June 2002: Multiple Comparisons (Rule 6.12)

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 highlights and amplifies it.

"Address Multiple Comparisons Before Starting a Study" (Rule 6.12)

"Develop a coherent strategy for dealing with multiple comparisons before starting a study. The strategy should include consideration of exploratory vs. confirmatory research, primary vs. secondary endpoints, and final analyses vs. interim analyses."

Escaping the Iron Claw of the Bonferroni Inequality

I focus on the Bonferroni inequality in this discussion because it is the most widely applied and has been studied the most intensively in the last ten years.

The "solution" to the multiple comparison problem is recommended to be multi-faceted. This section adds to the material in the text by a somewhat more formal outline of how to deal with the issue. The text did not capture the exciting work that has been going on in multiple comparisons in the last twenty years. Some truly innovative procedures have been developed with a synergistic effort to escape the iron claw of the Bonferroni inequality. Two strategies are particularly notable. The first attempts to sharpen the Bonferroni inequality; the second redefines the problem. I discuss each in turn.

The references are still relevant. The following strategies can–and should–be followed simultaneously.

1. An adjustment for multiple comparisons. The challenge is to define the "family" over which to make these adjustments. Some sensible families are classes of endpoints such as immune responses, various metrics for characterizing particles, and so forth.

   The usual Bonferroni approach can be improved by more recent strategies such as the Holm procedure and Hochberg procedures as referenced in Sankoh et al. (1997) or Proschan and Waclawiw (2000). Their work builds on papers by Šidák (1971) and Simes (1986).

   The Hochberg procedure is confusingly called a *step-down procedure* in Benjamini and Hochberg (1995) and a *step-up procedure* in Sankoh and Dubey (1997). It all depends on how the P-values are arranged. The Holm procedure is then the opposite. The key is whether you start with the largest P-value or the smallest P-value. To illustrate, the P-values

Table 1: Illustration of Bonferroni, Holm and Hochberg Adjustments for Multiple Comparisons.

Observed P-value	Bonferroni adjustment	k	k$\times P$-value	Holm adjustment	Hochberg adjustment
0.081	0.324	1	0.081	0.081	0.081
0.024	0.096	2	0.048	0.060	0.048
0.020	0.080	3	0.060	0.060	0.048
0.005	0.020	4	0.020	0.020	0.020

in Table 1 have been taken from Sankoh and Dubey (1997). There are 4 P-values and the Bonferroni adjustment multiplies every P-value by 4. The Holm and Hochberg adjustments both start from the adjusted values (k$\times P$-value) in Column 4 of the table. The Holm procedure starts from the smallest adjusted value and works its way up. The Hochberg procedure starts from the largest value and works its way down. Both procedures have the requirement that the original ranking must be maintained. Holm starts from 0.020 in Column 4. The next value is 0.060 and not significant. Hence all other values are not significant. The observed value of 0.048 in Column 4 is then adjusted up to 0.060 in order to maintain the ordering. Hochberg starts with 0.081 in Column 4. Since it is not significant the next value is 0.048. This is significant, hence all subsequent comparisons are significant and adjusted to maintain the raw P-value ordering. This requires that 0.060 in Column 4 be reduced(!) to 0.048. In this case, Bonferroni declares one comparison significant, Holm likewise, and Hochberg declares three comparisons to be significant. The Hochberg procedure is preferred because it also maintains the same per-experiment error rate as the Bonferroni and Holm procedures.

Table 1 illustrates that it's confusing to talk about step-down and step-up procedures. If the observed P-values had been arranged from lowest value to highest the terms would have been reversed. It would seem better to speak of the Holm procedure as small-large and the Hochberg procedure as large-small.

2. Hochberg (see Benjamini and Hochberg, 1995) made a fundamental contribution to the multiple comparison problem by defining the *False Discovery Rate* (FDR). Their work can be linked to to a seminal paper by Sorić (1989). Rather than fixing the Type I error rate he proposed fixing the rejection region. This makes sense in situations where scientists have a good idea about the significance of outcomes. The Hochberg approach has found particular usefulness in situations where there are many multiple comparisons such as in microarray analysis with hundreds or even thousands of comparisons. Storey (2002) has sharpened the Hochberg procedure.

3. As mentioned the bioassay approach removes the pairwise comparison of doses.

4. A standard distinction in the statistical literature is between exploratory and confirmatory analyses. The latter are hypotheses suggested by previous studies, proposed on the basis of putative mechanisms, suggested by extrapolation from human to animal studies, implicit in the identification of criteria pollutants, or other mechanisms. The endpoints can then be divided into exploratory and confirmatory analyses. It is usual practice not to adjust for multiple comparisons in the confirmatory analyses. Hence, the investigators should be encouraged to divide the analyses into these two categories. Ordinarily, the confirmatory class of analyses will have fewer endpoints than the exploratory class.

5. Results can be synthesized. For example, if an analysis shows that trends are similar in male and female rats, the results can often be combined to provide a single, species-specific estimate. A (parametric) analysis can usually incorporate a test of interaction or non-parallelism. If the interaction is not significant the results can be combined-on the basis of Ockham's razor, the principle of parsimony. A more liberal criterion can be used, for example, the interaction must not be significant at the 10% or 20% level. Another synthesis approach involves analyzing the endpoints as a multivariate response. This, again, automatically adjusts for the multiplicity of endpoints. The drawback is that there is a loss of power if a single significant endpoint is grouped with a large number of non-significant ones.

6. Schweder and Spjøtvoll (1983) developed a very useful graphical technique detecting null hypotheses. A slight modification is proposed here to make the graphs comparable to scree plots in factor analysis that essentially deal with the same issue, that is, given a factor analysis involving an unknown number of latent traits how many traits are there? A scree plot of eigenvalues is one approach to assessing that; see for example, Fisher and van Belle, 1993. Here I show a scree-plot for P-values.

   Suppose there are n tests resulting in P-values, $p_1, p_2, \cdots, p_n$. Let N_p be the number of P-values greater than p. A plot of N_p against the ordered P-values produces a plot as in Figure 1. A line drawn through the scree-part of the plot intersects the ordinate in the expected number of true null hypotheses. Thus this technique provides one way of assessing the number of true null hypotheses in a set of comparisons. In Figure 1 we estimate that there are about 220 true null hypotheses and, therefore, about 80 non-null hypotheses. Since the hypotheses are not independent it's not clear which hypotheses to select. However, the figure does suggest that there are non-null hypotheses in the study.

   To my knowledge this approach has not been tested by simulation studies.

7. The final solution to the multiple comparison problem is that the results must be coherent and make physiological sense. This is a qualitative cri-

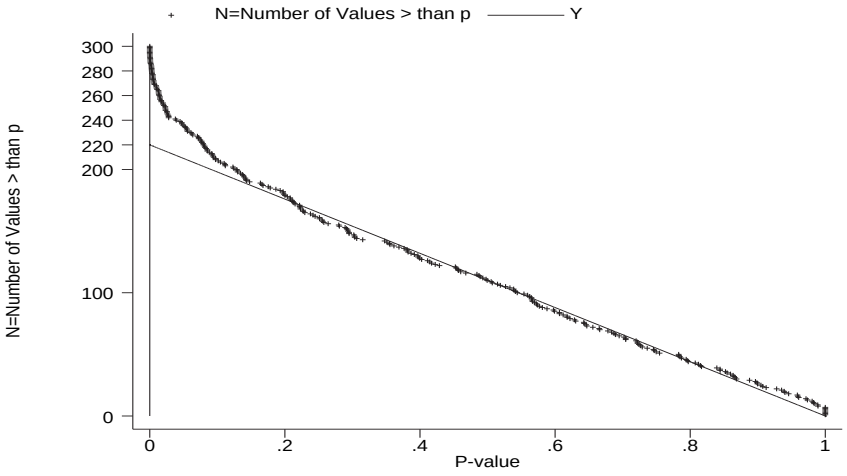

Figure 1: Scree plot of 300 P-values from a study by Crane et al. (2002). Line through these values drawn by eye.

terion that requires collegial discussion by all the collaborators (including the statistician). There is the constant danger of over-interpretation of data so the enthusiasm of the pattern-detector must be tempered by the skepticism of colleagues.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, **57:** 289-300.

Crane, P., Hightower, A., van Belle, G., Jolley, L. and Larson, E.B. (2002). Differential item functioning in the Mini-Mental State Examination: evidence of test item bias from two large longitudinal cohort studies. Submitted to the *Journal of the American Medical Association.*

Fisher, L.D. and van Belle, G. (1993). *Biostatistics: A Methodology for the Health Sciences*. John Wiley and Sons, New York, NY.

Proschan, M.A. and Waclawiw, M.A. (2000). Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials*, **21:** 527-539.

Sankoh, A.J., Huque, M.F. and Dubey, S.D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, **16:** 2529–2542.

Schweder, T. and Spjotvøll, E. (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika*, **69:** 493-502.

Šidák. Z. (1971). On multivariate normal probabilities of rectangles: their dependence on correlations. *Annals of Mathematical Statistics*, **39:** 1425-1434.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73:** 751-754.

Sorić, B. (1986). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, **84:** 608-610.

Storey, J.D. (2002). *False Discovery Rates: Theory and Applications to DNA Microarrays*. Ph.D. Dissertation, Stanford University. May be accessed at: http://www-stat.stanford.edu/ jstorey/papers/thesis.pdf

Responses

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