July 2002: Assess independence, equal variance, and normality—in that order (Rule 1.4)

Rules of the month are numbered in accordance with the numbering in the book. Thus, Rule 1.1 refers to the first rule in Chapter 1. And so on. These comments do not repeat the material in the book but highlight and amplify it. A rule is stated—as found in the book—and then discussed.

Statement of Rule 1.4

"Classical hypothesis tests assume that the observations are (1) independent, (2) all come from a population with the same variance, and, for parametric tests, (3) follow a normal distribution. The most important (in terms of maintaining an assumed Type I error level: the probability of rejecting the null hypothesis when it is true—see introduction to Chapter 2 for a fuller discussion of errors) are the first, then the second, then the third."

Further Comments on the Normality Assumption

A recent paper by Lumley et al. (2002) makes some important points regarding the normality assumption in *t*-tests and regression situations, providing a good springboard for further reflection. The paper gives a thorough review of the literature and is worth reading for that purpose alone. Their motivating example is the annualized medical cost in the Washington State Basic Health Plan that has a markedly skewed distribution. The costs are expressed in dollars; it is unhelpful to discuss the data in, say, logarithmic units. (Transformations have become less important with the introduction of link function models by Nelder and Wedderburn (1972).)

1. Based on simulations with very skewed data the authors show that with sample sizes of 500 the skewness is overcome. With more symmetrically distributed residuals the "central limit theorem effect" is much more rapid.

2. It's important to note that the central limit theorem effect only applies to inferences about associations. It does not apply to prediction intervals, which are very dependent on the original distribution.

3. The authors also point out that homogeneity of variance is a more important assumption than normality. However, they show that heteroscedasticity must be large in order to cause biased inferences. This applies to regression situations as well.

4. It should be noted that the paper goes beyond the usual normality assumption of regression conditional on the predictor variable. The

authors deal with the situation where the sample slope is essentially non-normally distributed.

Lumley et al. have provided a service in their analysis of the role of normality—and to a lesser extent the role of heteroscedasticity—in *t*-tests and regression. It would be exciting to have them do further work on the effect of dependence, such as autocorrelation, in these inference situations.

References

Lumley, T., Diehr, P., Emerson, S. and Chen L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23:** 151-169.

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135:** 370-384.

Responses

This section is intended to contain reader comments and perhaps responses from me. It provides a forum for discussion and further reflection.