Other Statistical Issues

VIII.A. SAMPLING

The problem is the failure to use independent random sampling (IRS). Except, perhaps, for commercial polling of political opinions and of marketing preferences, the social and behavioral sciences most typically do not use IRS. However, all inferential statistical procedures assume IRS. Philosophically, IRS is necessary for the sample to be representative of the population. Specifically, without IRS we do not have random variables. In particular, IRS is necessary if the sample mean is to be an unbiased estimator of the population mean.

1. Is it correct to say that all claims of statistical significance, or lack of it, are nonsense without IRS? (I am referring here to the claim of significance, not to the theory under study.)

2. Are there methods of inferential statistics that do not require IRS?

3. Usually our nonrandom data come from self-selected cohorts of participants (i.e., people already aggregated, some of whom will agree to be participants). Is there any way to take advantage of having data from three or four different samples of nonrandom data?

4. For nonrandom data, would we be more true to our methods if we returned to the rhetorical and editorial norms of earlier decades of this century and merely report sample statistics without claims of significance?

I suspect the question is referring to judgment or convenience samples such as the students in a beginning psychology class at a university. It is a leap of faith to claim that such samples reflect the entire population of adults. Only more careful sampling can demonstrate this, but certainly it is more likely to be true for innate cognitive processes than for learned social behavior.

Even here, claims of statistical significance are not entirely meaningless, if it is remembered what population is being referred to. That is, if one took repeated samples of students in beginning psychology classes (assuming the characteristics of incoming freshmen do not change), one could make predictions about the outcome of such samples. The error, of course, is to then generalize to a much larger population, ignoring possibly large sample biases, as well as to ignore the cluster effects of the sample.

2. There is a group of statisticians who develop models and test them and estimate the parameters of the model who claim that, for this purpose, the method of sampling does not matter unless it is related to the model. The problem is that it is often not easy to tell whether the sampling method impacts on the model, and different researchers will not agree on this. If one is comfortable with the assumption that the sample selection is unrelated, one can use the sample data. However, in estimating the sampling error of the estimated parameters, it is still necessary to take account of the cluster effects in sampling and not to use simple random sampling error estimates, as is often done.

3. If data are available from multiple judgment or convenience samples, it is certainly possible to say more than from a single such sample. Although issues of sample bias are still as troublesome as ever, comparing the results among the multiple samples does allow for an estimate of sampling variance that accounts for cluster effects. The variance estimate may be biased (most likely downward) if the sample clusters chosen are more similar to each other than would be a random sample of clusters, but at least one has minimum estimates of variance.

4. Obviously, incorrect estimates of sampling variance, generally too low, are worse than no estimates at all. Having said that, I think it is useful to provide sampling error estimates if they are qualified, often as minimum estimates. They are especially useful in identifying nonsignificant results because it is unlikely, but not impossible, that such results would become important.

Professor Seymour Sudman
University of Illinois

1. The phrase independent random sampling is a fuzzy one and not typically used by samplers. Instead, the usual delineation between various forms of probability samples is between simple random samples in which each element of the population has an equal, independent chance of selection, and more complex designs such as cluster samples in which clusters are chosen independently by probability methods, but units within clusters are not independent. From such probability samples, it is, of course, possible to compute sampling variances that are typically larger than those for simple random samples because of homogeneity within cluster.
There are, of course, lots of sampling references. A nice one to look at is Groves (1989).

REFERENCE


Editor: I had not heard the term IRS either (except in the scary governmental sense). If the underlying issue in this question is the objection to student samples, I believe that issue has been covered in an argument of external versus internal validity (quickly, that external validity, or the desire to generalize to participant populations in "effects" research, probably necessitates large, heterogeneous, representative samples, whereas for internal validity—that is, the desire to attribute causality in the lab toward the goal of "theory testing" research—homogeneous samples, like student populations, may not just suffice, they may be quite smart, in that their very homogeneity should lend clarity to the analysis, e.g., in a reduction of mean square error terms; cf. Calder, Phillips, & Tybout, 1981). One simply needs caution, or specificity, in defining one’s population; for example, as unrepresentative as undergraduates or graduate students may be on certain properties, my guess is that the distribution among them as to who uses Crest or Colgate likely resembles the general market shares. To push the argument further, if one is truly claiming to be concerned with generalizing results to some huge, nebulous population of consumers, let us consider that the marketplace has been global for some time now, so representation would necessitate herculean data collection efforts.


REFERENCES


VIII.B. REGRESSION WITH CORRELATED VARIABLES (MULTICOLLINEARITY)

Suppose that previous research has shown some variable to be associated with a dependent variable (e.g., market orientation and company performance). You think there may be other related activities that are also associated with performance (like production efficiency or selling). The independent variables might all be correlated together or not. The question is, "Are all these activities associated with the dependent variable?" Should you put all the independent variables into a regression together and look at the individual significance of each, or run individual regressions and look at whether each is significant? Or both? Does the answer depend on collinearity?

Professor Greg Allenby
Ohio State University

If your goal is to measure association, then simple correlations (between the dependent variable and each independent variable in turn) will do the trick. However, I believe you have confused a number of issues in your question. First, is collinearity a bad thing? Second, how do you deal with it?

As for collinearity, you must keep in mind that in the analysis of any firm or market data, collinearity will be present if the firm or the market "works." For example, suppose a firm finds different sales responses in different sales territories. Some areas like a particular brand and other areas do not. If the firm is paying attention to the data, then it probably is allocating more resources (more sales people, more advertising, more of anything) to the sales areas that perform well and pulling resources out of the areas with little "bang for the buck." The analysis of data from this firm will have strong collinearity, making it difficult to determine what is driving the increased sales.

The analysis of this data cannot proceed in a simple fashion. At a superficial level, all variables (sales people, advertising dollars, etc.) are related to sales, as indicated by their simple correlations. However, the whole point of multiple regression is to assess the contribution of association of one variable (x1) to another (y) holding fixed other covariates (x2, ..., xn). Therefore, if the goal is to assess the incremental contribution of x1, then multiple regression must be used and the simple regressions are worthless.

More fundamentally, however, I suspect that this question is not about association but rather about causation. If so, then there are only two routes available—either experimentation or modeling based on a set of primitive constructs (e.g., economic modeling). If the establishment of causal statements is the goal, then multiple regression is not the answer either.

Professor Robert Meyer
University of Pennsylvania

In general, the most conservative approach would be to throw all the independent variables into the model simultaneously and look at the partial correlation coefficients. If there is collinearity, you might make a case for looking at partial con-
trolls under some assumed causal (ordered) structure. The question is somewhat vague, so depending on what you mean by "related," I could also see reporting each variable's effect over a bounded range—from its smallest (multiple regression partial) to its largest (its univariate effect).

Editor: The questioner may find interesting Mason and Perreault (1991), whose study allowed them to conclude that perhaps multicollinearity is not as problematic as one might think, except under rather extreme conditions (i.e., small sample sizes, very high intercorrelations among the predictors, and rather low overall $R^2$s). Olkin and Finn (1995) may also be useful, because they describe confidence intervals for partial correlation coefficients and other tools that may begin to tease apart some of the researcher's underlying questions.

REFERENCES


VIII.C. DEALING WITH OUTLIERS

What should one do about the outliers for a regression or analysis of variance (which I define as a small percentage [ < 5%] of participants whose actual responses deviate greatly from their estimated responses). With least squares estimation, these few participants may exert a pronounced effect on estimates and significance tests. Should one merely report data on the original sample and forget about outliers? Should one report results based on the entire sample as well as results based on the sample minus the few outliers? What happens if the results are different? (My limited personal experience is that strange interactions may sometimes be due to outliers.) Should one remove outliers and report data without the few outliers; if so, how should one report the data? Should one dispose of the data; pray that one does not get a few strange participants; and collect, reanalyze, and report the new data?

Professor Edward Malthouse
Northwestern University

I find the following discussion on outliers to be excellent (from Neter, Kutner, Nachtsheim, & Wasserman, 1996):

Frequently in regression analysis applications, the data set contains some cases that are outlying or extreme; that is, the observations for these cases are well separated from the remainder of the data. These outlying cases may involve large residuals and often has dramatic effects on the fitted least squares regression function. It is therefore important to study the outlying cases carefully and decide whether they should be retained or eliminated, and if retained, whether their influence should be reduced in the fitting process and/or the regression model should be revised. (p. 368)

A first step is to examine whether an outlying case is the result of a recording error, breakdown of a measurement instrument, or the like. … If an outlying influential case is not clearly erroneous, the next step should be to examine the adequacy of the model. … Discarding of outlying influential cases that are not clearly erroneous and that cannot be accounted for by model improvements should be done only rarely, such as when the model is not intended to cover the special circumstances related to the outlying cases. … An alternative to discarding outlying cases that is less severe is to dampen the influence of these cases. (pp. 416–417)

If we know the observations are erroneous, clearly we should drop them. If we know they are legitimate, then there are two outstanding questions: (a) How do we identify influential observations? and (b) How do we dampen the influence of extreme cases? The answers to these questions are not straightforward. In response to the first, I strongly encourage analysts to study univariate descriptive statistics of each variable to be used in an analysis, including minimum, maximum, median, outer percentiles, and boxplots. All of these are easily generated by commercial software packages. The maximum value being much greater than the upper percentiles indicates there are some observations that may be problematic (likewise for the minimum). There are also a wide variety of diagnostics available in commercial software packages such as Cooke’s distance. Consult any good book on linear models for discussion of these methods and their limitations. These two steps should find many of the problematic cases, but not necessarily all. Studying marginal distributions can reveal many outliers, but an observation can be an outlier in the joint distribution of variables without being an outlier in any of the marginal distributions (e.g., Tuft, 1983, p. 14). Venables and Ripley (1994) warned, “It can be difficult or even impossible to spot outliers in multivariate or highly structured data” (p. 204).

There are several possible responses to the second question. If analysts can identify problematic cases, they can reduce their influence by Winsorizing or trimming them. Another alternative is to use a robust estimator such as least trimmed squares regression; see Neter et al. (1996) or Venables and Ripley (1994) for further discussion.

REFERENCES

Editor: My only advice is to be conservative about lopping off participants whose scores look different. Just as humans are not particularly good random number generators, we are not that good at detecting true outliers versus simply extra big (or small) numbers that are annoyingly different from the majority of the other data points (especially when the data are multivariate; Rousseew & van Zomeren, 1990). The statistical question is one of robustness—is the datum different enough that it actually has impact or influence on the subsequent computations (cf. Cook & Weisberg, 1982; Hadi & Simonoff, 1993)?

REFERENCES


VIII.D. WHY DO WE NOT USE MORE NONPARAMETRIC METHODS?

From my point of view, the analytical tools proposed by consumer psychologists and marketing scientists are dominated by parametric approaches. Why are the equivalent nonparametric methodologies that are also available used far less often even by scientists (apart from the end user, who, of course, will wait until computer programs are available).

Most of the multivariate techniques of data analysis (e.g., in regression analysis) impose more or less restrictive distributional assumptions about the data to be analyzed. Is there a general guideline, what one should do, if these assumptions are violated? What is recommended, if the distribution of data is a priori unknown?

Professor Edward Malthouse
Northwestern University

My answer focuses on explaining why nonparametric regression and nonparametric multivariate analysis models are not widely used. I do not discuss other classes of nonparametric methods such as hypothesis tests based on ranks, runs, and so on.

I believe there are two main reasons why consumer psychologists and marketing scientists predominantly use classical parametric approaches. The first is that classical parametric approaches are very well suited for the problems in the field. Data in this field typically have a low signal:noise ratio, and the linear structure imposed by most classical models is appropriate. Relations between variables are often fairly linear, at least in the region of interest. When theory suggests a nonlinear relation, classical models can be easily modified by a skilled analyst. Whenever there is a dependent variable, for example, regression problems interpretability is usually at least as important as raw predictive power, and classical linear models are good at this also. Interpretability is not so important for many problems in which nonparametric methods have blossomed, such as handwriting recognition. Psychologists and marketers often have small sample sizes, particularly those who analyze experimental data, and nonparametric methods usually require large samples. Psychologists and marketers often do confirmatory analyses in which statistical inference is very important. Classical linear models are good at this, too, although many of the methods that I mention later offer some statistical inference.

The second main reason that parametric methods dominate is the availability of software. Most of the marketers I know stick very close to Statistical Analysis System (SAS), Statistical Procedures for the Social Sciences (SPSS), and maybe their trusted old FORTRAN code that does nonlinear optimization to estimate any “off the beaten path” maximum likelihood problems they encounter. For the most part, analysts’ expectations are set by the software they use. It is usually easier to modify the data (e.g., perform variable transformations) so that a classical method offered by SAS or SPSS can be used to research, code, and debug software or acquire, install, and learn a new software package, such as S-Plus, offering more modern approaches. In the case of regression, finding appropriate transformations, Winsorizations, and interactions between variables will certainly get the job done, but some of the modern methods mentioned later require much less work. When I build predictive models for direct marketing, I often end up using a classical model, but I use modern methods to get to know my data and find transformations. Both SAS and SPSS have started introducing modules that fit nonparametric models. When experimenting with nonparametric models is as easy and convenient as, say, trying different factor analysis rotations or clustering routines is today, I believe that consumer psychologists and marketers will begin to use them more.

I now mention some modern models that I have found to be very useful. For those interested in reading more about these models, I highly recommend Venables and Ripley (1994), which gives an excellent introduction to each of the methods and others. The book gives many examples using the S-Plus software, but knowledge of S-Plus is not required. SAS and SPSS are beginning to offer software to fit the models described.
Generalized linear models (GLMs). This class of model gives one possible answer to the second part of the question regarding the distributional assumptions of classical models. GLMs allow the analyst to specify a wide range of distributions for the dependent variable. They also allow the analyst to model easily dependent variables with noninterval measurement levels (e.g., dichotomous, polychotomous, ordinal, and counts). SAS recently introduced Proc Genmod to fit GLMs. The programs GLIM and S-Plus also fit GLMs. See McCullagh and Nelder (1983) or Dobson (1983) for details.

Generalized additive models (GAMs). Tukey made a strong case for using smoothers as a powerful exploratory data analysis tool. Examining a smoothed version of a scatterplot can improve the interpretability greatly and reveal nonlinear relations. GAMs extend smoothing to allow for multiple predictor variables and build on the ideas of GLMs. They are very interpretable and allow the analyst to understand the nature of the relation between a dependent variable and several predictors. There is a large amount of statistical inference available. An SAS representative I know tells me they will introduce a GAM proc soon. S-Plus also fits them. See Hastie and Tibshirani (1990) or Green and Silverman (1994) for further reading.

Tree-based models. For example, CART and CHIAD have a number of advantages over linear models. Provided there is a small number of predictor variables, tree-based solutions are very easy to understand and highly interpretable. I find that nontechnical people can understand a tree-based model much more easily than regression parameter estimates. The output is a set of discrete groups (a segmentation), and managers find it easier to think discretely than continuously (as for regression). Tree-based models are very robust to extreme predictor-variable values, and thus, the danger of having an outlier affect the estimates is much less than for linear models (outlying dependent-variable values are still a problem, though). Tree-based approaches build excellent predictive models. For this reason, they are favored by some direct marketers. S-Plus's tree function, SPSS's Answer Tree software, and SAS's data-mining module (unfortunately not part of SAS/Stat) all fit tree-based models. See Breiman, Friedman, Olshen, and Stone (1984) for details.

There are several possible answers to the second question. As mentioned earlier, GLMs provide one way to modify the distributional assumptions of classical models. A second approach is to modify the data before the analysis with transformations so that the distributional assumptions are more closely met (e.g., variance-stabilizing transformations in regression). A quote is appropriate here: "All models are wrong but some are useful" (Box, 1979, p. 202). I suggest paying attention to the assumptions that can affect the usefulness of a model. If the data violate a certain assumption, will the estimate still be useful? Could the violation lead to an incorrect conclusion? Learn the idiosyncrasies of your modeling approach. What data characteristics affect the stability and usefulness of your model? As an example, outliers can have a strong effect on the final estimates of most models. After dropping outliers, the estimates can change greatly. The outliers can lead to an incorrect conclusion and therefore affect the usefulness of the model. (Also see Malthouse's response to the question on outliers, VIII.C., for a related discussion.)

REFERENCES


Professor Donald Lehmann
Columbia University

A few words on this issue follow: It is important to recognize that regression analysis makes no distributional assumptions; it merely computes the average conditional effect of different independent variables. The assumption underlying the statistical tests commonly performed is that the errors are independently normally distributed. Because error estimates are available after analysis, it is easy to test for departures from normality. In practice, most unimodal distributions (e.g., t) lead to similar results. More serious are correlations among errors that often arise from mis underspecified models and affect both standard errors and, in some cases, the mean of the estimated coefficients. See Johnston (1984) for more detail.

REFERENCE