9

Multivariate Analysis

The value of an analysis based on simultaneous observations on several variables—for example, height, weight, blood pressure, and cholesterol level—is that it can be used to detect subtle changes that might not be detectable, except with very large, prohibitively expensive samples, were you to consider only one variable at a time.

In this chapter we consider four approaches to the analysis of multivariate data: via the nonparametric combination of univariate tests; by parametric means, utilizing canonical forms and the properties of the multivariate normal; by permutation means utilizing essentially the same statistics as are used in the parametric approach but obtaining reference values from a permutation distribution; and by means of a nonparametric runs test.

We also consider methods for analyzing repeated measures.

9.1 Nonparametric Combination of Univariate Tests

To obtain a test that will take full advantage of the multivariate approach, we follow in the footsteps of Pesarin [1990, 2001] and harness several of the ideas we've developed previously—univariate statistics for optimally exposing differences among groups of metric, ordinal, or categorical observations, the use of ranks to place diverse observations on a single common scale, the Fisher omnibus statistic, and the permutation test.

As is the case with all the methods considered in this chapter, all the observations on a single experimental unit are maintained as a single indivisible vector. Labels are applied to and exchanged among these vectors, and not among the individual observations.

Let $\boldsymbol{X}$ denote the original $n \times K$ matrix of multivariate observations. Corresponding to $\boldsymbol{X}$ is a $1 \times K$ vector $\boldsymbol{T}_0 = \boldsymbol{T}(\boldsymbol{X})$ of univariate statistics. In a clinical trial, for example, some of the observations might relate to the occurrence or nonoccurrence of certain side effects, some might be the values

of certain blood chemistries, and others might relate to quality of life. The corresponding univariate statistics might include Pearson's chi-square, several t-statistics, and several Pitman correlations.

Permuting the labels on the observation vectors yields a new matrix $\boldsymbol{X}'$ and a new vector $\boldsymbol{T}' = \boldsymbol{T}(\boldsymbol{X}')$ of univariate statistics. To obtain a single summary statistic encompassing the information provided by all the observations, we proceed as follows:

1. Generate a large number N of permutations of $\boldsymbol{X}$ and thus obtain N vectors of univariate test statistics $\boldsymbol{T}_i$, $i = 1, \ldots, N$.
2. Rank the $N+1$ values of each single-variable test statistic separately. The rank should be related to the extent to which the statistic favors the alternative. For example, if large values of T_{ik} are to be expected when the principal hypothesis concerning the kth variable is false, then,
$$R_{ik} = R(T_{ik}) = \sum_h I(T_{hk} \leq T_{ik}], \quad \text{for } i = 0, \ldots, N,$$
where the indicator function $I[\mathrm{E}]$ takes values 1 or 0 according to whether the event E is true or false.
3. Combine the ranks of the K individual univariate tests using Fisher's omnibus statistic
$$U_i = -\sum_{k=1}^{K} \log\left[\frac{N + 0.5 - R_{ik}}{N+1}\right]; i = 1, \ldots, N.$$
4. Determine from the individual permutation distributions the marginal significance level of each of the single-variable statistics for the original nonpermuted observations,
$$p_k = \frac{0.5 + \sum_{m=1}^{N} I[T_{mk} \geq T_{0k}]}{N+1}, \quad k = 1, \ldots, K.$$
5. Combine these values into a single statistic
$$U_0 = -\sum_{k=1}^{K} \log[p_k] = -\sum_{k=1}^{K} \log\left[\frac{N + 0.5 - R_{0k}}{N+1}\right];$$
note that R_{0k} can take any value in the range 0 to N.
6. Determine the significance level of the combined test,
$$p = \frac{0.5 + \sum_{m=1}^{N} I[U_m \geq U_0]}{N+1}.$$

Liptak's or Tippett's combining functions, described in Section 5.2.1, can be employed in preference to Fisher's.

The range of application of this method is very general. If H_i, A_i denote the hypothesis, alternative associated with the ith variable, respectively, then the multivariate hypothesis you are testing is H_1 and H_2 and $\ldots H_J$ and the