CHAPTER TWELVE

Key Concepts

- univariate analysis
- multivariate analysis (MANOVA)
- multivariate general linear models
- discriminant analysis, discriminant function
- logistic, or logit transformation, logistic regression
- survival time, singly and progressively censored data, survivorship function
- hazard function, proportional hazards, Cox’s regression model, life-table method, Kaplan–Meier method
- permutation test, randomization test, Fisher’s exact test
- resampling, bootstrap, jackknife, cross-validation
- construction data set, training data set, validation data set
Some Specialized Techniques

SYMBOLS AND ABBREVIATIONS

\[ S(t) \] survivorship function (cumulative survival rate)

We have presented in the previous chapters basic concepts that should serve as building blocks for further study. These concepts have been illustrated by describing some of the common statistical methods found in the scientific literature. It would, however, be impossible to cover in a single book all the statistical techniques that are used in genetic and epidemiologic research. In this chapter we briefly familiarize you with some of the more specialized techniques of statistical analysis. Although the choice of which techniques to include and which to exclude is somewhat arbitrary, our aim has been to cover a few of the advanced methods of analysis that are more frequently encountered in research articles.

MULTIVARIATE ANALYSIS

Interest often centers on the simultaneous analysis of several response variables rather than a single response variable. Let us suppose, for example, that a study is designed to determine the effect of a treatment or allele on the following variables: diastolic blood pressure, serum cholesterol, and body weight. In particular, let us suppose that the purpose of the study is to compare the means of these three response variables in a treated group to the corresponding means of these three variables in a control group, or in a group of persons carrying a particular allele and a group not carrying that allele. If one focused on a single variable (e.g., diastolic blood pressure), then one of the methods described earlier could be used to analyze the data. All the methods we have described so far are univariate methods, in that only one random response variable (variates) is involved. (Sometimes multiple regression, in which there is more than one predictor variable, is also called a multivariate method. This terminology, however, is incorrect if there is only one response variable.) If, on the other hand, all three variables
are analyzed simultaneously as response variables, then the analysis is termed multivariate. For each of the univariate methods described earlier, there is an analogous multivariate method. For example, the multivariate analogue of Student’s $t$-test for comparing the means of two groups is called Hotelling’s $T^2$-test, named after the American statistician Harold Hotelling (1895–1973). Similarly, we can have multivariate regression analysis, multivariate analysis of variance (MANOVA), and multivariate analysis of covariance. Earlier, we discussed longitudinal and repeated measures data. These are special types of multivariate data and we must consider this multivariate aspect of the data in the statistical analysis. A unified approach to multivariate analysis for comparing group means is provided by multivariate general linear models.

You may wonder what the advantage is of performing a multivariate analysis rather than performing a set of univariate analyses. Why not, in our example, simply perform three $t$-tests: one for diastolic blood pressure, one for serum cholesterol, and one for body weight? Multivariate analysis has two advantages. First, it helps overcome the problem that, as the number of statistical tests performed at a given significance level increases, so does the probability (under $H_0$) of finding at least one significant result. Suppose there are 20 response variables and we perform 20 $t$-tests at the 5% significance level (i.e., we decide beforehand to declare a significant finding if any one of the 20 $p$-values is less than 0.05). As discussed in Chapter 8, if there are no pairwise differences between the means, by chance alone we should expect one of the tests to yield a significant result. If, however, we first perform a multivariate test, which takes account of the fact that 20 comparisons are being made, then we can appropriately control the overall probability of a type I error to be, for example, 0.05. When we perform 20 $t$-tests each at the 5% significance level, the overall probability of a type I error (i.e., the probability of rejecting the null hypothesis that all 20 pairs of means are equal when in fact they are equal) is much larger than 0.05. We discussed in Chapter 8 how we can use the Bonferroni method to obtain an upper bound for that probability, or Šidák’s method when the tests are independent; multivariate tests take advantage of the fact that the various response variables are not independent to increase power.

A second advantage of multivariate tests is that they are more sensitive to detecting group differences that depend on certain relationships among the response variables. This can best be seen by considering a simple situation in which there are just two groups and two response variables, $y_1$ and $y_2$, illustrated graphically in Figure 12.1. Here we have graphed, for two samples of 10 study units each, the values $y_1$ and $y_2$ observed on each study unit. It is clear from this scatter diagram that the two groups are completely separate, and a multivariate test of these data would show a highly significant difference between the two groups. But if we were to perform a $t$-test on the 20 values of $y_1$, or on the 20 values of $y_2$, neither result
Figure 12.1  Scatterplot of the variables $y_1$ and $y_2$ measured on the ten study units in each of two groups, indicated • and ●.

would be very significant because there is almost complete overlap between the two groups on each variable singly.

**DISCRIMINANT ANALYSIS**

Let us consider the situation in which it is unknown to which of two groups or populations a person belongs. For example, we may wish to know whether or not a woman is a carrier of the sex-linked hemophilia gene and hence has a risk of bearing a son with hemophilia. Suppose we have laboratory data available on the woman and we wish to use the information to classify her. Using a procedure known as discriminant analysis, it is possible to determine a mathematical function for this purpose, from data on a set of previously classified women. Thus, we would obtain two samples of women, one of women known to carry the disease allele (so-called obligate carriers) and one of women known not to carry that allele. (If a woman has two hemophiliac sons, for example, she is an obligate carrier; if she has no relatives with hemophilia, on the other hand, we can be virtually certain she does not carry the hemophilia allele.) A blood sample is taken from each woman and a set of relevant measurements, such as clotting-factor levels, are determined. The result of a discriminant analysis applied to these data is a discriminant function that can be used to help classify a woman whose maternal uncle (but no other relative), for example, has hemophilia. In the case of hemophilia A, if we let

$$y_1 = \log \text{(clotting factor XIII coagulant activity level)}$$

and

$$y_2 = \log \text{(clotting factor XIII-related antigen level)},$$
the discriminant function derived from these two variables is (approximately) \(3y_1 - 2y_2\). If this function is applied to the 20 points plotted in Figure 12.1, for example, the two groups are found to be distinct, with no overlapping of their ranges.

Discriminant analysis can also be used to classify individuals into one of several disease categories, based on vital signs, laboratory data, or both. There will usually be errors associated with such classifications, and we try to develop discriminant functions that will correctly classify individuals with a high probability. Because we cannot know whether a particular individual is classified correctly, we often estimate the probability of an individual belonging to each population. The higher the probability associated with a person belonging to a particular disease category, the more confidence we have that we can correctly classify that person.

**LOGISTIC REGRESSION**

In fitting a statistical model to a set of data, sometimes the response variable is dichotomous, whereas the predictor variables are continuous, discrete, or both. In the simplest situation, we would have one response and one predictor variable. For example, the response variable may be success or failure after a treatment, and we wish to model this response (the proportion of successes or failures) at selected doses of some treatment. In slightly more complex situations, we may want to model the proportion of failures (e.g., the proportion of a population with disease) in terms of suspected risk factors such as age, weight, and blood pressure. In cases such as these, the cumulative distributions tend to be S-shaped, or tilted S-shaped, as in Figure 12.2. This characteristic shape arises because failures often occur infrequently at low levels of the independent variable(s), then there is a range in which the failure rate increases rapidly, and finally there is a range in which

![Figure 12.2](image)

*Figure 12.2* Example of a curve depicting the cumulative proportion of failures as a function of a predictor variable \(x\).
most of the failures have occurred and so additional failures occur less frequently again.

A family of mathematical equations that has a shape resembling that of Figure 12.2 is given by the equation

\[
y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}
\]

or, equivalently,

\[
y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.
\]

With further algebraic manipulation, this is the same as

\[
\frac{y}{1 - y} = e^{\beta_0 + \beta_1 x}
\]

or, taking natural logarithms of both sides of this equation,

\[
\log_e \left[ \frac{y}{1 - y} \right] = \beta_0 + \beta_1 x.
\]

Thus, we have a transformation that converts the curve in Figure 12.2 into a straight line. This transformation is called the logistic or logit transformation; that is, the logistic transformation, or logit, of \( y \) is \( \log_e [y/(1 - y)] \). This is the basis of a logistic regression model, in which the logit of the response random variable \( Y \) (a proportion) is regressed on the predictor variable \( x \); the model is

\[
\log_e \left[ \frac{Y}{1 - Y} \right] = \beta_0 + \beta_1 x + \varepsilon
\]

where \( \varepsilon \) is a random error. There may also be several predictor variables, in which case the model is of the form:

\[
\log_e \left[ \frac{Y}{1 - Y} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon.
\]

A variety of computer programs are available for obtaining maximum likelihood estimates of the parameters \( \beta_0, \beta_1, \ldots, \beta_p \) of this model, and for testing hypotheses about them using the likelihood ratio criterion. The use of logistic regression models is fairly common in the medical literature, especially with \( Y \)
representing the probability of disease, so that $Y/(1 - Y)$ gives the odds in favor of
disease. Thus the natural logarithm of the odds is estimated by the regression func-
tion $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$, and the odds are estimated by $e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}$.
Now suppose, for example, that $x_1 = 1$ if there is exposure to some environmental
factor (or a particular allele is present), and $x_1 = 0$ if there is no such exposure
(that allele is absent). Then the odds ratio for exposed (carriers) versus unexposed
(non-carriers) is estimated as

$$\frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{e^{\beta_0 + \beta_2 x_2 + \ldots + \beta_p x_p}} = e^{\beta_1}.$$ 

Thus $\beta_1$ is the natural logarithm of the odds ratio for exposed versus unexposed
(carriers versus noncarriers) in this example. Recall that odds ratios are particularly
useful statistics for summarizing the results of case–control studies (Chapter 3). We
cannot estimate the probability of disease in such studies; but, by letting $Y$ be the
proportion of cases in the study, logistic regression can be used to find the odds
ratios for several different kinds of exposures and/or genotypes ($x_1, x_2, \text{etc.}$). The
 correponding estimated regression function (i.e., $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$)
can also be used as a discriminant function to help classify future persons into one
of the two classes (which in this instance are ‘disease’ and ‘no disease’).

To illustrate, suppose a case–control study is conducted where cases are persons
with cardiovascular disease and controls are persons who do not have this
disease. Let $Y = 1$ for cases and $Y = 0$ for controls and consider the predictor vari-
ables: age in years, gender (female $= 1$, male $= 0$), blood pressure (normal $= 0$,
high $= 1$) body mass index (weight/height$^2$ in kg/m$^2$) and history of type 2 diabetes
(no $= 0$, yes $= 1$). A typical analysis of the logistic regression model for the data might
be as shown in Table 12.1. We see that the coefficients in the logistic regression
model are significantly different from 0 for age, gender, blood pressure and history
of type 2 diabetes, but not for body mass index. The odds ratios are correspondingly
significantly greater than 1 except for body mass index. For example, the odds ratio

<table>
<thead>
<tr>
<th>Parameter</th>
<th>d.f.</th>
<th>Estimate</th>
<th>SE</th>
<th>Chi-square</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-7.0906</td>
<td>0.7828</td>
<td>82.04</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>0.0846</td>
<td>0.0120</td>
<td>49.77</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>0.3967</td>
<td>0.0748</td>
<td>28.14</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>1</td>
<td>0.3937</td>
<td>0.1027</td>
<td>14.70</td>
<td>0.0001</td>
</tr>
<tr>
<td>Body mass index</td>
<td>1</td>
<td>0.0027</td>
<td>0.0110</td>
<td>0.06</td>
<td>0.8049</td>
</tr>
<tr>
<td>History type 2 diabetes</td>
<td>1</td>
<td>0.2939</td>
<td>0.0817</td>
<td>12.95</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
### Odds ratio analysis

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>OR</th>
<th>95% CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>1.09</td>
<td>1.06 to 1.11</td>
</tr>
<tr>
<td>Gender (male vs female)</td>
<td>2.21</td>
<td>1.65 to 2.96</td>
</tr>
<tr>
<td>Blood pressure (high vs normal)</td>
<td>2.20</td>
<td>1.47 to 3.29</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>1.00</td>
<td>0.98 to 1.02</td>
</tr>
<tr>
<td>History type 2 diabetes (yes vs no)</td>
<td>1.80</td>
<td>1.31 to 2.48</td>
</tr>
</tbody>
</table>

for persons with high blood pressure is 2.20 with 95% confidence limits 1.47 to 3.29. Therefore, in this sample the odds of having cardiovascular disease is estimated to be 2.2 times greater if a person has high blood pressure compared to a person with normal blood pressure. Moreover, the odds ratio for age is 1.09, suggesting that the odds ratio increases by 9% for every year a person’s age increases. Note that each odds ratio is adjusted for other factors in the model.

### ANALYSIS OF SURVIVAL TIMES

In some studies, especially clinical trials, the response variable of interest may be the amount of time from some initial observation until the occurrence of an event, such as recurrence of disease, death, or some other type of failure. This time from initial observation until failure is called the survival time. Statistical analysis of a group of survival times usually focuses on the probability of surviving a given length of time, or on the mean or median survival time.

A distinguishing feature of survival data is the fact that the distribution of survival times is often skewed and far from normal. Furthermore, the exact survival times of some of the study units may be unknown. For example, a group of subjects may all enter a study at the same time, but some may not have ‘failed’ by the end of the study, or they may be lost to follow-up at some point in the study. In such cases the survival times are said to be censored. If a study is conducted so that subjects are observed until a pre-specified proportion (e.g., 60%) have failed, or if all subjects are observed for a fixed period (e.g., 5 years) and some subjects have not failed by the end of that period, the resulting survival times for the survivors are said to be singly censored. In most clinical studies, however, patients are recruited into the study over time, and each patient is observed for a different length of time. Then, if some of the patients have not failed by the end of the study, the resulting survival times are said to be progressively censored.

A distribution of survival times can be characterized by one of three functions: (1) the survivorship function, (2) the probability density function, and (3) the hazard function.
The survivorship function \( S(t) \) is defined as the probability that a study unit survives longer than time \( t \); thus, if \( T \) is the random variable denoting survival time,

\[
S(t) = P(T > t).
\]

\( S(t) \) is also known as the *cumulative survival rate*, and the graph of \( S(t) \) is called the *survival curve* (Figure 12.3). At any time \( t \), \( S(t) \) gives the proportion still surviving at time \( t \). Recall that the cumulative distribution function of \( T \) is given by \( F(t) = P(T \leq t) \). Hence it follows that

\[
F(t) = 1 - S(t).
\]

The corresponding density function is the probability density function \( f(t) \) of the survival time. Areas under this curve represent the probability of failure in intervals of time.

The *hazard function* of survival time \( T \) is the density of failure at a particular time, given that there has been survival until that time. The hazard function is also known as the *instantaneous failure rate*, the *force of mortality*, or the *conditional failure rate*. It can be thought of as the ‘instantaneous’ probability of failure at a particular time given there has been survival until that time. Because time is a continuous variable, however, it is a probability density. It is equal to \( f(t)/S(t) \). In 1972 Sir David Cox, a British statistician, introduced a method of analyzing survival times based on the assumption that the effect of each of the predictors \( x_1, x_2, \ldots, x_p \) is to multiply the whole hazard function by a certain amount. The underlying model is therefore called a *proportional hazards model* or sometimes simply *Cox’s regression model*. Specifically, denoting the hazard function \( h(t) \), the model can be written as
\[ h(t) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} \]

where \( h_0(t) \) is the hazard when all the \( x \) variables equal zero. The regression coefficients \( \beta_1, \beta_2, \ldots, \beta_p \) are estimated by a maximum likelihood method that does not depend on the shape of \( h(t) \) or \( h_0(t) \), and the estimates measure the effect of each predictor on the hazard function. If, for example, \( x_2 \) is the amount of a particular food eaten, then the hazard function is multiplied by \( e^{\beta_2} \) for each unit of that food eaten; \( \beta_2 > 0 \) would imply that the food has a harmful effect (increasing the hazard), while \( \beta_2 < 0 \) would imply a beneficial effect (decreasing the hazard).

Just as for logistic regression, there are computer programs for obtaining maximum likelihood estimates of the parameters and for testing hypotheses about them, in large samples, using the likelihood ratio criterion.

**ESTIMATING SURVIVAL CURVES**

We shall describe two methods of estimating survival curves: (1) the life-table method and (2) the Kaplan–Meier method. In the life-table method, the survival times are first grouped into fixed intervals such as months or years. Let \( n_i \) be the number of study units surviving at the beginning of the \( i \)th interval, \( d_i \) the number of failures in the \( i \)th interval, and \( c_i \) the number of censored survival times in the \( i \)th interval. Then the probability that a study unit that has survived to the beginning of the \( i \)th interval will survive through to the end of that interval is estimated as

\[ s_i = \frac{n_i - d_i - c_i/2}{n_i - c_i/2} \]

(It is assumed that the censored individuals leave randomly throughout the interval, so that on an average only half of them are present during the interval.) The overall probability of surviving until the end of the \( k \)th interval is estimated as the product of the probabilities of surviving through each of the first \( k \) intervals, that is,

\[ s_1 s_2 \ldots s_k. \]

The Kaplan–Meier method of estimating survival curves uses the exact failure times rather than grouping the survival times into intervals. Denote the ranked times of failure or censoring for the \( m \) subjects in a group by \( t_1 < t_2 < \ldots < t_i < \ldots < t_m \). Let \( u_i \) be the number of units surviving at time \( t_i \), and \( f_i \) the number that fail at time \( t_i \). A unit with survival time censored at time \( t_i \) is assumed to survive up to and
including time \( t_i \). Then the probability that a study unit that has survived to time \( t_{i-1} \) will survive to time \( t_i \) is estimated as

\[
q_i = \frac{u_i}{u_i + f_i}.
\]

As before, the overall probability of surviving to time \( t_k \) is estimated as

\[q_1q_2 \cdots q_k.
\]

Expressions are available for the standard deviations of each of these estimates of the survival times.

To illustrate the computations for estimating Kaplan–Meier survival curves, suppose that two interventions are investigated in a sample of 64 patients who are suffering from a disease with a short survival expectancy. Persons were randomly assigned to intervention A or intervention B so that \( m = 32 \) patients are allocated to each intervention. Further suppose the patients are followed until death, until lost to follow-up or until the end of study after 3 years (36 months) of follow-up. The time each patient was followed is shown for the first six events (months until patient

<table>
<thead>
<tr>
<th>Table 12.2</th>
<th>First six events for each of the two intervention groups</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(i) Intervention A</strong></td>
<td><strong>(ii) Intervention B</strong></td>
</tr>
<tr>
<td><strong>Months</strong></td>
<td><strong>No. Left</strong></td>
</tr>
<tr>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>14</td>
<td>26</td>
</tr>
</tbody>
</table>
died or survival time censored) in the first column of Table 12.2. For example, the first patient in intervention group A died at 2 months, so

$$q_1 = \frac{u_1}{u_1 + f_1} = \frac{31}{31 + 1} = 0.9688.$$  

Similarly,

$$q_2 = \frac{u_2}{u_2 + f_2} = \frac{30}{30 + 1} = 0.9677,$$

so that the estimated probability of survival at 5 months is $q_1 q_2 = 0.9688 \times 0.9677 = 0.9375$. The estimated survival curves for all 36 months are shown in Figure 12.4; in this figure, circles indicate the times that patients were censored. On employing a test for the equality of the two curves known as the log rank test, we find a chi-square of 6.03 with 1 d.f. and hence $p = 0.014$, suggesting the survival profile is better with intervention B.

**Figure 12.4**  Kaplan–Meier survival curves for two intervention groups.
PERMUTATION TESTS

Throughout this book, we have described methods of testing hypotheses that assume particular sampling distributions for the test statistic when the null hypothesis is true. These distributions can often be justified on the basis of having large samples of independent observations. But if the samples are small and/or the observations are not independent, these methods may not be valid. In Chapter 8 we described Bayesian methods that have been proposed to alleviate these problems, especially when a large number of tests are performed relative to the sample size, but to use these methods we also need to assume prior distributions for all the model parameters. In this section we describe a general frequentist method of testing that makes no distributional assumptions, but rather depends on proper experimental design. These tests are known as permutation tests, or randomization tests, and can be employed to compare means, for example in most sampling plans or experimental designs. Here we describe the method for completely randomized and stratified designs, and for the special case of a $2 \times 2$ contingency table.

COMPLETELY RANDOMIZED DESIGN

For brevity, we introduce the basic idea for a completely randomized design with only two treatment groups. Suppose subjects are randomly assigned to one of the two treatment groups, denoted A and B, in such a manner that $m$ subjects are assigned to group A and $n$ to group B. There are a total of

$$M = \binom{m+n}{m} = \frac{(m+n)!}{m!n!}$$

ways that this can be accomplished, and we say that there are $M$ permutations of the $m+n$ subjects with $m$ in group A and $n$ in group B. Let $\bar{X}_i$ and $\bar{Y}_i$ represent the sample means of a variable of interest in groups A and B, respectively, and let $D = \bar{X} - \bar{Y}$ represent the difference in the means. Provided the distribution of the random variable is the same in groups A and B, including their having the same mean, we could exchange any of the observations in A for observations in B without affecting the null sampling distribution of $D$. Now, let $d_{\text{obs}} = \bar{x} - \bar{y}$ denote the difference in the means for the data observed on execution of the experiment. Once the data are observed, we can calculate the $M$ values of $d$ corresponding to each permutation of the observed sample data. If the subjects are in fact randomly assigned to treatment groups, each permutation of the data is equally likely and thus has a priori probability $1/M$ of being the experimental outcome for the given set of data. Therefore, if we arrange the $M$ values of $d$ in order from smallest to largest, we can produce an empirical cumulative probability distribution (known as
the permutation distribution) from which we can determine the exact probability of a random experiment resulting in a value of $d$ as extreme as, or more extreme than, the $d_{\text{obs}}$ found in the permutation corresponding to the experimental outcome. Thus the $p$-value for the test is the proportion of the $M$ values of $d$ that are as extreme as, or more extreme, than the observed $d_{\text{obs}}$. When calculating the $p$-value we must include both extreme tails for a two-sided alternative hypothesis and the appropriate tail for one-sided alternative. Because we can find the exact probability of observing specific outcomes under the null hypothesis, the permutation test is often referred to as an exact test.

As a numerical example, consider an experiment conducted in a small pilot study conducted in planning a larger clinical trial to investigate two treatments, A and B, aimed at lowering serum glucose levels as measured by hemoglobin A1c (%). Eleven diabetic patients were available and they were randomly assigned to receive either A or B, with six receiving A and five receiving B. Interest was in testing the null hypothesis that the mean for A is equal to the mean for B against the one-directional alternative that the mean for A is less than that for B. Suppose the following HbA1c values were observed:

\[
\begin{align*}
\text{A: } & \quad 6.0, 5.5, 5.3, 6.2, 6.7, 5.1 \\
\text{B: } & \quad 7.3, 6.9, 7.6, 8.1, 6.5
\end{align*}
\]

Here, $\bar{x} = (6.0 + 5.5 + 5.3 + 6.2 + 6.7 + 5.1)/6 = 5.80$, $\bar{y} = (7.3 + 6.9 + 7.6 + 8.1 + 6.5)/5 = 7.28$, and $d_{\text{obs}} = \bar{x} - \bar{y} = 5.80 - 7.28 = -1.48$. For these data there are

\[
M = \frac{11!}{6!5!} = 462
\]

permutations of the data possible, but to perform the test we need to identify only those permutations of the data that lead to values of $d$ as extreme as, or more extreme than, $d_{\text{obs}} = -1.48$ in the direction that favors the alternative hypothesis. There is only one permutation that is more extreme in this context and that is the one obtained by exchanging the 6.7 (the largest value) in group A for the 6.5 (the smallest value) in group B. There are therefore two permutations as or more extreme, leading to $p = 2/462 = 0.004329$ and the conclusion that the mean was significantly lower in the patients that received treatment A.

As the number of treatments and/or the total sample size increases, the number of permutations becomes larger and it becomes more difficult to identify the extreme permutations. Fortunately, we can employ an asymptotically equivalent test by sampling from the permutation distribution. We re-randomize the subjects and their now known data to the treatment groups a large number of times to
produce a very close approximation to the exact permutation distribution. Let \( R \) = the number of replications produced this way – for example, we choose \( R = 10,000 \). For each replication, we calculate \( d \) and, in exactly the same way determine the proportion of these replications as extreme as, or more extreme than, the \( d_{\text{obs}} \) calculated from the original experimental results. This results in a close approximation to the exact \( p \)-value so long as \( R \times p \) is large enough. For example, if \( p = 0.05 \), \( R = 1600 \) will result in a 95% confidence interval from about 0.04 to 0.06.

### RANDOMIZED BLOCK (STRATIFIED) DESIGNS

In this section, we continue our discussion of permutation tests by again considering just two treatments, but in a randomized block design with only two subjects per block, as in designs where each block is an individually matched pair of subjects. The members of each pair are randomly assigned to either treatment A or treatment B. Suppose \( m \) pairs of subjects are available for an experiment so that \( m \) subjects are randomly assigned to receive A and \( m \) are randomly assigned to receive B. In this setting there is a total of \( M = 2^m \) ways of assigning the subjects to treatments. Each of these permutations a priori is equally likely under the randomization plan if there is no difference in the effects of treatments A and B. As before, we let \( X \) and \( Y \) denote the outcome random variable for subjects who receive A and B, respectively. Note that for each pair we can calculate the difference \( x - y \), and then average these individual differences; or equivalently, we can average \( x \) and \( y \) and then calculate the difference between the averages, as was done in the context of a completely randomized design. Accordingly, we can produce the permutation distribution and determine exact \( p \)-values just as for a completely randomized design. As before, when the number of permutations is large, we can sample from the permutation distribution. The main difference required for a stratified design is that the permutation must be performed within the strata – two permutations within each block in the above example – with the result that the number of permutations is different in the two cases. By permuting within strata we satisfy what is known as the exchangeability requirement necessary for any permutation test to be valid. Permutation tests will often increase the power of non-independent genetic tests, for example to obtain genome-wide significance levels; but if there is differential population stratification among the groups to be compared, identifying the strata and performing the permutations within strata is required for validity. On the other hand, because the permutation test tests the equality of the two or more group distributions, not just their means, there is no need to identify the strata if the stratification is the same in all the groups to be compared – though doing so and preforming the permutations within strata would usually lead to a more powerful test.
FISHER’S EXACT TEST

Another version of the permutation test is known as Fisher’s exact test and is widely used to analyze categorical data arranged in a $2 \times 2$ table with small numbers. We illustrate Fisher’s exact test by assuming the response is binary and interest is in comparing the proportions of subjects in two independent samples whose responses can be tabulated in the same category. The data structure can be summarized as in Table 12.3.

Table 12.3 Categorical data structure for comparing a binary response in two independent samples

<table>
<thead>
<tr>
<th>Response</th>
<th>A</th>
<th>B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$a$</td>
<td>$b$</td>
<td>$a + b$</td>
</tr>
<tr>
<td>0</td>
<td>$c$</td>
<td>$d$</td>
<td>$c + d$</td>
</tr>
<tr>
<td>Total</td>
<td>$a + c$</td>
<td>$b + d$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

The null hypothesis is that the proportion of subjects in group A who respond ‘1’ is equal to the proportion in group B who respond ‘1’. The labeling of the responses as 1 or 0 is completely arbitrary (the response can be success or fail, yes or no, improved or did not improve, etc.) but it is useful in establishing a link to our previous discussion of the permutation test. Thus, on recognizing that the proportion who respond 1 is just the average of the 0s and 1s, $\bar{x}$ for group A and $\bar{y}$ for group B, it is easy to see that Fisher’s exact test employs the same strategy as that invoked in the permutation test for the equality of two means described above. Although the notation is different, we calculate $M$, the number of ways the $N$ subjects can be assigned to the treatment groups, in the same way we did before; that is, the $N$ subjects can be assigned to groups A and B with $a + c$ and $b + d$ subjects, respectively, in $M$ possible ways, where

$$M = \binom{N}{a + c} = \frac{N!}{(a + c)! (b + d)!}.$$

Once the data are collected, we could list the $M$ permutations and for each calculate $d = \bar{x} - \bar{y}$ and determine from the permutation distribution the significance of $d_{obs}$.

Fisher’s exact test assumes not only that the column totals $a + c$ and $b + d$ are fixed, but also in addition that the row totals $a + b$ and $c + d$ are fixed. We then recognize that once any one of the values $a$, $b$, $c$ and $d$ is known, the remaining three are determined by the constraints imposed by the fixed margins assumption.
and this limits the number of distinct tables possible. It can be shown that the permutation probability of each such table is equal to

$$\frac{(a+b)}{a} \left( \frac{c+d}{c} \right) = \frac{[(a+b)! (a+c)! (b+d)! (c+d)!]}{N! a! b! c! d!},$$

which is the hypergeometric probability distribution function, to be used to find the exact null hypothesis probability of observing the specific set of frequencies $a, b, c$ and $d$.

To illustrate, we consider the set of observed data that are summarized in Table 12.4. The null hypothesis of interest was 'the proportion of 1s with treatment A is equal to the proportion of 1s with treatment B'; that is, $H_0: \pi_A = \pi_B$. We noted in Chapter 9 that if this hypothesis is true and the expected frequency in each cell of the table is 5 or more, we could perform a chi-square test. The chi-square test is not appropriate here because all the expected values are less than five. With the margins fixed as observed, the number of subjects in group A who respond 1 could be $a = 0, 1, \ldots, 8$. Once $a$ is determined, the other 3 observations are also determined under the constraints of fixed margins. Hence, there are 9 distinct

<table>
<thead>
<tr>
<th>Table 12.4</th>
<th>Illustrative data: Fisher’s exact test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 12.5</th>
<th>Probability distribution for $2 \times 2$ tables with margins fixed at the values shown in Table 12.4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table</strong></td>
<td><strong>a</strong></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>
tables that could have been observed, one for each of the 9 possible values of \( a \). We calculate the exact probability of observing each possible table with the margins fixed at the values shown in Table 12.4 using the above null hypothesis probability for the corresponding table. These probabilities are given in Table 12.5. We illustrate how this table is used to test the null hypothesis for each of three \textit{a priori} pre-specified research hypotheses, only one of which would be used in a given experiment. We consider first the two one-sided alternative research hypotheses \( \pi_A > \pi_B \) and \( \pi_A < \pi_B \), then the two-sided alternative \( \pi_A \neq \pi_B \).

1. \( \pi_A < \pi_B \). To obtain the \( p \)-value for this research hypothesis, we consider outcomes that would be extreme in the sense that the alternative hypothesis would be more likely than the null if the null is in fact false; here, such values would be associated with larger estimated values of \( \pi_A \) and so probabilities in the lower part of Table 12.5. In Table 12.4, \( a = 7 \) and values as extreme or even more extreme would be \( a = 7 \) or \( a = 8 \). The \( p \)-value for the outcome \( a = 7 \) would then be \( p = 0.01185 + 0.00037 = 0.01222 \) and it would be appropriate to reject the null hypothesis and conclude that \( \pi_A > \pi_B \).

2. \( \pi_A > \pi_B \). For this research hypothesis, outcomes that tend to refute the null hypothesis are associated with small values of \( a \) and hence and the probabilities in the upper part of Table 12.5. In Table 12.4, \( a = 7 \) and values as extreme or even more extreme would be \( a = 0, 1, 2, 3, 4, 5, 6 \) or \( 7 \). The \( p \)-value for would then be \( p = 0.00004 + 0.00296 + \ldots + 0.01185 = 0.98778 \) so it would be inappropriate to reject the null hypothesis. We would conclude that the null hypothesis is plausible in light of these data.

3. \( \pi_A \neq \pi_B \). Here, extreme outcomes in either direction tend to cast doubt on the null hypothesis. We see that \( 0.00004 + 0.00296 = 0.00300 < 0.01222 \) so that \( a = 0 \) or \( 1 \) are even more extreme in the opposite direction than \( a = 7 \) or \( 8 \). Therefore, the two-sided \( p \)-value for the observed outcome would be \( 0.00300 + 0.01222 = 0.01522 \) and would lead to rejection of the null hypothesis.

**RESAMPLING METHODS**

A variety of methods have been proposed for making statistical inferences that involve treating a sample of data as a ‘population’ and then repeatedly selecting samples (\textit{resampling}) from that population in order to estimate the sampling distribution or one or more parameters of interest. These methods do not require any distributional assumptions, such as the data or estimators have an underlying normal distribution, and are therefore \textit{distribution-free}. They are simple in concept but can be tedious to implement in some applications. They are especially useful in situations where we may be interested in a estimating or testing hypotheses
about a parameter in an application where a theory has not been developed for that purpose; or a theory has been developed but the resulting methods may not have desirable statistical properties; or a theory has been developed but the resulting computational formulas are complicated and difficult to use; or, in some instances, where a usual ‘parametric’ estimation may be impossible. By resampling a large number of subsamples from the original sample and estimating the parameters of interest by (known or intuitive) estimators from each subsample, we can determine the sampling distribution of the estimators and investigate the accuracy, precision and other properties of these estimators. Their main advantage is that they provide robust estimates of standard errors and confidence limits for population parameters.

**BOOTSTRAP RESAMPLING**

*Bootstrap* resampling is an approach to statistical inference that obtained its name from the notion of leveraging an initial sample of data to gain information for making inferences that would ordinarily require a much larger data resource, analogous to ‘pulling oneself up by one’s bootstraps’. Suppose we have a random sample of size \( n \) from a population and we wish to use these data to estimate a population parameter, complete with a 95% confidence interval. Bootstrap resampling or ‘bootstrapping’ involves taking \( B \) random replicate samples of the same size \( n \) from this original sample of size \( n \). However, we sample with replacement. In other words, when forming each sample, after randomly selecting each observation it is put back into the sample before randomly selecting the next observation. In this way we can select samples that are the same size as the original sample without all the replicate samples being identical, so that we have in a sense ‘pulled up’ more information than first appeared to be available in the original sample. The number of replicate samples is somewhat arbitrary, but \( B = 10,000 \) is often used and appears to provide suitable results for most problems. We describe a simple use of bootstrapping to give the basic idea, but keep in mind that there are many more complex applications where the result is not known beforehand.

We first explain how bootstrapping can be used to estimate the bias of an estimator. Recall that in Chapter 6 we said if the average of the sample estimates for all possible samples equals the value of the parameter being estimated, we say the estimator is unbiased; otherwise it is biased. The bias of an estimator is the difference between the mean of the estimator, i.e. the mean of the estimator’s distribution, and the parameter being estimated. We also stated in Chapter 6 that if we used the sample size \( n \) as the divisor when estimating the population variance \( \sigma^2 \), the estimator is biased by a factor \((n - 1)/n\). We could verify this result using bootstrapping as follows. We first estimate the variance of the sample using the
divisor $n$ instead of $n - 1$. We then select $B$ replicate samples with replacement of size $n$ from the original sample and again use the divisor $n$ instead of $n - 1$ to estimate the variance from each of these $B$ bootstrap samples. If we now average these $B$ estimates, we find that this average is larger than the estimate obtained (using the same divisor, $n$) from the original sample, by a factor close to $(n - 1)/n$ (when $B$ is large). This means that the original estimate of the variance was too small. If the bias of any estimate is not close to zero, we should correct the estimate by subtracting the bias from it.

Once we have estimates of a parameter of interest from each of the $B$ replicate samples, we can plot these estimates as a histogram or a cumulative plot to characterize the general properties of this sampling distribution. In particular, we can find percentiles of this empirical distribution. For example, the 5th and 95th percentiles of this distribution would form a 90% confidence interval for the unknown parameter.

**JACKKNIFE RESAMPLING**

The *jackknife* is another variation of resampling and is similar to bootstrapping, with the main difference being in the way the subsamples are selected. Given a sample of size $n$, the jackknife method is based upon $n$ subsamples each of size $n - 1$, where each subsample is obtained by leaving out one of the original sample observations but no observation is left out of more than one subsample. These $n$ subsamples are then used exactly as the $B$ subsamples were used in bootstrapping to estimate bias and construct confidence intervals.

**CROSS-VALIDATION**

*Cross-validation* is a strategy that is often employed to investigate whether a specific statistical analysis that has been conducted using one set of data produces confirmatory results when it is applied to other, independent sets of data. Rather than conduct completely independent studies for this purpose, a popular practice in investigations that are sufficiently large is to separate the original data set into two or more independent subsets and then replicate or otherwise ‘test’ the analysis in the different subsets. Ideally, the original data would be randomly allocated to the subsets.

In discriminant analysis, for example, one criterion used to evaluate the ability of the estimated discriminant function to reliably classify individuals into one of the two or more populations is to use the estimated function to classify subjects whose population status is known and then tabulate the percentage classified correctly for each population. But if we do this using the same population samples that were
used to construct the discriminant function in the first place, the estimated function is, in a sense, ‘drawn to’ the data used in its construction and, therefore tends to perform better than it would if used to classify the individuals in an independent sample. To cross-validate, we could randomly divide the original data set into two subsets and use one subset to construct the discriminant function and the second to evaluate its performance. The data set used to construct the function is sometimes referred to as the *training data set* and that used for the evaluation is sometimes called the *validation data set*. Further strategies may be employed analogous to bootstrapping and jackknifing. For example we could repeat the process many times so that different subsets are selected as *construction data sets*. An investigation of the stability of the estimated function as different construction sets are used would be informative about the general applicability of the function in practice.

**SUMMARY**

1. Multivariate analysis is the simultaneous analysis of several dependent variables. It allows for appropriate control of the probability of a type I error when many dependent variables are involved; and it can sometimes detect group differences that are not obvious when the variables are examined individually. Every univariate method of analysis has a multivariate analogue.

2. The purpose of discriminant analysis is to find a function of several variables that can help classify a study unit as coming from a particular population.

3. Logistic regression is used to model a proportion of a population (e.g., the proportion with a disease), or a probability, as a function of one or more independent variables. The logistic transformation changes a tilted S-shaped curve into a straight line. The estimated regression coefficients can be interpreted as the logarithms of odds ratios. The estimated regression function can also be used as a discriminant function (e.g., to help classify persons as having a disease or not).

4. Survival analysis is used when the dependent variable is a survival time (i.e., the time to a well-defined event, such as death). The distribution of survival times, which is usually skewed, may be characterized by a probability density function, a survivorship function (the complement of the cumulative distribution function), or a hazard function.

5. The hazard function gives the density of failure at a particular time, given there has been survival up to that time. In the proportional hazards model (Cox’s regression model) it is assumed that the effect of each independent variable is to cause the whole hazard function to be increased or decreased multiplicatively.
6. Survival data are usually censored either singly (if every unit has been observed for the same amount of time) or progressively (if the units have been observed for different lengths of time). Two methods of estimating survival curves (that allow for censoring) are the life-table method and the Kaplan–Meier method.

7. Permutation tests rely on appropriate randomization when a study is conducted to obtain exact significance levels; when applied to nonrandomized studies they assume exchangeability under the null hypothesis of the data being permuted. Fisher’s exact test permutes the data in a contingency table in such a way that the marginal totals remain the same.

8. Bootstrap samples are obtained from a given sample by sampling with replacement. A large number of such samples can be used to estimate empirically the sampling distribution of any statistic. The jackknife creates \( n \) samples of size \( n - 1 \) from a sample of size \( n \) by leaving out each of the observations one at a time. Cross-validation divides the original data set into two or more independent subsets so that independent subsets can be used to perform and evaluate the analysis.

FURTHER READING

Everitt, B.S. (1989) *Statistical Methods for Medical Investigations*. Oxford University Press, New York. (This book gives a good overview of some of the topics in this chapter. Only a limited mathematical background is required to understand the material.)


PROBLEMS

1. An analysis is carried out to study the effects of three treatments on total serum cholesterol in patients with elevated cholesterol levels. The statistical model underlying the analysis included age of the patient as a predictor variable. The resulting analysis is called

A. analysis of covariance
B. multivariate analysis
C. discriminant analysis
D. logistic regression analysis
E. survival analysis
2. An experiment was conducted in which patients were randomly assigned to either an active treatment or a placebo. After 3 months of treatment, data were obtained for four variables: total serum cholesterol, serum triglyceride, systolic blood pressure, and diastolic blood pressure. A statistical analysis was carried out to test the null hypothesis that the treatment had no effect on any of the four variables. The resulting analysis is called

A. univariate analysis  
B. discriminant analysis  
C. logistic regression analysis  
D. survival analysis  
E. multivariate analysis

3. An investigator studied two groups of patients: one group with confirmed coronary heart disease and a second in which overt coronary heart disease was not present. Total serum cholesterol, serum triglyceride, systolic blood pressure, and diastolic blood pressure were determined for each patient. The investigator wished to derive from these data a mathematical function that would help decide whether a patient with unknown coronary heart disease status, but on whom these four variables had been determined, has coronary heart disease. An appropriate statistical method for doing this is

A. univariate analysis  
B. discriminant analysis  
C. survival analysis  
D. analysis of variance  
E. analysis of covariance

4. Which of the following is an advantage of multivariate analysis?

A. The computations for it are simpler.  
B. It requires fewer assumptions.  
C. It allows for proper control of the type I error when tests are performed on many response variables.  
D. It avoids the requirement of randomization.  
E. It always provides a more powerful test than a set of separate univariate analyses.

5. An investigator is studying the probability of disease in relation to several suspected risk-factor variables. A plot of the proportions with disease in
various categories of each of the risk-factor variables indicates that each of the cumulative distributions is shaped like a tilted S. This suggests that the investigator should consider a

A. univariate analysis
B. discriminant analysis
C. analysis of covariance
D. logistic regression analysis
E. none of the above

6. An investigator reported that the data from a study were analyzed using the Kaplan–Meier method. The investigator was most likely studying

A. multivariate data
B. survival data
C. discrete data
D. uncensored data
E. none of the above

7. An investigator wishes to estimate the instantaneous probability that a patient will die, given that the patient has survived a given amount of time since an operation. In other words, the investigator is interested in estimating the following function of time since the operation

A. probability density function
B. survivorship function
C. hazard function
D. cumulative distribution function
E. none of the above

8. A study of survival times of patients receiving coronary bypass operations is terminated while some of the patients are still surviving. For purposes of analysis, the survival times of these patients are said to be

A. discrete
B. multivariate
C. censored
D. terminated
E. none of the above

9. A researcher wishes to develop a statistical model to predict serum cholesterol levels based on a knowledge of five measures of dietary intake. The method for developing such a model can be described as

A. multiple regression analysis
B. multivariate analysis
C. discriminant analysis  
D. analysis by Cox’s regression model  
E. survival analysis

10. An analysis is performed in which the proportion of persons with a disease in a group is divided by the proportion without the disease. A multiple regression analysis is carried out on the logarithm of the resulting ratio. This is an example of a general method known as

A. correlation analysis  
B. multivariate analysis  
C. survival analysis  
D. logistic regression analysis  
E. censored data analysis

11. A statistician is faced with the analysis of a set of data comprising measurements of three continuous dependent variables, observed in an experiment that used a factorial arrangement of treatments in a completely randomized design. Based on this information, the most appropriate method of analysis is

A. discriminant analysis  
B. paired t-test  
C. multivariate analysis of variance  
D. survival analysis  
E. proportional hazard function analysis

12. Cox’s proportional hazards model is used to investigate relationships between survival time and a set of

A. discriminant functions  
B. percentiles  
C. prognostic factors  
D. censored data  
E. cumulative distribution functions

13. All the following are multivariate statistical techniques except

A. Hotelling’s $T^2$-test  
B. MANOVA  
C. discriminant analysis  
D. Student’s t-test  
E. multivariate general linear models
14. Subjects are recruited into a study over time as they come out of intensive care from a particular operation, and the study is terminated after 30% of the subjects have relapsed. The survival time to relapse is said to be

A. a logistic regression
B. progressively censored
C. missing
D. a maximum likelihood estimate
E. multivariate

15. A randomized, double-blind clinical trial was conducted to study the effect of a drug for lowering blood pressure versus a placebo control. The response variables of interest were systolic and diastolic blood pressure. Based on this information, the statistical analysis requires a technique appropriate for

A. data with missing endpoints
B. censored data
C. multivariate response
D. categorical response
E. noncompliance