discrete DV context relate to the error term, \( e_i \). Greene (1997, pp. 873M–882) described these problems. (He provided citations in Footnote 2 on page 874 that describes situations when it may be possible to use a linear model even with a discrete DV; these circumstances tend to depend on the specific characteristics of the sample.) Given the availability of software and relative ease of implementation to estimate models with dichotomous DVs, it would be much simpler to use a logit model.

For a simple applied discussion of the logit model, see Hair, Anderson, Tatham, and Black (1995, pp. 129–132). This book also provides SPSS and Statistical Analysis System (SAS) programs to estimate logit models (pp. 712 and 720, respectively).

**REFERENCES**


---

Professor Robert Meyer
University of Pennsylvania

If one is interested in testing hypotheses, never. On the other hand, if one's goal is just to get consistent estimates of effect sizes under the assumption that the data are generated by a linear probability model, then ANOVA is fine. But it is not clear when this would arise. Also, ANOVA does not require the DV to be measured on a metric scale—the requirements focus on the behavior of the error structure (which usually are satisfied only with metric scales). I might add that the only exception would be if one has truly small sample sizes, where the large sample properties of maximum likelihood and ANOVA do not really hold—in this case, I would probably do an ANOVA and not attempt to report significance tests.

Professors Eric Bradlow and Wes Hutchinson
University of Pennsylvania

The biggest problem with using standard ANOVA for dichotomous data is that the violations of the ANOVA assumptions will generally result in statistical tests that are too conservative. This is because the tests have less statistical power than the log-linear weighted least squares and maximum likelihood tests used in the standard approaches to modeling categorical data in a contingency table. Most commonly used statistical packages have procedures that estimate the right model, conduct the right tests, and have input statements and output formats similar to those of ANOVA (e.g., Proc Catmod in SAS). It is important to note that the ANOVA model is linear, and the standard model for categorical data is log linear. Thus, the main effects and interactions have different meanings in the two analyses. “Crossover” type interactions affect the interaction terms in both models. However, data generated from a simple main effects linear model will have a nonzero interaction term when analyzed with a log-linear model, and, conversely, data generated from a log-linear (or multiplicative) main effects model will have a nonzero interaction term when analyzed by a linear model. The classic reference on categorical data is Bishop, Fienberg, and Holland (1975), and a good reference on the assumptions made by ANOVA and what to do when they are violated is Miller (1986).

**REFERENCES**


Editor: Other good references on the issues and solutions to this kind of problem include Aldrich and Nelson (1984) and Long (1997). As in Neelamegham's previous exposition, take \( P_i = Pr(y_i = 1) \), and model not \( P_i \), but the ratio \( P_i/(1 - P_i) \), or more specifically its natural logarithm:

\[
\log \left[ \frac{P_i}{(1 - P_i)} \right] = \Sigma \beta_k X_{ik}
\]

we can solve for \( P_i \) and obtain (Long, 1997, pp. 51, 266):

\[
P_i = \exp (\Sigma \beta_k X_{ik})/[1 + \exp (\Sigma \beta_k X_{ik})]
\]

from this logistic function, the \( P_i \)s of which, no matter the values of the predictors (the \( X_{ik} \)s), or the model's \( \Sigma \beta_k X_{ik} \) terms, have the nice property of being constrained to 0 through 1 (Aldrich & Nelson, 1984, p. 32).

Logit (and log-linear) models are probably universally considered superior for binary DVs. They are not difficult to implement or understand. You may find of interest the discussion on Question II.G. in this special issue.

**REFERENCES**

II.G. HOW DO I INTERPRET LOGIT PARAMETERS?

I would like to know how to interpret the output from logit analysis. I am working on a 2 x 2 factorial design that blocks on sex of the respondent. Sometimes the DV will be dichotomous; other times it will be polytomous. I plan to use the SPSS log-linear procedure to perform the analysis—for example, LOGLINEAR DV (1,2) BY SEX (1,2) A (1,2) B (1,2)/DESIGN. I have not run the experiment yet, so I do not have any data to examine. From what I can tell, however, the output from logit analysis is not simple to interpret. Any advice would be appreciated.

Editor: Be not afeared! Interpreting parameters from logit models is no more complicated (or simple) than doing so for b weights in OLS (i.e., regular ol') regressions.

To work with a concrete example, say your DV is brand choice (0 = no, 1 = yes, the focal brand was selected); gender (coded, 1 = male, 2 = female); Factor A was a between-subjects manipulated variable representing whether the consumer was exposed to a print advertisement for the brand that featured a verbal presentation (0) or numeric information (1), and Factor B was a between-subjects measured segmentation variable, say a median split identifying the consumer's level of expertise in the category (1 = novice, 2 = expert).

If you ran your model and obtained results something like

\[
\text{Predicted choice} = .07 \text{ gender} + .78 \text{ ad} - .54 \text{ expertise} + .36 \text{ (ad \times expert)}
\]

you would conclude that type of ad, expertise, and their confluence contributed significantly to your prediction and understanding of the consumers' choices, and that there were no gender differences. Specifically, the numeric ad is more effective than the verbal one in contributing to consumers choosing this focal brand; as consumers become more expert in the category, they are less likely to choose this focal brand; and the combinations of experts and numeric ads, and novices and verbal ads, were most effective in enhancing this brand's selection.

I like to think of these coefficients as "likelihoods" (not using the word in any technical sense). The coefficients are clearly not "probabilities," for, as in this example, they can be negative. But "likely" captures the meaning of "probably" without the restriction of a range from zero to one. Therefore, a large (significant) negative coefficient indicates an unlikely effect (e.g., it was less likely that experts chose this brand; or more precisely, for your data, this brand was less frequently chosen among the experts compared with novices), a large positive coefficient indicates a pattern in the data that is very likely, or a combination that was found with relatively greater frequency in your data (e.g., the numeric ad seemed to contribute to respondents choosing this brand, or more of your respondents chose the brand when they had been exposed to the numeric ad, compared with respondents who had seen the verbal ad), and a small (insignificant) coefficient indicates an effect that is no more or less likely than chance (chance is not necessarily 50-50, but depends on the proportions of sample respondents with each combination of independent variable characteristics).

By way of background, three observations may help. First, imagine a smaller example than yours—a 2 x 2 table, crossing brand choice (no-yes) with gender (male-female). We may test a hypothesis of independence (i.e., no gender brand preference differences) with a chi-square, where the expected frequencies were computed in the usual manner:

\[
\frac{(n_{ii}) (n_{jj})}{n_{++}}.
\]

Note that this formula is multiplicative. Much of statistics is linear, meaning the terms are added (or subtracted) but not multiplied (or divided); it usually simplifies the optimization procedures, and it probably fits theoretical conceptualizations better as well. To transform this model to a linear function, we take the natural logarithm of both sides:

\[
\ln(e_{ij}) = \ln(n_{ii}) + \ln(n_{ij}) - \ln(n_{++}),
\]

with the result that the equation is now linear (i.e., additive) in the log scale—hence, it is log linear. (When there are many predictor terms, the model is "long linear." @)

Second, in this 2 x 2 example, we can discuss an "odds ratio" (OR), a direct comparison of the odds that a person will choose the brand given that he is male (\(O_{yes,male}/O_{no,male}\)), versus the odds that the person will choose the brand given that she is female (\(O_{yes,female}/O_{no,female}\)). If gender has no influence on brand choice, these two odds will be approximately the same, so a ratio of the two, the OR, \(OR = ([O_{yes,male}/O_{no,male}]/[O_{yes,female}/O_{no,female}]) = ([O_{yes,male}]/[O_{no,male}])/([O_{yes,female}]/[O_{no,female}])\) should be about one. (Or, better, compute Yule’s Q = (OR – 1)/(OR + 1), which is bounded between -1 and 1, therefore interpreted similarly to a correlation coefficient; Feinberg, 1981.)

Third, now we are ready to understand the logit. This first point gives us a sense about why logit models are in a log scale, and this second point gives us a sense of how comparisons are made. A logit, then, is a log odds ratio: \(\ln(O_{brand,chosen}, given predictors/O_{not chosen, given predictors})\). Essentially, the logic of a logit is a logistic prediction that says, "for each combination of the predictor variables (gender, numeric or verbal ad, expert or novice), where is the brand choice more (and less)
CONTINUOUS AND DISCRETE VARIABLES

likely" or observed with greater or lesser frequency in these data (cf. Iacobucci & McGill, 1990).

There are many good introductions to log-linear models and logit models, including Aldrich and Nelson (1984), Demaris (1992), Feinberg (1981), Kennedy (1983), Knoke and Burke (1980), and Menard (1995). Bishop et al. (1975) is helpful for more complicated matters.

REFERENCES


II.H. TREATING AN INDIVIDUAL DIFFERENCE PREDICTOR AS CONTINUOUS OR CATEGORICAL

Suppose I want to test an interaction between a situational manipulation and an individual difference variable. The individual difference variable is measured on a continuous scale. A simple example would be one in which the DV is a measure of information search, and the independent variables are time pressure (two levels) and need for cognition (expressed as a score between one and seven). There are different ways to test the interaction. A common practice is to do a median split on the individual difference measure (e.g., high need for cognition [NFC] and low NFC participants), and analyze the resulting 2 x 2. However, (a) any split is arbitrary (e.g., I could also do a three-way split and keep only the very high and very low NFC participants), and (b) any split results in a loss of information.

Is it better to retain the continuous measure? I assume that I should run a regression analysis with (in this case) time pressure, NFC, and the interaction term as predictors. The interaction is significant if the beta on the interaction terms is significantly different from zero. A significant interaction would mean that the NFC slopes in the high time pressure and the low time pressure conditions are different. But, how do I continue?: (a) Can I do something akin to simple effects tests?, (b) How do I report the results (verbally and graphically), (c) What do I do if I have two individual difference measures? (or, is there a world between ANOVA [which I know a bit] and LISREL [which I do not know at all]?, and (d) NFC and other individual difference measures are measured with error. In an ANOVA context nobody ever seems to worry about that. Still, I suspect that the measurement error makes it less likely to discover a true effect. Is there any way I can take this into account and increase the likelihood to uncover the true effect?

What if my situational variable is like in the previous problem? For example, I am interested in how high and low NFC consumers react to emotional versus informational ads. I want to use real ads, which have been scored as informational or emotional by a number of judges. Ads are nested in ad type, which can be between-subjects or within-subjects. Is it still possible (or advisable) to treat NFC as a continuous variable? How does it work?

Professor Alice Tybout
Northwestern University

The primary basis for my reaction is philosophical. If one's theorizing leads to predicting an interaction that is expected to follow a particular form, then follow-up contrasts are an important aspect of testing the theory. Furthermore, such follow-up contrasts cannot be conducted without some arbitrary "cutting" of the continuous variable into discrete categories. Because ANOVA requires that such cuts be made at the outset of the analysis, I feel that using ANOVA makes for a simpler presentation of the data than does regression. If one were to use regression, the overall analysis would be performed treating the variable as continuous, but the continuous variable would need to be cut before any follow-up contrasts on a significant interaction could be conducted (see Professor Neelamegham's discussion later). Also, as a practical matter, a researcher who relies on ANOVA seems more likely to conduct and report the follow-up contrasts because they can be specified as part of the basic data analysis. It is my observation that researchers using regression sometimes simply report the significance of effects in the basic model and fail to perform further analyses.

The typical argument against the ANOVA approach is the one that you mention: Any split is arbitrary and results in the loss of information. This is certainly true, but as I indicate earlier, splitting the data is unavoidable if follow-up contrasts on an interaction are to be conducted, and I believe that it is absolutely essential to conduct and report follow-up contrasts to support a theoretical interpretation of results. When cutting the data, a median split is a conservative approach. If the variable has an effect, but it is a subtle one, the effect may go undetected. Other splits, such as upper or lower quartiles, may increase the likelihood of finding difference if the pattern is linear and the sample size is large, but even this approach may fail to detect a nonlinear relation. And, unfortunately, if the
pattern is nonlinear, it also may be undetected by regression, which assumes linearity.

One approach that I use to address these problems is to begin with a simple plot of the data. This information helps me decide what cuts to make. If the data look linear, I tend to take the conservative median split approach, especially if my sample size is relatively small. If no effects emerge, but the pattern is as anticipated and my sample size permits, I may explore more extreme splits, such as contrasting the top versus the bottom third of the sample. However, I cannot recall any instance when this approach actually helped obtain significance because of the tradeoff with sample size.

As outlined previously, I prefer an ANOVA presentation (both as an author and as a reviewer), but I have encountered reviewers who have the opposite preference. When this occurs, the simplest solution seems to be to conduct the second analysis requested by the reviewer and report it as a footnote. You can see an example of this approach in my article with Joan Meyers-Levy on context effects (Meyers-Levy & Tybout, 1997, Footnote 3).

REFERENCE


Professor Ramya Neelamegham
University of Colorado

This question raises issues regarding the nature of an interaction and simple effects tests for continuous data in a regression modeling framework. To express the model analytically, let us define a few terms. Let Y denote the DV (e.g., information search). Let XI denote the categorical independent variable (i.e., time pressure). Let the continuous individual difference measure (e.g., need for cognition) be denoted by Z. We wish to ascertain if there is an interaction between XI and Z and detect the size and nature of such an interaction.

In the ANOVA, the data are fit via a model that tests for the main effects of XI, Z, and their interaction. If the F statistic for the interaction is significant, the experimental behavioral researcher knows to examine plots of cell means and to run further statistics to test the simple effects. The analogous procedure within regression may be less familiar, so it is worth reviewing. To determine whether there is an interaction between XI and Z, estimate the following two models:

\[ Y = a + b_1XI + b_2Z \]  
\[ Y = a + b_1XI + b_2Z + b_3XZ. \]

A t test for the statistical significance of \( b_3 \), or an F test for the difference in the multiple Rs for the two models, reveals whether there is an interaction between XI and Z.

To determine the strength of the interaction between XI and Z, examine the difference in the squared multiple correlation for Equations 1a and 1b. This difference reflects the strength of the interaction. If, for example, \( R^2(1a) = .74 \) and \( R^2(1b) = .96 \), then the strength of the interaction is \( .96 - .74 = .22 \); that is, the interaction effect accounts for 22% of the variance in the Y variable. For our example, this would mean that the interaction between need for cognition and time pressure accounts for 22% of the variance in information search.

The nature of the interaction is understood in ANOVA via testing simple effects. To conduct an analysis in the regression context akin to simple effect tests, we wish to see the change in the effect of XI on Y at different levels of Z. In other words, we want to know, What happens to \( b_1 \) at different levels of Z? Does \( b_1 \) remain the same or does it change at high, medium, and low values of Z? To answer these questions, we calculate \( b_1 \) at different values of Z. In certain cases, theory may suggest the values of Z that are most relevant. In the absence of such theoretical bases, it is possible to calculate the value of \( b_1 \) at low, medium, and high values of Z, where low may be defined as 1 standard deviation below the mean of Z, medium could be the mean value for Z, and high could be 1 standard deviation above the mean of Z (e.g., Jaccard, Turrisi, & Wan, 1990).

Following Jaccard, Turrisi, and Wan (1990), we calculate \( b_1 \) and the standard error of \( b_1 \) at the low, medium, and high values of Z using the following equations:

\[ b_{1@Z} = b_1 + b_2Z \]  


standard error of \( b_{1@Z} = \sqrt{\sigma_r^2 + Z^2\sigma_b^2 + 2Z\text{cov}(b_1,b_2)}. \]  

Please see Jaccard, Turrisi, and Wan for further details.

For example, if we estimate the model specified in Equation 1b and obtain the following parameters

\[ Y = 2.5 + (1.5)XI + (.7) Z + (-.8) X^2Z \]

and assume further that the mean of Z is 4 and the standard deviation is 2.5. Then, we have the low, medium, and high values of Z to be 1.5, 4, and 6.5, respectively. The values of \( b_1 \) (as per Equation 2a) at these three values of Z are 2.55, 4.3, and 6.05, respectively (i.e., \( b_1 = 1.5 + .7(1.5); 4.3 = 1.5 + .7(4); 6.05 = 1.5 + .7(6.5) \)). Similarly, it is possible to calculate the standard error of \( b_1 \) at each of these levels of Z using the computer output from the regression analysis. Assume we found these to be .98, 1.8, and .98, respectively. These calculations can be presented in the following table. We examine the t statistics (\( b_1 \) over its standard error), to determine whether they are statistically significant.
from zero and also whether the nature of the change in slope at different values of NfC. In this example, the slope increases as NfC increases. These t tests are analogous to simple main effects analysis in ANOVA.

<table>
<thead>
<tr>
<th>Need for Cognition</th>
<th>( b_1 ) (Effect of Time Pressure on Info Search)</th>
<th>SE</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>2.55</td>
<td>.98</td>
<td>2.60</td>
</tr>
<tr>
<td>Medium</td>
<td>4.30</td>
<td>1.8</td>
<td>2.39</td>
</tr>
<tr>
<td>High</td>
<td>6.15</td>
<td>.98</td>
<td>6.28</td>
</tr>
</tbody>
</table>

Results could either be reported in table form, as done earlier, or in a graphical manner by calculating the value of \( Y \) (information search) at different values of NfC and time pressure. In the following, we calculate the value of \( Y \), as given in Equation 3, at time pressure equals zero and one, and \( Z \) at high, medium, and low values. For example, as depicted in the graph in Figure 1,

For time = 0 and high NfC,
\[ y = 2.5 + (1.5)0 + (.7)6.5 + (-.8)0 = 7.05 \]

For time = 1 and low NfC,
\[ y = 2.5 + (1.5)1 + (.7)1.5 + (-.8)1.5 = 3.85 \]

In the case of more than one individual difference variable, we can carry out exactly the same analysis as previously carried out. We modify the model to include \( Z_1 \) and now an additional variable, \( Z_2 \):

\[ Y = a + b_1 X_1 + b_2 Z_1 + b_3 X_1 \times Z_1 + b_4 Z_2 + b_5 X_1 \times Z_2 \quad (4) \]

A t test for \( b_5 \) reveals whether there is a significant interaction between \( X_1 \) and \( Z_2 \). We can compare the multiple correlation between this model and a model that does not include the interaction term to detect the strength of the interaction term. All other analyses can be carried out in the same manner as for the case of a single individual difference variable. For examples of models with multiple continuous moderator variables, see Jaccard, Turrisi, and Wan (1990).

Regarding your questions on the impact of measurement error, you may wish to refer to Maddala (1989, chap. 11) for a discussion of the effects of measurement error in a linear regression context. A great book to learn more about the effects of measurement error in the context of interactions is Jaccard and Wan (1996).

The effects of measurement error in a single variable in a linear regression context are easily illustrated. We also present a methodology that allows us to deal with such errors and uncover the true effects. Consider a linear regression model with one independent variable (\( X \)) that is measured with error.

In other words, we wish to estimate the following model:

\[ Y = a + b_1 X_{\text{true}} + e \quad (5a) \]

However, we do not have measures for \( X_{\text{true}} \). We have error-prone measures for \( X_{\text{obs}} \) denoted by \( X_{\text{obs}} \). If we denote the measurement error by \( \delta \), we obtain the following relation:

\[ X_{\text{obs}} = X_{\text{true}} + \delta \quad (5b) \]

Substituting for observables in Equation 5a, we obtain

\[ Y = a + b_1 (X_{\text{obs}} - \delta) + e = a + b_1 X_{\text{obs}} - b_1 \delta + e \quad (5c) \]

It can be shown that the coefficient estimate \( b_1 \) in Equation 5c underestimates \( b_1 \) in Equation 5a (Maddala, 1989, p. 381). The extent of underestimation depends on the ratio of the variance of the measurement error to the variance of the true \( X \) variable (i.e. \( \sigma_\delta^2 / \sigma_{X_{\text{true}}}^2 \)). Unless we have an estimate of this ratio, it is not possible to uncover the true effects. One way to obtain an estimate for this ratio is to use factor analysis. Note that this is only possible if we have multiple indicators for \( X_{\text{true}} \). Such data are typically available for individual difference measures such as NfC. We can use the structure of either exploratory or confirmatory factor analysis to obtain an estimate of \( \sigma_\delta^2 / \sigma_{X_{\text{true}}}^2 \). Neelamegham and Jain (1999) demonstrated this procedure in the context of discrete choice models. The methodology they suggest works as follows:

1. Estimate a factor analysis model (exploratory or confirmatory) using all the available indicators to obtain a measure of the unobservable \( X_{\text{true}} \).
2. Use the output of factor analysis to obtain an estimate of the variance of the latent factor (this is the \( X_{\text{true}} \) variable such as need for cognition) and the variance of the measurement error (\( \sigma_\delta^2 \)).
3. Account for this measurement error in the estimation of the model of interest.

If I understand your question regarding a more complicated model correctly, let us say you have four levels of advertising (emotional ads 1 and 2, rational ads 1 and 2, and say...
these are between-subjects levels of a factor for simplicity),
time pressure (two levels), and a continuous individual differ-
ence variable. You can still treat NfC as a continuous variable.
The model independent variables are (a) time pressure
dummy denoted by XI; (b) ad dummy variables: D1, D2, D3;
and (c) NfC variable denoted by Z.

As an example, if we were to assume no interaction be-
tween time pressure and advertising dummy variables (just to
keep the presentation here simple), we would obtain the fol-
lowing model:

\[
Y = a + b_1X_1 + b_2D_1 + b_3D_2 + b_4D_3 + b_5Z +
Z(b_6X_1 + b_7D_1 + b_8D_2 + b_9D_3) + e
\]  

(6)

The difference between this model and the model specified
in Equation 1b is only additional dummy variables and in-
teraction terms. All the analysis steps remain the same as
before: (a) We conduct tests of significance of interactions
between NfC and the time pressure and advertising dummies exactly as before, using t tests on the b weights; (b)
comparison of multiple correlations between nested models
allows us to ascertain the strength of different interactions;
(c) to examine the nature of interactions and report these
verbally and graphically, first express Z in mean deviation
form and estimate the model specified in Equation 6. It is
then possible to construct a table with different values of b1
when Z is at the mean level and XI, D1, D2 and D3 take
values of zero or one; (d) next, calculate the low value of Z
and subtract this score from each value of Z; in effect, we
are transforming Z and creating a new variable. Now, esti-
mate Equation 6 with this transformed Z and XI, D1, D2
and D3 taking different values; again we obtain a series of
and calculate b1 when Z is at the low level and XI, D1, D2
and D3 take values of zero or one; (e) conduct a similar
analysis using the high score for Z. Thus, we have three se-
ries of b1 values for different levels of Z, D1, D2, D3 and
XI; (f) at each stage, in a similar fashion, we can calculate
the values of Y for different values of Z and different levels
of XI with D1, D2, and D3. Although this analysis would
definitely be more laborious, the logic remains the same as
for the simpler case cited earlier (cf. Jaccard, Turrisi, &

REFERENCES

Jaccard, James, Turrisi, Robert, & Wan, Choi K. (1990). Interaction effects in

Jaccard, James, & Wan, Choi K. (1996). LISREL approaches to interaction

Macmillan.

experience goods: An econometric model and analysis. Journal of Mar-
keting Research, 36, 373–386.

Professor Julie Irwin
University of Pennsylvania

This question includes two topics: explicating interactions
among continuous predictors and splitting data into parts. The
two topics are linked by common statistical practice:
Most researchers know how to interpret interactions among
categorical predictors but feel less comfortable interpreting
interactions when one or both predictors are continuous. As
the questioner notes, a common temptation is to divide the
continuous variables (often at their median) and then run the
regression (or ANOVA) on the categorical data. The interpre-
tation of interactions is addressed in other sections of this spe-
cial issue (e.g., Question VI.E.), so this answer will
concentrate on the common practice of rendering continuous
data discrete.

Philosophical issues

The true underlying nature of personality traits was at one
time a prevailing interest in psychology, and the issue of
whether (and which) personality traits were classifiable in
continuous versus categorical terms has enjoyed especially
heated debate. The statistical treatment of a personality trait is
a statement about the researcher’s beliefs about how those
traits appear in nature. Probably most traits are truly continu-
ous, although there have been convincing arguments that cer-
tain traits may exhibit discrete characteristics (Gangestad &
Snyder, 1985). The fact that individual difference measures
usually are measured on a continuous scale in the first place
seems to be a signal to leave them as they are. It is a bit odd to
measure NfC (for instance) on a continuous scale and then
convert it to a binary measure. Certainly, the researcher would
need to support such a choice theoretically (using evidence
that a continuous construct has been proven to be incorrect—
arguably a difficult task).

Statistical Issues

Reduction of power for detecting main effects. Reg-
ardless of whether dichotomization reflects the true state of
the world, there are some statistical difficulties associated
with dividing continuous data into groups. For main effects
tests, dichotomizing variables severely reduces power (e.g.,
Cohen, 1978, 1983; Humphreys & Fleishman, 1974; Peters &
Van Voorhis, 1940). In the simplest case, in which variables Y
and X are sampled from a bivariate normal distribution, the
correlation \( r \) between \( Y \) and \( X \) is reduced to .798* \( r \) when \( X \) is
split into two equal-sized groups (Cohen, 1983; Peters & Van
Voorhis, 1940). The statistical reasons for this reduction are
fairly intuitive: The values of \( X \) close to the median are espe-
cially likely to end up on one or the other side of the median.
purely by error. When X is dichotomized, these values “count” as instances of their category just as much as do the extreme values of X, thus obscuring the true relation between X and Y.

**Spurious significance levels for interaction models.**

The reduction of power to detect main effects is problematic, but not fatal. Far more damaging is the effect of dichotomization in models with interactions. Dichotomizing the continuous components of an interaction can lead to biased estimates of both the main effects and the interaction, and the bias induces Type 1 error (e.g., Maxwell & Delaney, 1993). Dichotomizing can lead to significant interaction coefficients in which there is in fact no interaction and spuriously augmented main effect coefficients.

This counterintuitive result is driven by correlations among the predictor variables and is worse the more correlated the predictors are (for details, see Maxwell & Delaney, 1993). It is highly likely that behavioral predictors in a regression model will be correlated, thus increasing the likelihood for Type 1 error. Obviously, this finding has serious implications for the extant literature and for the judgment of current research findings.

**Summary and Recommendations**

Regardless of one’s philosophical taste for dichotomization, the statistical issues are clear cut. Dichotomizing continuous variables is likely to lead to biased estimates. These biases are not always conservative (i.e., dichotomizing does not always reduce power). Thus, it makes more sense for researchers to leave continuous variables continuous, especially when interactions are involved.

**REFERENCES**


Editor: We all know that ANOVA and regression are special cases of the general linear model, but we often operate as if regression is from Mars and ANOVA is from Venus. Each model is formulated to illuminate slightly different aspects of research questions that share the form, What is the nature of my predictors’ impact on my DV? Although both models perform optimally under the condition of having a continuous DV, ANOVA applications are those with discrete, categorical predictors or factors, and regression applications are those with continuous explanatory variables. In ANOVA, we compare means and assess relative sources of variance; in regression, we examine slopes and absolute amounts of variance explained.

In our field, the discipline in which a researcher was trained goes a long way to explain whether their natural orientations and intuitions come in the ANOVA or regression form. (When you dream, do you dream in 2 x 2 plots of means, or betas?) Behavioral researchers with a psychological orientation tend to feel more comfortable in the ANOVA framework, whereas researchers with an economic background tend to prefer regression—these are the respective models on which we were weaned in graduate school, and familiarity lends a great deal to confidence and facility with a method. This implicit correlation is also related to the kind of research question posed most often by the researcher—the psychologist conducts laboratory or survey experiments seeking to compare sources of variation to refute theoretical predictions versus the economist who models survey or secondary data bases to assess how much variability has been explained to make statements about effects estimations and predictions.

My comment regarding one’s confidence with the tools of one’s trade is not intended as a “feel good” observation—confidence in one’s methodological and statistical abilities helps the researcher communicate clearly and precisely to the intended audience. In our collective efforts to progress our science in both journal articles and conference presentations, we must understand our analytical framework to convey our methods and results clearly. In doing so, we must be sensitive to the approaches that are most likely familiar to the recipient audience. Behavioral researchers versed in ANOVA who are receiving findings in a less familiar framework will have to spend their cognitive resources on the methods, or else they may not comprehend or, worse, might find suspicious, the intended substantive theoretical advance.

Familiarity does not dominate if another method is proven superior—then we all must take notice and learn the new methodology. However, when we take continuous variables and treat them as discrete factors, we usually do so because (a) we care more about the comparison between two extreme means from a theory testing perspective, and the shape of the functional form of the points in between seems superfluous; and (b) we believe that in making a continuous scale dichotomous, the inherent “loss of information” (e.g., the reduction in the correlation cited by Irwin) only makes our subsequent statistical tests more conservative, so if we err, we do so at our own peril. (Though on this point, results in Maxwell &
Delaney, 1993, suggest caution against using two or more dichotomized variables in the same model.

Powers of tests are affected by several things, including the effect size being studied, which may be truncated as Irwin described. In addition, we usually acknowledge that power is affected by sample size or, specifically here, degrees of freedom. Neither of these tests is inherently weaker on this criterion: The test of the difference between two means in ANOVA or the slope coefficient in regression use a single degree of freedom in the numerator and comparable error degrees of freedom (depending, of course, on the other terms in the models).

Ultimately, the trade-off shapes up to be one of the continuous predictor's somewhat greater statistical power versus the dichotomous predictor's somewhat greater ease of interpretation and diagnosticity regarding theory testing. Defenders of the former would argue that a test using a dichotomous treatment cannot be adequately diagnostic if it is not sufficiently powerful. Defenders of the latter would argue that even if one were to work within the continuous framework, at some point the data need to be dichotomized in some arbitrary manner anyway to demonstrate the nature of the results (e.g., in plots of means).

Perhaps the two camps can meet if we more often used the analysis of covariance (ANCOVA) model, which is, of course, a natural blending of ANOVA's categorical factors and regression's continuous scales. Whether one thinks of this model as an ANOVA with some continuous variables thrown in (perhaps also interacting with the factors) or as a regression that includes some dummy variables (to represent the factors), it can powerfully address research questions like that posed here. After results—main effects and interactions—are deemed significant or not, follow-up analyses may proceed by examining contrasts on main effects and simple effects in interactions, as is commonly done in ANOVA, or as presented nicely here by Neelamegham in regression terms—whichever vantage is most likely to seem compelling to the intended audience.

Finally, although whatever information we obtain from one method (e.g., regression slopes) can be coaxed from the other (e.g., ANOVA means), it can mean doing some serious convolutions and gyrations. A nice introductory book that helps translate between ANOVA and regression is Allen (1979). Additional references on obtaining ANOVA-like results from a regression model, in addition to Jaccard, Turrisi, and Wan (1990), include Jaccard and Wan (1995) and Jaccard, Wan, and Turrisi (1990).

REFERENCES


A related question follows: I examined an ANCOVA with age as a covariate (a linear relation was assumed), obtained no significant relation between age and the DV (inferences about physician accountability), and concluded that there was no relation. Afterward, I was rereading the literature and on reflection felt that the relation might be approximated better by a step function (less than 25 years vs. 25 and up). Well, the effect of age is now significant. This result would seem to suggest that one ought to carefully consider the expected functional form of the relation when one does significance tests.

Editor: This scenario is a very nice example of a theoretically-driven distinction in this matter of whether a variable should be treated as continuous or discrete. It would be nice to have a sense of these particular data (e.g., a scatterplot) to understand why the continuous treatment did not yield the relation that the binary variable did.