Some of the Most Common Questions Asked of Statisticians - Our Favorite Answers and Recommended Readings

Allison, David B., Gorman, Bernard S.[1]

Contents

[1] *Genetic, Social & General Psychology Monographs*, 119(2), 1992:

ABSTRACT. We addressed the following 16 of the most common questions or concerns encountered when consulting with applied researchers: detecting and managing outliers; handling missing data; multiple comparisons and family wise alpha rate protection; the disadvantages of dichotomization; the nature of the general linear model, writing the results section; determining the number of factors to retain in factor analysis; power analysis; parametric versus nonparametric statistics; appropriate numbers of predictor variables for use in multiple regression; variable selection procedures in multiple regression; interpreting beta weights; graphic display of data; interpreting interaction effects; alternative analyses for pretest-posttest control group designs; and choosing between multivariate and univariate procedures. We offer brief responses, not exhaustive theoretical expositions, but, we believe they will help fellow consultants, teachers, and researchers to answer their own questions and those of their consultees, students, and associates.

0. Introduction ↑

IN DISCUSSIONS about research design and statistical analysis with applied researchers in the social, behavioral, and medical sciences, we have observed common questions or concerns among many different consultees. We have written this article for researchers with some background in statistics and research methods, who prefer to be and should be involved in both the conceptualization and conduct of the analyses and often wish to be directed to readings that are relevant and comprehensible to someone with only modest statistical training. Our purpose in this article was not to introduce new statistical techniques but to offer some guidance on commonly confronted issues.

Herein, we have described the 16 most commonly encountered questions or concerns, offered brief responses, and listed our favorite sources recommended to clients. These responses are intended to be brief and not exhaustive theoretical expositions. As such, interested readers will often find themselves wanting more detail than we have provided on particular questions. This is where our "favorite" sources come in. Sources have been selected using several criteria: brevity, recency, and, most important, readability. As brevity is a consideration, articles or chapters have generally been preferred over books. In some cases, an exceptionally useful piece of software has also been cited.

1. How do I find out if I have outliers and what do I do if I have them? ↑

To be able to interpret results of statistical analyses unambiguously, one needs to insure that the results are not distorted by the presence of outliers. Outliers are cases (usually subjects) that come from different populations than do most of the other cases in the sample. Some investigators also consider unusual observations that result from erroneous data generation, collection, or transcription procedures to be outliers. The general principle governing outlier detection methods is that extreme scores occur rarely. Therefore, cases with extreme scores are likely to be outliers. However, it is important to note that it is not possible to definitively determine whether an extreme case is actually an outlier or "just an extreme case." We will return to this point shortly.

The first step in checking for outliers is to examine the univariate distributions. This can easily be accomplished by obtaining frequency distributions, histograms, and box plots from

most major statistical packages. Cases that stray far from other cases in graphic displays or are more than three standard deviations from the mean are suspect as univariate outliers.

After checking for univariate outliers, the next step is to check for bivariate outliers. Bivariate outliers are cases that stray far from the "swarm" of other cases in two-dimensional space. They can frequently be identified visually by examining scatter plots. Numerically, just as a case more than three standard deviations from the mean may be a univariate outlier, cases with standardized residuals from the regression line may be bivariate outliers. Again, most major statistical packages will produce scatter plots and compute standardized residuals. It is important to remember that even though a case is not a univariate outlier on either variable, it can still be a bivariate outlier. For example, imagine a survey respondent who reports being 12 years old and earning $30,000 per year. Neither figure alone is unusual but the combination is quite rare.

One may also wish to check for multivariate outliers. We focus here on outliers in the context of multiple regression, as this is probably by far the most used multivariate statistical technique. However, most of these procedures are germane to other analytic strategies. Three types of outliers can occur in multiple regression: outliers on the criterion, outliers among the predictors, and outliers that have undue influence on the regression equation. When checking for outliers on the criterion, one simply needs to compute the multiple regression equation and calculate standardized residuals with any standard statistics package. As in the bivariate case, cases with standardized residuals greater than three (absolute value) are possible outliers. Stevens (1984) offers a more sophisticated but still readable treatment of this topic including significance tests for outliers on the criterion.

Regarding outliers on the predictors (or for that matter, any set of variables), one can obtain a measure of the distance of each case from the centroid of all cases. The centroid is a multivariate average of all variables (i.e., the coordinates describing the center of a "swarm" in the p-variate hyperspace, where p is the number of predictors). The measure of distance is referred to as Mahalanobis' D[sup2] and is a generalization of a z-score. Tabachnick and Fidell (1989) point out that Mahalanobis' D[sup2] can be interpreted as a chi-square with p degrees of freedom and recommend testing each D[sup2] at p = .001. Cases with significant D[sup2]s are considered potential outliers.

Finally, Cook's Distance (CD) is a measure of how much the regression equation would change if the case under consideration were dropped. CD is a measure of the joint influence of any case on both the criterion variable and the predictors. In this way, CD may be the most important indicator of potential outliers, because it indicates how much a case will influence the results of analysis. Cook and Weisberg (1982) suggest that a CD of about 1.0 be considered large.

Detecting potential outliers is all well and good. However, what does one do with them once detected? First, we recommend careful checking of the raw data for entry, transcription, coding, and transformation errors. This often accounts for a substantial amount of presumed outliers. If outliers are not merely the result of some data handling error, four broad options are available: (a) ignore them; (b) eliminate them; (c) transform them (e.g. Winsorize, see Cook & Weisberg, 1982); or (d) perform the analyses both with and without the outliers.

As we mentioned earlier, there is no sure way to tell if an outlier is actually from a different population or just extreme. Therefore, we do not favor elimination of cases. Rather, of the four options, we decidedly favor the fourth alternative. In the event that both sets of analyses give essentially the same results, this can be mentioned and results obtained with the outliers can be reported. In our experience, this is almost always the case with samples of any

reasonable size. In the event that different results are obtained with and without the outliers, then both can be reported.

Regarding sources, virtually all the plots and statistics we have been discussing can be easily obtained from any of the major statistical packages (e.g., SPSS, SAS, BMDP) and we therefore recommend their use in detecting outliers. For investigators using factor analysis, a program by Comrey (1985) may be helpful in identifying multivariate outliers that may affect the factor solution. For readings, we highly recommend Stevens (1984) as a thorough and comprehensible source. Other helpful sources are Barnette (1978) and Johnson (1985).

2. How do I handle missing data? ⇧

Missing data is one of the most common problems encountered in research. Although survey and archival studies are probably most prone to this, the problem can also be present in experimental research (Welch, Frank, & Costello, 1983). The appropriate handling of missing data involves a two-stage decision process.

In the first stage, one must decide whether or not the data are believed to be missing at random. The assumption that data are missing randomly, that is, that "the available data and missing data for each item [variable] are each random subsets of the data for the complete sample" (Hertel, 1976, p. 460) is essential to the use of "imputation" methods described below. Imputation can be defined as the estimation of a missing value and the subsequent use of that estimate in statistical analyses.

Although there is no sure way to determine if data are missing randomly, two heuristics are available. One is a rule of thumb suggesting that if "too many" data are missing from any one variable, the data should not be assumed to be missing randomly. In this case, the variable in question should be dropped if at all possible. What is "too much" missing data is debatable, but Hertel (1976) suggested a 15% cut-off point. That is, if 15% or more of subjects are missing data on any one variable, then the variable may be excluded from the analysis.

A somewhat more sophisticated method, devised by Cohen and Cohen (1983), consists of dummy coding a new variable for the presence or absence of missing data on the variable in question. This dummy variable can then be entered into correlations or regression equations as a predictor of other variables. To the extent that the dummy variable is correlated with other variables, data cannot be assumed to be missing randomly.

In the event that some data are missing but are believed to be missing randomly, several methods of handling this problem are available. The simplest methods involve deleting subjects having missing data. Listwise deletion entails excluding a subject from any analysis in which he or she is missing a value for any variable involved in the analysis. The advantage of this method is its simplicity.

The disadvantage is that, if the amount of missing data is at all substantial and multivariate procedures are in use, it will result in a substantial loss of subjects and, consequently, power.

In pairwise deletion, a correlation or variance-covariance matrix is computed "by using for each pair of variables (Xi, Xj) as many cases as have values for both variables" (Cohen & Cohen, 1983, p. 278). Multivariate procedures can then be performed on the resulting matrix. This method has the advantage of not losing any data. The disadvantage is that it is possible for the resulting matrix to be somewhat "ill-conditioned" or even a matrix that could not possibly occur with real data. This is particularly likely if data are missing in nonrandom ways. Pairwise deletion should be used with extreme caution, particularly if more than a small portion of data is missing.

Missing data imputation methods are generally superior alternatives to deletion. There are three primary methods for the imputation of missing data. In the first method, referred to as "mean imputation," one simply enters the mean value of a variable for any subject who is missing data on that variable. Although this method is the simplest of the three, it is not recommended, because it will artificially reduce the variability around the mean and potentially attenuate observed relationships among variables in the study. On the other hand, it is a conservative approach, inasmuch as relationships that are not strong will not be found. The reader might adopt this procedure when a bit of "extra" conservatism is desired.

The second method is referred to as "random imputation" or "sequential hot deck imputation" (Little & Rubin, 1987). Although there are several variations of this procedure, the basic method entails randomly ordering the records in one's data file and then assigning "to any missing score the value of the nearest preceding available nonmissing score for that item [variable]" (Hertel, 1976, p. 470). The advantage of this method is that it does not affect the variances of individual variables in a systematic way. However, since it does introduce more random variance ("noise"), it can also attenuate relationships between variables.

The third method is regression estimation. In this method, one computes a regression equation with one or more variables as predictor(s) and the variable with missing data as the criteria. The resulting equation is then used to predict what values the missing data would have taken were it not missing, and these values are imputed. The advantage of this method is that it provides the most accurate estimates of missing values. There are two disadvantages. The first is computational complexity. The second is that when there is a distinction between independent (or predictor) variables and dependent (or criterion) variables, one cannot be used to estimate the other as this would artificially inflate the research findings (Raymond, 1986).

Cohen and Cohen's (1983) dummy coding can be helpful in this process. Specifically, Cohen and Cohen suggest that one might use imputation methods but assess the relations among variables with imputed values after partialing out the effects of missing data with the dummy coded variable. Cohen and Cohen (1983), Hertel (1976), and Raymond (1986) are excellent sources on this topic.

3. If I am doing multiple comparisons, should I use some adjustment procedure to protect my family wise alpha rate and if so which one? ↑

The appropriate handling of the multiple comparison situation has caused considerable consternation among researchers and statisticians alike. In part, this may be because there are two distinct issues at the heart of the controversy, one epistemological, the other statistical. We have tried to graphically depict the decision making process at hand via the decision tree in Figure 1.

On the epistemological level, one must decide if one wishes to protect/control the family wise alpha level. The problem is simple (although the solution is not). Using Fisherian statistics, when we reject the null hypothesis at the prespecified alpha level (e.g., p = .05), we are stating that there is only a 5% chance that we have made a Type I error. If the appropriate assumptions are met, our statement would be correct for any one comparison. However, suppose that in a given study, we are testing 20 comparisons. The probability of making at least one Type I error becomes 1-(.95[sup20]) or .64. Moreover, the expected number of Type I errors is 1.0; many find this unacceptably high.

In contrast, others (e.g., Saville, 1990) argue that

The natural unit is the comparison, not the experiment. An experiment is no more a natural unit than a program consisting of several projects. Clearly, it is unsatisfactory to have the size

of the experiment, or the number of experiments in a project, influencing the probability of detecting a particular pairwise difference. (p.177)

In other words, changing the probabilities associated with a particular hypothesis test because the researcher is testing other hypotheses seems not only irrelevant, but also punishes researchers for ambitious multifactorial studies.

This issue is of great concern to many applied researchers who are equally concerned with Type II and Type I errors (Davis & Gaito, 1984). Neither the logic of the "nonmultiple comparisonist" (e.g., Saville, 1990) nor the classic "multiple comparisonist" (e.g., Tukey, 1977a) is flawed. The only advice we can provide is that each individual researcher must ultimately "cut the Gordian knot," as Saville (1991,p. 167) states, and simply choose an epistemological position. Darlington (1990,chap. 11) presents a cogent discussion of the philosophical issues involved. In the event that the researcher chooses not to "protect" the family wise alpha rate, we recommend simply using the nominal alpha rate for each comparison or using the unrestricted least significant difference (LSD; see Saville, 1990).

In contrast, if one decides to protect the family wise alpha rate, a second set of decisions has to be made. Most often, multiple comparison procedures (MCPs) are discussed in the context of analysis of variance (ANOVA), because this is historically where they were developed (Klockars & Sax, 1986). However, multiple comparisons occur in other situations as well. An illustrative situation occurred when Weiss et al. (1980) studied the behavioral effects of artificial food dyes on the disruptive behavior of 22 children. Weiss et al. conducted randomized single-subject alternating treatments designs (Barlow & Hersen, 1984) with each child. Data were analyzed via nonparametric randomization tests (Edgington, 1987). As 10 dependent variables were separately examined for each of 22 subjects, 220 comparisons were made, raising the familywise alpha rate to 1 - (.95220) or .999987. This figure would hardly be acceptable to anyone concerned about familywise alpha inflation.

In this situation (where tests are truly independent and not part of an ANOVA), the primary MCP available is the Bonferroni correction (Darlington, 1990). The Bonferroni correction is probably the easiest to use. One simply divides the nominal alpha level (e.g., .05) by the total number of comparisons being made. In the Weiss et al. (1980) case, if .05 was the chosen alpha level, the per comparison alpha level would be set at .05/220 or .00023. Although this is an extremely stringent test, for one of the 22 children, 5 of the 10 dependent variables did have associated p values that exceeded the specified level. Thus, for that one child, Weiss et al. could confidently state that the food dyes did have an effect.

Another situation occurs when one is testing differences among means in ANOVA-type designs. Here, there are many procedures available. Jaccard, Becker, and Wood, (1984) thoroughly reviewed all major MCPs that hold the familywise alpha rate at or below the nominal alpha level in terms of a variety of Type I and Type II errors. They review the major alternatives and provide recommendations for between-groups designs, within-groups designs, and mixed designs, under both optimal conditions (equal ns, normality, homogeneity of variance) and suboptimal conditions (unequal ns, nonnormality, heterogeneity of variance). The Newman-Keuls test, the least significant difference test (LSD), the restricted LSD, and Duncan's test were not recommended. Specific recommendations varied with the situation. We refer the reader to Jaccard et al. (1984).

4. If I have a continuous independent variable, should I do a median split and compare "low" and "high" groups to each other? ↑

This is one of the few issues on which there is generally a clear and simple answer: No! In responding to this question, we are reminded of Einstein's oft quoted remark, "Keep things as

simple as possible, but not simpler." The desire to make things simple generally seems to underlie the intention to dichotomize (or otherwise make categorical) continuous variables. However, the outcome of this procedure does not make things simpler at all for four major reasons.

First, dichotomizing continuous variables can drastically lower statistical power (1--beta; or the likelihood of rejecting the null hypothesis when it is actually false). A detailed discussion and proof of this phenomenon can be found in Cohen (1983). Cohen showed that dichotomizing one variable (e.g., the independent) at the mean results in power losses equivalent to discarding 38% of one's subjects. Even greater losses in power occur when variables are dichotomized at points above or below the mean or when the dependent variable is also dichotomized. Given the low power typically available in many researches (Ross), 1990), any practice that unnecessarily lowers power further seems unconscionable. Compensation for this practice would necessitate collecting data on a substantially greater number of subjects. In the long run this can hardly be considered "simpler."

Second, dichotomizing continuous variables into "high" and "low" groups based on median, mean, or other splits of one's sample data in no way insures that groups defined as "high" or "low" correspond to groups so labeled in other studies or in the population overall. Imagine, for example, performing a median split on the distribution of IQ scores among members of the Mensa Society, and labeling the lower half of the distribution as the "low IQ group." The absurdity of this example may be obvious. However, the practice of dichotomization is quite common among other highly selected and nonrepresentative samples (e.g. college students). Thus, dichotomization makes the interpretation of one's basic constructs less simple.

Third, dichotomizing two continuous predictor variables and treating them as independent variables in an ANOVA model potentially creates a nonorthogonal design. If the predictor variables are correlated, ANOVA cell sizes will be unequal, creating interpretive difficulties. Main effects can no longer simply be added together. Although this is equally true in regression analyses, the decision making and interpretive process involved in multiple regression more explicitly acknowledges and was designed for this colinearity (Humphreys & Fleishman, 1974).

Finally, with multiple predictors, dichotomization can play havoc with interaction effects among the predictors. Veiel (1988) has shown quite eloquently that both the magnitude and direction of interaction effects can be greatly dependent on and distorted by dichotomization. Furthermore, the type and degree of distortion will vary with different and essentially arbitrary cut points.

In sum, as Humphreys (1978b) stated, research on individual differences requires correlational analysis, not ANOVA. Although regression and multiple regression may seem more complex, it can be shown that ANOVA is a special case of regression (Pirot & Lustig, 1984; see next section). In the long run, the use of regression/correlation analysis will make research with continuous variables far simpler.

Unfortunately, in our experience, the belief in the appropriateness of dichotomization seems fairly resistent to change. So that skeptical readers may convince themselves (or their research associates), that these are not the opinion of just one statistician, we have cited several excellent discussions of these issues by several different authors (Cohen, 1983; Falzer, 1974; Humphreys, 1978a, 1978b; Humphreys & Fleishman, 1974; Veiel, 1988).

At this point, the reader may wonder "Are there ever situations in which dichotomization is warranted or better?" Our opinion is that the answer is yes when there is a strong theoretical rationale, and the distribution can be shown to be significantly bimodal by an appropriate

statistical test (e.g., Hartigan & Hartigan, 1985). In this case, nature has "dichotomized" the distribution for us and the sample should be split at the nadir, between the two modes, not at the median or any other arbitrarily selected value. We also believe that these situations occur rarely and that many dichotomies based on theoretical distinctions between groups are not supported by inspection of data that prove to be continuous and unimodal. For an example of this argument, see Eysenck (1970) on the value of a depressed/nondepressed dichotomy.

Finally, it should be noted that dichotomizing data after the fact should not be confused with the use of the "extreme groups design" to increase power. Under many circumstances, the latter is an appropriate and powerful method for testing weak to moderate relationships with expensive measures (Feldt, 1961). It should also be noted that even in the extreme groups design, the analysis of variance approach is still inferior to regression/correlation (Abrahams & Alf, 1978; Alf & Abrahams, 1975).

5. What do you mean when you say that ANOVA, regression, t tests, discriminant function analysis, and so on, are really all the same? ↑

These techniques differ on whether variables are categorical or continuous and on how many independent and dependent variables there are. However, they are all subsumed under the general linear model and can be seen as special cases of canonical correlation (Knapp, 1978). Baggaley (1981) intelligibly portrays the relationships among these techniques (and several others) graphically, whereas Tabachnick and Fidell (1989,chap. 13) provide an excellent verbal description of the general linear model.

Canonical correlation analysis (CCA) is a multivariate technique for relating a set of p independent or predictor variables to g dependent or criteria variables. They may be either categorical or continuous (Share, 1984). Multiple regression is a special case of CCA when there is only one dependent variable. When one has only one independent variable, multiple regression becomes simple bivariate regression/correlation. If the one independent variable is a dichotomous categorical variable (e.g., gender), the resulting r is a point-biserial correlation coefficient, rpb. The significance test for this correlation is equivalent to the significance test for a t test and rpb can be converted to t (Rosenthal & Rosnow, 1991). It is well known that F = t2 (in the case of the pooled variance t). Thus, this t test can be seen as a special case of ANOVA with one between-groups factor having two levels. In turn, the ANOVA is a special case of multiple analysis of variance (MANOVA) in which there is only one dependent variable. Finally, we come full circle when we see MANOVA as a special case of CCA in which all independent variables are categorical.

We find that understanding these connections helps one understand the meaning and output of various statistical analyses. In addition, it helps researchers to be more flexible in their choice of data analytic strategy rather than relying rigidly on only one technique.

6. How should I write my results? ↑

This question really requires three different responses at three different levels. First, there is the response at the broadest level, really a response to the question "How do I do good scientific writing?", most often asked by graduate students tackling one of their first research projects. For this we recommend some combination of several sources. One is the APA Publication Manual (American Psychological Association, 1983). In addition, both Rosenthal and Rosnow (1991,Appendix A) and Kidder and Judd (1986,chap. 17) have sections on writing research reports. Checklists of good reporting practices (an excellent example is Maher, 1978) can be helpful reminders, even to experienced researchers.

Often consultees are comfortable writing research reports in general. But on a more specific level, they ask what constitutes good or appropriate "statistical writing" in the results section. Again, the APA Publication Manual is a good source, but our favorite may well be Bailar and Mosteller (1988). Although their article is nominally aimed at medical journals, their suggestions and discussions apply equally well to educational and psychological research.

Finally, researchers are often unsure about how to describe specific statistical techniques and their output. Although this can happen with any technique, it occurs most often with multivariate statistical techniques (e.g., ANOVA, MANOVA, multiple regression, discriminant analysis, etc.). Here, we have found Tabachnick and Fidell (1989) to be an excellent source. At the end of each chapter discussing a specific multivariate technique, a sample report of a hypothetical analysis is presented. The reports are invariably well written, clear, comprehensible, and detailed without being verbose.

7. If I'm doing a factor analysis, how should I decide how many factors to retain? ↑

This is a question that has been the subject of much thought. Although research on this issue will undoubtedly continue, some consensus appears to be developing. At this time, Zwick and Velicer (1986) have published what may be the most comprehensive and up-to-date work in this area. They conducted a Monte Carlo comparison of five of the most common rules for determining the numbers of meaningful factors or components. The five rules were Horn's parallel analysis (PA), Velicer's minimum average partial (MAP), Cattell's scree test, Bartlett's chi-square test, and the Guttman-Kaiser eigenvalue greater than 1.0 rule (K1).

Zwick and Velicer (1986) summarize their findings in the following:

. . . PA was clearly the most frequently accurate method followed by MAP and scree. The tendency of K1 to overestimate was marked. The K1 method never underestimated. The Bartlett test was quite inaccurate and variable . . . (p. 439)

One of the most important points is which rules not to use. Given the inaccuracy of the Bartlett test, it is clearly not recommended. Of even greater note is the performance of the K1 rule. Despite its empirical shortcomings (Zwick & Velicer, 1986) and the fact that it has been shown to be theoretically unsound (Cliff, 1988), it is probably the most commonly used rule. We suspect, like Zwick and Velicer (1986), that this occurs because it is the default procedure in SPSSx, BMDP, and SAS. Readers are cautioned against the blind use of this rule.

In terms of what rules to use, obviously the PA criterion is empirically sound. However, two practical concerns may mitigate against its use. First, the theoretical rationale for the method is complex and may be difficult to communicate to research associates and consumers. Moreover, to the best of our knowledge, no major statistical package incorporates PA (although programs which perform PA have been written; Velicer, Fava, Zwick, & Harrop, 1988). Thus PA may not be accessible to many users.

In contrast, MAP not only performs quite well under most circumstances but has recently been included in an easily accessible user-friendly statistics package (Gorsuch, 1990). Thus our primary recommendation is use of the MAP criterion with Gorsuch's program. For researchers without access to this program, the scree test is simple and easy to apply and is still fairly accurate, particularly when component saturation (the magnitude of the loading of each variable on a component) is high (Zwick & Velicer, 1986). Thus, the scree test represents a good "backup" method.

8. How many subjects do I need ? Do I have enough power? ↑

This is undoubtedly the single most commonly asked question of statistical consultants (Kraemer, 1985). At the risk of being flippant, we have been tempted to respond "Whatever your question about power analysis is, the answer is 82." Although this is clearly an oversimplification, it actually approximates an appropriate response in many (but by no means all) applied psychological researches. Our rationale is as follows. Power is a direct function of sample size (n), alpha levels (alpha), and effect size (delta). Conceptually, effect size is defined as the impact of one variable (or set of variables) on another variable (or set of variables). If any three of these parameters are held constant, the fourth is determined.

Most applied researchers are willing to operate at alpha = .05. Most researchers would also like the power of their investigations to be at least .99; that is, they would like to have a 99% chance of rejecting the null hypothesis if it is in fact false. However, sample sizes for this power level are usually prohibitively large. In practice, Cohen (1988) recommends a power level of .80 and most researchers find this acceptable. The third parameter, effect size, is what Lipsey (1990) refers to as "the problematic parameter." It is on this parameter that the calculation of the required n usually hangs. Although there are many indicators of effect size, we find Friedman's (1982) rm, conceptually equivalent to a product moment correlation coefficient, to be easily interpretable. When we ask applied researchers if they are interested in finding small effects or primarily moderate to large effects, most select the latter. Cohen (1988) defines a moderate effect as equivalent to rm of .30.

Entering Friedman's power tables with a: = .05, rm = .30, and power = .80, we find that the required sample size is 82. The reader may now see the basis for our earlier flippant response. Moreover, the reader may discern our favorite source in this area. Friedman (1982) provides a single table that will easily answer many questions about power analysis and requires the reader to perform little or no calculation. Clearly, we do not mean to imply that Friedman's article will answer all questions, but it is certainly an excellent starting point.

Readers interested in a more general but still brief and readable introduction to power calculations may find Muenz (1989) or Kraemer and Thieman (1987) quite helpful. Bird and Hall (1986) present an excellent source for researchers planning studies involving protected post-hoc comparisons. Bartko, Pulver, and Carpenter (1988) introduce extremely simple nomograms for power analyses involving either paired or independent sample t tests. Finally, for readers interested in the full range and complexity of power calculations, there is no better source than the classic Cohen (1988).

Recently, some helpful software has been developed. Here no one or two sources can be recommended since no program calculates power in all relevant situations. Readers may wish to consult Goldstein (1989) for a review of software and also consider some programs made available since Goldstein's review (e.g. Allison & Gorman, in press; Borenstein, Cohen, Rothstein, Pollack, & Kane, 1990; Darlington, 1990, Appendix 3; Dupont & Plummer, 1990; NCSS, 1991; Rothstein, Borenstein, Cohen, & Pollack, 1990). Researchers requiring frequent power calculations for diverse designs might maintain a potpourri of power software.

9. Which should I use, a parametric or nonparametric test? ↑

Most classical statistical tests are based on the highly useful assumptions that the data have been randomly sampled and are normally distributed. However, researchers are often faced with data sets that contain discrete measurements that are not normally distributed or are not randomly sampled. The critical issue in deciding whether to use parametric or nonparametric tests is whether the assumptions of the parametic tests can be met. Under the leadership of Kendall (1962) and Siegel (1956), interest arose in the use of nonparametric statistics.

According to Siegel and others (Edgington, 1969; Gaito, 1970; Gibbons, 1971, Marascuilo & McSweeny, 1977), nonparametric statistics can provide useful alternatives to parametric tests and, in some cases, provide unique analyses that could simply not be achieved with traditional parametric methods.

Proponents argue that nonparametric tests require fewer assumptions about distributions, especially those of normality and equal variance within groups. Most advocates of nonparametric tests agree that, whereas parametric tests should be used with data that are "truly numeric" (i.e., data that fit interval and ratioscale measurement), because nonparametric tests require only ranking and/or counting, they may be more appropriate for nominal and ordinal scale data.

Given the fact that desk calculators were hardly affordable in the 1950s, early proponents of nonparametrics believed that the less laborious hand computations required by nonparametric tests was an advantage but this is of minor importance in the age of personal computers. McSweeny and Katz (1978) stated that since many nonparametric tests require only rank orders, they are less sensitive to outliers. Some tests of contingency tables and some tests of ordinal data can be performed only as nonparametric tests.

Randomization tests provide a useful class of nonparametric tests (Edgington, 1969, 1987). In general, these tests provide all possible sortings and permutations of an observed data set. By tallying the occurrence of data patterns, the exact probability of specific data patterns can be assessed. Then the pattern of scores in the actual data set can be compared to the frequency of possible chance sortings. Thus, a researcher can offer a statement about the likelihood that the observed data pattern occurred by chance alone.

With very small data sets (i.e. ' 10 observations) it is possible to list all permutations by hand. The amount of computation needed for sorting large data sets can be prohibitively extensive. However, with the advent of faster computers and sampling techniques in which a sample of some but not all possible combinations is used, the task becomes more manageable.

Edgington (1980) and Harwell (1988) stated that when data are randomly sampled, parametric tests are most powerful. However, when a population has not been randomly sampled, even when the subjects are randomly assigned to groups, they argue that nonparametric randomization tests provide more power. Harwell (1988) stated that nonparametric tests do well in controlling for Type I error with non-normal distributions and are more powerful than parametric tests when nonnormality is present. Gaito (1970) added that one can often make probability statements that are "exact" in nonnormal distributions regardless of shape. Furthermore, for samples with as few as 6 cases, there may be no alternative to using a nonparametric test.

Given the claims of the lack of restrictions and flexibility of nonparametric tests, it might be concluded that they should be used under all circumstances. However, it can be shown that the use of nonparametric tests may have serious drawbacks. Gaito (1970), and Harwell (1988) voiced a strong series of warnings against the unselective use of nonparametric tests. Cohen (1965) concurs and does so with such flair that we suspect many readers of his section titled "Nonparametric Nonpanacea" will find it as entertaining as informative.

The reasons for these authors' reluctance toward nonparametric methods are several-fold. First, a large body of research has demonstrated that t and F tests are fairly robust to assumption violations, especially if sample sizes are equal and large (> 30 or so). Moreover, it cannot be said that all distribution-free tests are insensitive to distribution shape differences. For example, the commonly used Wilcoxon/Mann-Whitney U test is more sensitive to skewness and kurtosis than is the t test. Although there have been attempts in recent years to

solve complex designs with nonparametric procedures, Cohen's (1965) argument that there are few tests available for complex designs still holds.

Most important, nonparametric tests often have lower power efficiencies than their parametric analogues. That is, all else being equal, more cases will be needed to reject the null hypothesis with a nonparametric test than with a parametric test. If the assumptions of a parametric test are met, then a corresponding nonparametric test will be less powerful. Finally, when estimates of parameters are needed, then parametric tests must be used.

It can also be shown that a middle ground can be achieved. For example, non-normal distributions can be transformed to normality (see Tukey, 1977b). They also can be ranked and analyzed with more traditional parametric techniques, substituting the ranks for interval and ratio scale scores. Harwell (1988) demonstrated that methods devised by Puri and Sen (1969,1971) provide a conservative method for analyzing complex designs. In these methods, rank orders are substituted for scores, and the data are submitted to parametric analyses. Summary statistics such as F ratios and t values based on the ranked data are then converted to proportions of explained variance measures and tested for significance by conservative chi-square statistics. In a similar vein, Rassmussen and Dunlap (1991) have shown that when data depart from normality, parametric analyses of transformed data result in fewer Type II errors than nonparametric analyses and fewer Type I errors than parametric analyses of raw (untransformed) data.

It appears that if sample sizes are large, if assumptions are not seriously violated, and if data are at least ordinal, then researchers should consider staying with traditional parametric tests. On the other hand, if the data are of nominal scale, the assumptions of parametric tests are severely violated, and specialized small-sample techniques are available, then a nonparametric test might be employed. Alternatively, a researcher might attempt to transform data to meet the assumptions of parametric tests and submit data to more traditional tests or use the Puri and Sen (1969,1971) approach.

We suggest that if the researcher is in doubt, parametric analyses of both raw and transformed data and nonparametric analyses be performed. As with the handling of outliers described earlier, if all analyses give essentially the same results, this can be mentioned and those results obtained with the parametric analysis of raw data can be reported. Again, in our experience, this is usually the case. In the event that different results are obtained with different analyses, results should be interpreted with considerable caution and all should be reported, if only parenthetically, in a footnote.

10. How many predictor variables can l use in multiple regression? ↑

There is no rule written in stone for this response. Theoretically, one can have as many as n--2 predictors, where n is the number of subjects. However, in practice such a rule would result in ridiculously low power (because only one degree of freedom would remain). Moreover, any regression weights and R[sup2] values obtained would be highly unstable.

Stevens (1986) provides an excellent discussion of the issue. He makes a convincing argument that a good rule of thumb is: "No more than one predictor for every 15 subjects." Stevens bases his argument on two main points of evidence. One is Herzberg's (1969) formula for estimating validity shrinkage. Validity shrinkage is defined as the difference in the R[sup2] based on sample data and the R[sup2] that would be obtained using the sample regression equation in the population. Stevens showed that when the sample R[sup2] is .50 (a reasonable estimate for applied research) validity shrinkage begins to become small (i.e. about 12% of the sample R[sup2]) when the ratio n/k is 15, and k is the number of predictors.

Finally, Stevens cites a study by Park and Dudycha (1974) showing that, assuming an R[sup2] of 50, when n/k >/= 15, there is a 90% probability that validity shrinkage will be less than 5%.

11. What are the differences among all these different types of variable selection procedures in multiple regression (e.g. forward, backward, stepwise, simultaneous, hierarchical, all subsets) and which should I use? ↑

Drawing heavily on an article by Hocking (1976), Rawlings (1988,chap. 7) offers an overview of the different purposes of regression, how these different purposes lead to different variable selection criteria, and how the various selection criteria work. Although there is little of Rawlings' presentation that cannot be found in Hocking, we find the former to be a more palatable offering to nonstatisticians. Other readable discussions can be found in Darlington (1968) and Wampold and Freund (1987).

The most important issue in determining which method to use is the reason that the analysis is being undertaken. Rawlings (1988) distinguishes six purposes of regression. On a broader level, we distinguish between three general purposes: a) description; b) model testing; and c) prediction/estimation.

When the object is simple description of the behavior of the response variable in a particular data set, there is little reason to be concerned about elimination of variables from the model, about causal relationships, or about the realism of the model. The best description of the response variable, in terms of minimum residual sum of squares, will be provided by the full model, and it is unimportant whether the variables are causally related or the model is realistic. (p.169)

Thus in this situation, "simultaneous entry" of all variables into the equation in a single step is appropriate. In practice, we rarely, if ever, encounter researchers with this goal in mind.

Often, researchers wish to use multiple regression to test theoretical models. For example, an investigator may believe that physical exercise improves mood solely by improving self-concept; that is, that self-concept completely mediates the relationship between exercise and mood. The investigator surveys 1,000 subjects and derives estimates of the degree to which they exercise, their self-concept, and their "average mood." Under these circumstances, hierarchical regression is the appropriate technique. In this procedure, the researcher decides in what order predictor variables will enter the regression equation based on the hypotheses being tested.

Given this example, the researcher would first enter self-concept into the regression equation (with mood as the criterion variable). Then, exercise would be entered into the equation. If exercise explained a significant amount of variance in mood when self-concept was in the model, the researcher's hypothesis would be disconfirmed. If exercise did not explain a significant amount of variance in mood when self-concept was in the model, the researchers hypothesis would be supported (confirmation is beyond correlational data).

Finally, some researchers are interested in prediction or estimation. For instance, obesity researchers usually need measures of adiposity (fatness) for their studies. Although adiposity can be measured very accurately by dissection, underwater weighing, or dual photon absorptiometry (DPA), such methods are impractical for many investigators. However, skinfold thickness often correlates with total adiposity and is quite practical to measure. A researcher who measures skinfolds at 10 to 20 sites could probably combine these measurements to yield a highly accurate estimate of total fatness. This could be tested by collecting the skinfold measures and correlating them with a more direct measure of adiposity

(e.g., from DPA). Two questions can then be answered in the regression analysis. Are all of the skinfolds necessary or can time and money be saved by dropping some of them without losing any accuracy in the estimation? What are the optimal weights for each of the skinfolds (predictor variables)?

In this situation, two broad alternatives are available--all-subsets regression and stepwise regression. All-subsets regression is exactly what it sounds like. All possible subsets of predictor variables are tried. The researcher may then select the best subset. "Best" is usually defined by a mathematical criterion (Rawlings, 1988, describes several) that tries to achieve the best "compromise" between maximizing R[sup2] and minimizing the number of predictors. One drawback to all-subsets regression is computational demand. As the number of predictors becomes large, computational time can become great even for modern high speed computers. Some computer programs (e.g., SAS) contain an algorithm called the "leaps and bounds" algorithm, which provides an approximation to all-subsets regression with less computational cost.

The major alternatives to all-subsets regression are the stepwise procedures, which include forward selection, backward elimination, and stepwise selection. In forward selection, the variable entered first is the variable with the largest zero-order correlation with the criterion. The next variable entered is that with the largest first-order partial correlation when the first predictor is already in the model. This is repeated until some prespecified "significance level to enter" fails to be reached.

Backward elimination is the reverse. All variables are simultaneously entered into the regression. Then the predictor whose removal would produce the least rise in the residual sum of squares is removed. This process is repeated until some prespecified "significance level to remove" is reached.

The procedure most often labeled "stepwise" is designed to take advantage of the effects that the addition or deletion of one variable can have on the contributions of other variables (Rawlings, 1988). It is a selection process that can switch from forward to backward and back at any step in which the addition or elimination of any predictor will enhance the model.

Two caveats should be mentioned regarding all subsets and stepwise procedures. First, these procedures often capitalize on chance relationships in the sample data and may "overfit" the data. Therefore, when they are used to generate prediction/regression equations it is essential that the resulting equations be validated with independent data (Rawlings, 1988). Second, the ordinary F tests of R[sup2] values are not applicable because a greater number of predictors have actually been "tried" than are included in the tested model (Henderson, & Denison, 1989). These two facts are often ignored. Consequently stepwise regression has been referred to as not only one of the most used but also one of the most misused statistical techniques (Henderson, & Denison).

12. How do I interpret beta weights from regression output? Are they the best measures of the "importance" of predictor variables? ⇑

There are actually two different terms that need to be distinguished here; raw score regression weights and standardized regression weights. The raw score regression weights are simply the optimal values by which to multiply predictor values to obtain estimates of criterion values. They are useful when the regression equation is actually being used to make predictions. However, as they are highly scale dependent, they are unsuitable as measures of the importance of a predictor variable. For example, if one were predicting weight from height, changing the measurement system from inches and pounds to meters and kilograms would

substantially alter the raw regression weight, although there would be no effect on the strength of the relationship.

The standardized regression weight (Beta') is the weight assigned to a predictor variable when all variables in the model have been transformed to unit variance. Although this measure is a reasonable indicator of the importance of predictor variables, it is not the only indicator, and not necessarily the best. Darlington (1968) discusses five possible indicators of the importance of a predictor variable. When predictor variables are uncorrelated, all five indicators are equivalent, and the squared zero-order correlation coefficient (r[sup2]) is probably the easiest to understand and communicate to others. When predictor variables are uncorrelated and causal modeling is at issue (e.g., in path analysis), the standardized regression weight, Beta', may be of greater interest because it demonstrates how many standard deviations the criterion variable will change given a one standard deviation change in the predictor variable.

When prediction is the issue and variables may be correlated, Beta', the standardized regression weight when all other predictors are in the model, is a good indicator of the "uniqueness" or "usefulness" and, therefore, perhaps the importance of the predictor variable. Darlington's (19683 article, though challenging reading, is a valuable source on this issue as well as on several other issues in multiple regression.

13. When should I display my data graphically and what is the best way to do so? ↑

Many computer graphic programs can produce high quality statistical graphics and several statistics packages, database management, and spreadsheet programs have graphics options. It is now possible to produce graphs rapidly and inexpensively. However, the easy availability of statistical graphics software presents a mixed blessing. On a positive note, graphs enable researchers and their readers to see data patterns and relationships that would be difficult, if not impossible, to see in tabular presentations (Anscombe, 1973). However, a poorly chosen or poorly drawn graph may also be difficult to read and may produce visual illusions that distort rather than enhance statistical presentations. It has often been said that a picture is worth 1,000 words. However, when it comes to statistical graphics, we might wish to add the corollary--not if it takes two thousand words to describe the picture!

Several questions should be raised before using statistical graphics. Perhaps the most important is whether the graph displays any unique information. A graph should not be a simple repetition of tabular information. For example, a simple bar chart that presents four group means from a single-factor analysis of variance would not provide any major advantage over a simple table of means. If, however, the graph also displayed error bars containing standard deviations or perhaps confidence limits around the means, then readers can easily understand the relative degree of overlap of the group distributions.

Graphs enable parallel information processing. That is, they present the viewer with many visual dimensions and features simultaneously. Graphs are especially useful in situations that require readers to grasp the meaning of a series of measurements that vary along a dimension in an orderly fashion. Among this class of serial data sets are time series data; as can be found in the single-subject design (Barlow, & Hersen, 1984) and other repeated measures designs (Morley & Adams, 1991); patterns of residuals in regression analysis (Cohen & Cohen, 1983); and the scree test in factor analysis (Gorsuch, 1983) that displays the size of characteristic roots as a function of their extraction order.

Graphs are extremely useful for displaying the shapes of statistical distributions. Although the histogram display has been a mainstay of statistical graphics for more than 100 years,

variations on the stem and leaf plot, the box and whiskers plot and quantile-quantile plots have recently proven to be remarkably useful in understanding distributional patterns.

The term "statistical map" brings to mind the notion of quantities that vary in both amount and density over geographic regions, such as the incidence of psychological disorders in a given state. Mapping, however, does not have to be confined to displaying purely geographic distributions. Maps can also be used to display data over derived statistical dimensions, as has been common practice in factor analysis and multidimensional scaling, where the regional placement of points may reveal much more information than simple tabular displays.

Analysis of variance is probably the most commonly used statistical method for evaluating designed experiments. Many interesting effects, especially in treatment outcome studies, are not found in the ANOVA main effects but, rather, in the interaction effects. Although a display of F ratios and their associated significance levels in an analysis of variance table may reveal the presence of an interaction effect, the meaning of an interaction can be best revealed by one or more two-dimensional plots of treatment means (Rosnow & Rosenthal, 1989; Winer, Brown, & Michels, 1991).

Once the choice of whether to use a graph has been made, the next series of choices forces the researcher to consider strategies for producing the most effective graph. Under the active leadership of John Tukey, William Chambers, and William Cleveland, a science of graphic presentation has started to emerge (Chambers, Cleveland, Kleiner, & Tukey, 1983; Cleveland, 1985; Cleveland & McGill, 1985; Kosslyn, 1985; Wainer, 1984; Wainer & Thissen, 1981). From their work, several principles stand out as essential to effective graphic representation.

According to these authors, the graph should use symbols and devices relevant to the task at hand. For example, if you wish to present a linear relationship, use lines that show the best fit. If you wish to emphasize a possible nonlinear or a curvilinear relationship, then display curves found by smoothing techniques or nonlinear regression. Where possible, provide evidence in the form of confidence bands or error bars that the curve or line is an appropriate fit. If regional clustering of points within a graph is to be detected, then use a scattergram or map.

A graphic presentation should strike a balance between simplicity and complexity. Researchers should place as much useful information as possible in a small space but also try to reduce irrelevant clutter. Tufte (1983) and Wainer (1984) warn users to eliminate "chart junk," which includes features such as icons, fancy plot symbols, and pictures that tend to make the graph look attractive but not necessarily informative.

In statistical graphics, "more" is not necessarily "better". Most computer graphic packages give users a wide range of choices about the density and number of grid lines and the number of curves and legends. However, too many lines produce a distracting overload. In general, grid lines should be reduced to the bare minimum necessary to convey useful information. Similarly, the presence of too many curves and their associated legends will turn the rapid simultaneous task of graph viewing into a tedious sequential task in which viewers have to painstakingly interpret each line in turn.

Line graphs, statistical maps, and bar charts require appropriately chosen vertical and horizontal axes and scales. It is important to pick a scale that presents a realistic picture of the variation in the data set. For example, a scale with a wide range will obscure data that varies over a narrow range. Conversely, a scale with a narrow range magnifies data that have little variation.

Studies of graphical perception indicate that people do best in situations that require one-dimensional judgments. They become decreasingly accurate at tasks that require them to form

two- and three-dimensional judgments. Cleveland (1985) found that simple one-dimensional comparisons of simple line, dot, and bar graphs are easier to perform than angular comparisons in pie charts, where subjects typically underestimate area. Some graphics packages boast of their ability to produce three-dimensional graphics but many readers will find it nearly impossible to extract useful information from them. Three-dimensional displays often produce reversible figure illusions and other perspective illusions.

Color can be used to delineate map areas and shading can often be used to display gradations in quantity. However, Wainer (1984) also found that the unfortunate combination of color and area representation can lead readers to form incorrect judgments. For example, suppose that the proportion of Republican voters was very high in Delaware but relatively low in New York. If high proportions of Republican voters were represented as dark black shading over the state of Delaware and low rates as white or gray shading over the state of New York, then readers would be drawn to the New York segment of the graph more than to the Delaware segment. Most probably, the Delaware data would be ignored. A simple placement of same-sized box graphs within each state might alleviate the problem.

In conclusion, graphs may enhance a statistical presentation, or they may distort it. A well-chosen graph should do more than simply repeat tabular data. While visually appealing, they may be misleading.

14. How do I interpret interaction effects? ↑

Rosnow and Rosenthal (1989) provide a very brief and readable description of how to display and interpret interaction effects in the analysis of variance (ANOVA). A more detailed account and an excellent exposition of the possible reasons for interaction effects are provided in their chapter devoted exclusively to interactions (Rosenthal & Rosnow, 1991). Defining an interaction effect is simple. An interaction effect is the multiplicative effect of two (or more) variables after controlling for the individual additive effects (i.e., main effects) of the independent variables.

In truth, there are many different types and definitions of interaction effects. Although the multiplicative variant is only one, it is by far the most commonly discussed. For an excellent and detailed (though admittedly demanding) article on the many types of interaction, see Southwood (1978). Conceptually, an interaction effect occurs when the effect of one variable depends on the level of another. It should be noted that in this context we are considering interactions among independent (predictor) variables (e.g., Treatment x Treatment interactions) and not Subject x Treatment interactions.

Describing and explaining a particular interaction is somewhat more difficult. One controversy surrounds graphical presentation in the ANOVA context. Rosenthal and Rosnow (1991; Rosnow & Rosenthal, 1989) describe methods for graphically displaying the residual means after removing main effects. However, Meyer (1991) disputed the necessity of Rosenthal and Rosnow's suggestions and argued for graphing raw means, which is the procedure advocated in the vast majority of standard ANOVA texts (e.g., Winer et al., 1991). It is our opinion that the disagreement stems from two distinct issues being conflated; the analysis of interaction effects and the interpretation of interaction effects. The former is essentially a statistical problem over which there is no disagreement. The latter is a problem in human factors and needs to be treated as such.

In some complex designs, residual means may be tabulated to reveal patterns (Rosenthal & Rosnow, 1991, F 375). Alternatively, we find that inspection of raw means often provides the viewer with rich information regarding such factors as whether the interaction is ordinal or

disordinal. Winer et al. (1991) extensively describe the "geometric" interpretation of plots of raw means. In short, interpretation of an interaction effect that has already been shown to be statistically significant is no longer a statistical problem but rather a problem in human perception. We believe this problem is best addressed empirically by researchers in graphical perception rather than analytically. In the meantime, we offer the simple suggestion that researchers plot interaction effects in the manner they find most aids their understanding.

An additional interpretive strategy was proposed by Mood (1950,p. 337). Mood suggested that the F ratios of main effect mean squares over interaction mean squares can be taken as indices of the relative magnitudes of these effects. Thus interaction effects can also be interpreted in relation to the other effects present.

Regarding explanation, Rosenthal and Rosnow (1991) provide an explanation of several different patterns of interaction and discuss possible explanations for these effects. These explanations include both substantive interpretations and interpretations of interactions as possible measurement artifacts.

An important consideration regarding possible artifactual interaction is scale of measurement. Many statistically significant interaction effects disappear if the dependent variable is subjected to a nonlinear monotonic transformation (e.g., a log transformation). Two considerations apply here. First, if dependent variables are measured on an ordinal scale rather than on a ratio or interval scale, then the scaling is arbitrary and any monotonic transformation is permissible. In that event, if a transformation eliminates nonadditivity, the data can be described more parsimoniously with only additive effects, and the interaction effect may be artifactual.

The second obvious consideration is whether a transformation can eliminate the nonadditivity. As Winer et al. (1991) state:

Not all interaction effects can be regarded as functions of the scale of measurement. In cases where profiles cross, or . . . have quite different shapes [i.e., disordinal interactions], transformations on the scale of measurement will not remove interaction effects. (p.445)

Finally, researchers using multiple regression should be aware that interaction terms can be analyzed and interpreted here as well. In regression analyses, interactions are usually coded as product terms and are entered into regression equations after partialing out the "main effects" of the individual predictor variables. An interaction occurs when the slope of the regression between two variables depends on the value assumed by a third. A comprehensible presentation of interaction effects in the context of multiple regression can be found in Darlington (1990,chap. 13).

There are differing perspectives on the proper interpretation of interaction effects. One extreme can be illustrated by the following quotation from Edmond Murphy (1982):

There is some peril that "interaction," really a device of accountancy--a statement of the extent to which such-and-such effects are nonadditive--will be reified. A perplexed accountant can always "balance his books" by ascribing discrepancies between credit and debit to "petty cash" or "miscellaneous small donations." If he is stupid as well as perplexed he may eventually come to believe that the terms refer to real phenomena, that is, he may reify them . . . Thus "interaction" is in its scientific aspect simply a confession of the inadequacy of our scientific model: a recognition that the addition of effects does not describe what happens . . . Beyond that, it tells us nothing about underlying processes of the interaction; it is a term without interpretable scientific meaning. I do not regard a measurement as a meaning. (p.167)

Although we are often inclined to agree with Murphy with respect to three-way and higher-order interactions, we see no evidence to support his statement as an overarching principle. Instead, we are inclined to agree with Rosenthal & Rosnow (1991), Darlington (1990), Southwood (1978), and the majority of the research community, that meaningful and scientific interpretations can be ascribed to many interaction effects.

15. If I have pretest-posttest control group design, should I do repeated measures ANOVA, an ANCOVA with pretest as the covariate, or something else ? ⇧

Cole (1988) and Huck and McLean (1975) render excellent discussions of this issue. Given the pretest-posttest control group design (i.e., split-plot design), one can discern at least four options: (a) ignore the pretest data and analyze the posttest scores only with a simple one-way ANOVA; (b) use a repeated measures ANOVA with one between-groups factor (treatment assignment) and one within groups factor (trials); (c) analyze change scores (e.g. posttest minus pretest) in a oneway ANOVA; or (d) analyze the data as an ANCOVA with one between groups factor (treatment assignment) and one covariate (pretest scores).

The first method, ignoring the pretest data, is generally not recommended for two reasons. First, it discards information, arguably the most precious commodity researchers have (Cohen, 1990). Second, it is generally the least powerful of all the approaches.

The repeated measures approach utilizes all available information and is more powerful than the posttest-only approach. However, it is frequently misinterpreted (Huck & McLean, 1975). The crucial effect in this analysis is the Treatment x Trials interaction. However, many researchers mistakenly interpret one of the main effects in this situation (Huck & McLean). Thus this method may not be ideal. Regarding the analysis of gain scores, it can be shown that the Treatment X Trials interaction in the repeated measures approach is mathematically equivalent to the main effect for treatment in the change-score approach (Huck & McLean). As this approach is more easily interpreted, it may be preferable to the repeated measures analysis.

The repeated measures ANOVA or gain score analyses are equivalent to the ANCOVA when the pretest-posttest correlation = 1.0. However, as this is rarely the case, the ANCOVA, which corrects posttest scores via the actual sample pretest-posttest correlation, provides a more accurate representation of treatment effects and is therefore usually more powerful (Huck & McLean, 1975). The superior power of the ANCOVA increases as the pretest-posttest correlation decreases. Thus, the ANCOVA is generally recommended as the preferred method of analyses in these designs. The one exception to this may be when power is already high and ease of interpretation is crucial. Although we are generally quite reluctant to recommend any procedure with inferior power, under these circumstances, the gain score approach may be preferred due to its more obvious meaning.

Two final points are noteworthy. First, when planning such experiments, we recommend considering Maxwell, Delaney, and Dill's (1984) "alternate ranks" procedure. This procedure assigns subjects to treatment conditions on the basis of rankings on pretest scores and will further increase the power of the subsequent ANCOVA. Second, if posthoc tests are to be done following an ANCOVA, additional procedures need to be used with covariate adjusted means (See Stevens, 1986, for a discussion).

16. If I have multiple independent and/or dependent variables, which should I use, several univariate tests or a multivariate test? ↑

Although the theory and methods for multivariate statistics have been available almost as long as those for univariate statistics, these methods have only been used regularly since the availability of easy to use statistical packages (e.g., SPSS, SAS, and BMDP). It is easy to find a number of examples of the use of multivariate statistical techniques in many of the journals in the applied areas of education, psychology, medicine, and other disciplines. Unfortunately, many applications indicate a lack of understanding of the relationship between univariate and multivariate procedures.

The most common example of this misunderstanding can be found when a researcher has a factorial design with one or more independent variables or factors and has several dependent measures. A common data analysis strategy applied to this design is to do an overall MANOVA and follow up each of the significant effects with univariate ANOVAs. Significant effects in each of these ANOVAs are then followed by the appropriate multiple comparisons and/or simple effects.

The (somewhat erroneous) rationale for this strategy is that the overall MANOVA provides a protection for the Type I error rate for the entire experiment so that researchers can do the univariate ANOVAs and feel assured that their significant results reflect the true rejection of false null hypotheses. To further complicate matters, other authors have extended this logic to multivariate analyses in general and recommend that a multivariate test always precede a number of univariate tests when the researcher has a number of measures to provide a control of the Type I error rate.

In an excellent article on interpreting the results of a MANOVA, Bray and Maxwell (1982) indicate that a MANOVA does not provide good protection for a number of ANOVAs. They argue that if one is primarily interested in the control of the Type I error rate for a set of univariate ANOVAs, then one should use a Bonferroni procedure or a variant recommended by Timm (1975). This variant entails using the Bonferroni procedure only after a significant MANOVA.

However, Bray and Maxwell (1982) state that the Bonferroni and Timm (1975) procedures ignore the relationships among the variables and therefore do not take into account the redundancies and suppressor effects among the multiple dependent measures. They recommend the use of a MANOVA if one is interested in these redundancies and/or suppressor effects.

Huberty and Morris (1989) have also pointed out that performing an overall MANOVA does not provide an overall protection for Type I error for subsequent ANOVAs and follow-up tests as is often thought. They recommend that if one is interested in group differences on a set of dependent measures that one perform separate ANOVAs and make an adjustment for Type I error by using either an additive (Bonferroni) inequality or a multiplicative inequality. They also recommend that one presents the correlations among the dependent variables so that these can be used in interpreting the overall results. It is interesting to note that those who use a MANOVA as a Type I error rate protection rarely examine all of the rich information about the redundancies and/or suppressor effects presented in the typical computer printout of these methods.

The key issue implicit in these discussions is the fact that the univariate and multivariate procedures address different statistical and research questions. A series of univariate ANOVAs addresses the issue of group differences on each dependent measure, ignoring that measure's relationship to the other dependent measures. A MANOVA addresses the question

of group differences on a composite or composites of the dependent measures (Share, 1984). The form of the composites depends on the multivariate procedure being used, that is, the weights for the composites are computed so that they meet some statistical criteria.

In MANOVA, the composites are called canonical variates and are formed using equations called discriminant functions. The weights of these functions are determined so that the groups are as different as they can be on those composites, that is, the weights maximize group differences on the composites. The weights for forming the composites for other multivariate techniques maximize other criteria (for details, see Tabachnick & Fidell, 1989).

Consider the following example. If one wanted to compare groups of unipolar depressives, bipolar depressives, and nondepressed controls on a number of clinical measures (e.g., anxiety, depression, aggressiveness, paranoia, etc.) the choice of a data analysis technique would depend on what question one wanted to answer. A series of univariate ANOVAs would address themselves to the question of whether these two groups were different on each of the dependent measures without regard to whether these measures were correlated. Huberty and Morris (1989) state that in this case one is considering the variables to be "conceptually independent," which is why they recommend providing a matrix of correlations among the measures so that one can use these correlations to qualify conclusions about the group differences.

A MANOVA on this same set of dependent measures would address the question of whether these two groups were different from each other on linear composites of these measures. These linear composites would be assumed to underlie these measures and constitute factors based on what these measures had in common in relationship to the group differences (e.g., psychopathology). When choosing to do a MANOVA, it is wise to choose variables that are believed to hang together and can be considered manifestations of some underlying factor or factors. In a loose sense, all multivariate methods perform a factor analysis in that they form linear composites and then perform the analyses (correlation, tests of differences) on these composites.

The real question is not whether you should use a univariate or multivariate statistical technique, but what information about your variables you wish to obtain that will address the hypotheses of your research study.

Conclusion

These responses will certainly not answer every question that a researcher facing a data set will have. Nor do they even represent responses to all the questions we are commonly asked, but space limits our presentation. However, we hope they will be helpful to researchers embarking on a data analysis, and we think the responses, and especially the suggested readings, will be helpful to fellow consultants, teachers, and researchers in answering the questions of their consultees, students, and associates. Finally, we do not intend for these suggestions to be taken as "written in stone." Like all fields of scientific inquiry, statistics is an evolving discipline, and the future will undoubtedly bring new insights.

Received December 14, 1992

We are grateful to Louis Hogopian, Ronald D. Franklin, Stanley Heshka, and Vincent C. Alfonso for their helpful comments on earlier drafts of this article.

Address correspondence to David B. Allison, Obesity Research Center, St. Luke's/ Roosevelt Hospital, Columbia University College of Physicians and Surgeons, 411 West 114th Street, Suite 3D, New York, New York 10025.

DIAGRAM: FIGURE 1. Decision tree for multiple comparison question.

REFERENCES

Abrahams, N. M., & Alf, E. F. (1978). Relative costs and statistical power in the extreme groups approach. Psychometrika, 43, 11-17.

Alf, E. E, & Abrahams, N. M. (1975). The use of extreme groups in assessing relationships. Psychometrika, 40, 563-572.

Allison, D. B., & Gorman, B. S. (in press). Power analysis for testing differences between dependent and independent correlations: Equations and software. Educational & Psychological Measurement.

American Psychological Association. (1983). Publication manual of the American Psychological Association (3rd ed.). Washington, DC: Author.

Anscombe, E J. (1973). Graphs in statistical analysis. American Statistician, 27, 17-21.

Baggaley, A. R. (1981). Multivariate analysis: An introduction for consumers of behavioral research. Evaluation Review, 5, 123-131.

Bailar, J. C., & Mosteller, E (1988). Guidelines for statistical reporting in articles for medical journals. Annals of Internal Medicine, 108, 266-273.

Barlow, D. H., & Hersen, M. (1984). Single case experimental designs: Strategies for studying behavior change. New York: Pergamon Press.

Barnette, V. (1978). The study of outliers: Purpose and model. Applied Statistics, 27, 242-250.

Bartko, J. J., Pulver, A. E., & Carpenter, W. T. (1988). The power of analysis: Statistical perspectives. Part 2. Psychiatry Research, 23, 301-309.

Bird, K. D., & Hall, W. (1986). Statistical power in psychiatric research. Australian and New Zealand Journal of Psychiatry, 20, 189-200.

Borenstein, M., Cohen, J., Rothstein, H. R., Pollack, S., & Kane, J. M. (1990). Statistical power analysis for one-way analysis of variance: A computer program. Behavior Research Methods, Instruments, and Computers, 22, 271-282.

Bray, J. H., & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. Review of Educational Research, 52, 340-367.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P A. (1983). Graphical methods for data analysis. Monterey, CA: Wadsworth.

Cleveland, W. S. (1985). The elements of graphing data. Monterey, CA: Wadsworth.

Cleveland, W. S., & McGill, R. (1985) Graphical perception and graphs for analyzing and presenting scientific data. Science, 229, 828-833.

Cliff, N. (1988). The eigenvalue-greater-than-one rule and the reliability of components. Psychological Bulletin, 103, 276-279.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), Handbook of clinical psychology (pp. 120-142). New York: McGraw-Hill.

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7, 249-253.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cole, D. A. (1988). Statistics for small groups: The power of the pretest. Journal of the Association for Persons with Severe Handicaps, 13, 142-146.

Comrey, A. L. (1985). A method for removing outliers to improve factor analytic results. Multivariate Behavioral Research, 20, 273-281.

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman Hall.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.

Darlington, R. B. (1990). Regression and linear models. New York: McGraw-Hill.

Davis, C., & Gaito, J. (1984). Multiple comparison procedures within experimental research. Canadian Psychology, 25, 1-13.

Dupont, W. D., & Plummer, W. D. (1990). Power and sample size calculations: A review and computer program. Controlled Clinical Trials, 11, 116-128.

Edgington, E. S. (1969). Statistical inference: The distribution-free approach. New York: McGraw-Hill.

Edgington, E. S. (1980). Randomization tests. New York: Marcel Dekker.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. Journal of Counseling Psychology, 34, 437-442.

Eysenck, H. J. (1970). The classification of depressive illnesses. British Journal of Psychiatry, 117, 241-250.

Falzer, P. R. (1974). Representative design and the general linear model. Speech Monographs, 41, 127-138.

Feldt, L. S. (1961). The use of extreme groups to test for the presence of a relationship. Psychometrika, 26, 307-316.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. Educational and Psychological Measurement, 42, 521-526.

Gaito, J. (1970). Nonparametric methods in psychological research. In E. F. Heerman & L. A. Breskamp (Eds.), Readings in statistics for the behavioral sciences (pp. 38-49). Englewood Cliffs NJ: Prentice-Hall.

Gibbons, J. D. (1971). Nonparametric statistical inference. New York: McGraw-Hill.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. American Statistician, 43, 253-260.

Gorsuch, R. L. (1983). Factor analysis. Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (1990). UniMult guide. Altadena, CA: Fuller Theological Seminary.

Hartigan, J. A., & Hartigan P. M. (1985). The Dip Test of Unimodality. Annals of Statistics, 13, 70-84.

Harwell, M. R. (1988). Choosing between parametric and nonparametric tests. Journal of Counseling and Development, 67, 35-38.

Henderson, D. A., & Denison, D. R. (1989). Stepwise regression in social and psychological research. Psychological Reports, 64, 251-257.

Hertel, B. R. (1976). Minimizing error variance introduced by missing data routines in survey analysis. Sociological Methods and Research, 4, 459-474.

Herzberg, P. A. (1969). The parameters of cross validation. Psychometrika, Monograph supplement, No. 16.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 302-308.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze a pretest-posttest design: A potentially confusing task. Psychological Bulletin, 82, 511-518.

Humphreys, L. G. (1978a). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. Journal of Educational Psychology, 70, 873-876.

Humphreys, L. G. (1978b). Research on individual differences requires correlational analysis, not ANOVA. Intelligence, 2, 1-4.

Humphreys, L. G., & Fleishman, A. I. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual difference variables. Journal of Educational Psychology, 66, 464-472.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. Psychological Bulletin, 96, 589-596.

Johnson, A. F. (1985). Beneath the technological fix: Outliers and probability statements. Journal of Chronic Diseases, 38, 957-961.

Kendall, M. G. (1962). Rank correlation methods (3rd ed). London: Griffin.

Kidder, L. H., & Judd, C. M. (1986). Research methods in social relations. New York: Holt, Rinehart, and Winston.

Klockars, A. J., & Sax, G. (1986). Multiple comparisons. Beverly Hills, CA: Sage.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. Psychological Bulletin, 85, 410-416.

Kosslyn, S. M. (1985). Graphics and human information processing. Journal of the American Statistical Association, 80, 499-512.

Kraemer, H. C. (1985). A strategy to teach the concept and application of power of statistical tests. Journal of Educational Statistics, 10, 173-195.

Kraemer, H. C., & Thieman, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Lipsey, M. W. (1990). Design sensitivity. Statistical power for experimental research. Newbury Park, CA: Sage.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: John Wiley.

Maher, B. A. (1978). A reader's, writer's, and reviewer's guide to assessing research reports in clinical psychology. Journal of Consulting and Clinical Psychology, 46, 835-838.

Marascuilo, L. A., & McSweeny, M. (1977). Non-parametric and distribution-free methods for the social sciences. Monterey, CA: Brooks/Cole.

Maxwell, S., Delaney, H., & Dill, C. (1984). Another look at ANCOVA versus blocking. Psychological Bulletin, 95, 136-147.

McSweeny, M., & Katz, B. M. (1978). Nonparametric statistics: Use and misuse. Perceptual and Motor Skills, 46, 1023-1032.

Meyer, D. L. (1991). Misinterpretation of interaction effects: A reply to Rosnow and Rosenthal. Psychological Bulletin, 110, 571-573.

Mood, A. M. (1950). Introduction to the theory of statistics. New York: McGraw-Hill.

Morley, S., & Adams, M. (1991). Graphical analysis of single-case time series data. British Journal of Clinical Psychology, 30, 97-115.

Muenz, L. R. (1989). Power calculations for statistical design. In N. Schneiderman, S. M. Weiss, & P. Kaufmann (Eds.), Handbook of research methods in cardiovascular behavioral medicine, (pp. 615-633). New York: Plenum Press.

Murphy, E. A. (1982). Biostatistics in medicine. Baltimore, MD: The Johns Hopkins University Press.

NCSS. (1991). Power analysis and sample size. Kaysville, UT: Author.

Park, C., & Dudycha, A. (1974). A cross validation approach to sample size determination for regression models. Journal of the American Statistical Association, 69, 214-218.

Pirot, M., & Lustig, S. D. (1984). The conceptual and mathematical unity of analysis of variance, simple correlation, and multiple regression. Perceptual and Motor Skills, 59, 751-756.

Puri, M. L., & Sen, P. K. (1969). A class of rank order tests for a general linear hypothesis. Annals of Mathematical Statistics, 40, 1325-1343.

Puri, M. L., & Sen, P. K. (1971). Nonparametric statistics in multivariate analysis. New York: Wiley.

Rasmussen, J. L., & Dunlap, W. P. (1991). Dealing with non-normal data: Parametric analysis of transformed data vs. nonparametric analysis. Educational & Psychological Measurement, 51, 809-820.

Rawlings, J. O. (1988). Applied regression analysis: A researcher's tool (Chapter 7, Model development: Selection of variables, pp. 168-191). Belmont, CA: Wadsworth & Brooks/Cole.

Raymond, M. R. (1986). Missing data in evaluation research. Evaluation and the Health Professions, 9, 395-420.

Rosenthal, R., & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. Psychological Bulletin, 105, 143-146.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Rothstein, H. R., Borenstein, M., Cohen, J., & Pollack, S. (1990). Statistical power analysis for multiple regression/correlation: A computer program. Educational and Psychological Measurement, 50, 819-830.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. The American Statistician, 44, 174-180.

Saville, D. J. (1991). Reply to Holland and Lea. American Statistician, 45, 166-167.

Share, D. L. (1984). Interpreting the output of multivariate analyses: A discussion of current approaches. British Journal of Psychology, 75, 349-362.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Southwood, K. E. (1978). Substantive theory and statistical interaction: Five models. American Journal of Sociology, 83, 1154-1203.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. Psychological Bulletin, 95, 334-344.

Stevens, J. P. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (1989). Using multivariate statistics. New York: Harper & Row.

Timm, N. H. (1975). Multivariate analysis: With applications in education and psychology. Monterey, CA: Brooks-Cole.

Tufte, E. R. (1983) The visual display of quantitative information. Chesire, CT: Graphics Press.

Tukey, J. W. (1977a). Some thoughts on clinical trials, especially problems of multiplicity. Science, 198, 679-684.

Tukey, J. W. (1977b). Exploratory data analysis. Reading, MA: Addison & Wesley.

Veiel, H. O. F. (1988). Base-rates, cut-points and interaction effects: The problem with dichotomized continuous variables. Psychological Medicine, 18, 703-710.

Velicer, W. E, Fava, J. L., Zwick, W. R., & Harrop, J. W. (1988). Program CAX (Component Analysis extended). Unpublished computer program, University of Rhode Island.

Wainer, H. (1984). How to display data badly The American Statistician, 38, 137-147.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. Annual Review of Psychology, 32, 191-241.

Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. Journal of Counseling Psychology, 34, 372-382.

Weiss, B., Williams, J. H., Margen, S., Abrams, B., Caan, B., Citron, L. J., Cox, C., McKibben, J., Ogar, D., & Schultz, S. (1980). Behavioral responses to artificial food colors. Science, 207, 1487-1489.

Welch, W. P., Frank, R. G., & Costello, A. J. (1983). Missing data in psychiatric research: A solution. Psychological Bulletin, 94, 177-180.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical principles in experimental design (3rd ea., pp. 333-342, 442-445). New York: McGraw-Hill.

Zwick, W. R., & Velicer, W. E (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-442.

DAVID B. ALLISON Obesity Research Center Columbia University College of Physicians and Surgeons BERNARD S. GORMAN Department of Psychology Hofstra University LOUIS H. PRIMAVERA Department of Psychology St. John's University