5

Multiple Tests

In this chapter we consider methods to control the overall error rate when multiple tests are performed and, if the tests are independent, methods to combine them.

5.1 Controlling the Overall Error Rate

One of the difficulties with clinical trials and other large-scale studies is that frequently so many variables are under investigation that one or more of them is practically guaranteed to be statistically significant by chance alone. If we perform 20 tests at the 5% level, we expect at least one significant result in twenty on the average. If the variables are related (and in most large-scale medical and sociological studies the variables have complex interdependencies), the number of falsely significant results could be many times greater.

David Freeman [1983] conducted a simulation study in which he generated 100 values each of 51 independent normally distributed variables. He designated one of the variables as the "dependent" variable and using a multiple regression technique found that 15 of the remaining variables made significant contributions as predictors at the 25% level. In a second multiple regression confined to these 15 variables, he found that 14 of the 15 made significant contributions as predictors at the 25% level and 6 of the 15 made contributions that were significant at 5% level. Clearly, "significance" in a multiple regression context is deceptive.

One way, and not a very good one, to ensure that the probability of making at least one Type I error is less than some predesignated value α is to make the k different comparisons each at level α/k. This method, attributed to Carl Bonferroni, is conservative, so that it can result in increased Type II error and, in consequence, has been widely criticized (see, for example, Perneger, 1998).

A better method, first described by Holm [1979], first orders the p-values from smallest to largest (or the corresponding standardized test statistics from largest to smallest). One begins with the most significant result

and decides whether to accept or reject. Obviously, once a hypothesis is accepted then all hypotheses with larger p values are accepted as well. If a hypothesis is rejected, then a new critical value is determined, and the next p-value inspected. Permutation procedures utilizing this step-down approach were developed independently by Westfall and Young [1993], Blair and Karniski [1994], and Troendle [1995]; a test based on the latter's work is described in the next subsection.

The chief weakness of the step-down procedure is its dependence on the rejection criteria used to test the smallest p-value, normally $p_{(1)} \leq \alpha/k$. An alternative developed by Hochberg [1988] begins with the largest p-value at the first step. If a hypothesis is rejected then all hypotheses with smaller p-values are rejected as well. If a hypothesis is accepted, then a new critical value is determined, and the next p-value inspected. Blair, Troendle, and Beck [1996] report that this step-up method, an example of which is provided in Section 5.1.2, is slightly more powerful than the step-down.

5.1.1 Standardized Statistics

As an alternative to the analytic step-up method of Dunnett and Tarnhane [1992], which requires a specific distribution and correlation structure, we may apply the following permutation method due to Troendle [1996]. Suppose we have measured k variables on each subject, and are now confronted with k test statistics $s_1, s_2, \ldots, s_k$. To make these statistics comparable, we need to standardize them and render them dimensionless, dividing each by its respective L_1 or L_2 norm. For example, if one variable, measured in centimeters, takes values like 144, 150, and 156, and the other, measured in meters, takes values like 1.44, 1.50, and 1.56, we might set $t_1 = s_1/4$ and $t_2 = s_2/0.04$.

Next, we order the standardized statistics by magnitude so that $t_{(1)} \leq \cdots \leq t_{(k)}$. Denote the corresponding hypotheses as $H_{(1)}, \ldots, H_{(k)}$. The probability that at least one of these statistics will be significant by chance alone at the α level is $1-(1-\alpha)^k$, which is approximately $k\alpha$. But once we have rejected one hypothesis (assuming it was false), there will be only $k-1$ true hypotheses to guard against rejecting.

Begin with $i = 1$ and

1. repeatedly resample the data (with or without replacement), estimating the cut-off value $\varphi(\alpha, k-i+1)$ such that $\alpha = \Pr\{T(k-i+1) \leq \varphi(\alpha, k-i+1)\}$, where $T(k-i+1)$ is the largest of the $k-i+1$ test statistics $t_{(1)} \leq \cdots \leq t_{(k-i+1)}$ for a given resample;
2. if $t_{(k-i+1)} \leq \varphi(\alpha, k-i+1)$, then accept all the remaining hypotheses $H_{(1)}, \ldots, H_{(k-i+1)}$ and STOP.
   Otherwise, reject $H_{(k-i+1)}$, increment i, and RETURN to step 1.