which lies at a distance d_i from the origin that is at the mean of the data points. This point also turns out to be the mean of the principal component values. Suppose, now, that we drop a line down into the plane that contains the axes corresponding to the first two principal components. This is indicated by the vertical dotted line in the figure. This point in the plane we could now project onto the value for the first and second principal components. These values, with lengths f_{i1} and f_{i2}, are the same as we would get by dropping perpendiculars directly from the point to those two axes. Again, we might assess the adequacy of the characterization of the data point by the first two principal components by comparing the length of its projection in the plane, g_i, with the length of the line from the origin to the original data point, d_i. If we compare the squares of these two lengths, each summed over all of the data points, and use the Pythagorean theorem again, the following results hold:

$$\begin{aligned}\frac{\sum_{i=1}^n g_i^2}{\sum_{i=1}^n d_i^2} &= \frac{\sum_{i=1}^n f_{i1}^2 + \sum_{i=1}^n f_{i2}^2}{\sum_{i=1}^n d_i^2} \\ &= \frac{\sum_{i=1}^n [(Y_{i1} - \overline{Y}_1)^2/(n-1)] + \sum_{i=1}^n [(Y_{i1} - \overline{Y}_2)^2/(n-1)]}{\sum_{i=1}^n d_i^2/(n-1)} \\ &= \frac{V_1 + V_2}{V}\end{aligned}$$

Using this equation, we see that the percent of the variability explained by the first two principal components is the ratio of the squared lengths of the projections onto the plane of the first two principal components divided by the squared lengths of the original data points about their mean. This also gives us a geometric interpretation of the total variance. It is the sum for all the data points of the squares of the distance between the point corresponding to the mean of the sample and the original data points. In other words, the first two principal components may be characterized as giving a plane for which the projected points onto the plane contain as high a proportion as possible of the squared lengths associated with the original data points. From this we see that the percent of variability explained by the first two principal components will be 100 if and only if all of the data points lie in some plane through the origin, which is the mean of the data.

The coefficients associated with the principal components are usually calculated by computer; in general, there is no easy formula to obtain them. Thus, the examples in this chapter will begin with the coefficients for the principal components and their variance. (There is an explicit solution when there are only two variables, and this is given in Problem 14.9.)

Example 14.1. We turn to the data of Table 14.1. Equations for the principal components are

$$\begin{aligned}Y_1 &= -0.6245X + 0.7809Y \\ Y_2 &= 0.7809X + 0.6245Y\end{aligned}$$

For the first data point, $(X, Y) = (-0.52, 0.60)$, the values are

$$\begin{aligned}Y_1 &= -0.6245 \times (-0.52) + 0.7809 \times 0.60 = 0.79 \\ Y_2 &= 0.7809 \times (-0.52) + 0.6245 \times 0.60 = -0.03\end{aligned}$$

If we compute all of the numbers, we find that the values for each of the 20 data points on the principal components are as given in Table 14.2.

Table 14.2 Data Point Values

Data		Principal Component Values		Data		Principal Component Values	
X	Y	Y_1	Y_2	X	Y	Y_1	Y_2
−0.52	0.60	0.79	−0.03	0.08	0.23	0.13	0.21
0.04	−0.51	−0.42	−0.28	−0.06	−0.59	−0.42	−0.42
1.29	−1.19	−1.74	0.26	1.25	−1.25	−1.76	0.19
−1.12	1.90	2.19	0.31	0.53	−0.45	−0.68	0.13
−1.02	0.31	0.88	−0.60	0.14	0.47	0.28	0.40
0.10	−1.15	−0.96	−0.64	0.48	−0.11	−0.39	0.31
−0.32	−0.13	0.10	−0.33	−0.61	1.04	1.20	0.17
0.08	−0.17	−0.18	−0.04	−0.47	0.34	0.56	−0.16
0.49	0.18	−0.16	0.50	0.41	0.29	−0.02	0.50
0.54	0.20	0.49	−0.29	−0.22	−0.00	0.13	−0.18

From these data we may compute the sample variance of Y_1 and Y_2 as well as the variance of X and Y. We find the following values:

$$V_1 = 0.861, \qquad V_2 = 0.123, \qquad \text{var}(X) = 0.411, \qquad \text{var}(Y) = 0.573$$

From these data we may compute the percent of variability explained by the two principal components, individually and together.

1. Percent of variability explained by the first principal component $= 100 \times 0.861/(0.411 + 0.573) = 87.5\%$.
2. Percent of variability explained by the second principal component $= 100 \times 0.123/(0.411 + 0.573) = 12.5\%$.
3. Percent of variability explained by the first two principal components $= 100 \times (0.861 + 0.123)/(0.411 + 0.573) = 100\%$.

We see that the first principal component of the data in Figure 14.4 contains a high proportion of the variability. This may also be seen visually by examining the plot while orienting your eyes so that the horizontal line is the direction of the first principal component. Certainly, there is much more variability in that direction than in direction 2, the direction of the second principal component.

14.5 USE OF THE COVARIANCE, OR CORRELATION, VALUES AND PRINCIPAL COMPONENT ANALYSIS

The coefficients of the principal components and their variances can be computed by knowing the covariances between the X_j's. One might think that as a general search for relationships among X_j's, the principal component will be appropriate as an exploratory tool. Sometimes, this is true. However, consider what happens when we have different scales of measurement. Suppose, for example, that among our units, one unit is height in inches and another is systolic blood pressure in mmHg. In principal component analysis we are adding the variability in the two variables. Suppose now that we change our measurement of height from inches to feet. Then the standard deviation of the height variable will be divided by 12 and the variance will be divided by 144. In the total variance the contribution of height will have dropped greatly.

Equivalently, the blood pressure contribution (and any other variables) will become much more important. Recomputing the principal components will produce a different answer. In other words, the measurement units are important in finding the principal component because the variance of any individual variable is compared directly to the variance of another variable without regard to whether or not the units are appropriate for the comparison. We reiterate: *The importance of a variable in principal component analysis changes with a change of scale of one or more of the variables*. For this reason, principal component analysis is most appropriate and probably has its best applications when all the variables are measured in the same units; for example, the X_j variables may be measurements of length in inches, with the variables being measurements of different parts of the body, and the covariances between variables such as arm length, leg length, and body length.

In some situations with differing units, one still wants to try principal component analyses. In this case, standardized variables are often used; that is, we divide each variable by its standard deviation. Each rescaled variable then has a variance of 1 and the covariance matrix of the new standardized variables is the correlation matrix of the original variables. The interpretation of the principal components is now less clear. If many of the variables are highly correlated, the first principal component will tend to pick up this fact; for example, with two variables, a high correlation means the variables lie along a line. The ellipse of concentration has one axis along the line; that direction gives us the direction of the first principal component. When standardized variables are used, since each variable has a variance of 1, the sum of the variances is p. In looking at the percent of variability explained, there is no need to compute the total variance separately; it is p, the number of variables. We emphasize that when the correlations are used, there should be some reason for doing this beside the fact that the variables do not have measurements in comparable units.

14.6 STATISTICAL RESULTS FOR PRINCIPAL COMPONENT ANALYSIS

Suppose that we have a sample of size n from a multivariate normal distribution with unknown covariances. Let $V_i(\text{pop})$ be the true (unknown) population value for the variance of the ith principal component when computed from the (unknown) true variances; let V_i be the variance of the principal components computed from the sample covariances. Then the following are true:

1.

$$\frac{V_i - V_i(\text{pop})}{V_i(\text{pop})\sqrt{2/(n-1)}}, \qquad i = 1, \ldots, p \tag{8}$$

for large n is approximately a standard normal, $N(0, 1)$, random variable. These variables are approximately statistically independent.

2. $100(1-\alpha)\%$ confidence intervals for $V_i(\text{pop})$ for large n are given by

$$\left(\frac{V_i}{1 + z_{1-\alpha/2}\sqrt{2/(n-1)}}, \frac{V_i}{1 - z_{1-\alpha/2}\sqrt{2/(n-1)}} \right) \tag{9}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile value of the $N(0, 1)$ distribution.

Further statistical results on principal component analysis are given in Morrison [1976] and Timm [1975].

Principal component analysis is a least squares technique, as were analysis of variance and multiple linear regression. Outliers in the data can have a large effect on the results (as in other cases where least squares techniques are used).

14.7 PRESENTING THE RESULTS OF A PRINCIPAL COMPONENT ANALYSIS

We have seen that principal component analysis is designed to explain the variability in data. Thus, any presentation should include:

1. The variance of the principal components
2. The percent of the total variance explained by each individual principal component
3. The percent of the total variance explained cumulatively by the first m terms (for each m)

It is also useful to know how closely each variable X_j is related to the values of the principal components Y_i; this is usually done by presenting the correlations between each variable and each of the principal components. Let

$$Y_i = a_{i1}X_1 + \cdots + a_{ip}X_p$$

The correlation between one of the original variables X_j and the kth principal component Y_i is given by

$$r_{jk} = \text{correlation of } X_j \text{ and } Y_k = \frac{a_{kj}\sqrt{V_k}}{s_j} \tag{10}$$

In this equation, V_i is the variance of the ith principal component, while s_j is the standard deviation of X_j. These results are summarized in Table 14.3.

By examining the variables that are highly correlated with a principal component, we can see which variables contribute most to the principal component. Alternatively, glancing across the rows for each variable X_j we may see which principal component has the highest correlation with the variable. An X_i that has the highest correlations with the first few principal components is contributing more to the overall variability than variables with small correlations with the first few principal components. In Section 14.9, several examples of principal component analysis are given, including an example of the use of such a summary table (Table 14.4).

Table 14.3 Summary of a Principal Component Analysis Using Covariances

	Correlation of the Principal Components and the X_j's				Standard
Variables	1	2	$\cdots$	p	Deviations of the X_j
X_1	$\frac{a_{11}\sqrt{V_1}}{s_1}$	$\cdots$	$\cdots$	$\frac{a_{p1}\sqrt{V_p}}{s_1}$	s_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_p	$\frac{a_{1p}\sqrt{V_1}}{s_p}$	$\cdots$	$\cdots$	$\frac{a_{pp}\sqrt{V_p}}{s_p}$	s_p
Variance of principal component	V_1	V_2	$\cdots$	V_p	
% of total variance	$\frac{100V_1}{V}$	$\cdots$	$\cdots$	$\frac{100V_p}{V}$	
Cumulative % of total variance	$\frac{100V_1}{V}$	$\frac{100(V_1+V_2)}{V}$	$\cdots$	1	

Table 14.4 Data for Example 14.2

Principal Component	Variance Explained	Percent of Total Variance	Cumulative Percent of Total Variance
1	7.82	41.1	41.1
2	4.46	23.5	64.6
3	1.91	10.1	74.7
4	0.88	4.6	79.4
5	0.76	4.0	83.3
6	0.56	2.9	86.3
7	0.45	2.4	88.6
8	0.38	2.0	90.7
9	0.35	1.9	92.5
10	0.31	1.6	94.1
11	0.19	1.0	95.1
12	0.18	0.9	96.1
13	0.16	0.8	96.9
14	0.14	0.7	97.7
15	0.13	0.7	98.3
16	0.10	0.5	98.9
17	0.10	0.5	99.4
18	0.06	0.3	99.7
19	0.05	0.3	100.0

14.8 USES AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a technique for explaining variability. Following are some of the uses of principal components:

1. Principal component analysis is a search for linear relationships for explaining variability in a multivariate sample. The first few principal components are important because they may summarize a large proportion of the variability. However, the understanding of which variables contribute to the variability is important only if most of the variance comes about because of important relationships among the variables. After all, we can increase the variance of a variable, say X_1, by increasing the error of measurement. If we have a phenomenally large error of measurement, the variance of X_1 will be much larger than the variances of the rest of the variables. In this case, the first principal component will be approximately equal to X_1, and the amount of variability explained will be close to 1. However, such knowledge is not particularly useful, since the variability in X_1 does not make X_1 the most important variable, but in this case, reflects a very poorly measured quantity. Thus, to decide that the first few principal components are important summary variables, you must feel that the relationships among them come from linear relationships which may shed some light on the data being studied.

2. In some cases the first principal component is relatively uninteresting, with more informative relationships being found in the next few components. One simple case comes from analyzing physical measurements of plants or animals to display species differences: the first principal component may simply reflect differences in size, and the next few components give the more interesting differences in shape.

3. We may take the first two principal components and plot the values for the first two principal components of the data points. We know that among all possible plots in only two dimensions, this one gives the best fit in one precise mathematical sense. However, it should be noted that other techniques of multivariate analysis give two-dimensional plots that are the best fit or most interesting in other precise mathematical senses (see Note 14.1).

4. In some situations we have many measurements of somewhat related variables. For example, we might have a large number of size measurements on different portions of the human body. It may be that we want to perform a statistical inference, but the large number of variables for the relatively small number of cases involved makes such statistical analysis inappropriate. We may summarize the data by using the values on the first few principal components. *If the variability is important* (!), we have then reduced the number of variables without getting involved in multiple comparison problems. We may proceed to statistical analysis. For example, suppose that we are trying to perform a discriminant analysis and want to use size as one of the discriminating variables. However, for each of a relatively small number of cases we may have many anthropometric measurements. We might take the first principal component as a variable to summarize all the size relationships. One of the examples of principal component analysis below gives a principal component analysis of physical size data.

14.9 PRINCIPAL COMPONENT ANALYSIS EXAMPLES

Example 14.2. Stoudt et al. [1970] consider measurements taken on a sample of adult females from the United States. The correlations among these measurements (as well as weight and age) are given in Table 11.21. The variance explained for each principal component is presented in Table 14.4.

These data are very highly structured. Only three (of 19) principal components explain over 70% of the variance. Table 14.5 summarizes the first three principal components. The

Table 14.5 Example 14.2: First Three Principal Components

	Correlation of the Principal Components and the Variables		
Variables	1	2	3
SITHTER	0.252	0.772	0.485
SITHTNORM	0.235	0.748	0.470
KNEEHT	0.385	0.722	−0.392
POPHT	0.005	0.759	−0.444
ELBOWHT	0.276	0.243	0.783
THIGHHT	0.737	−0.007	0.204
BUTTKN	0.677	0.476	−0.348
BUTTPOP	0.559	0.411	−0.444
ELBOWBR	0.864	−0.325	−0.033
SEATBR	0.832	−0.050	0.096
BIACROM	0.504	0.350	−0.053
CHEST	0.890	−0.228	−0.018
WAIST	0.839	−0.343	−0.106
ARMGTH	0.893	−0.267	0.068
ARMSKIN	0.733	−0.231	0.124
INFRASCA	0.778	−0.371	0.056
HT	0.251	0.923	−0.051
WT	0.957	−0.057	0.001
AGE	0.222	−0.488	−0.289
Variance of principal components	7.82	4.46	1.91
Percent of total variance	41.1	23.5	10.1
Cumulative percent of total variance	41.1	64.6	74.7

first component, in the direction of greatest variation, is associated heavily with the weight variables. The highest correlation is with weight, 0.957. Other variables associated with size—such as chest and waist measurements, arm girth, and skinfolds—also are highly correlated with the first principal component. The second component is most closely associated with physical length measurements. Height is the most highly correlated variable. Other variables with correlations above 0.7 are the sitting heights (normal and erect), knee height, and popliteal height.

Since we are working with a correlation matrix, the total variance is 19, the number of variables. The average variance, in fact the exact variance, per variable is 1. Only these first three principal components have variance greater than 1. The other 16 directions correspond to a variance of less than 1.

Example 14.3. Reeck and Fisher [1973] performed a statistical analysis of the amino acid composition of protein. The mole percent of the 18 amino acids in a sample of 207 proteins was examined. The covariances and correlations are given in Table 14.6. The diagonal entries and numbers above them give the variances and covariances; the lower numbers are the correlations. The mnemonics are:

Asp	Aspartic acid	Met	Methionine
Thr	Threonine	Ile	Isoleucine
Ser	Serine	Leu	Leucine
Glu	Glutamic acid	Tyr	Tyrosine
Pro	Proline	Phe	Phenylalanine
Gly	Glycine	Trp	Tryptophan
Ala	Alanine	Lys	Lysine
Cys/2	Half-cystine	His	Histidine
Val	Valine	Arg	Arginine

The principal component analysis applied to the data produced Table 14.7, where k is the dimension of the subspace used to represent the data and C is the proportion of the total variance accounted for in the best k-dimensional representation.

In contrast to Example 14.2, eight principal components are needed to account for 70% of the variance. In this example there are no simple linear relationships (or directions) that account for most of the variability. In this case the principal component correlations are not presented, as the results are not very useful.

14.10 FACTOR ANALYSIS

As in principal component analysis, factor analysis looks at the relationships among variables as expressed by their correlations or covariances. While principal component analysis is designed to model and explain as much of the variability as possible, factor analysis seeks to explain the relationships among the variables. The assumption of the model is that the relationships may be explained by a few unobserved variables, which will be called *factors*. It is hoped that fewer factors than the original number of variables will be needed to explain the relationships among the variables. Thus, conceptually, one may simplify the understanding of the correlations between the variables.

It is difficult to present the technique without having the model and many of the related issues discussed first. However, it is also difficult to understand the related issues without examples. Thus, it is suggested that you read through the material about the mathematical model, go through the examples, and then with this understanding, reread the material about the mathematical model.

Table 14.6 Example 14.3: Reeck and Fisher [1973] Covariance/Correlation Matrix[a]

	Asp	Thr	Ser	Glu	Pro	Gly	Ala	Cys/2	Val	Met	Ile	Leu	Tyr	Phe	Trp	Lys	His	Arg
Asp	6.5649	0.2449	0.7879	−1.5329	−1.9141	−1.8328	−1.7003	−0.4974	−0.1374	0.0810	0.6332	−1.0855	0.6413	0.1879	0.3873	0.7336	0.0041	−1.5633
Thr	0.0517	3.4209	1.3998	−1.3341	−0.3531	−0.7752	−0.6428	0.4468	0.3603	−0.3502	0.1620	−1.2836	0.1804	−0.0978	0.1114	−0.3348	−0.2594	−0.8938
Ser	0.1219	0.2999	6.3687	−1.6465	0.1876	−0.8922	−1.3593	−0.3123	0.6659	−0.6488	−0.3738	−1.1125	0.4403	0.0432	0.2552	−1.6972	−0.3025	−1.4289
Glu	−0.1789	−0.2157	−0.1951	11.1880	−0.5866	−2.1665	−0.7732	−0.1443	−1.5346	0.0002	−0.3804	1.6210	−1.1824	−0.6684	−0.6778	0.0192	−0.3154	0.1169
Pro	−0.3566	−0.0911	−0.0355	−0.0837	4.3891	1.4958	−0.4259	1.0159	−0.7017	−0.4171	−0.8453	−0.9980	−0.0868	−0.1187	0.1163	−0.7021	−0.1612	0.4801
Gly	−0.2324	−0.1362	−0.1149	−0.2105	0.2320	9.4723	1.2857	0.1737	−0.3883	−0.4226	−0.2812	−2.3936	−0.8971	−0.7784	−0.2637	−1.0861	−0.2526	−0.0037
Ala	−0.2417	−0.1266	−0.1962	−0.0842	−0.0741	0.1522	7.5371	−2.1250	0.8498	0.1810	−0.4183	1.2480	−1.3374	−0.4320	−0.5219	−1.1641	−0.2730	0.0701
Cys/2	−0.0717	0.0892	−0.0457	−0.0159	0.1790	0.0208	−0.2857	7.3393	−1.3667	−0.4788	−1.3959	−2.3443	0.5408	−0.6282	0.1136	0.2727	−0.7482	0.1447
Val	−0.0275	0.1001	0.1356	−0.2357	−0.1721	−0.0648	0.1590	−0.2592	3.7885	−0.0632	0.5700	0.2767	−0.1348	−0.2303	−0.2792	−0.7921	−0.0632	−0.8223
Met	0.0294	−0.1759	−0.2388	0.0001	−0.1849	−0.1275	0.0612	−0.1642	−0.0302	1.1589	0.2493	0.2438	−0.1397	0.2060	−0.0159	0.1715	0.1457	0.0945
Ile	0.1426	0.0505	−0.0855	−0.0656	−0.2328	−0.0527	−0.0879	−0.2974	0.1690	0.1337	3.0023	−0.1857	−0.2785	−0.0870	−0.1296	0.2361	−0.0829	−0.3956
Leu	−0.1701	−0.2786	−0.1770	0.1946	−0.1912	−0.3122	0.1825	−0.3474	0.0571	0.0928	−0.0430	6.2047	−1.0362	0.2515	−0.2332	−0.6337	0.3951	1.0593
Tyr	0.1605	0.0625	0.1119	−0.2267	−0.0266	−0.1869	−0.3123	0.1280	−0.0444	−0.0832	−0.1031	−0.2667	0.1823	0.1659	0.9201	−0.5061	0.0855	0.1436
Phe	0.0525	−0.0379	0.0123	−0.1431	−0.0406	−0.1811	−0.1126	−0.1660	−0.0847	0.1370	−0.0360	0.0723	0.2262	1.9512	0.2223	−0.8382	0.3434	0.1796
Trp	0.1576	0.0628	0.1054	−0.2113	0.0579	−0.0893	−0.1982	0.0437	−0.1495	−0.0154	−0.0780	−0.0976	0.1823	0.1659	0.9201	−0.5061	0.0855	0.1436
Lys	0.1061	−0.0670	−0.2491	0.0021	−0.1241	−0.1307	−0.1571	0.0373	−0.1507	0.0590	0.0505	−0.0942	0.0733	−0.2223	−0.1954	7.2884	−0.1830	−1.0898
His	0.0014	−0.1194	−0.1020	−0.0803	−0.0655	−0.0699	−0.0847	−0.2351	−0.0276	0.1152	−0.0408	0.1350	0.0314	0.2093	0.0759	0.0577	1.3795	0.2280
Arg	−0.3068	−0.2430	−0.2847	0.0176	0.1152	−0.0006	0.0128	0.0269	−0.2124	0.0441	−0.1148	0.2138	−0.0882	0.0646	0.0753	−0.2030	0.0976	3.9550

[a]Diagonal and upper entries are variances and covariances. Below the diagonal are the correlations.

Table 14.7 Principal Component Analysis Data

k	C	k	C	k	C
1	0.13	7	0.66	13	0.90
2	0.26	8	0.70	14	0.93
3	0.37	9	0.75	15	0.95
4	0.46	10	0.79	16	0.98
5	0.55	11	0.83	17	1.00
6	0.61	12	0.86	18	1.00

We now turn to the model. We observe jointly distributed random variable $X_1, \ldots, X_p$. The assumption is that each X is a linear sum of the factors plus some remaining residual variability. That is, the model is the following:

$$\begin{array}{ccccccccccccc} X_1 & = & E(X_1) & + & \lambda_{11}F_1 & + & \lambda_{12}F_2 & + & \cdots & + & \lambda_{1k}F_k & + & e_1 \\ \vdots & & \vdots & & \vdots & & \vdots & & & & \vdots & & \vdots \\ X_p & = & E(X_p) & + & \lambda_{p1}F_1 & + & \lambda_{p2}F_2 & + & \cdots & + & \lambda_{pk}F_k & + & e_p \end{array} \tag{11}$$

In this model, each X_i is equal to its expected value, plus a linear sum of k factors and a term for residual variability. This looks like a series of multiple regression equations; each of the variables X_i is regressed on the variables $F_1, \ldots, F_k$. There are, however, major differences between this model and the multiple regression model of Chapter 11. Observations and assumptions about this model are the following:

1. The factors F_j are *not* observed; only the $X_1, \ldots, X_p$ are observed, although the X_i variables are expressed in terms of these smaller number of factors F_j.
2. The e_i (which are also unobserved) represent variability in the X_i not explained by the factors. We do *not* assume that these residual variability terms have the same distribution.
3. Usually, the number of factors k is unknown and must be determined from the data. We shall first consider the model and the analysis where the number of factors is known; later, we consider how one might search for the appropriate number of factors.

Assumptions made in the model, in addition to the linear equations given above, are the following:

1. The factors F_j are standardized; that is, they have mean zero and variance 1.
2. The factors F_j are uncorrelated with each other, and they are uncorrelated with the e_i terms. See Section 14.12 for a relaxation of this requirement.
3. The e_i's have mean zero and are uncorrelated with each other as well as with the F_j's. They may have different variances.

It is a fact that if p factors F are allowed, there is no need for the residual variability terms e_i. One can reproduce any pattern of covariances or correlations using p factors when p variables X_i are observed. This, however, is not very useful because we have summarized the p variables which were observed with p unknown variables. Thus, in general, we will be interested in k factors, where k is less than p.

Let ψ_i be the variance of e_i. With the assumptions of the model above, the variance of each X_i can be expressed in terms of the coefficients λ_{ij} of the factors and the residual variance ψ_i.

The equation giving the relationship for k factors is

$$\text{var}(X_i) = \lambda_{i1}^2 + \cdots + \lambda_{ik}^2 + \psi_i \tag{12}$$

In words, the variance of each X_i is the sum of the squares of the coefficients of the factors, plus the variance of e_i. The variance of X_i has two parts. The sum of the coefficients λ_{ij} squared depends on the factors; the factors contribute in common to all of the X_i's. The e_i's correlate only with their own variable X_i and not with other variables in the model. In particular, they are uncorrelated with all of the X_i's except for the one corresponding to their index. Thus, we have broken down the variance into a part related to the factors that each variable has in common, and the unique part related to the residual variability term. This leads to the following definition.

Definition 14.5. $c_i = \sum_{j=1}^{k} \lambda_{ij}^2$ is called the *common part of the variance* of X_i, c_i is also called the *communality* of X_i, ψ_i is called the *unique* or *specific part of the variance* of X_i, and ψ_i is also called the *uniqueness* or *specificity*.

Although factor analysis is designed to explain the relationships between the variables and not the variance of the individual variables, if the communalities are large compared to the specificities of the variables, the model has also succeeded in explaining not only the relationships among the variables but the variability in terms of the common factors.

Not only may the variance be expressed in terms of the coefficients of the factors, but the covariance between any two variables may also be expressed by

$$\text{cov}(X_i, X_j) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{ik}\lambda_{jk} \qquad \text{for } i \neq j \tag{13}$$

These equations explain the relationships among the variables. If both X_i and X_j have variances equal to 1, this expression gives the correlation between the two variables. There is a standard name for the coefficients of the common factors.

Definition 14.6. The coefficients λ_{ij} are called the *factor loadings* or *loadings*. λ_{ij} represents the loading of variable X_i and factor F_j.

In general, $\text{cov}(X_i, F_j) = \lambda_{ij}$. That is, λ_{ij} is the covariance between X_i and F_j. If X_i has variance 1, for example if it is standardized, then since F_j has variance 1, the factor loading is the correlation coefficient between the variable and the factor.

We illustrate the method by two examples.

Example 14.4. We continue with the measurement data of U.S. females of Example 14.2. A factor analysis with three underlying factors was performed on these data. Since we are trying to explain the correlations between the variables, it is useful to examine the fit of the model by comparing the observed and modeled correlations. We do this by examining the residual correlation.

Definition 14.7. The *residual correlation* is the observed correlation minus the fitted correlation from the factor analysis model.

Table 14.8 gives the residual correlations below the diagonal; on the diagonal are the estimated uniquenesses, the part of the (standardized) variance not explained by the three factors.

A rule of thumb is that the correlation has been explained reasonably when the residual is less than 0.1 in absolute value. This is convenient because it is easy to scan the residual matrix for a zero after a decimal point. Of course, depending on the purpose, more stringent requirements may be considered.

Table 14.8 Residual Correlations: Example 14.4

		STHTER 1	STHTNORM 2	KNEEHT 3	POPHT 4	ELBOWHT 5
STHTER	1	0.034				
STHTNORM	2	0.002	0.151			
KNEEHT	3	−0.001	0.001	0.191		
POPHT	4	0.001	0.002	0.048	0.276	
ELBOWHT	5	−0.001	−0.011	0.011	−0.004	0.474
THIGHHT	6	−0.009	0.004	0.003	−0.076	0.035
BUTTKN	7	−0.002	0.000	−0.016	−0.056	−0.021
BUTTPOP	8	−0.002	0.011	−0.042	−0.064	−0.035
ELBOWBR	9	0.000	0.013	−0.004	0.014	−0.010
SEATBR	10	−0.002	0.013	0.016	−0.041	0.020
BIACROM	11	0.004	−0.005	−0.000	0.014	−0.089
CHEST	12	0.003	0.004	0.003	0.030	−0.015
WAIST	13	0.005	−0.004	0.002	0.032	0.006
ARMGTH	14	−0.001	−0.004	0.004	−0.009	0.003
ARMSKIN	15	−0.005	0.016	0.025	−0.012	−0.004
INFRASCA	16	−0.002	0.006	0.020	0.016	0.004
HT	17	0.000	−0.001	−0.000	0.003	0.008
WT	18	−0.000	−0.009	−0.004	−0.005	0.008
AGE	19	0.002	0.024	0.003	0.024	−0.042

		THIGHHT 6	BUTTKN 7	BUTTPOP 8	ELBOWBR 9	SEATBR 10
THIGHHT	6	0.499				
BUTTKN	7	0.062	0.251			
BUTTPOP	8	0.040	**0.136**	0.425		
ELBOWBR	9	−0.012	−0.017	−0.016	0.158	
SEATBR	10	0.035	0.070	0.010	−0.016	0.338
BIACROM	11	0.049	−0.035	−0.039	0.012	−0.042
CHEST	12	−0.038	−0.044	−0.017	0.036	−0.056
WAIST	13	−0.067	−0.023	−0.021	0.037	−0.029
ARMGTH	14	0.005	0.005	0.007	−0.014	0.008
ARMSKIN	15	0.048	0.019	0.021	−0.030	0.047
INFRASCA	16	0.004	−0.025	−0.007	−0.003	−0.030
HT	17	−0.003	−0.001	0.001	0.004	−0.014
WT	18	0.017	0.009	−0.004	−0.011	0.019
AGE	19	**−0.172**	−0.056	−0.034	0.078	0.002

		BIACROM 11	CHESTGRH 12	WSTGRTH 13	RTARMGRH 14	RTARMSKN 15
BIACROM	11	0.679				
CHEST	12	0.072	0.148			
WAIST	13	−0.008	0.032	0.172		
ARMGTH	14	−0.014	−0.014	−0.031	0.134	
ARMSKIN	15	−0.053	−0.041	−0.046	0.075	0.487
INFRASCA	16	−0.010	0.013	0.003	0.013	**0.171**
HT	17	0.002	−0.000	−0.002	−0.001	0.003
WT	18	−0.003	0.000	0.004	0.009	−0.030
AGE	19	**−0.106**	0.033	0.105	−0.017	−0.012

		INFRASCA 16	HT 17	WT 18	AGE 19
INFRASCA	16	0.317			
HT	17	0.002	0.056		
WT	18	−0.018	0.001	0.057	
AGE	19	−0.017	0.016	−0.034	0.770

Table 14.9 Factor Loadings for a Three-Factor Model: Example 14.4

		Factor Loadings (Pattern)[a]		
Variable	Number	Factor 1	Factor 2	Factor 3
SITHTER	1		0.346	0.920
SITHTNORM	2		0.332	0.859
KNEEHT	3		0.884	0.146
POPHT	4	−0.271	0.801	
ELBOWHT	5	0.222	−0.120	0.680
THIGHHT	6	0.672	0.125	0.181
BUTTKN	7	0.436	0.741	
BUTTPOP	8	0.339	0.679	
ELBOWBR	9	0.914		
SEATBR	10	0.781	0.171	0.150
BIACROM	11	0.344	0.390	0.225
CHEST	12	0.916	0.114	
WAIST	13	0.898		−0.126
ARMGTH	14	0.929		
ARMSKIN	15	0.714		
INFRASCA	16	0.823		
HT	17		0.804	0.538
WT	18	0.929	0.265	0.103
AGE	19	0.328	−0.124	−0.328
VP		7.123	3.632	2.628
Proportion var.		0.375	0.191	0.138
Cumulative var.		0.375	0.566	0.704

[a]Loadings less than 0.1 have been omitted.

In this example there are four large absolute values of residuals (−0.172, 0.171, 0.136, and −0.106). This suggests that more factors are needed. (In Problem 14.10 we consider analysis of these data with more factors.) The factor loadings are presented in Table 14.9. Loadings below 0.1 in absolute value are omitted, making it easier to see which variables are related to which factors. In this example the first factor has high loadings on weight and bulk measurements (variables 14, 18, 12, 9, 13, 16, 10, 15, and 6) and might be called a *weight* factor. The second factor has high loadings on length or height measurements (variables 3, 17, 4, 7, and 8) and might be considered a *height* factor. The third factor seems to be a *sitting height* factor.

The variables have been reordered so that variables loading on the same factor appear together. When this is done, clusters of correlated variables often appear, which may be appreciated visually by replacing correlations by symbols or colors. Figure 14.7 is a graph of the correlation data from Table 11.21 using circles whose radius is proportional to the correlation, shaded light gray for positive correlations and dark gray for negative correlations.

The sum of the squares of loadings for a factor (VP) is the portion of the sum of the X_i variances (the total variance) that is explained by the factor. The table also gives this as a proportion of the total and as a cumulative proportion of the total. In all, these factors explain 70% of the variability in the measurements.

Example 14.5. As a second example, consider coronary artery disease patients with left main coronary artery disease. This patient group was discussed in Chaitman et al. [1981]. In this factor analysis, 12 variables were considered and four factors were used with 357 cases. The factor analysis was based on the correlation matrix. The variables and their mnemonics (names) are:

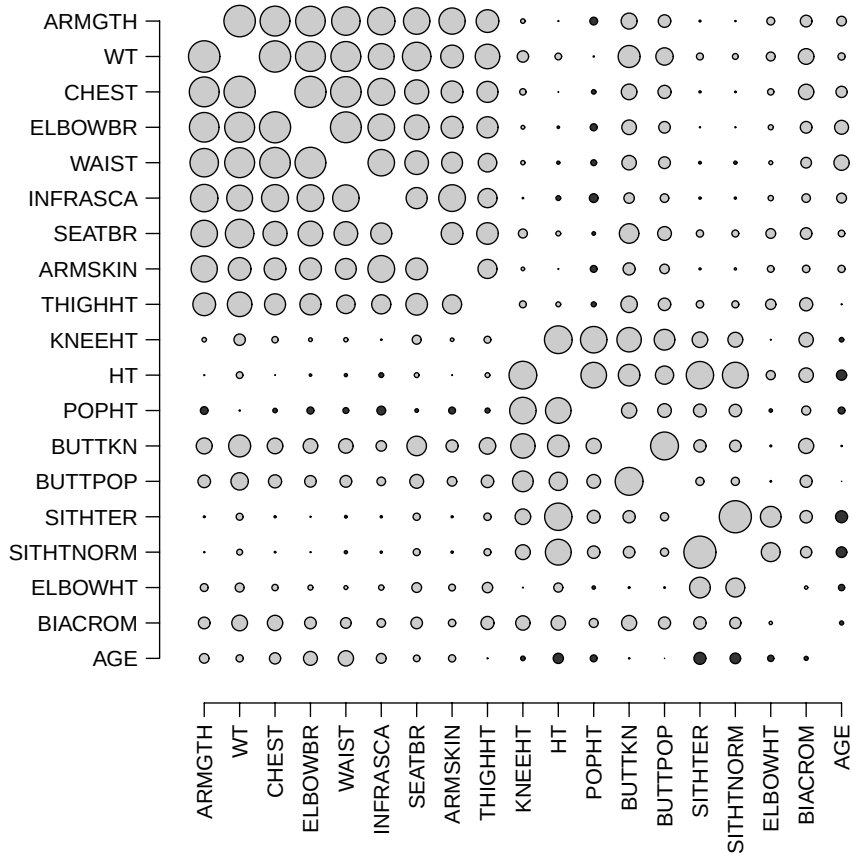


Figure 14.7 Correlations for Example 14.4. The radius of the circle is proportional to the absolute value of the correlation. Light gray circles indicate positive correlations; dark gray circles, negative. (Data from Stoudt et al. [1970].)

- *SEX*: 0 = male, 1 = female.
- *PREVMI*: 0 = history of prior myocardial infarction, 1 = no such history.
- *FEPCHEP*: time in weeks since the first episode of anginal chest pain; this analysis was restricted to patients with anginal chest pain.
- *CHCLASS*: severity of impairment due to angina (chest pain); ranging from I (mildly impaired) to IV (any activity is limited; almost totally bedridden).
- *LMCA*: the percent diameter narrowing of the left main coronary artery; this analysis was restricted to 50% or more narrowing.
- *AGE*: in years.
- *SCORE*: the amount of impairment of the pumping chamber (left ventricle) of the heart; score ranges from 5 (normal) to 30 (not attained).
- *PS70*: the number of proximal (near the beginning of the blood supply) segments of the coronary arteries with 70% or more diameter narrowing.
- *LEFT*: this variable (and RIGHT) tells if the right artery of the heart carries as much blood as normal. LEFT (dominant) implies that the right coronary artery carries little blood; 8.8% of these cases fell into this category. Code: LEFT = 1 (left dominant); LEFT = 0 otherwise.

Table 14.10 Correlations (as the Bottom Entry in Each Cell) and the Residual Correlations (as the Top Entry) in Each Cell[a]

	SEX	PREMI	FEPCHEP	CHCLASS	LMCA	AGE
SEX	0.933					
	1.000					
PREVMI	0.053	0.802				
	0.040	1.000				
FEPCHEP	−0.013	−0.043	0.714			
	−0.002	−0.161	1.000			
CHCLASS	0.056	−0.000	−0.001	0.796		
	0.073	−0.117	0.217	1.000		
LMCA	0.010	0.049	0.005	−0.037	0.989	
	0.012	0.036	0.041	0.004	1.000	
AGE	−0.026	0.019	0.012	−0.001	0.024	0.727
	−0.013	−0.107	0.286	0.227	0.065	1.000
SCORE	0.000	−0.001	−0.000	0.000	0.000	0.000
	0.030	−0.427	0.143	0.185	0.019	0.175
PS70	−0.028	−0.057	−0.027	0.062	−0.016	0.013
	−0.054	−0.188	0.129	0.087	−0.034	0.044
LEFT	0.015	−0.011	−0.015	0.025	0.011	−0.005
	−0.027	−0.022	0.014	0.099	0.063	0.064
RIGHT	0.009	−0.007	−0.009	0.015	0.006	−0.003
	0.054	0.017	−0.033	−0.062	−0.049	−0.077
NOVESLS	0.000	0.000	0.000	−0.000	0.000	0.000
	−0.033	−0.183	0.206	0.014	−0.034	0.130
LVEDP	0.014	0.023	0.001	0.024	0.019	−0.015
	0.020	−0.072	0.119	0.135	0.041	0.109

	SCORE	PS70	LEFT	RIGHT	NOVESLS	LVEDP
SCORE	0.021					
	1.000					
PS70	0.001	0.514				
	0.198	1.000				
LEFT	−0.000	−0.004	0.281			
	0.007	0.004	1.000			
RIGHT	−0.000	−0.004	0.002	0.175		
	−0.041	−0.013	−0.767	1.000		
NOVESLS	0.000	0.000	0.000	0.000	0.000	
	0.284	0.693	−0.071	0.073	1.000	
LVEDP	0.000	−0.025	−0.007	−0.004	0.000	0.930
	0.175	0.029	0.068	−0.086	0.063	1.000

[a]The diagonal entry on top is the estimated uniqueness for each variable. Four factors were used.

- *RIGHT*: there are three types of dominance of the coronary arteries: LEFT above, unbalanced (implicitly coded when LEFT = 0 and RIGHT = 0), and RIGHT. Right dominance is the usual case and occurs when the right coronary artery carries a usual amount of blood. 85.8% of these cases are right dominant: RIGHT = 1; otherwise, RIGHT = 0.
- *NOVESLS*: the number of diseased vessels with $\geq$ 70% stenosis or narrowing of the three major arterial branches above and beyond the left main disease.
- *LVEDP*: the left ventricular end diastolic pressure. This is the pressure in the heart when it is relaxed between beats. A damaged or failing heart has a higher pressure.

Table 14.11 Factor Loadings: Example 14.5

	Factor[a]			
	1	2	3	4
SEX				
PREVMI	−0.103	−0.396	−0.174	
FEPCHEP	0.152		0.535	
CHCLASS			0.125	0.428
LMCA				
AGE				0.502
SCORE		0.108	0.981	0.158
PS70		0.683	0.117	
LEFT	−0.818			0.124
RIGHT	0.917			−0.121
NOVESLS		0.980	0.166	
LVEDP			0.143	0.215
VP[b]	1.525	1.487	1.210	0.872
Proportion var.	0.127	0.124	0.101	0.073
Cumulative var.	0.127	0.251	0.352	0.425

[a] Loadings below 0.100 are omitted.
[b] VP is the portion of sum of squares explained by the factor.

Factor analysis is designed primarily for continuous variables. In this example we have many discrete variables, and even dummy or indicator variables. The analysis is considered more descriptive or explanatory in this case.

Examining the residual values in Table 14.10, we see a fairly satisfactory fit; the maximum absolute value of a residual is 0.062, but most are much smaller. Examination of the uniqueness diagonal column on top shows that the number of vessels diseased, NOVESLS, and SCORE are explained essentially by the factors (uniqueness = 0.000). Some other variables retain almost all of their variability: SEX (uniqueness = 0.993) and LMCA (uniqueness = 0.989). Since we have explained most of the relationships among the variables without using the variability of these factors, SEX and LMCA must be weakly related to the other factors. This is readily verified by looking at the correlation matrix; the maximum absolute correlation involving either of the variables is $r = 0.073$, $r^2 = 0.005$. They explain $\frac{1}{2}$ of 1% or less of the variability in the other variables.

Let us now look at the factor loading (or correlation) values in Table 14.11. The first factor has heavy loadings on the two *dominance* variables. This factor could be labeled a dominance factor. The second factor looks like a *coronary artery disease* (CAD) *factor*. The third is a heart attack, a *ventricular function* factor. The fourth might be labeled a *history* variable.

The first factor exists largely by definition; if LEFT = 1, then RIGHT = 0, and vice versa. The second factor is also expected; if proximal segments are diseased, the arteries are diseased. The third factor makes biological sense. A damaged ventricle often occurs because of a heart attack. The factor with moderate loadings on AGE, FEPCHEP, and CHCLASS is not as clear.

14.11 ESTIMATION

Many methods have been suggested for estimation of the factor loadings and the specificities, that is, the coefficients λ_{ij} and the variance of the residual term e_i. Consider equation (11) and suppose that we change the scale of X_i. Effectively, this is the same as looking at a new variable cX_i; the new value is the old value multiplied by a constant. Multiplying through the equations of equation (11) by the constant, and remembering that we have restricted the factors

to have variance 1, we see that factor loading should be multiplied by the same factor as X_i. Only one method of estimation has this property, which also implies that we can use either the covariance matrix or correlation matrix as input to the estimation. This method is the maximum likelihood method; it is our method of choice. The method seems to give the best fit, where fit is examined as described below. There are drawbacks to the method. There can be multiple possible solutions, and software may not converge to the best solution, particularly if the best solution involves a communality of 1.00 for some variable (the "Heywood case"). The examples in this chapter are fairly well behaved, and essentially the same solution was obtained with the programs BMDP and R. For a review of other methods, we recommend the book by Gorsuch [1983]. This book, which is cited extensively below, contains a nice review of many of the issues of factor analysis. Two shorter volumes are those of Kim and Mueller [1983, 1999].

14.12 INDETERMINACY OF THE FACTOR SPACE

There appears to be something magical about factor analysis; we are estimating coefficients of variables that are not even observed. It is difficult to imagine that one can estimate this at all. In point of fact, it is not possible to estimate the F_i uniquely, but one can estimate the F_i up to a certain indeterminacy. It is necessary to describe this indeterminacy in mathematical terms.

Mathematically, the factors are unique except for possible linear combinations. Geometrically, suppose that we think of the factors (e.g., a model with $k = 2$) as corresponding to values in a plane. Let this plane exist in three-dimensional space. For example, the subspace corresponding to the two factors (i.e., the plane) might be the plane of the paper of this book. Within this three-dimensional space, factor analysis would determine which plane contains the two factors. However, any two perpendicular directions in the factor plane would correspond to factors that equally well fit the data in terms of explaining the covariances or correlations between the variables. Thus, we have the factors identified up to a certain extent, but we are allowed to rotate them within a subspace.

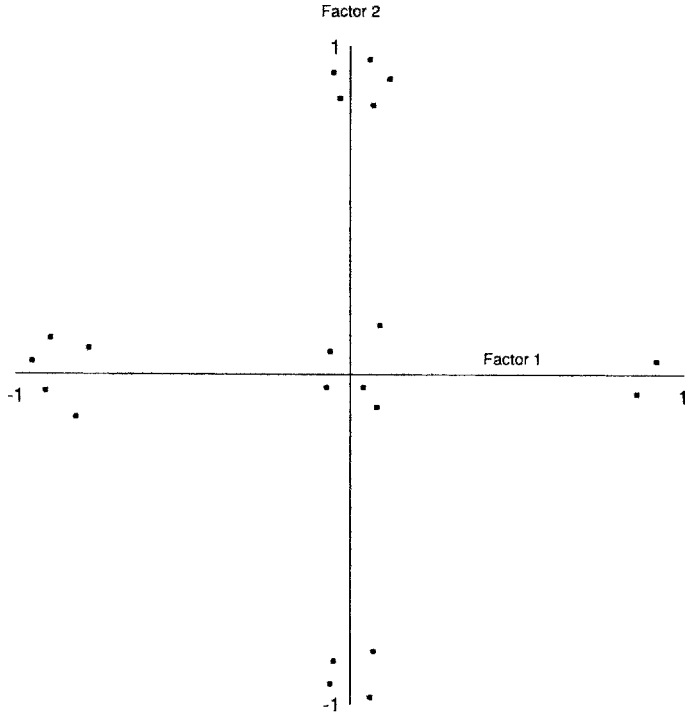
This indeterminacy allows one to "fiddle" with different combinations of factors (i.e., rotations) so that the factors are considered "easy to interpret." As discussed at some length below, one of the strengths and weaknesses of factor analysis is the possibility of finding factors that represent some abstract concept. This task is easiest when the factors are associated with some subset of the variables. That is, one would like factors that have high loadings (in terms of absolute value) on some subset of variables and very low (near zero in absolute value) loadings on the rest of the variables. In this case, the factor is closely associated with the subset of the variables that have large absolute loadings. If these variables have something in common conceptually (e.g., they are all measures of blood pressure) or in a psychological study they all seem to be related to aggressive behavior, one might then identify the specific factor as a blood pressure factor or an aggression factor.

Another complication in the literature of factor analysis is related to the choice of a specific basis in the factor subspace. Suppose for the moment that we are dealing with the correlations among the X_i's. In this case, as we saw before, the loadings on the factors are correlations of the factor with the variable. Thus each loading will be in absolute value less than or equal to 1. It will be easy to interpret our factors if the absolute value is near zero or near 1. Consider Figure 14.8(a) and (b), plots of the loadings on factors 1 and 2, with a separate point for each of the variables X_i. In Figure 14.8(a) there is a very nice pattern. The variables corresponding to points on the factor 1 axis of ± 1 or on the factor 2 axis of ± 1 are variables associated with each of the factors. The variables plotted near zero on both factors have little relationship to the two factors; in particular, factor 1 would be associated with the variables having points near ± 1 along its axis, including variables 1 and 10 as labeled. This would be considered a very nice loading pattern, and easy to interpret, having the simple structure as described above. In Figure 14.8(b) we see that if we look at the original factors 1 and 2, it is difficult to interpret

the data points, but should we rotate by θ as indicated in the figure, we would have factors easy to interpretation (i.e., each factor associated with a subset of the X_i variables). By looking at such plots and then drawing lines and deciding on the angle θ visually, we have what is called *visual rotation*. When the factor subspace contains a variety of factors (i.e., $k > 2$), the situation is not as simple. If we rotate factors 1 and 2 to find a simple interpretation, we will have altered the relationship between factors 1 and 2 and the other factors, and thus, in improving the relationship between 1 and 2 to have a simple form, we may weaken the relationship between 1 and 5, for example. Visual rotation of factors is an art that can take days or even weeks. The advantage of such rotation is that the mind can weigh the different trade-offs. One drawback of visual rotation is that it may be rotated to give factors that conform to some pet hypothesis. Again, the naming and interpretation of factors are discussed below. Thus, visual rotation can take an enormous amount of time and is subject to the biases of the data analyst (as well as to his or her creativity).

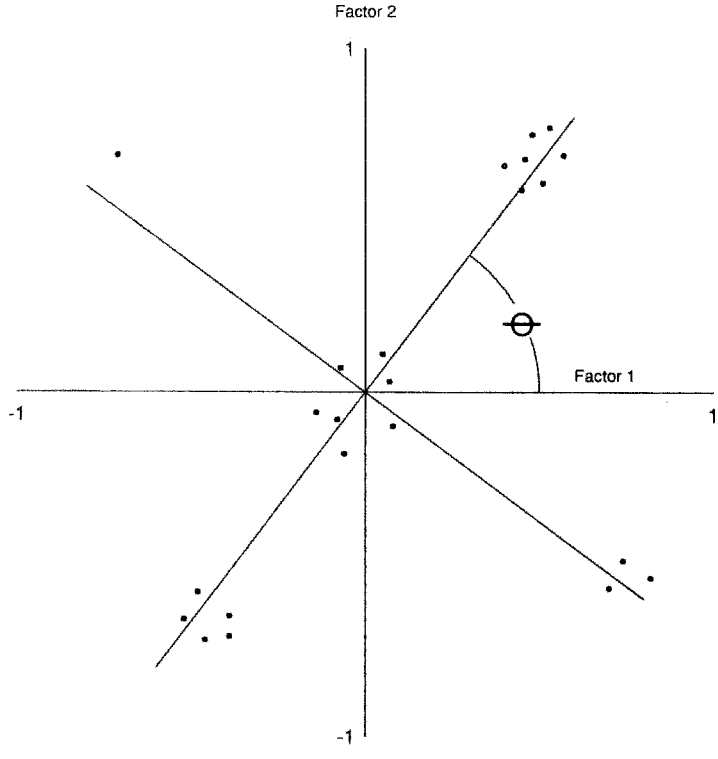
Because of the time constraints for analysis, the complexity of the rotation, and the potential biases, considerable effort has been devoted to developing analytic methods of rotating the factors to get the best rotation. By *analytic* we mean that there is an algorithm describing whether or not a particular rotation for all of the factors is desirable. The computer software, then, finds the best orientation.

Note 14.2 describes two popular criteria, the *varimax method* and the *quartimax method*. A factor analysis is said to have a *general factor* if there is a factor that is associated with all or almost all of the variables. The varimax method can be useful but does not allow general factors and should not be used when such factors may occur. Otherwise, it is considered one of the most



a. Very good loading pattern.
All loadings with absolute value near zero or one.

Figure 14.8 Two-factor loading patterns. (*Continued overleaf*)



b. This pattern suggests rotating
the factors by the angle Θ to have a simple structure.

Figure 14.8 *continued*

satisfactory methods. (In fact, factor analysis was developed in conjunction with the study of intelligence. In particular, one of the issues was: Does intelligence consist of one general factor or a variety of uncorrelated factors corresponding to different types of intelligence? Another alternative model for intelligence is a general factor plus other factors associated with some subset of measures of performance thought to be associated with intelligence.)

The second popular method is the quartimax method. This method, in contrast to the varimax method, tends to have one factor with large loadings on all the variables and not many large loadings among the rest of the factors. In the examples of this chapter we have used the varimax method. We do not have the space to get into all the issues involved in the selection of a rotation method.

Returning to visual rotation, suppose that we have the pattern shown in Figure 14.9. We see that there are no perpendicular axes for which the loadings are 1 or -1, but if we took two axes corresponding to the dashed lines, the interpretation might be simplified. Factors corresponding to the two dashed lines are no longer uncorrelated with each other, and one may wonder to what extent they are "separate" factors. Such factors are called *oblique factors,* the word *oblique* coming from the geometric picture and the fact that in geometry, oblique lines are lines that do not intersect at a right angle. There are a number of analytic methods for getting oblique rotations, with snappy names such as *oblimax*, *biquartimin*, *binormamin*, and *maxplane*. References to these may be found in Gorsuch [1983]. If oblique axes or bases are used, the formulas for the variance and covariances of the X_i's as given above no longer hold. Again, see Gorsuch for more in-depth consideration of such issues.

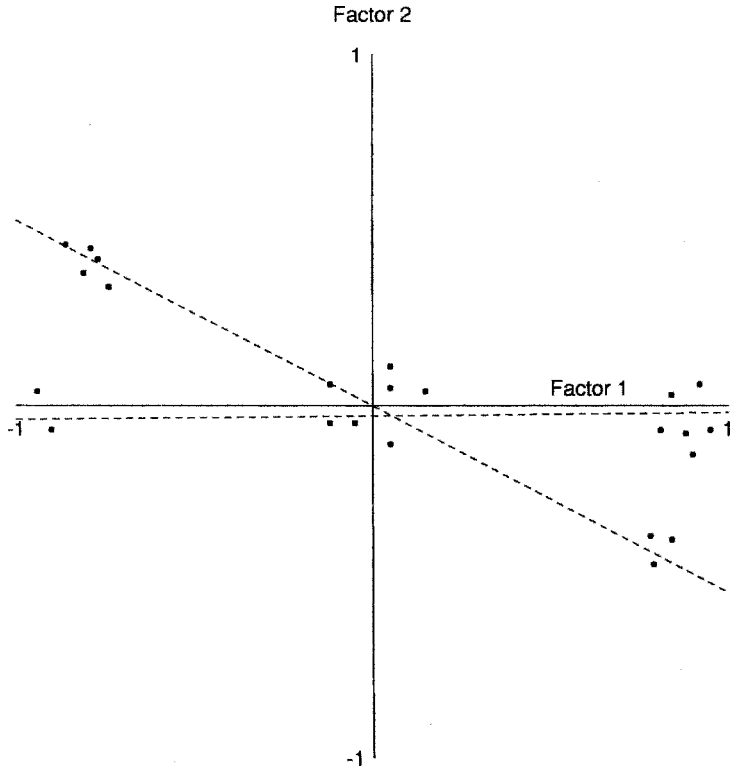


Figure 14.9 Orthogonal and oblique axes for factor loadings.

To try a factor analysis it is not necessary to be expert with every method of estimation and rotation. An exploratory data analysis may be performed to see the extent to which things simplify. We suggest the use of the maximum likelihood estimation method for estimating the coefficients λ_{ij}, where the rotation is performed using the varimax method unless one large general factor is suspected to occur.

Example 14.6. We return to Examples 14.4 and 14.5 and examine plots of the correlations of the variables with the factors. Figure 14.10 shows the plots for Example 14.4, where the numbers on the plot correspond to the variable numbers in Table 14.9.

The plot for factors 2 and 3 looks reasonable (absolute values near 0 or 1). The other two plots have in-between points making interpretation of the factors difficult. This, along with the large residuals mentioned above, suggests trying an analysis with a few more factors.

The plots for Example 14.5 are given in Figure 14.11. These plots suggest factors fairly easy of interpretation, with few, if any, points with moderate loadings on several factors. The interpretation of the factors, discussed in Example 14.5, was fairly straightforward.

14.13 CONSTRAINED FACTOR ANALYSIS

In some situations there are physical constraints on the factors that affect the fitting and interpretation of the factor analysis model. One important application of this sort is in the study of air pollution. Particulate air pollution consists of small particles of smoke, dust, or haze, typically 10 μm in size or smaller. These particles come from a relatively small number of sources, such

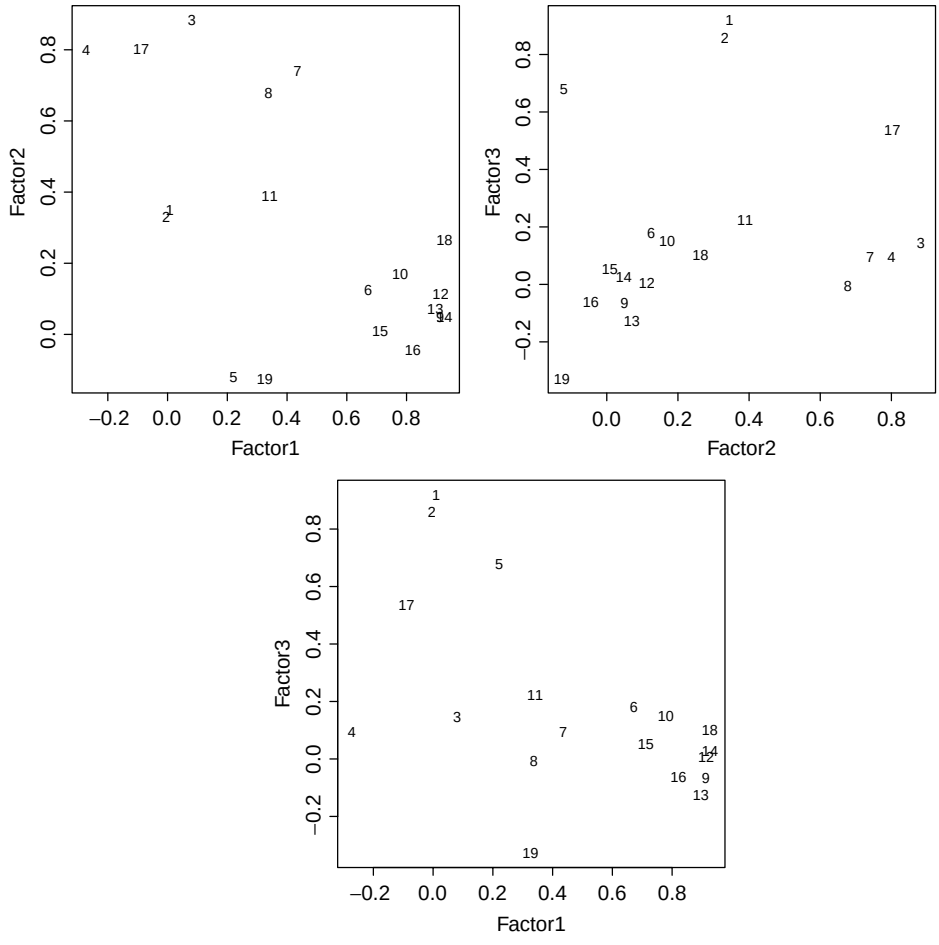


Figure 14.10 Factor loadings for Example 14.4.

as car and truck exhaust, smoke from fireplaces, road dust, and chemical reactions between gases in the air. Particles from different sources have differing distributions of chemical composition, so the chemical composition of particles in the air will be approximately an average of those for each source, weighted according to that source's contribution to overall pollution. That is, we have a factor analysis model in which the factor loadings λ represent the contribution of each source to overall particulate air pollution, the factors F characterize the chemical composition of each source, and the uniquenesses c_i are due largely to measurement error.

In this context the factor analysis model is modified slightly by removing the intercept in each of the regression models of equation (11). Rather than constraining each factor to have zero mean and unit variance, we constrain all the coefficients F and λ to be nonnegative. That is, a source cannot contain a negative amount of some chemical element and cannot contribute a negative concentration of particles. These physical constraints reduce the rotational indeterminacy of the model considerably. On the other hand, it is not reasonable to require that factors are orthogonal to each other, so that oblique rotations must be considered, restoring some of the indeterminacy.

The computation is even more difficult than for ordinary factor analysis, and specialized software is needed [Paatero, 1997, 1999; Henry, 1997]. The full data are needed rather than just a correlation or covariance matrix.

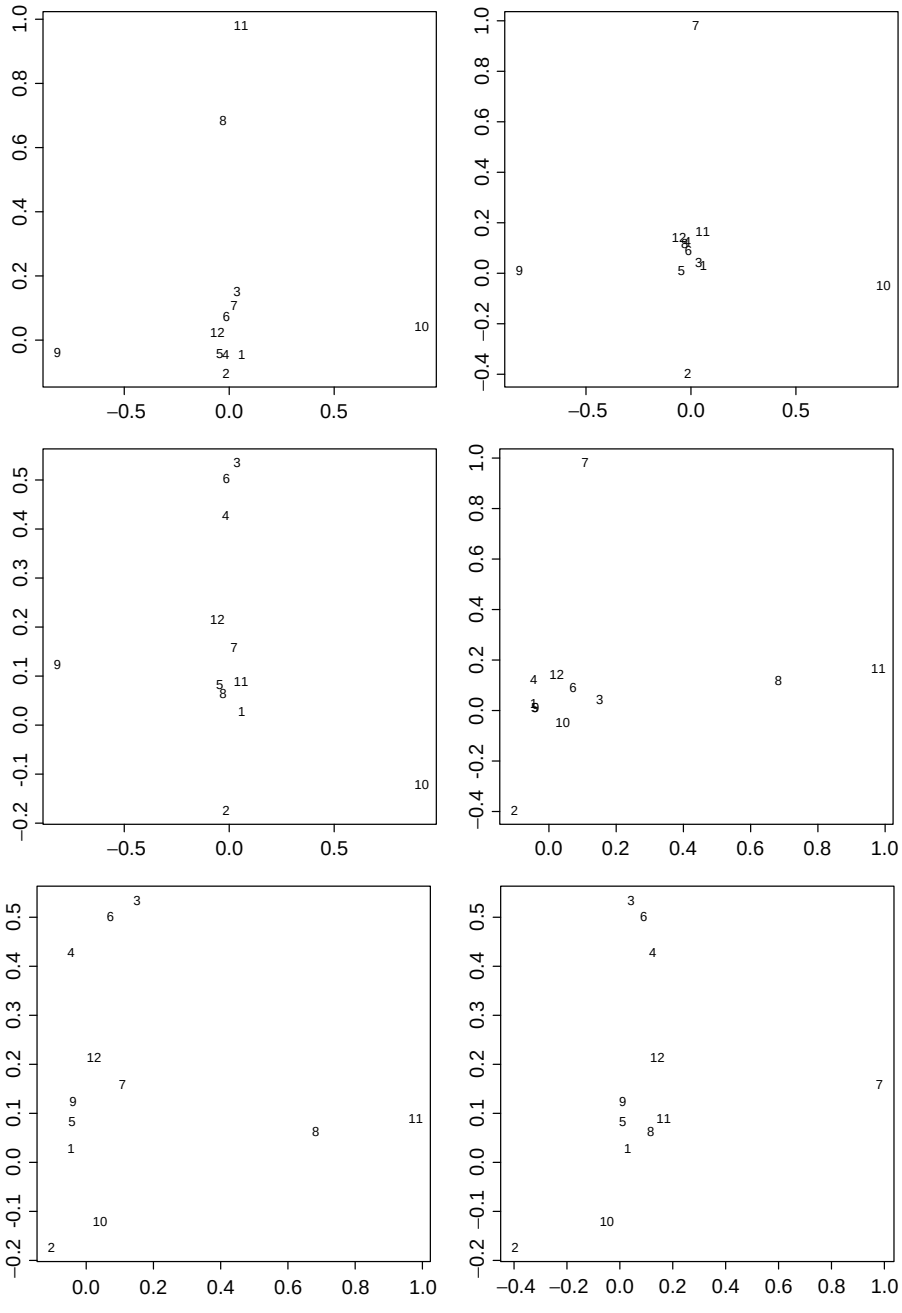


Figure 14.11 Factor loadings for Example 14.5.

Example 14.7. In February 2000, the U.S. Environmental Protection Agency held a workshop on source apportionment for particulate air pollution [U.S. EPA, 2000]. The main part of the workshop was a discussion of two constrained factor analysis methods which were used to investigate fine particulate air pollution from Phoenix, Arizona. Data were available for 981 days, from March 1995 through June 1998, on concentrations of 44 chemical elements and on carbon content, divided into organic carbon and elemental carbon.

The UNMIX method [Henry, 1997] gave a five-factor model:

Source	Concentration ($\mu g/m^3$)
Vehicles	4.7
Secondary aerosol	2.6
Soil	1.8
Diesel	1.2
Vegetative burning	0.7
Unidentified	1.6

and the PMF method [Paatero, 1997] gave a six-factor model:

Source	Concentration ($\mu g/m^3$)
Motor vehicles	3.5
Coal-fired power	2.1
Soil	1.9
Smelter	0.5
Biomass burning	4.4
Sea salt	0.1

Some of these factors were expected and their likely composition known a priori, such as vehicle exhaust with large amounts of both organic and elemental carbon, and soil with aluminium and silicon. Others were found and interpreted as a result of the analysis; the diesel source had both the elemental carbon characteristic of diesel exhaust and the maganese attributed to fuel additives. The secondary aerosol source in the UNMIX results probably corresponds to the coal-fired power source of PMF and perhaps some of the other burning; it would consist of sulfate and nitrate particles formed by chemical reactions in the atmosphere.

The attributions of fine particles to combustion, soil, and chemical reactions in the atmosphere were reasonably consistent between these methods, but separating different types of combustion proved much more difficult. This is probably a typical case and illustrates that the indeterminacy in the basic factor analysis model can partly, but not entirely, be overcome by substantive knowledge.

14.14 DETERMINING THE NUMBER OF FACTORS

In this section we consider what to do when the number of factors is unknown. Estimation methods of factor analysis begin with knowledge of k, the number of factors. But this number is usually not known or hypothesized. There is no universal agreement on how to select k; below we examine a number of ways of doing this. The first step is always carried out.

1. Examine the values of the residual correlations. In this section we suppose that we are trying to model the correlations between variables rather than their covariances. Recall that with maximum likelihood estimation, fitting one is the same as fitting the other. In looking at the residual correlations, as done in Examples 14.4 and 14.5, we may feel that we have done a good job if all of the correlations have been fit to within a specified difference. If the residual correlations reveal large discrepancies, the model does not fit.

2. There are statistical tests *if* we can assume that multivariate normality holds and we use the maximum likelihood estimation method. In this case, there is an asymptotic chi-square test for any hypothesized fixed number of factors. Computation of the test statistic is complex

and given in Note 14.3. However, it is available in many statistical computer programs. One approach is to look at successively more factors until the statistic is not statistically significant; that is, there are enough factors so that one would not reject at a fixed significance level the hypothesis that the number of factors is as given. This is analogous to a stepwise regression procedure. If we do this, we are performing a stepwise procedure, and the true and nominal significance levels differ (as usual in a stepwise analysis).

3. Looking at the roots of the correlation matrix:

a. If the correlations are arranged in a square pattern or matrix, as usually done, this pattern is called a *correlation matrix*. Suppose that we perform a principal component analysis and examine the variances of the principal components $V_1 \geq V_2 \geq \cdots \geq V_p$. These values are called the *eigenvalues* or *roots* of the correlation matrix. If we have the correlation matrix for the entire population, Guttman [1954] showed that the number of factors, k, must be greater than or equal to the number of roots greater than or equal to 1. That is, the number of factors in the factor analytic model must be greater than or equal to the number of principal components whose variance is greater than or equal to 1. Of course, in practice we do not have the population correlation matrix but an estimate. The number of such roots greater than or equal to 1 in a sample may turn out to be smaller or larger. However, because of Guttman's result, a reasonable starting value for k is the number of roots greater than or equal to 1 for the sample correlation matrix. For a thorough factor analysis, values of k above and below this number should be tried and the residual patterns observed. The number of factors in Examples 14.4 and 14.5 was chosen by this method.

b. *Scree* is the name for the rubble at the bottom of a cliff. The scree test plots the variances of the principal components. If the plot looks somewhat like Figure 14.12, one looks to separate the climb of the cliff from the scree at the bottom of the cliff. We are directed to pick the cliff, components 1, 2, 3, and possibly 4, rather than the rubble. A clear plastic ruler is laid across the bottom points, and the number of values above the line is the number of important factors. This advice is reasonable when a sharp demarcation can be seen, but often the pattern has no clear breakpoint.

c. Since we are interested in the correlation structure, we might plot as a function of k (the number of factors) the maximum absolute value of all the residuals of the estimated

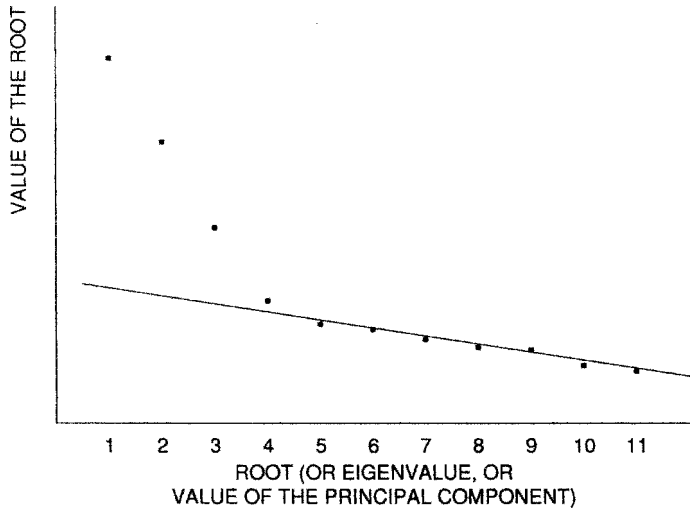


Figure 14.12 Plot for the scree test.

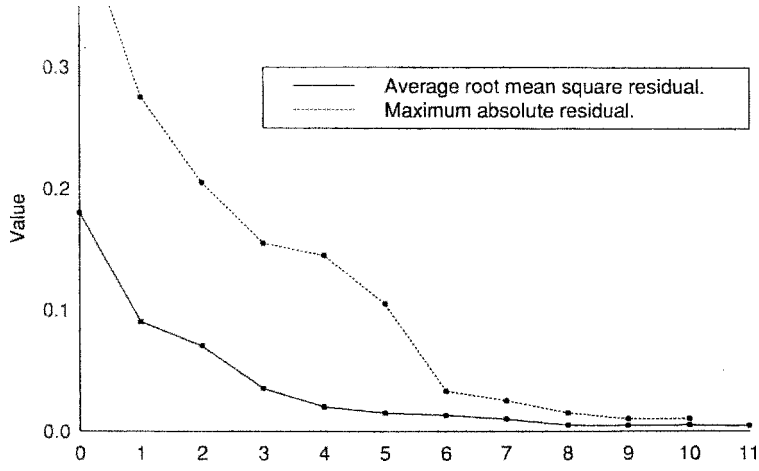


Figure 14.13 Plot of the maximum absolute residual and the average root mean square residual.

correlations. Another useful plot is the square root of the sum of the squares of all of the residual correlations divided by the number of such residual correlations, which is $p(p-1)/2$. If there is a break in the plots of the curves, we would then pick k so that the maximum and average squared residual correlations are small. For example, in Figure 14.13 we might choose three or four factors. Gorsuch suggests: "In the final report, interpretation could be limited to those factors which are well stabilized over the range which the number of factors may reasonably take."

14.15 INTERPRETATION OF FACTORS

Much of the debate about factor analysis stems from the naming and interpretation of factors. Often, after a factor analysis is performed, the factors are identified with concepts or objects. *Is a factor an underlying concept or merely a convenient way of summarizing interrelationships among variables*? A useful word in this context is *reify*, meaning to convert into or to regard something as a concrete thing. Should factors be reified?

As Gorsuch states: "A prime use of factor analysis has been in the development of both the theoretical constructs for an area and the operational representatives for the theoretical constructs." In other words, a prime use of factor analysis requires reifying the factors. Also, "The first task of any research program is to establish empirical referents for the abstract concepts embodied in a particular theory."

In psychology, how would one deal with an abstract concept such as aggression? On a questionnaire a variety of possible "aggression" questions might be used. If most or all of them have high loadings on the same factor, and other questions thought to be unrelated to aggression had low loadings, one might identify that factor with aggression. Further, the highest loadings might identify operationally the questions to be used to examine this abstract concept.

Since our knowledge is of the original observations, without a unique set of variables loading a factor, interpretation is difficult. Note well, however, that there is no law saying that one must interpret and name any or all factors.

Gorsuch makes the following points:

1. "The factor can only be interpreted by an individual with extensive background in the substantive area."

2. "The summary of the interpretation is presented as the factor's name. The name may be only descriptive or it may suggest a causal explanation for the occurrence of the factor. Since the name of the factor is all most readers of the research report will remember, it should be carefully chosen." *Perhaps it should not be chosen at all in many cases.*
3. "The widely followed practice of regarding interpretation of a factor as confirmed solely because the post-hoc analysis 'makes sense' is to be deplored. Factor interpretations can only be considered hypotheses for another study."

Interpretation of factors may be strengthened by using cases from other populations. Also, collecting other variables thought to be associated with the factor and including them in the analysis is useful. They should load on the same factor. Taking "marker" variables from other studies is useful in seeing whether an abstract concept has been embodied in more or less the same way in two different analyses.

For a perceptive and easy-to-understand discussion of factor analysis, see Chapter 6 in Gould [1996], which deals with scientific racism. Gould discusses the reification of intelligence in the Intelligence Quotient (IQ) through the use of factor analysis. Gould traces the history of factor analysis starting with the work of Spearman. Gould's book is a cautionary tale about scientific presuppositions, predilections, and perceptions affecting the interpretation of statistical results (it is not necessary to agree with all his conclusions to benefit from his explanations). A recent book by McDonald [1999] has a more technical discussion of reification and factor analysis. For a semihumorous discussion of reification, see Armstrong [1967].

NOTES

14.1 Graphing Two-Dimensional Projections

As noted in Section 14.8, the first two principal components can be used as plot axes to give a two-dimensional representation of higher-dimensional data. This plot will be best in the sense that it shows the maximum possible variability. Other multivariate graphical techniques give plots that are "the best" in other senses.

Multidimensional scaling gives a two-dimensional plot that reproduces the distances between points as accurately as possible. This view will be similar to the first two principal components when the data form a football (ellipsoid) shape, but may be very different when the data have a more complicated structure. Other *projection pursuit techniques* specifically search for views of the data that reveal holes, clusters, lines, and other departures from an ellipsoidal shape. A relatively nontechnical review of this concept is given by Jones and Sibson [1987].

Rather than relying on a single two-dimensional projection, it is also possible to display animated sequences of projections on a computer screen. The projections can be generated by random rotations of the data or by projection pursuit methods that attempt to show "interesting" projections. The free computer program GGobi (*http://www.ggobi.org*) implements many of these techniques.

Of course, more sophisticated searches performed by computer mean that more caution in interpretation is needed from the analyst. Substantial experience with these techniques is needed to develop a feeling for which graphs indicate real structure as opposed to overinterpreted noise.

14.2 Varimax and Quartimax Methods of Choosing Factors in a Factor Analysis

Many analytic methods of choosing factors have been developed so that the loading matrix is easy to interpret, that is, has a simple structure. These many different methods make the factor analysis literature very complex. We mention two of the methods.

1. *Varimax method.* The varimax method uses the idea of maximizing the sum of the variances of the squares of loadings of the factors. Note that the variances are high when the λ_{ij}^2 are near 1 and 0, some of each in each column. In order that variables with large communalities are not overly emphasized, weighted values are used. Suppose that we have the loadings λ_{ij} for one selection of factors. Let θ_{ij} be the loadings for a different set of factors (the linear combinations of the old factors). Define the weighted quantities

$$\gamma_{ij} = \theta_{ij} \Big/ \sqrt{\sum_{j=1}^{m} \lambda_{ij}^2}$$

The method chooses the θ_{ij} to maximize the following:

$$\sum_{j=1}^{k} \left[\frac{1}{p} \sum_{i=1}^{p} \gamma_{ij}^4 - \frac{1}{p^2} \left(\sum_{i=1}^{p} \gamma_{ij}^2 \right)^2 \right]$$

Some problems have a factor where all variables load high (e.g., general IQ). Varimax should not be used if a general factor may occur, as the low variance discourages general factors. Otherwise, it is one of the most satisfactory methods.

2. *Quartimax method.* The quartimax method works with the variance of the square of all p_k loadings. We maximize over all possible loadings θ_{ij}:

$$\max_{\theta_{ij}} \left[\sum_{i=1}^{p} \sum_{j=1}^{k} \theta_{ij}^4 - \frac{1}{pm} \left(\sum_{i=1}^{p} \sum_{j=1}^{k} \theta_{ij}^2 \right) \right]$$

Quartimax is used less often, since it tends to include one factor with all major loadings and no other major loadings in the rest of the matrix.

14.3 Statistical Test for the Number of Factors in a Factor Analysis When $X_1, \ldots, X_p$ Are Multivariate Normal and Maximum Likelihood Estimation Is Used

This note presupposes familiarity with matrix algebra. Let A be a matrix and A' denote the transpose of A; if A is square, let $|A|$ be the determinant of A and $\mathrm{Tr}(A)$ be the trace of A. Consider a factor analysis with k factors and estimated *loading matrix*

$$\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{nk} \end{pmatrix}$$

The test statistic is

$$X^2 = \left(n - 1 - \frac{2p+5}{6} - \frac{2k}{3} \right) \log_e \left(\frac{|\Lambda\Lambda' + \psi|}{|S|} \right) \mathrm{Tr}(S(\Lambda\Lambda' + \psi)^{-1})p$$

where S is the sample covariance matrix, ψ a diagonal matrix where $\psi_{ii} = s_i - (\Lambda\Lambda')_{ii}$, and s_i the sample variance of X_i. If the true number of factors is less than or equal to k, X^2 has a chi-square distribution with $[(p-k)^2 - (p+k)]/2$ degrees of freedom. The null hypothesis of only k factors is rejected if X^2 is too large.

One could try successively more factors until this is not significant. The true and nominal significance levels differ as usual in a stepwise procedure. (For the test to be appropriate, the degrees of freedom must be > 0.)

PROBLEMS

The first four problems present principal component analyses using correlation matrices. Portions of computer output (BMDP program 4M) are given. The coefficients for principal components that have a variance of 1 or more are presented. Because of the connection of principal component analysis and factor analysis mentioned in the text (when the correlations are used), the principal components are also called *factors* in the output. With a correlation matrix the coefficient values presented are for the standardized variables. You are asked to perform a subset of the following tasks.

- **(a)** Fill in the missing values in the "variance explained" and "cumulative proportion of total variance" table.
- **(b)** For the principal component(s) specified, give the percent of the total variance accounted for by the principal component(s).
- **(c)** How many principal components are needed to explain 70% of the total variance? 90%? Would a plot with two axes contain most (say, $\geq$ 70%) of the variability in the data?
- **(d)** For the case(s) with the value(s) as given, compute the case(s) values on the first two principal components.

14.1 This problem uses the psychosocial Framingham data in Table 11.20. The mnemonics go in the same order as the correlations presented. The results are presented in Tables 14.12 and 14.19. Perform tasks (a) and (b) for principal components 2 and 4, and task (c).

14.2 Measurement data on U.S. females by Stoudt et al. [1970] were discussed in this chapter. The same correlation data for adult males were also given (Table 14.14). The principal

Table 14.12 Problem 14.1: Variance Explained by Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	4.279180	0.251716
2	1.633777	0.347821
3	1.360951	?
4	1.227657	0.500092
5	1.166469	0.568708
6	?	0.625013
7	0.877450	0.676627
8	0.869622	0.727782
9	0.724192	0.770381
10	0.700926	0.811612
11	0.608359	?
12	0.568691	0.880850
13	0.490974	0.909731
14	?	0.935451
15	0.386540	0.958189
16	0.363578	0.979576
17	?	?

[a]The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.13 Problem 14.1: Principal Components

		Unrotated Factor Loadings (Pattern) for Principal Components				
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
TYPEA	1	0.633	−0.203	0.436	−0.049	0.003
EMOTLBLE	2	0.758	−0.198	−0.146	0.153	−0.005
AMBITIOS	3	0.132	−0.469	0.468	−0.155	−0.460
NONEASY	4	0.353	0.407	−0.268	0.308	0.342
NOBOSSPT	5	0.173	0.047	0.260	−0.206	0.471
WKOVRLD	6	0.162	−0.111	0.385	−0.246	0.575
MTDISSAG	7	0.499	0.542	0.174	−0.305	−0.133
MGDISSAT	8	0.297	0.534	−0.172	−0.276	−0.265
AGEWORRY	9	0.596	0.202	0.060	−0.085	−0.145
PERSONWY	10	0.618	0.346	0.192	−0.174	−0.206
ANGERIN	11	0.061	−0.430	−0.470	−0.443	−0.186
ANGEROUT	12	0.306	0.178	0.199	0.607	−0.215
ANGRDISC	13	0.147	−0.181	0.231	0.443	−0.108
STRESS	14	0.665	−0.189	0.062	−0.053	0.149
TENSION	15	0.771	−0.226	−0.186	0.039	0.118
ANXSYMPT	16	0.594	−0.141	−0.352	0.022	0.067
ANGSYMPT	17	0.723	−0.242	−0.256	0.086	−0.015
	VP[a]	4.279	1.634	1.361	1.228	1.166

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor loading matrix corresponding to that factor. The VP is the variance explained by the factor.

component analysis gave the results of Table 14.15. Perform tasks (a) and (b) for principal components 2, 3, and 4, and task (c).

14.3 The Bruce et al. [1973] exercise data for 94 sedentary males are used in this problem (see Table 9.16). These data were used in Problems 9.9 to 9.12. The exercise variables used are DURAT (duration of the exercise test in seconds), $VO_{2\ MAX}$ [the maximum oxygen consumption (normalized for body weight)], HR [maximum heart rate (beats/min)], AGE (in years), HT (height in centimeters), and WT (weight in kilograms). The correlation values are given in Table 14.17. The principal component analysis is given in Table 14.18. Perform tasks (a) and (b) for principal components 4, 5, and 6, and task (c) (Table 14.19). Perform task (d) for a case with DURAT $= 600$, $VO_{2\ MAX} = 38$, HR $= 185$, AGE $= 29$, HT $= 165$, and WT $= 71$. (*N.B.*: Find the value of the *standardized* variables.)

14.4 The variables are the same as in Problem 14.3. In this analysis 43 active females (whose individual data are given in Table 9.14) are studied. The correlations are given in Table 14.21. the principal component analysis in Tables 14.22 and 14.23. Perform tasks (a) and (b) for principal components 1 and 2, and task (c). Do task (d) for the two cases in Table 14.24 (use standard variables). See Table 14.21.

Problems 14.5, 14.7, 14.8, 14.10, 14.11, and 14.12 consider maximum likelihood factor analysis with varimax rotation (from computer program BMDP4M). Except for Problem 14.10, the number of factors is selected by Guttman's root criterion (the number of eigenvalues greater than 1). Perform the following tasks as requested.

Table 14.14 Problem 14.2: Correlations

		STHTER 1	STHTHL 2	KNEEHT 3	POPHT 4	ELBWHT 5
STHTER	1	1.000				
STHTHL	2	0.873	1.000			
KNEEHT	3	0.446	0.443	1.000		
POPHT	4	0.410	0.382	0.798	1.000	
ELBWHT	5	0.544	0.454	−0.029	−0.062	1.000
THIGHHT	6	0.238	0.284	0.228	−0.029	0.217
BUTTKNHT	7	0.418	0.429	0.743	0.619	0.005
BUTTPOP	8	0.227	0.274	0.626	0.524	−0.145
ELBWELBW	9	0.139	0.212	0.139	−0.114	0.231
SEATBRTH	10	0.365	0.422	0.311	0.050	0.286
BIACROM	11	0.365	0.335	0.352	0.275	0.127
CHESTGRH	12	0.238	0.298	0.229	0.000	0.258
WSTGRTH	13	0.106	0.184	0.138	−0.097	0.191
RTARMGRH	14	0.221	0.265	0.194	−0.059	0.269
RTARMSKN	15	0.133	0.191	0.081	−0.097	0.216
INFRASCP	16	0.096	0.152	0.038	−0.166	0.247
HT	17	0.770	0.717	0.802	0.767	0.212
WT	18	0.403	0.433	0.404	0.153	0.324
AGE	19	−0.272	−0.183	−0.215	−0.215	−0.192

		THIGH-HT 6	BUTT-KNHT 7	BUTT-POP 8	ELBW-ELBW 9	SEAT-BRTH 10
THIGHHT	6	1.000				
BUTTKNHT	7	0.348	1.000			
BUTTPOP	8	0.237	0.736	1.000		
ELBWELBW	9	0.603	0.299	0.193	1.000	
SEATBRTH	10	0.579	0.449	0.265	0.707	1.000
BIACROM	11	0.303	0.365	0.252	0.311	0.343
CHESTGRH	12	0.605	0.386	0.252	0.833	0.732
WSTGRTH	13	0.537	0.323	0.216	0.820	0.717
RTARMGRH	14	0.663	0.342	0.224	0.755	0.675
RTARMSKN	15	0.480	0.240	0.128	0.524	0.546
INFRASCP	16	0.503	0.212	0.106	0.674	0.610
HT	17	0.210	0.751	0.600	0.069	0.309
WT	18	0.684	0.551	0.379	0.804	0.813
AGE	19	−0.190	−0.151	−0.108	0.156	0.043

		BIACROM 11	CHESTGRH 12	WSTGRTH 13	RTARMGRH 14	RTARMSKN 15
BIACROM	11	1.000				
CHESTGRH	12	0.418	1.000			
WSTGRTH	13	0.249	0.837	1.000		
RTARMGRH	14	0.379	0.784	0.712	1.000	
RTARMSKN	15	0.183	0.558	0.552	0.570	1.000
INFRASCP	16	0.242	0.710	0.727	0.667	0.697
HT	17	0.381	0.189	0.054	0.139	0.060
WT	18	0.474	0.885	0.821	0.849	0.562
AGE	19	−0.261	0.062	0.299	−0.115	−0.039

		INFRASCP 16	HT 17	WT 18	AGE 19
INFRASCP	16	1.000			
HT	17	−0.003	1.000		
WT	18	0.709	0.394	1.000	
AGE	19	0.045	−0.270	−0.058	1.000

Table 14.15 Problem 14.2: Variance Explained by the Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	7.839282	0.412594
2	4.020110	0.624179
3	1.820741	0.720007
4	1.115168	0.778700
5	0.764398	0.818932
6	?	0.850389
7	0.475083	?
8	0.424948	0.897759
9	0.336247	0.915456
10	?	0.931210
11	0.252205	0.944484
12	?	0.955404
13	0.202398	0.966057
14	0.169678	0.974987
15	0.140613	0.982388
16	0.119548	?
17	0.117741	0.994872
18	0.055062	0.997770
19	0.042365	1.000000

[a]The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.16 Exercise Data for Problem 14.3

		Univariate Summary Statistics	
	Variable	Mean	Standard Deviation
1	DURAT	577.10638	123.83744
2	$VO_{2\ MAX}$	35.63298	7.51007
3	HR	175.39362	18.59195
4	AGE	49.78723	11.06955
5	HT	177.39851	6.58285
6	WT	79.00000	8.71286

Table 14.17 Problem 14.3: Correlation Matrix

		DURAT	$VO_{2\ MAX}$	HR	AGE	HT	WT
DURAT	1	1.000					
$VO_{2\ MAX}$	2	0.905	1.000				
HR	3	0.678	0.647	1.000			
AGE	4	−0.687	−0.656	−0.630	1.000		
HT	5	0.035	0.050	0.107	−0.161	1.000	
WT	6	−0.134	−0.147	0.015	−0.069	0.536	1.000

Table 14.18 Problem 14.3: Variance Explained by the Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	3.124946	0.520824
2	1.570654	?
3	0.483383	0.863164
4	?	0.926062
5	?	0.984563
6	0.092621	1.000000

[a]The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.19 Problem 14.3: Principal Components

		Unrotated Factor Loadings (Pattern) for Principal Components	
		Factor 1	Factor 2
DURAT	1	0.933	−0.117
$VO_{2\ MAX}$	2	0.917	−0.120
HR	3	0.832	0.057
AGE	4	−0.839	−0.134
HT	5	0.128	0.860
WT	6	−0.057	0.884
	VP[a]	3.125	1.571

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor loading matrix corresponding to that factor. The VP is the variance explained by the factor.

Table 14.20 Exercise Data for Problem 14.4

		Univariate Summary Statistics	
	Variable	Mean	Standard Deviation
1	DURAT	514.88372	77.34592
2	$VO_{2\ MAX}$	29.05349	4.94895
3	HR	180.55814	11.41699
4	AGE	45.13953	10.23435
5	HT	164.69767	6.30017
6	WT	61.32558	7.87921

Table 14.21 Problem 14.4: Correlation Matrix

		DURAT	$VO_{2\ MAX}$	HR	AGE	HT	WT
DURAT	1	1.000					
$VO_{2\ MAX}$	2	0.786	1.000				
HR	3	0.528	0.337	1.000			
AGE	4	−0.689	−0.651	−0.411	1.000		
HT	5	0.369	0.299	0.310	−0.455	1.000	
WT	6	0.094	−0.126	0.232	−0.042	0.483	1.000

Table 14.22 Problem 14.4: Variance Explained by the Principal Components[a]

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	3.027518	?
2	1.371342	0.733143
3	?	?
4	0.416878	0.918943
5	?	0.972750
6	?	1.000000

[a]The variance explained by each factor is the eigenvalue for that factor. Total variance is defined as the sum of the diagonal elements of the correlation (covariance) matrix.

Table 14.23 Problem 14.4: Principal Components

		Unrotated Factor Loadings (Pattern) for Principal Components	
		Factor 1	Factor 2
DURAT	1	0.893	−0.201
$VO_{2\ MAX}$	2	0.803	−0.425
HR	3	0.658	0.162
AGE	4	−0.840	0.164
HT	5	0.626	0.550
WT	6	0.233	0.891
	VP[a]	3.028	1.371

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor loading matrix corresponding to that factor. The VP is the variance explained by the factor.

Table 14.24 Data for Two Cases, Problem 14.3

	Subject 1	Subject 2
DURAT	660	628
$VO_{2\ MAX}$	38.1	38.4
HR	184	183
AGE	23	21
HT	177	163
WT	83	52

a. Examine the residual correlation matrix. What is the maximum residual correlation? Is it < 0.1? < 0.5?

b. For the pair(s) of variables, with mnemonics given, find the fitted residual correlation.

c. Consider the plots of the rotated factors. Discuss the extent to which the interpretation will be simple.

d. Discuss the potential for naming and interpreting these factors. Would you be willing to name any? If so, what names?

e. Give the uniqueness and communality for the variables whose numbers are given.

f. Is there any reason that you would like to see an analysis with fewer or more factors? If so, why?

g. If you were willing to associate a factor with variables (or a variable), identify the variables on the shaded form of the correlations. Do the variables cluster (form a dark group), which has little correlation with the other variables?

14.5 A factor analysis is performed upon the Framingham data of Problem 14.1. The results are given in Tables 14.25 to 14.27 and Figures 14.14 and 14.15. Communalities were obtained from five factors after 17 iterations. The communality of a variable is its squared multiple correlation with the factors; they are given in Table 14.26. Perform tasks (a), (b)

Table 14.25 Problem 14.5: Residual Correlations

		TYPEA 1	EMOTLBLE 2	AMBITIOS 3	NONEASY 4	NOBOSSPT 5	WKOVRLD 6
TYPEA	1	0.219					
EMOTLBLE	2	0.001	0.410				
AMBITIOS	3	0.001	0.041	0.683			
NONEASY	4	0.003	0.028	−0.012	0.635		
NOBOSSPT	5	−0.010	−0.008	0.001	−0.013	0.964	
WKOVRLD	6	0.005	−0.041	−0.053	−0.008	0.064	0.917
MTDISSAG	7	0.007	−0.010	−0.062	−0.053	0.033	0.057
MGDISSAT	8	0.000	0.000	0.000	0.000	0.000	0.000
AGEWORRY	9	0.002	0.030	0.015	0.017	0.001	−0.017
PERSONWY	10	−0.002	−0.010	0.007	0.007	−0.007	−0.003
ANGERIN	11	0.007	−0.006	−0.028	0.005	−0.018	0.028
ANGEROUT	12	0.001	0.056	0.053	0.014	−0.070	−0.135
ANGRDISC	13	−0.011	0.008	0.044	−0.019	−0.039	0.006
STRESS	14	0.002	−0.032	−0.003	0.018	0.030	0.034
TENSION	15	−0.004	−0.006	−0.016	−0.017	0.013	0.024
ANXSYMPT	16	0.004	−0.026	−0.028	−0.019	0.009	−0.015
ANGSYMPT	17	−0.000	0.018	−0.008	−0.012	−0.006	0.009

		MTDISSAG 7	MTDISSAT 8	AGEWORRY 9	PERSONWY 10	ANGERIN 11	ANGEROUT 12
MTDISSAG	7	0.574					
MGDISSAT	8	0.000	0.000				
AGEWORRY	9	0.001	−0.000	0.572			
PERSONWY	10	−0.002	0.000	0.001	0.293		
ANGERIN	11	0.010	−0.000	0.015	−0.003	0.794	
ANGEROUT	12	0.006	−0.000	−0.006	−0.001	−0.113	0.891
ANGRDISC	13	−0.029	−0.000	0.000	0.001	−0.086	0.080
STRESS	14	−0.017	−0.000	−0.015	0.013	0.022	−0.050
TENSION	15	0.004	−0.000	−0.020	0.007	−0.014	−0.045
ANXSYMPT	16	0.026	−0.000	0.037	−0.019	0.011	−0.026
ANGSYMPT	17	0.004	−0.000	−0.023	0.006	0.012	0.049

		ANGRDISC 13	STRESS 14	TENSION 15	ANXSYMPT 16	ANGSYMPT 17
ANGRDISC	13	0.975				
STRESS	14	−0.011	0.599			
TENSION	15	−0.005	0.035	0.355		
ANXSYMPT	16	−0.007	0.015	0.020	0.645	
ANGSYMPT	17	0.027	−0.021	−0.004	−0.008	0.398

Table 14.26 Problem 14.5: Communalities

1	TYPEA	0.7811
2	EMOTLBLE	0.5896
3	AMBITIOS	0.3168
4	NONEASY	0.3654
5	NOBOSSPT	0.0358
6	WKOVRLD	0.0828
7	MTDISSAG	0.4263
8	MGDISSAT	1.0000
9	AGEWORRY	0.4277
10	PERSONWY	0.7072
11	ANGERIN	0.2063
12	ANGEROUT	0.1087
13	ANGRDISC	0.0254
14	STRESS	0.4010
15	TENSION	0.6445
16	ANXSYMPT	0.3555
17	ANGSYMPT	0.6019

Table 14.27 Problem 14.5: Factors (Loadings Smaller Than 0.1 Omitted)

		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
TYPEA	1	0.331	0.185	0.133	0.753	0.229
EMOTLBLE	2	0.707	0.194		0.215	
AMBITIOS	3				0.212	0.515
NONEASY	4	0.215	0.105	0.163	0.123	−0.516
NOBOSSPT	5		0.101		0.142	
WKOVRLD	6				0.281	
MTDISSAG	7		0.474	0.391	0.178	
MGDISSAT	8		0.146	0.971	−0.143	
AGEWORRY	9	0.288	0.576			
PERSONWY	10	0.184	0.799	0.138	0.127	
ANGERIN	11	0.263			−0.238	0.272
ANGEROUT	12	0.128	0.179		0.196	−0.148
ANGRDISC	13	0.117			0.102	
STRESS	14	0.493	0.189		0.337	
TENSION	15	0.753	0.193		0.190	
ANXSYMPT	16	0.571	0.138			
ANGSYMPT	17	0.748	0.191			
	VP[a]	2.594	1.477	1.181	1.112	0.712

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

(TYPEA, EMOTLBLE) and (ANGEROUT, ANGERIN), (c), (d), and (e) for variables 1, 5, and 8, and tasks (f) and (g). In this study, the TYPEA variable was of special interest. Is it associated particularly with one of the factors?

14.6 This question requires you to do the fitting of the factor analysis model. Use the Florida voting data of Problem 9.34 available on the Web appendix to examine the structure of

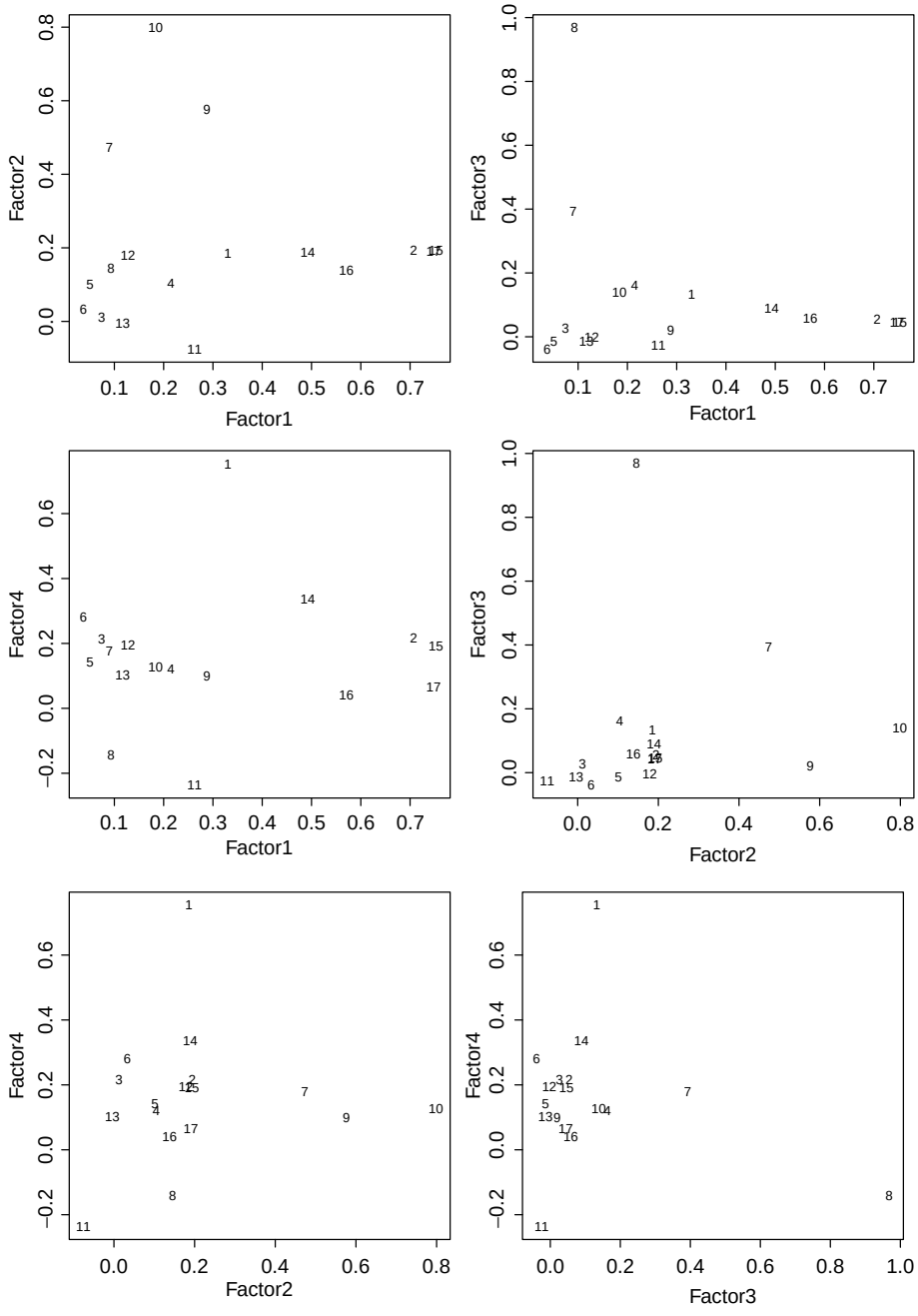


Figure 14.14 Problem 14.5, plots of factor loadings.

voting in the two Florida elections. As the counties are very different sizes, you will need to convert the counts to proportions voting for each candidate, and it may be useful to use the logarithm of this proportion. Fit models with one, two, or three factors and try to interpret them.

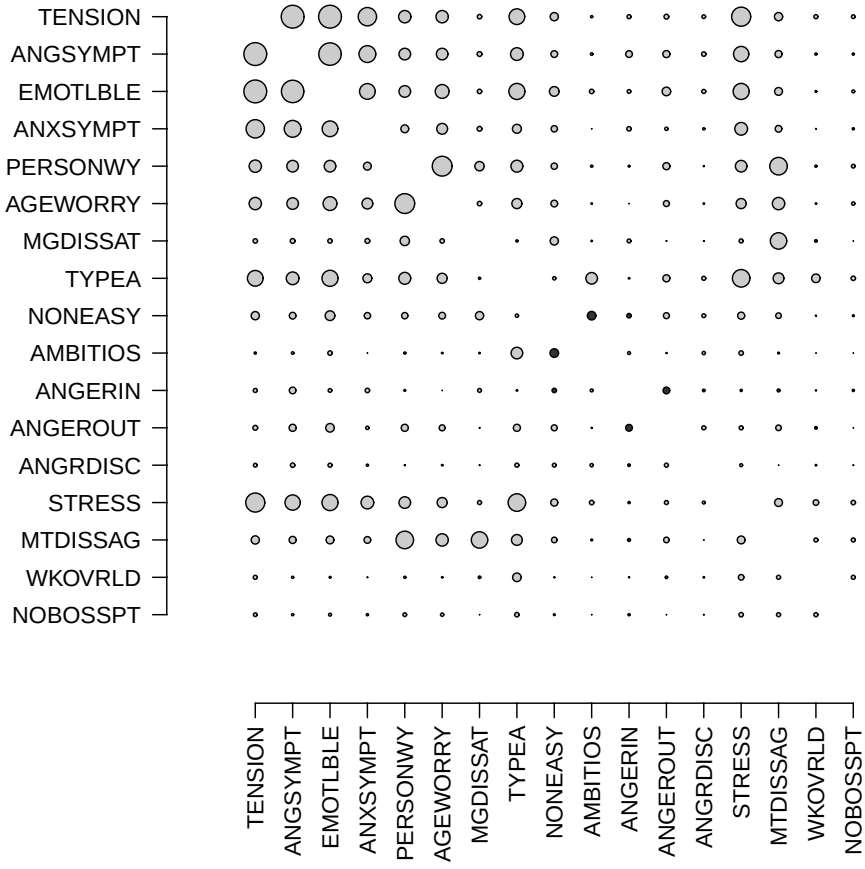


Figure 14.15 Shaded correlation matrix for Problem 14.5.

14.7 Starkweather [1970] performed a study entitled "Hospital Size, Complexity, and Formalization." He states: "Data on 704 United States short-term general hospitals are sorted into a set of dependent variables indicative of organizational formalism and a number of independent variables separately measuring hospital size (number of beds) and various types of complexity commonly associated with size." Here we used his data for a factor analysis of the following variables:

- *SIZE:* number of beds.
- *CONTROL:* a hospital was scored: 1 proprietary control; 2 nonprofit community control; 3 church operated; 4 public district hospital; 5 city or county control; 6 state control.
- *SCOPE* (of patient services): "A count was made of the number of services reported for each sample hospital. Services were weighted 1, 2, or 3 according to their relative impact on hospital operations, as measured by estimated proportion of total operating expenses."
- *TEACHVOL:* "The number of students in each of several types of hospital training programs was weighted and the products summed. The number of paramedical students

Table 14.28 Problem 14.7: Correlation Matrix

		SIZE 1	CONTROL 2	SCOPE 3	TEACHVOL 4	TECHTYPE 5	NONINPRG 6
SIZE	1	1.000					
CONTROL	2	−0.028	1.000				
SCOPE	3	0.743	−0.098	1.000			
TEACHVOL	4	0.717	−0.040	0.643	1.000		
TECHTYPE	5	0.784	−0.034	0.547	0.667	1.000	
NONINPRG	6	0.523	−0.051	0.495	0.580	0.440	1.000

Table 14.29 Problem 14.7: Communalities[a]

1	SIZE	0.8269
2	CONTROL	0.0055
3	SCOPE	0.7271
4	TEACHVOL	0.6443
5	TECHTYPE	1.0000
6	NONINPRG	0.3788

[a] Communalities obtained from two factors after eight iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.30 Problem 14.7: Residual Correlations

		SIZE 1	CONTROL 2	SCOPE 3	TEACHVOL 4	TECHTYPE 5	NONINPRG 6
SIZE	1	0.173					
CONTROL	2	0.029	0.995				
SCOPE	3	0.013	−0.036	0.273			
TEACHVOL	4	−0.012	0.012	−0.014	0.356		
TECHTYPE	5	−0.000	0.000	−0.000	−0.000	0.000	
NONINPRG	6	−0.020	−0.008	−0.027	0.094	−0.000	0.621

was weighted by 1.5, the number of RN students by 3, and the number of interns and residents by 5.5. These weights represent the average number of years of training typically involved, which in turn constitute a rough measure of the relative impact of students on hospital operations."

- *TECHTYPE:* types of teaching programs. The following scores were summed: 1 for practical nurse training program; 2 for RN; 3 for medical students; 4 for interns; 5 for residents.
- *NONINPRG:* noninpatient programs. Sum the following scores: 1 for emergency service; 2 for outpatient care; 3 for home care.

The results are given in Tables 14.28 to 14.31, and Figures 14.16 and 14.17. The factor analytic results follow. Perform tasks (a), (c), (d), and (e) for 1, 2, 3, 4, 5, and 6, and tasks (f) and (g).

Table 14.31 Problem 14.7: Factors (Loadings 14.31 Smaller Than 0.1 Omitted)

		Factor 1	Factor 2
SIZE	1	0.636	0.650
CONTROL	2		
SCOPE	3	0.357	0.774
TEACHVOL	4	0.527	0.605
TECHTYPE	5	0.965	0.261
NONINPRG	6	0.312	0.530
	VP[a]	1.840	1.743

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

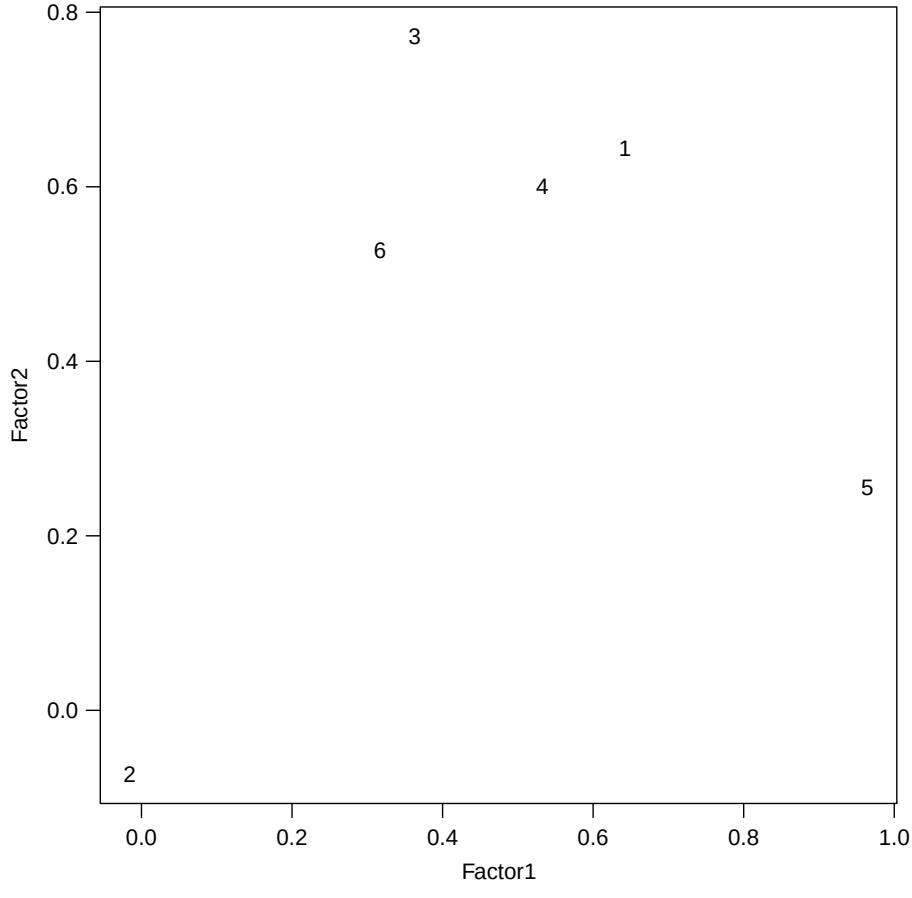


Figure 14.16 Problem 14.7, plot of factor loadings.

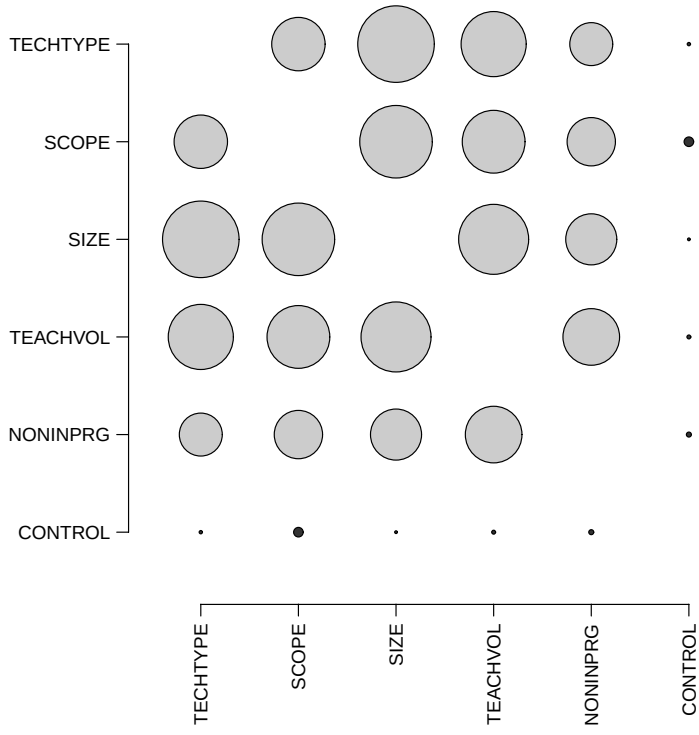


Figure 14.17 Shaded correlation matrix for Problem 14.7.

Table 14.32 Problem 14.8: Residual Correlations

		DURAT	$VO_{2\ MAX}$	HR	AGE	HT	WT
DURAT	1	0.067					
$VO_{2\ MAX}$	2	0.002	0.126				
HR	3	−0.005	−0.011	0.678			
AGE	4	0.004	0.011	−0.092	0.441	6	
HT	5	−0.006	0.018	−0.021	0.0106	0.574	
WT	6	0.004	−0.004	−0.008	0.007	0.605	0.301

14.8 This factor analysis examines the data used in Problem 14.3, the maximal exercise test data for sedentary males. The results are given in Tables 14.32 to 14.34 and Figures 14.18 and 14.19. Perform tasks (a), (b) (HR, AGE), (c), (d), and (e) for variables 1 and 5, and tasks (f) and (g).

14.9 Consider two variables, X and Y, with covariances (or correlations) given in the following notation. Prove parts (a) and (b) below.

	Variable	
Variable	**1**	**2**
X	*a*	*c*
Y	*c*	*b*

Table 14.33 Problem 14.8: Communalities[a]

1	DURAT	0.9331
2	$VO_{2\ MAX}$	0.8740
3	HR	0.5217
4	AGE	0.5591
5	HT	0.4264
6	WT	0.6990

[a]Communalities obtained from two factors after six iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.34 Problem 14.8: Factors

		Factor 1	Factor 2
DURAT	1	0.962	0.646
$VO_{2\ MAX}$	2	0.930	−0.092
HR	3	0.717	
AGE	4	−0.732	−0.154
HT	5		0.833
WT	6		0.833
	VP[a]	2.856	1.158

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

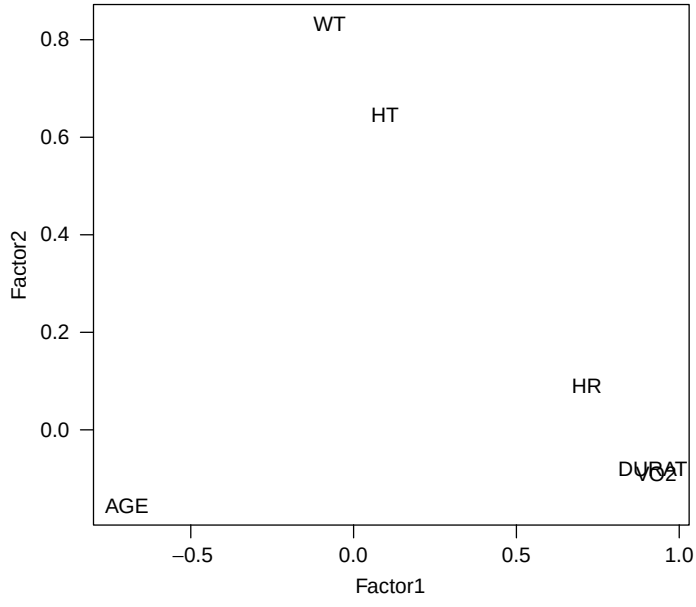


Figure 14.18 Problem 14.8, plot of factor loadings.

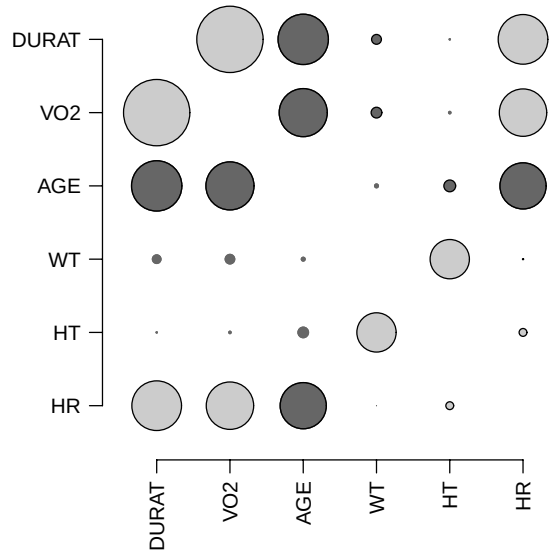


Figure 14.19 Shaded correlation matrix for Problem 14.8.

(a) We suppose that $c \neq 0$. The variance explained by the first principal component is

$$V_1 = \frac{(a+b) + \sqrt{(a-b)^2 + 4c^2}}{2}$$

The first principal component is

$$\sqrt{\frac{c^2}{c^2 + (V_1 - a)^2}} X + \frac{c}{|c|} \sqrt{\frac{(V_1 - a)^2}{c^2 + (V_1 - a)^2}} Y$$

(b) Suppose that $c = 0$. The first principal component is X if $a \geq b$, and is Y if $a < b$.

(c) The introduction to Problems 9.30–9.33 presented data on 20 patients who had their mitral valve replaced. The systolic blood pressure before and after surgery had the following variances and covariance:

	SBP	
	Before	**After**
Before	349.74	21.63
After	21.63	91.94

Find the variance explained by the first and second principal components.

14.10 The exercise data of the 43 active females of Problem 14.4 are used here. The findings are given in Tables 14.35 to 14.37 and Figures 14.20 and 14.21. Perform tasks (a), (c), (d), (f), and (g). Problem 14.8 examined similar exercise data for sedentary males.

Table 14.35 Problem 14.10: Residual Correlations

		DURAT	$VO_{2\ MAX}$	HR	AGE	HT	WT
DURAT	1	0.151					
$VO_{2\ MAX}$	2	0.008	0.241				
HR	3	0.039	−0.072	0.687			
AGE	4	0.015	0.001	−0.013	0.416		
HT	5	−0.045	0.013	−0.007	−0.127	0.605	
WT	6	0.000	0.000	0.000	−0.000	0.000	0.000

Table 14.36 Problem 14.10: Communalities[a]

1	DURAT	0.8492
2	$VO_{2\ MAX}$	0.7586
3	HR	0.3127
4	AGE	0.5844
5	HT	0.3952
6	WT	1.0000

[a]Communalities obtained from two factors after 10 iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.37 Problem 14.10: Factors

		Factor 1	Factor 2
DURAT	1	0.907	0.165
$VO_{2\ MAX}$	2	0.869	
HR	3	0.489	0.271
AGE	4	−0.758	−0.102
HT	5	0.364	0.513
WT	6		0.997
	VP[a]	2.529	1.371

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor.

Which factor analysis do you feel was more satisfactory in explaining the relationship among variables? Why? Which analysis had the more interpretable factors? Explain your reasoning.

14.11 The data on the correlation among male body measurements (of Problem 14.2) are factor analyzed here. The computer output gave the results given in Tables 14.38 to 14.40 and Figure 14.22. Perform tasks (a), (b) (POPHT, KNEEHT), (STHTER, BUTTKNHT), (RTARMSKN, INFRASCP), and (e) for variables 1 and 11, and tasks (f) and (g). Examine the diagonal of the residual values and the communalities. What values are on the diagonal of the residual correlations? (The diagonals are the 1–1, 2–2, 3–3, etc. entries.)

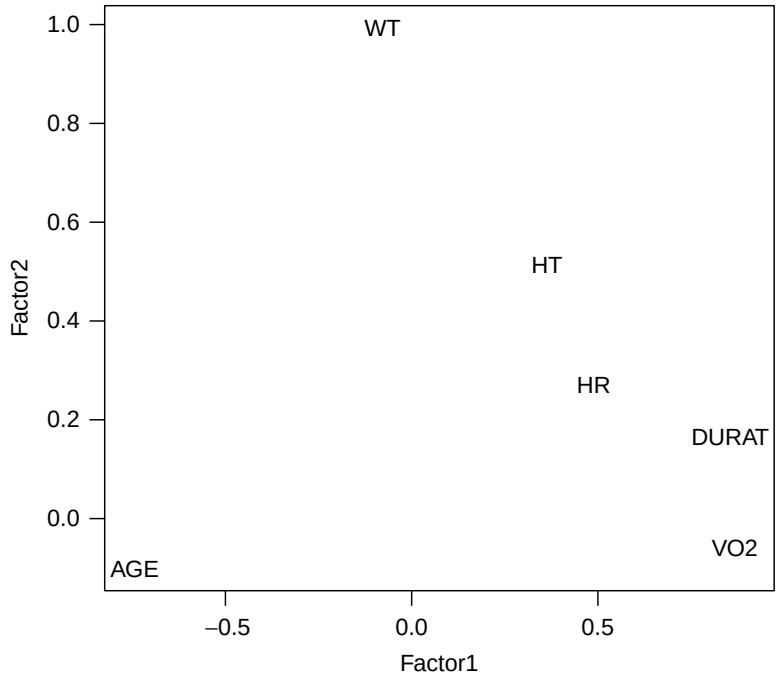


Figure 14.20 Problem 14.10, plot of factor loadings.

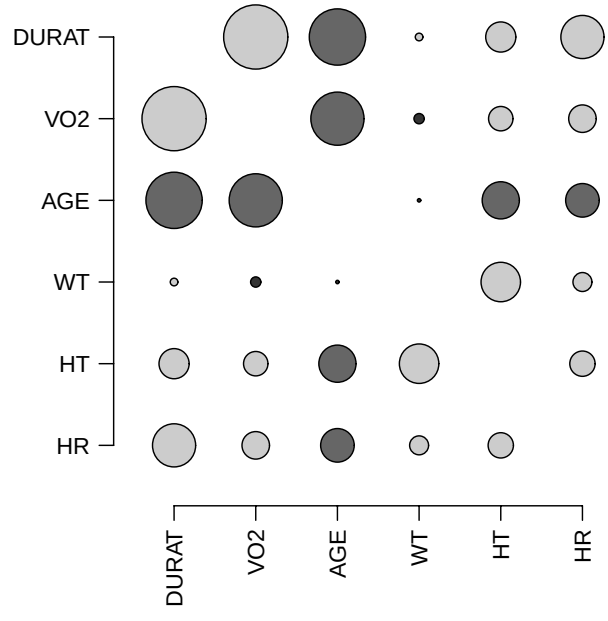


Figure 14.21 Shaded correlation matrix for Problem 14.10.

Table 14.38 Problem 14.11: Residual Correlations

		STHTER 1	STHTNORM 2	KNEEHT 3	POPHT 4	ELBWHT 5
STHTER	1	0.028				
STHTNORM	2	0.001	0.205			
KNEEHT	3	0.000	−0.001	0.201		
POPHT	4	0.000	−0.006	0.063	0.254	
ELBWHT	5	−0.001	−0.026	−0.012	0.011	0.519
THIGHHT	6	−0.003	0.026	0.009	−0.064	−0.029
BUTTKNHT	7	0.001	−0.004	−0.024	−0.034	−0.014
BUTTPOP	8	−0.001	0.019	−0.038	−0.060	−0.043
ELBWELBW	9	−0.001	0.008	0.007	−0.009	0.004
SEATBRTH	10	−0.002	0.023	0.015	−0.033	−0.013
BIACROM	11	0.006	−0.009	0.009	0.035	−0.077
CHESTGRH	12	−0.001	0.004	−0.004	0.015	−0.007
WSTGRTH	13	0.001	−0.004	−0.002	0.008	0.006
RTARMGRH	14	0.002	0.011	0.012	−0.006	−0.021
RTARMSKN	15	−0.002	0.025	−0.002	−0.012	0.009
INFRASCP	16	−0.002	0.003	−0.009	−0.002	0.020
HT	17	−0.000	0.001	−0.003	−0.003	0.007
WT	18	0.000	−0.007	0.001	0.004	0.007
AGE	19	−0.001	0.006	0.010	−0.014	−0.023

		THIGHHT 6	BUTTKNHT 7	BUTTPOP 8	ELBWELBW 9	SEATBRTH 10
THIGHHT	6	0.462				
BUTTKNHT	7	0.012	0.222			
BUTTPOP	8	0.016	0.076	0.409		
ELBWELBW	9	0.032	−0.002	0.006	0.215	
SEATBRTH	10	0.023	0.020	−0.017	0.007	0.305
BIACROM	11	−0.052	−0.019	−0.027	0.012	−0.023
CHESTGRH	12	−0.020	−0.013	−0.011	0.025	−0.020
WSTGRTH	13	−0.002	0.006	0.009	−0.006	−0.009
RTARMGRH	14	0.009	0.000	0.013	0.011	−0.017
RTARMSKN	15	0.038	0.039	0.015	−0.019	0.053
INFRASCP	16	−0.025	0.008	−0.000	−0.022	0.001
HT	17	0.005	0.005	0.005	0.000	−0.001
WT	18	−0.004	−0.005	−0.007	−0.006	0.004
AGE	19	−0.012	−0.010	−0.014	0.011	0.007

		BIACROM 11	CHESTGRH 12	WSTGRTH 13	RTARMGRH 14	RTARMSKN 15
BIACROM	11	0.684				
CHESTGRH	12	0.051	0.150			
WSTGRTH	13	−0.011	0.000	0.095		
RTARMGRH	14	−0.016	−0.011	−0.010	0.186	
RTARMSKN	15	−0.065	−0.011	0.009	0.007	0.601
INFRASCP	16	−0.024	−0.005	0.014	−0.022	0.199
HT	17	−0.008	0.000	−0.003	−0.005	0.004
WT	18	0.006	0.002	0.002	0.006	−0.023
AGE	19	−0.015	−0.006	−0.002	0.014	−0.024

		INFRASCP 16	HT 17	WT 18	AGE 19
INFRASCP	16	0.365			
HT	17	0.003	0.034		
WT	18	−0.003	0.001	0.033	
AGE	19	−0.022	0.002	0.002	0.311

Table 14.39 Problem 14.11: Communalities[a]

1	STHTER	0.9721
2	STHTNORM	0.7952
3	KNEEHT	0.7991
4	POPHT	0.7458
5	ELBWHT	0.4808
6	THIGHHT	0.5379
7	BUTTKNHT	0.7776
8	BUTTPOP	0.5907
9	ELBWELBW	0.7847
10	SEATBRTH	0.6949
11	BIACROM	0.3157
12	CHESTGRH	0.8498
13	WSTGRTH	0.9054
14	RTARMGRH	0.8144
15	RTARMSKN	0.3991
16	INFRASCP	0.6352
17	HT	0.9658
18	WT	0.9671
19	AGE	0.6891

[a]Communalities obtained from four factors after six iterations. The communality of a variable is its squared multiple correlation with the factors.

Table 14.40 Problem 14.11: Factors (Loadings Smaller Than 0.1 Omitted)

		Factor 1	Factor 2	Factor 3	Factor 4
		Unrotated[a]			
STHTER	1	0.100	0.356	0.908	−0.104
STHTNORM	2	0.168	0.367	0.795	
KNEEHT	3	0.113	0.875	0.128	
POPHT	4	−0.156	0.836	0.133	
ELBWHT	5	0.245	−0.151	0.617	−0.131
THIGHHT	6	0.675	0.131	0.114	−0.230
BUTTKNHT	7	0.308	0.819	0.100	
BUTTPOP	8	0.188	0.742		
ELBWELBW	9	0.873			0.131
SEATBRTH	10	0.765	0.209	0.247	
BIACROM	11	0.351	0.298	0.213	−0.242
CHESTGRH	12	0.902	0.137	0.118	
WSTGRTH	13	0.892			0.323
RTARMGRH	14	0.873			−0.198
RTARMSKN	15	0.625			
INFRASCP	16	0.794			
HT	17		0.836	0.507	−0.098
WT	18	0.907	0.308	0.218	−0.049
AGE	19		−0.135	−0.160	0.801
	VP[a]	6.409	3.964	2.370	0.978

[a]The VP for each factor is the sum of the squares of the elements of the column of the factor pattern matrix corresponding to that factor. When the rotation is orthogonal, the VP is the variance explained by the factor

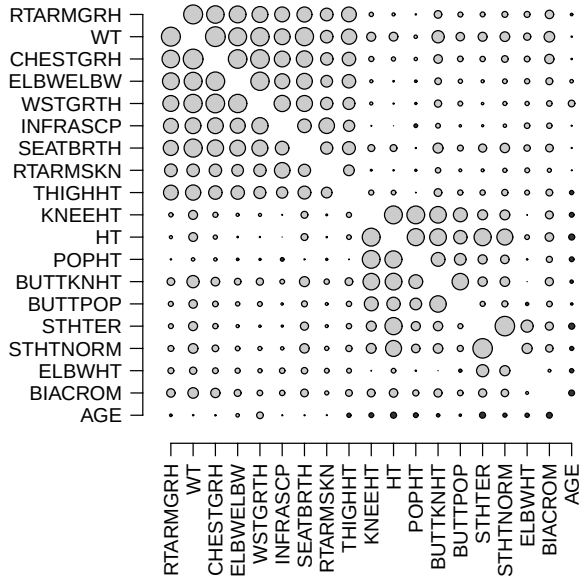


Figure 14.22 Shaded correlation matrix for Problem 14.11.

REFERENCES

Armstrong, J. S. [1967]. Derivation of theory by means of factor analysis, or, Tom Swift and his electric factor analysis machine. *American Statistician* **21**: 17–21.

Bruce, R. A., Kusumi, F., and Hosmer, D. [1973]. Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *American Heart Journal*, **85**: 546–562.

Chaitman, B. R., Fisher, L., Bourassa, M., Davis, K., Rogers, W., Maynard, C., Tyros, D., Berger, R., Judkins, M., Ringqvist, I., Mock, M. B., Killip, T., and participating CASS Medical Centers [1981]. Effects of coronary bypass surgery on survival in subsets of patients with left main coronary artery disease. Report of the Collaborative Study on Coronary Artery Surgery. *American Journal of Cardiology*, **48**: 765–777.

Gorsuch, R. L. [1983]. *Factor Analysis*. 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Gould, S. J. [1996]. *The Mismeasure of Man*. Revised, Expanded Edition. W.W. Norton, New York.

Guttman, L. [1954]. Some necessary conditions for common factor analysis. *Psychometrika*, **19**(2): 149–161.

Henry, R. C. [1997]. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* **37**: 525–530.

Jones, M. C., and Sibson, R. [1987]. What is projection pursuit? *Journal of the Royal Statistical Society, Series A*, **150**: 1–36.

Kim, J.-O., and Mueller, C. W. [1999]. *Introduction to Factor Analysis: What It Is and How to Do It.* Sage University Paper 13. Sage Publications, Beverly Hills, CA.

Kim, J.-O., and Mueller, C. W. [1983]. *Factor Analysis: Statistical Methods and Practical Issues.* Sage University Paper 14. Sage Publications, Beverly Hills, CA.

McDonald, R. P. [1999]. *Test Theory: A Unified Treatment. Lawrence* Erlbaum Associates, Mahwah, NJ.

Morrison, D. R. [1990]. *Multivariate Statistical Methods*, 3rd ed. McGraw-Hill, New York.

Paatero, P. [1997]. Least squares formulation of robust, non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, **37**: 23–35.

Paatero, P. [1999]. The multilinear engine: a table-driven least squares program for solving multilinear problems, including *n*-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, **8**: 854–888.

Reeck, G. R., and Fisher, L. D. [1973]. A statistical analysis of the amino acid composition of proteins. *International Journal of Peptide Protein Research*, **5**: 109–117.

Starkweather, D. B. [1970]. Hospital size, complexity, and formalization. *Health Services Research*, Winter, 330–341. Used with permission from the Hospital and Educational Trust.

Stoudt, H. W., Damon, A., and McFarland, R. A. [1970]. *Skinfolds, Body Girths, Biacromial Diameter, and Selected Anthropometric Indices of Adults: United States, 1960–62*. Vital and Health Statistics. Data from the National Survey. Public Health Service Publication 1000, Series 11, No. 35. U.S. Government Printing Office, Washington, DC.

Timm, N. H. [2001]. *Applied Multivariate Analysis*. Springer-Verlag, New York.

U.S. EPA [2000]. *Workshop on UNMIX and PMF as Applied to* $PM_{2.5}$. National Exposure Research Laboratory, Research Triangle Park, NC. *http://www.epa.gov/ttn/amtic/unmixmtg.html*.

CHAPTER 15

Rates and Proportions

15.1 INTRODUCTION

In this chapter and the next we want to study in more detail some of the topics dealing with counting data introduced in Chapter 6. In this chapter we want to take an epidemiological approach, studying populations by means of describing incidence and prevalence of disease. In a sense this is where statistics began: with a numerical description of the characteristics of a state, frequently involving mortality, fecundity, and morbidity. We call the occurrence of one of those outcomes an *event*. In the next chapter we deal with more recent developments, which have focused on a more detailed modeling of survival (hence also death, morbidity, and fecundity) and dealt with such data obtained in experiments rather than observational studies. An implication of the latter point is that sample sizes have been much smaller than used traditionally in the epidemiological context. For example, the evaluation of the success of heart transplants has, by necessity, been based on a relatively small set of data.

We begin the chapter with definitions of incidence and prevalence rates and discuss some problems with these "crude" rates. Two methods of standardization, direct and indirect, are then discussed and compared. In Section 15.4, a third standardization procedure is presented to adjust for varying exposure times among individuals. In Section 15.5, a brief tie-in is made to the multiple logistic procedures of Chapter 13. We close the chapter with notes, problems, and references.

15.2 RATES, INCIDENCE, AND PREVALENCE

The term *rate* refers to the amount of change occurring in a quantity with respect to time. In practice, *rate* refers to the amount of change in a variable over a specified time interval divided by the length of the time interval.

The data used in this chapter to illustrate the concepts come from the Third National Cancer Survey [National Cancer Institute, 1975]. For this reason we discuss the concepts in terms of incidence rates. The *incidence* of a disease in a fixed time interval is the number of new cases diagnosed during the time interval. The *prevalence* of a disease is the number of people with the disease at a fixed time point. For a chronic disease, incidence and prevalence may present markedly different ideas of the importance of a disease.

Consider the Third National Cancer Survey [National Cancer Institute, 1975]. This survey examined the incidence of cancer (by site) in nine areas during the time period 1969–1971.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

The areas were the Detroit SMSA (Standard Metropolitan Statistical Area); Pittsburgh SMSA, Atlanta SMSA, Birmingham SMSA, Dallas–Fort Worth SMSA, state of Iowa, Minneapolis–St. Paul SMSA, state of Colorado, and the San Francisco–Oakland SMSA. The information used in this chapter refers to the combined data from the Atlanta SMSA and San Francisco–Oakland SMSA. The data are abstracted from tables in the survey. Suppose that we wanted the rate for all sites (of cancer) combined. The rate per year in the 1969–1971 time interval would be simply the number of cases divided by 3, as the data were collected over a three-year interval. The rates are as follows:

$$\text{Combined area :} \quad \frac{181{,}027}{3} = 60{,}342.3$$

$$\text{Atlanta :} \quad \frac{9{,}341}{3} = 3{,}113.7$$

$$\text{San Francisco–Oakland :} \quad \frac{30{,}931}{3} = 10{,}310.3$$

Can we conclude that cancer incidence is worse in the San Francisco–Oakland area than in the Atlanta area? The answer is "yes and no." Yes, in that there are more cases to take care of in the San Francisco–Oakland area. If we are concerned about the chance of a person getting cancer, the numbers would not be meaningful. As the San Francisco–Oakland area may have a larger population, the number of cases per number of the population might be less. To make comparisons taking the population size into account, we use

$$\text{incidence per time interval} = \frac{\text{number of new cases}}{\text{total population} \times \text{time interval}} \tag{1}$$

The result of equation (1) would be quite small, so that the number of cases per 100,000 population is used to give a more convenient number. The rate per 100,000 population per year is then

$$\text{incidence per 100,000 per time interval} = \frac{\text{number of new cases}}{\text{total population} \times \text{time interval}} \times 100{,}000$$

For these data sets, the values are:

$$\text{Combined area :} \quad \frac{181{,}027 \times 100{,}000}{21{,}003{,}451 \times 3} = 287.3 \text{ new cases per } 100{,}000 \text{ per year}$$

$$\text{Atlanta :} \quad \frac{9{,}341 \times 100{,}000}{1{,}390{,}164 \times 3} = 224.0 \text{ new cases per } 100{,}000 \text{ per year}$$

$$\text{San Francisco-Oakland :} \quad \frac{30{,}931 \times 100{,}000}{3{,}109{,}519 \times 3} = 331.6 \text{ new cases per } 100{,}000 \text{ per year}$$

Even after adjusting for population size, the San Francisco–Oakland area has a higher overall rate.

Note several facts about the estimated rates. The estimates are binomial proportions times a constant (here 100,000/3). Thus, the rate has a standard error easily estimated. Let N be the total population and n the number of new cases; the rate is $n/N \times C$ ($C = 100{,}000/3$ in this example) and the standard error is estimated by

$$\sqrt{C^2 \frac{1}{N} \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

or

$$\text{standard error of rate per time interval} = C\sqrt{\frac{1}{N} \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

For example, the combined area estimate has a standard error of

$$\frac{100{,}000}{3}\sqrt{\frac{1}{21{,}003{,}451}\frac{181{,}027}{21{,}003{,}451}\left(1-\frac{181{,}027}{21{,}003{,}451}\right)}=0.67$$

As the rates are assumed to be binomial proportions, the methods of Chapter 6 may be used to get adjusted estimates or standardized estimates of proportions.

Rates computed by the foregoing methods,

$$\frac{\text{number of new cases in the interval}}{\text{population size} \times \text{time interval}}$$

are called *crude* or *total rates*. This term is used in distinction to *standardized* or *adjusted rates*, as discussed below.

Similarly, a *prevalence rate* can be defined as

$$\text{prevalence} = \frac{\text{number of cases at a point in time}}{\text{population size}}$$

Sometimes a distinction is made between *point prevalence* and *prevalence* to facilitate discussion of chronic disease such as epilepsy and a disease of shorter duration, for example, a common cold or even accidents. It is debatable whether the word *prevalence* should be used for accidents or illnesses of short duration.

15.3 DIRECT AND INDIRECT STANDARDIZATION

15.3.1 Problems with the Use of Crude Rates

Crude rates are useful for certain purposes. For example, the crude rates indicate the load of new cases per capita in a given area of the country. Suppose that we wished to use the cancer rates as epidemiologic indicators. The inference would be that it was likely that environmental or genetic differences were responsible for a difference, if any. There may be simpler explanations, however. Breast cancer rates would probably differ in areas that had differing gender proportions. A retirement community with an older population will tend to have a higher rate. To make fair comparisons, we often want to adjust for the differences between populations in one or more factors (covariates). One approach is to find an index that is adjusted in some fashion. We discuss two methods of adjustment in the next two sections.

15.3.2 Direct Standardization

In direct standardization we are interested in adjusting by one or more variables that are divided (or naturally fall) into discrete categories. For example, in Table 15.1 we adjust for gender and for age divided into a total of 18 categories. The idea is to find an answer to the following question: Suppose that the distribution with regard to the adjusting factors was not as observed, but rather, had been the same as this other (reference) population; what would the rate have been? In other words, we apply the risks observed in our study population to a reference population.

In symbols, the adjusting variable is broken down into I cells. In each cell we know the number of events (the numerator) n_i and the total number of individuals (the denominator) N_i:

Level of adjusting factor, i:	1	2	$\cdots$	i	$\cdots$	I
Proportion observed in study population:	$\frac{n_1}{N_1}$	$\frac{n_2}{N_2}$	$\cdots$	$\frac{n_i}{N_i}$	$\cdots$	$\frac{n_1}{N_1}$

Table 15.1 Rate for Cancer of All Sites for Blacks in the San Francisco–Oakland SMSA and Reference Population

	Study Population n_i/N_i		Reference Population M_i	
Age	Females	Males	Females	Males
<5	8/16,046	6/16,493	872,451	908,739
5–9	6/18,852	7/19,265	1,012,554	1,053,350
10–14	6/19,034	3/19,070	1,061,579	1,098,507
15–19	7/16,507	6/16,506	971,894	964,845
20–24	16/15,885	9/14,015	919,434	796,774
25–29	27/12,886	19/12,091	755,140	731,598
30–34	28/10,705	18/10,445	620,499	603,548
35–39	46/9,580	25/8,764	595,108	570,117
40–44	83/9,862	47/8,858	650,232	618,891
45–49	109/10,341	108/9,297	661,500	623,879
50–54	125/8,691	131/8,052	595,876	558,124
55–59	120/6,850	189/6,428	520,069	481,137
60–64	102/5,017	158/4,690	442,191	391,746
65–69	119/3,806	159/3,345	367,046	292,621
70–74	75/2,264	154/1,847	300,747	216,929
75–79	44/1,403	72/931	224,513	149,867
80–84	28/765	51/471	139,552	84,360
>85	25/629	26/416	96,419	51,615
Subtotal	974/169,123	1,188/160,984	10,806,804	10,196,647
Total	2,162/330,107		21,003,451	

Source: National Cancer Institute [1975].

Both numerator and denominator are presented in the table. The crude rate is estimated by

$$C\frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i}$$

Consider now a *standard or reference population*, which instead of having N_i persons in the ith cell has M_i.

	Reference Population					
Level of adjusting factor	1	2	⋯	i	⋯	I
Number in reference population	M_1	M_2	⋯	M_i	⋯	M_I

The question now is: If the study population has M_i instead of N_i persons in the ith cell, what would the crude rate have been? We cannot determine what the crude rate was, but we can estimate what it might have been. In the ith cell the proportion of observed deaths was n_i/N_i. If the same proportion of deaths occurred with M_i persons, we would expect

$$n_i^* = \frac{n_i}{N_i} M_i \text{ deaths}$$

Thus, if the adjusting variables had been distributed with M_i persons in the ith cell, we estimate that the data would have been:

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Expected proportion of cases:	$\frac{n_1 M_1/N_1}{M_1}$	$\frac{n_2 M_2/N_2}{M_2}$	$\cdots$	$\frac{n_i^*}{M_i}$	$\cdots$	$\frac{n_I M_I/N_I}{M_I}$

The *adjusted rate*, r, is the crude rate for this estimated standard population:

$$r = \frac{C \sum_{i=1}^{I} n_i M_i/N_i}{\sum_{i=1}^{I} M_i} = \frac{C \sum_{i=1}^{I} n_i^*}{\sum_{i=1}^{I} M_i}$$

As an example, consider the rate for cancer for all sites for blacks in the San Francisco–Oakland SMSA, adjusted for gender and age to the total combined sample of the Third Cancer Survey, as given by the 1970 census. There are two gender categories and 18 age categories, for a total of 36 cells. The cells are laid out in two columns rather than in one row of 36 cells. The data are given in Table 15.1.

The crude rate for the San Francisco–Oakland black population is

$$\frac{100{,}000}{3} \frac{974 + 1188}{169{,}123 + 160{,}984} = 218.3$$

Table 15.2 gives the values of $n_i M_i/N_i$.

The gender- and age-adjusted rate is thus

$$\frac{100{,}000}{3} \frac{193{,}499.42}{21{,}003{,}451} = 307.09$$

Note the dramatic change in the estimated rate. This occurs because the San Francisco–Oakland SMSA black population differs in its age distribution from the overall sample.

The variance is estimated by considering the denominators in the cell as fixed and using the binomial variance of the n_i's. Since the cells constitute independent samples,

$$\begin{aligned} \operatorname{var}(r) &= \operatorname{var}\left(C \sum_{i=1}^{I} \frac{n_i M_i}{N_i} \Big/ \sum_{i=1}^{I} M_i \right) \\ &= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \left(\frac{M_i}{N_i} \right)^2 \operatorname{var}(n_i) \end{aligned}$$

Table 15.2 Estimated Number of Cases per Cell ($n_i M_i/N_i$) if the San Francisco–Oakland Area Had the Reference Population Age and Gender Distribution

Age	Females	Males	Age	Females	Males
<5	434.97	330.59	55–59	9,110.70	14,146.69
5–9	322.26	382.74	60–64	8,990.13	13,197.41
10–14	334.64	172.81	65–69	11,476.21	13,909.34
15–19	412.14	350.73	70–74	9,962.91	18,087.20
20–24	926.09	511.66	75–79	7,041.03	11,590.14
25–29	1,582.24	1,149.65	80–84	5,107.79	9,134.52
30–34	1,622.98	1,040.10	>85	3,832.23	3,225.94
35–39	2,857.51	1,629.30			
40–44	5,472.45	3,283.80	Subtotal	85,029.16	108,470.26
45–49	6,972.58	7,247.38	Total	193,499.42	
50–54	8,570.30	9,080.26			

$$= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \left(\frac{M_i}{N_i}\right)^2 N_i \frac{n_i}{N_i} \left(1 - \frac{n_i}{N_i}\right)$$

$$= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \frac{M_i}{N_i} \frac{n_i M_i}{N_i} \left(1 - \frac{n_i}{N_i}\right)$$

where $M_\cdot = \sum_{i=1}^{I} M_i$.

If n_i/N_i is small, then $1 - n_i/N_i \doteq 1$ and

$$\text{var}(r) \doteq \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \frac{M_i}{N_i} \left(\frac{n_i M_i}{N_i}\right) \tag{2}$$

We use this to compute a 95% confidence interval for the adjusted rate computed above. Using equation (2), the standard error is

$$\begin{aligned} \text{SE}(r) &= \frac{C}{M_\cdot} \sqrt{\sum_{i=1}^{I} \frac{M_i}{N_i} \left(\frac{n_i M_i}{N_i}\right)} \\ &= \frac{100{,}000}{3} \frac{1}{21{,}003{,}451} \left(\frac{872{,}451}{16{,}046} 434.97 + \cdots\right)^{1/2} \\ &= 7.02 \end{aligned}$$

The quantity r is approximately normally distributed, so that the interval is

$$307.09 \pm 1.96 \times 7.02 \quad \text{or} \quad (293.3, 320.8)$$

If adjusted rates are estimated for two different populations, say r_1 and r_2, with standard errors $\text{SE}(r_1)$ and $\text{SE}(r_2)$, respectively, equality of the adjusted rates may be tested by using

$$z = \frac{r_1 - r_2}{\sqrt{\text{SE}(r_1)^2 + \text{SE}(r_2)^2}}$$

The $N(0,1)$ critical values are used, as z is approximately $N(0,1)$ under the null hypothesis of equal rates.

15.3.3 Indirect Standardization

In indirect standardization, the procedure of direct standardization is used in the opposite direction. That is, we ask the question: What would the mortality rate have been for the study population if it had the same rates as the population reference? That is, we apply the observed risks in the reference population to the study population.

Let m_i be the number of deaths in the reference population in the ith cell. The data are:

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Observed proportion in reference population:	$\frac{m_1}{M_1}$	$\frac{m_2}{M_2}$	$\cdots$	$\frac{m_i}{M_i}$	$\cdots$	$\frac{m_I}{M_I}$

where both numerator and denominators are presented in the table. Also,

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Denominators in study population:	N_1	N_2	$\cdots$	N_i	$\cdots$	N_I

The estimate of the rate the study population would have experienced is (analogous to the argument in Section 15.3.2)

$$r_{\text{REF}} = \frac{C \sum_{i=1}^{I} N_i (m_i / M_i)}{\sum_{i=1}^{I} N_i}$$

The crude rate for the study population is

$$r_{\text{STUDY}} = \frac{C \sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i}$$

where n_i is the observed number of cases in the study population at level i. Usually, there is not much interest in comparing the values r_{REF} and r_{STUDY} as such, because the distribution of the study population with regard to the adjusting factors is not a distribution of much interest. For this reason, attention is usually focused on the *standardized mortality ratio* (SMR), when death rates are considered, or the *standardized incidence ratio* (SIR), defined to be

$$\text{standardized ratio} = s = \frac{r_{\text{STUDY}}}{r_{\text{REF}}} = \frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i m_i / M_i} \tag{3}$$

The main advantage of the indirect standardization is that the SMR involves only the total number of events, so you do not need to know in which cells the deaths occur for the study population. An alternative way of thinking of the SMR is that it is the observed number of deaths in the study population divided by the expected number if the cell-specific rates of the reference population held.

As an example, let us compute the SIR of cancer in black males in the Third Cancer Survey, using white males of the same study as the reference population and adjusting for age. The data are presented in Table 15.3. The standardized incidence ratio is

$$s = \frac{8793}{7474.16} = 1.17645 = 1.18$$

One reasonable question to ask is whether this ratio is significantly different from 1. An approximate variance can be derived as follows:

$$s = \frac{O}{E} \qquad \text{where} \quad O = \sum_{i=1}^{I} n_i = n. \quad \text{and} \quad E = \sum_{i=1}^{I} N_i \left(\frac{m_i}{M_i} \right)$$

The variance of s is estimated by

$$\text{var}(s) = \frac{\text{var}(O) + s^2 \text{ var}(E)}{E^2} \tag{4}$$

The basic "trick" is to (1) assume that the number of cases in a particular cell follows a Poisson distribution and (2) to note that the sum of independent Poisson random variables is Poisson. Using these two facts yields

$$\text{var}(O) \doteq \sum_{i=1}^{I} n_i = n \tag{5}$$

Table 15.3 Cancer of All Areas Combined, Number of Cases, Black and White Males by Age and Number Eligible by Age

Age	Black Males		White Males			
	n_1	N_1	m_1	M_1	$\frac{N_i m_i}{M_i}$	$\left(\frac{N_i}{M_i}\right)^2 m_i$
<5	45	120,122	450	773,459	69.89	10.85
5–9	34	130,379	329	907,543	47.26	6.79
10–14	39	134,313	300	949,669	42.43	6.00
15–19	45	112,969	434	837,614	58.53	7.89
20–24	49	86,689	657	694,670	81.99	10.23
25–29	63	71,348	688	647,304	75.83	8.36
30–34	84	57,844	724	533,856	78.45	8.50
35–39	129	54,752	1,097	505,434	118.83	12.87
40–44	318	57,070	2,027	552,780	209.27	21.61
45–49	582	56,153	3,947	559,241	396.31	39.79
50–54	818	48,753	6,040	503,163	585.23	56.71
55–59	1,170	42,580	8,711	432,982	856.65	84.24
60–64	1,291	33,892	10,966	352,315	1,054.91	101.48
65–69	1,367	27,239	11,913	261,067	1,242.97	129.69
70–74	1,266	17,891	11,735	196,291	1,069.59	97.49
75–79	788	9,827	10,546	138,532	748.10	53.07
80–84	461	4,995	6,643	78,044	425.17	27.21
>85	244	3,850	3,799	46,766	312.75	25.75
Total	8,793	1,070,700	81,006	8,970,730	7,474.16	708.53

and

$$\operatorname{var}(E) \doteq \operatorname{var}\left(\sum_{i=1}^{I} \frac{N_i}{M_i} m_i\right) = \sum_{i=1}^{I} \left(\frac{N_i}{M_i}\right)^2 m_i \tag{6}$$

The variance of s is estimated by using equations (4), (5), and (6):

$$\operatorname{var}(s) = \frac{n_\cdot + s^2 \sum (N_i/M_i)^2 m_i}{E^2}$$

A test of the hypothesis that the population value of s is 1 is obtained from

$$z = \frac{s-1}{\sqrt{\operatorname{var}(s)}}$$

and $N(0, 1)$ critical values.

For the example,

$$\sum_{i=1}^{I} n_i = n_\cdot = 8793$$

$$E = \sum_{i=1}^{I} \frac{N_i}{M_i} m_i = 7474.16$$

$$\text{var}(E) \doteq \sum_{i=1}^{I} \left(\frac{N_i}{M_i}\right)^2 m_i = 708.53$$

$$\text{var}(s) \doteq \frac{8793 + (1.17645)^2 \times 708.53}{(7474.16)^2} = 0.000174957$$

From this and a standard error of $s \doteq 0.013$, the ratio is significantly different from one using

$$z = \frac{s - 1}{\text{SE}(s)} = \frac{0.17645}{0.013227} = 13.2$$

and $N(0, 1)$ critical values.

If the reference population is much larger than the study population, $\text{var}(E)$ will be much less than $\text{var}(O)$ and you may approximate $\text{var}(s)$ by $\text{var}(O)/E^2$.

15.3.4 Drawbacks to Using Standardized Rates

Any time a complex situation is summarized in one or a few numbers, considerable information is lost. There is always a danger that the lost information is crucial for understanding the situation under study. For example, two populations may have almost the same standardized rates but may differ greatly within the different cells; one population has much larger values in one subset of the cells and the reverse situation in another subset of cells. Even when the standardized rates differ, it is not clear if the difference is somewhat uniform across cells or results mostly from one or a few cells with much larger differences.

The moral of the story is that whenever possible, the rates in the cells used in standardization should be examined individually in addition to working with the standardized rates.

15.4 HAZARD RATES: WHEN SUBJECTS DIFFER IN EXPOSURE TIME

In the rates computed above, each person was exposed (eligible for cancer incidence) over the same length of time (three years, 1969–1971). (This is not quite true, as there is some population mobility, births, and deaths. The assumption that each person was exposed for three years is valid to a high degree of approximation.) There are other circumstances where people are observed for varying lengths of time. This happens, for example, when patients are recruited sequentially as they appear at a medical care facility. One approach would be to restrict the analysis to those who had been observed for at least some fixed amount of time (e.g., for one year). If large numbers of persons are not observed, this approach is wasteful by throwing away valuable and needed information. This section presents an approach that allows the rates to use all the available information if certain assumptions are satisfied.

Suppose that we observe subjects over time and look for an event that occurs only once. For definiteness, we speak about observing people where the event is death. Assume that over the time interval observed, if a subject has survived to some time t_0, the probability of death in a short interval from t_0 to t_1 is almost $\lambda(t_1 - t_0)$. The quantity λ is called the *hazard rate, force of mortality*, or *instantaneous death rate*. The units of λ are deaths per time unit.

How would we estimate λ from data in a real-life situation? Suppose that we have n individuals and begin observing the ith person at time B_i. If the person dies, let the time of death be D_i. Let the time of last contact be C_i for those people who are still alive. Thus, the time we are observing each person at risk of death is

$$O_i = \begin{cases} C_i - B_i & \text{if the subject is alive} \\ D_i - B_i & \text{if the subject is dead} \end{cases}$$

An unbiased estimate of λ is

$$\text{estimated hazard rate} = \widehat{\lambda} = \frac{\text{number of observed deaths}}{\sum_{i=1}^{n} O_i} = \frac{L}{\sum_{i=1}^{n} O_i} \tag{7}$$

As in the earlier sections of this chapter, $\widehat{\lambda}$ is often normalized to have different units. For example, suppose that $\widehat{\lambda}$ is in deaths per day of observation. That is, suppose that O_i is measured in days. To convert to deaths per 100 observation years, we use

$$\widehat{\lambda} \times 365 \frac{\text{days}}{\text{year}} \times 100$$

As an example, consider the paper by Clark et al. [1971]. This paper discusses the prognosis of patients who have undergone cardiac (heart) transplantation. They present data on 20 transplanted patients. These data are presented in Table 15.4. To estimate the deaths per year of exposure, we have

$$\frac{12 \text{ deaths}}{3599 \text{ exposure days}} \frac{365 \text{ days}}{\text{year}} = 1.22 \frac{\text{deaths}}{\text{exposure year}}$$

To compute the variance and standard error of the observed hazard rate, we again assume that L in equation (7) has a Poisson distribution. So conditional on the total observation period, the variability of the estimated hazard rate is proportional to the variance of L, which is estimated by L itself. Let

$$\widehat{\lambda} = \frac{CL}{\sum_{i=1}^{n} O_i}$$

where C is a constant that standardizes the hazard rate appropriately.

Table 15.4 Stanford Heart Transplant Data

i	Date of Transplantation	Date of Death	Time at Risk in Days (*if alive)[a]
1	1/6/68	1/21/68	15
2	5/2/68	5/5/68	3
3	8/22/68	10/7/68	46
4	8/31/68	—	608*
5	9/9/68	1/14/68	127
6	10/5/68	12/5/68	61
7	10/26/68	—	552*
8	11/20/68	12/14/68	24
9	11/22/68	8/30/69	281
10	2/8/69	—	447*
11	2/15/69	2/25/69	10
12	3/29/69	5/7/69	39
13	4/13/69	—	383*
14	5/22/69	—	344*
15	7/16/69	11/29/69	136
16	8/16/69	8/17/69	1
17	9/3/69	—	240*
18	9/14/69	11/13/69	60
19	1/3/70	—	118*
20	1/16/70	—	104*

[a] Total exposure days = 3599, $L = 12$.

Then the standard error of $\widehat{\lambda}$, $\mathrm{SE}(\widehat{\lambda})$, is approximately

$$\mathrm{SE}(\widehat{\lambda}) \doteq \frac{C}{\sum_{i=1}^{n} O_i}\sqrt{L}$$

A confidence interval for λ can be constructed by using confidence limits (L_1, L_2) for $E(L)$ as described in Note 6.8:

$$\text{confidence interval for } \lambda = \left(\frac{CL_1}{\sum_{i=1}^{n} O_i}, \frac{CL_2}{\sum_{i=1}^{n} O_i}\right)$$

For the example, a 95% confidence interval for the number of deaths is (6.2–21.0). A 95% confidence interval for the hazard rate is then

$$\left(\frac{6.2}{3599} \times 365, \frac{21.0}{3599} \times 365\right) = (0.63, 2.13)$$

Note that this assumes a constant hazard rate from day of transplant; this assumption is suspect. In Chapter 16 some other approaches to analyzing such data are given.

As a second more complicated illustration, consider the work of Bruce et al. [1976]. This study analyzed the experience of the Cardiopulmonary Research Institute (CAPRI) in Seattle, Washington. The program provided medically supervised exercise programs for diseased subjects. Over 50% of the participants dropped out of the program. As the subjects who continued participation and those who dropped out had similar characteristics, it was decided to compare the mortality rates for men to see if the training prevented mortality. It was recognized that subjects might drop out because of factors relating to disease, and the inference would be weak in the event of an observed difference.

The interest of this example is in the appropriate method of calculating the rates. All subjects, *including the dropouts*, enter into the computation of the mortality for active participants! The reason for this is that had they died during training, they would have been counted as active participant deaths. Thus, training must be credited with the exposure time or observed time when the dropouts were in training. For those who did not die and dropped out, the date of last contact *as an active participant* was the date at which the subjects left the training program. (Topics related to this are dealt with in Chapter 16).

In summary, to compute the mortality rates for active participants, all subjects have an observation time. The times are:

1. O_i = (time of death − time of enrollment) for those who died as active participants
2. O_i = (time of last contact − time of enrollment) for those in the program at last contact
3. O_i = (time of dropping the program−time of enrollment) for those who dropped whether or not a subsequent death was observed

The rate $\widehat{\lambda}_A$ for active participants is then computed as

$$\widehat{\lambda}_A = \frac{\text{number of deaths observed during training}}{\sum_{\text{all individuals}} O_i} = \frac{L_A}{\sum O_i}$$

To estimate the rate for dropouts, only those who drop out have time at risk of dying as a dropout. For those who have died, the time observed is

$$O_i' = (\text{time of death} - \text{time the subject dropped out})$$

For those alive at the last contact,

$$O_i' = (\text{time of last contact} - \text{time the subject dropped out})$$

The hazard rate for the dropouts, $\widehat{\lambda}_D$, is

$$\widehat{\lambda}_D = \frac{\text{number of deaths observed during dropout period}}{\sum_{\text{dropouts}} O_i'} = \frac{L_D}{\sum O_i'}$$

The paper reports rates of 2.7 deaths per 100 person-years for the active participants based on 16 deaths. The mortality rate for dropouts was 4.7 based on 34 deaths.

Are the rates statistically different at a 5% significance level? For a Poisson variable, L, the variance equals the expected number of observations and is thus estimated by the value of the variable itself. The rates $\widehat{\lambda}$ are of the form

$$\widehat{\lambda} = CL \qquad (L \text{ the number of events})$$

Thus, $\text{var}(\widehat{\lambda}) = C^2 \text{ var}(L) \doteq C^2 L = \widehat{\lambda}^2/L$.

To compare the two rates,

$$\text{var}(\widehat{\lambda}_A - \widehat{\lambda}_D) = \text{var}(\widehat{\lambda}_A) + \text{var}(\widehat{\lambda}_D) = \frac{\widehat{\lambda}_A^2}{L_A} + \frac{\widehat{\lambda}_D^2}{L_D}$$

The approximation is good for large L.

An approximate normal test for the equality of the rates is

$$z = \frac{\widehat{\lambda}_A - \widehat{\lambda}_D}{\sqrt{\widehat{\lambda}_A^2/L_A + \widehat{\lambda}_D^2/L_D}}$$

For the example, $L_A = 16, \widehat{\lambda}_A = 2.7$, and $L_D = 34, \widehat{\lambda}_D = 4.7$, so that

$$\begin{aligned} z &= \frac{2.7 - 4.7}{\sqrt{(2.7)^2/16 + (4.7)^2/34}} \\ &= -1.90 \end{aligned}$$

Thus, the difference between the two groups was not statistically significant at the 5% level.

15.5 MULTIPLE LOGISTIC MODEL FOR ESTIMATED RISK AND ADJUSTED RATES

In Chapter 13 the linear discriminant model or multiple logistic model was used to estimate the probability of an event as a function of covariates, $X_1, \ldots, X_n$. Suppose that we want a direct adjusted rate, where $X_1(i), \ldots, X_n(i)$ was the covariate value at the midpoints of the ith cell. For the study population, let p_i be the adjusted probability of an event at $X_1(i), \ldots, X_n(i)$. An adjusted estimate of the probability of an event is

$$\widehat{p} = \frac{\sum_{i=1}^{I} M_i p_i}{\sum_{i=1}^{I} M_i}$$

where M_i is the number of reference population subjects in the ith cell. This equation can be written as

$$\widehat{p} = \sum_{i=1}^{I} \left(\frac{M_i}{M.} p_i \right)$$

where $M. = \sum_{i=1}^{I} M_i$.

If the study population is small, it is better to estimate the p_i using the approach of Chapter 13 rather than the direct standardization approach of Section 15.3. This will usually be the case when there are several covariates with many possible values.

NOTES

15.1 More Than One Event per Subject

In some studies, each person may experience more than one event: for example, seizures in epileptic patients. In this case, each person could contribute more than once to the numerator in the calculation of a rate. In addition, exposure time or observed time would continue beyond an event, as the person is still at risk for another event. You need to check in this case that there are not people with "too many" events; that is, events "cluster" in a small subset of the population. A preliminary test for clustering may then be called for. This is a complicated topic. See Kalbfleisch and Prentice [2002] for references. One possible way of circumventing the problem is to record the time to the second or kth event. This builds a certain robustness into the data, but of course, makes it not possible to investigate the clustering, which may be of primary interest.

15.2 Standardization with Varying Observation Time

It is possible to compute standardized rates when the study population has the rate in each cell determined by the method of Section 15.4; that is, people are observed for varying lengths of time. In this note we discuss only the method for direct standardization.

Suppose that in each of the i cells, the rates in the study population is computed as CL_i/O_i, where C is a constant, L_i the number of events, and O_i the sum of the times observed for subjects in that cell. The adjusted rate is

$$\frac{\sum_{i=1}^{I} (M_i/L_i)\, O_i}{\sum_{i=1}^{I} M_i} = \frac{C \sum_{i=1}^{I} M_i \widehat{\lambda}_i}{M.} \qquad \text{where} \quad \widehat{\lambda}_i = \frac{L_i}{O_i}$$

The standard error is estimated to be

$$\frac{C}{M.} \sqrt{\sum_{i=1}^{I} \left(\frac{M_i}{O_i} \right) L_i}$$

15.3 Incidence, Prevalence, and Time

The *incidence* of a disease is the rate at which new cases appear; the *prevalence* is the proportion of the population that has the disease. When a disease is in a steady state, these are related via the average duration of disease:

$$\text{prevalence} = \text{incidence} \times \text{duration}$$

That is, if you catch a cold twice per year and each cold lasts a week, you will spend two weeks per year with a cold, so $2/52$ of the population should have a cold at any given time.

This equation breaks down if the disease lasts for all or most of your life and does not describe transient epidemics.

15.4 Sources of Demographic and Natural Data

There are many government sources of data in all of the Western countries. Governments of European countries, Canada, and the United States regularly publish vital statistics data as well as results of population surveys such as the Third National Cancer Survey [National Cancer Institute, 1975]. In the United States, the National Center for Health Statistics (*http://www.cdc.gov/nhcs*) publishes more than 20 series of monographs dealing with a variety of topics. For example, Series 20 provides natural data on mortality; Series 21, on natality, marriage, and divorce. These reports are obtainable from the U.S. government.

15.5 Binomial Assumptions

There is some question whether the binomial assumptions (see Chapter 6) always hold. There may be "extrabinomial" variation. In this case, standard errors will tend to be underestimated and sample size estimates will be too low, particularly in the case of dependent Bernoulli trials. Such data are not easy to analyze; sometimes a logarithmic transformation is used to stabilize the variance.

PROBLEMS

15.1 This problem will give practice by asking you to carry out analyses similar to the ones in each of the sections. The numbers from the National Cancer Institute [1975] for lung cancer cases for white males in the Pittsburgh and Detroit SMSAs are given in Table 15.5.

Table 15.5 Lung Cancer Cases by Age for White Males in the Detroit and Pittsburgh SMSAs

	Detroit		Pittsburgh	
Age	Cases	Population Size	Cases	Population Size
<5	0	149,814	0	82,242
5–9	0	175,924	0	99,975
10–14	2	189,589	1	113,146
15–19	0	156,910	0	100,139
20–24	5	113,003	0	68,062
25–29	1	113,919	0	61,254
30–34	10	92,212	7	53,289
35–39	24	90,395	21	55,604
40–44	101	108,709	56	70,832
45–49	198	110,436	148	74,781
50–54	343	98,756	249	72,247
55–59	461	82,758	368	64,114
60–64	532	63,642	470	50,592
65–69	572	47,713	414	36,087
70–74	473	35,248	330	26,840
75–79	365	25,094	259	19,492
80–84	133	12,577	105	10,987
>85	51	6,425	52	6,353
Total	3271	1,673,124	2480	1,066,036

(a) Carry out the analyses of Section 15.2 for these SMSAs.

(b) Calculate the direct and indirect standardized rates for lung cancer for white males adjusted for age. Let the Detroit SMSA be the study population and the Pittsburgh SMSA be the reference population.

(c) Compare the rates obtained in part (b) with those obtained in part (a).

15.2 **(a)** Calculate crude rates and standardized cancer rates for the white males of Table 15.5 using black males of Table 15.3 as the reference population.

(b) Calculate the standard error of the indirect standardized mortality rate and test whether it is different from 1.

(c) Compare the standardized mortality rates for blacks and whites.

15.3 The data in Table 15.6 represent the mortality experience for farmers in England and Wales 1949–1953 as compared with national mortality statistics.

Table 15.6 Mortality Experience Data for Problem 15.3

Age	National Mortality (1949–1953) Rate per 100,000/Year	Population of Farmers (1951 Census)	Deaths in 1949–1953
20–24	129.8	8,481	87
25–34	152.5	39,729	289
35–44	280.4	65,700	733
45–54	816.2	73,376	1,998
55–64	2,312.4	58,226	4,571

(a) Calculate the crude mortality rates.

(b) Calculate the standardized mortality rates.

(c) Test the significance of the standardized mortality rates.

(d) Construct a 95% confidence interval for the standardized mortality rates.

(e) What are the units for the ratios calculated in parts (a) and (b)?

15.4 Problems for discussion and thought:

(a) Direct and indirect standardization permit comparison of rates in two populations. Describe in what way this can also be accomplished by multiway contingency tables.

(b) For calculating standard errors of rates, we assumed that events were binomially (or Poisson) distributed. State the assumption of the binomial distribution in terms of, say, the event "death from cancer" for a specified population. Which of the assumptions is likely to be valid? Which is not likely to be invalid?

(c) Continuing from part (b), we calculate standard errors of rates that are population based; hence the rates are not samples. Why calculate standard errors anyway, and do significance testing?

15.5 This problem deals with a study reported in Bunker et al. [1969]. Halothane, an anesthetic agent, was introduced in 1956. Its early safety record was good, but reports of massive hepatic damage and death began to appear. In 1963, a Subcommittee on the National Halothane Study was appointed. Two prominent statisticians, Frederick Mosteller and Lincoln Moses, were members of the committee. The committee designed a large cooperative retrospective study, ultimately involving 34 institutions

Table 15.7 Mortality Data for Problem 15.5

	Number of Operations			Number of Deaths		
Physical Status	Total	Halothane	Cyclopropane	Total	Halothane	Cyclopropane
Unknown	69,239	23,684	10,147	1,378	419	297
1	185,919	65,936	27,444	445	125	91
2	104,286	36,842	14,097	1,856	560	361
3	29,491	8,918	3,814	2,135	617	403
4	3,419	1,170	681	590	182	127
5	21,797	6,579	7,423	314	74	101
6	11,112	2,632	3,814	1,392	287	476
7	2,137	439	749	673	111	253
Total	427,400	146,200	68,169	8,783	2,375	2,109

that completed the study. "The primary objective of the study was to compare halothane with other general anesthetics as to incidence of fatal massive hepatic necrosis within six weeks of anesthesia." A four-year period, 1959–1962, was chosen for the study. One categorization of the patients was by physical status at the time of the operation. Physical status varies from good (category 1) to moribund (category 7). Another categorization was by mortality level of the surgical procedure, having values of low, middle, high. The data in Table 15.7 deal with middle-level mortality surgery and two of the five anesthetic agents studied, the total number of administrations, and the number of patients dying within six weeks of the operation.

(a) Calculate the crude death rates per 100,000 per year for total, halothane, and cyclopropane. Are the crude rates for halothane and cyclopropane significantly different?

(b) By direct standardization (relative to the total), calculate standardized death rates for halothane and cyclopropane. Are the standardized rates significantly different?

(c) Calculate the standardized mortality rates for halothane and cyclopropane and test the significance of the difference.

(d) The calculations of the standard errors of the standardized rates depend on certain assumptions. Which assumptions are likely not to be valid in this example?

15.6 In 1980, 45 SIDS (sudden infant death syndrome) deaths were observed in King County. There were 15,000 births.

(a) Calculate the SIDS rate per 100,000 births.

(b) Construct a 95% confidence interval on the SIDS rate per 100,000 using the Poisson approximation to the binomial.

(c) Using the normal approximation to the Poisson, set up the 95% limits.

(d) Use the square root transformation for a Poisson random variable to generate a third set of 95% confidence intervals. Are the intervals comparable?

(e) The SIDS rate in 1970 in King County is stated to be 250 per 100,000. Someone wants to compare this 1970 rate with the 1980 rate and carries out a test of two proportions, $p_1 = 300$ per 100,000 and $p_2 = 250$ per 100,000, using the binomial distributions with $N_1 = N_2 = 100{,}000$. The large-sample normal approximation is used. What part of the Z-statistic: $(p_1 - p_2)$/standard error$(p_1 - p_2)$ will be right? What part will be wrong? Why?

Table 15.8 Heart Disease Data for Problem 15.7

Gender	Age	Epileptics: Person-Years at Risk	New and Nonfatal IHD Cases	Incidence in General Population per 100,000/year
Male	30–39	354	2	76
	40–49	303	2	430
	50–59	209	3	1291
	60–69	143	4	2166
	70+	136	4	1857
Female	30–39	534	0	9
	40–49	363	1	77
	50–59	218	3	319
	60–69	192	4	930
	70+	210	2	1087

15.7 Annegers et al. [1976] investigated ischemic heart disease (IHD) in patients with epilepsy. The hypothesis of interest was whether patients with epilepsy, particularly those on long-term anticonvulsant medication, were at less than expected risk of ischemic heart disease. The study dealt with 516 cases of epilepsy; exposure time was measured from time of diagnosis of epilepsy to time of death or time last seen alive.

(a) For males aged 60 to 69, the number of years at risk was 161 person-years. In this time interval, four IHD deaths were observed. Calculate the hazard rate for this age group in units of 100,000 persons/year.

(b) Construct a 95% confidence interval.

(c) The expected hazard rate in the general population is 1464 per 100,000 persons/year. How many deaths would you have expected in the age group 60 to 69 on the basis of the 161 person-years experience?

(d) Do the number of observed and expected deaths differ significantly?

(e) The raw data for the incidence of ischemic heart disease are given in Table 15.8. Calculate the expected number of deaths for males and the expected number of deaths for females by summing the expected numbers in the age categories (for each gender separately). Treat the total observed as a Poisson random variable and set up 95% confidence intervals. Do these include the expected number of deaths? State your conclusion.

(f) Derive a formula for an indirect standardization of these data (see Note 15.2) and apply it to these data.

15.8 A random sample of 100 subjects from a population is divided into two age groups, and for each age group the number of cases of a certain disease is determined. A reference population of 2000 persons has the following age distribution:

	Sample		Reference Population
Age	**Total Number**	**Number of Cases**	**Total Number**
1	80	8	1000
2	20	8	1000

(a) What is the crude case rate per 1000 population for the sample?

(b) What is the standard error of the crude case rate?

(c) What is the age-adjusted case rate per 1000 population using direct standardization and the reference population above?

(d) How would you test the hypothesis that the case rate at age 1 is not significantly different from the case rate at age 2?

15.9 The data in Table 15.9 come from a paper by Friis et al. [1981]. The mortality among male Hispanics and non-Hispanics was as shown.

Table 15.9 Mortality Data for Problem 15.9

	Hispanic Males		Non-Hispanic Males	
Age	Number	Number of Deaths	Number	Number of Deaths
0–4	11,089	0	51,250	0
5–14	18,634	0	120,301	0
15–24	10,409	0	144,363	2
25–34	16,269	2	136,808	9
35–44	11,050	0	106,492	46
45–54	6,368	7	91,513	214
55–64	3,228	8	70,950	357
65–74	1,302	12	34,834	478
75+	1,104	27	16,223	814
Total	79,453	56	772,734	1,920

(a) Calculate the crude death rate among Hispanic males.

(b) Calculate the crude death rate among non-Hispanic males.

(c) Compare parts (a) and (b) using an appropriate test.

(d) Calculate the SMR using non-Hispanic males as the reference population.

(e) Test the significance of the SMR as compared with a ratio of 1. Interpret your results.

15.10 The data in Table 15.10, abstracted from National Center for Health Statistics [1976], deal with the mortality experience in poverty and nonpoverty areas of New York and Seattle.

(a) Using New York City as the "standard population," calculate the standardized mortality rates for Seattle taking into account race and poverty area.

(b) Estimate the variance of this quantity and calculate 99% confidence limits.

(c) Calculate the standardized death rate per 100,000 population.

Table 15.10 Mortality Data for Problem 15.10

		New York City		Seattle	
Area	Race	Population	Death Rate per 1000	Population	Death Rate per 1000
Poverty	White	974,462	9.9	29,016	22.9
	All others	1,057,125	8.5	14,972	12.5
Nonpoverty	White	5,074,379	11.6	434,854	11.7
	All other	788,897	6.4	51,989	6.5

(d) Interpret your results.

(e) Why would you caution a reviewer of your analysis about the interpretation?

15.11 In a paper by Foy et al. [1983] the risk of getting *Mycoplasma* pneumonia in a two-year interval was determined on the basis of an extended survey of schoolchildren. Of interest was whether children previously exposed to *Mycoplasma pneumoniae* had a smaller risk of recurrence. In the five- to nine-year age group, the following data were obtained:

	Exposed Previously	**Not Exposed Previously**
Person-years at risk	680	134
Number with *Mycoplasma* pneumonia	7	8

(a) Calculate 95% confidence intervals for the infection rate per 100 person-years for each of the two groups.

(b) Test the significance of the difference between the infection rates.

***(c)** A statistician is asked to calculate the study size needed for a new prospective study between the two groups. He assumes that $\alpha = 0.05$, $\beta = 0.20$, and a two-tailed, two-sample test. He derives the formula

$$\lambda_2 = \sqrt{\lambda_1} - \frac{2.8}{\sqrt{n}}$$

where λ_i is the two-year infection rate for group i and n is the number of persons per group. He used the fact that the square root transformation of a Poisson random variable stabilizes the variance (see Section 10.6). Derive the formula and calculate the infection rate in group 2, λ_2 for $\lambda_1 = 10$ or 6, and sample sizes of 20, 40, 60, 80, and 100.

15.12 In a classic paper dealing with mortality among women first employed before 1930 in the U.S. radium dial–painting industry, Polednak et al. [1978] investigated 21 malignant neoplasms among a cohort of 634 women employed between 1915 and 1929. The five highest mortality rates (observed divided by expected deaths) are listed in Table 15.11.

(a) Test which ratios are significantly different from 1.

(b) Assuming that the causes of death were selected without a particular reason, adjust the observed p-values using an appropriate multiple-comparison procedure.

(c) The painters had contact with the radium through the licking of the radium-coated paintbrush to make a fine point with which to paint the dial. On the basis of this

Table 15.11 Mortality Data for Problem 15.12

Ranked Cause of Death	Observed Number	Expected Number	Ratio
Bone cancer	22	0.27	81.79
Larynx	1	0.09	11.13
Other sites	18	2.51	7.16
Brain and CNS	3	0.97	3.09
Buccal cavity, pharynx	1	0.47	2.15

information, would you have "preselected" certain malignant neoplasms? If so, how would you "adjust" the observed p-value?

15.13 Consider the data in Table 15.12 (from Janerich et al. [1974]) listing the frequency of infants with Simian creases by gender and maternal smoking status.

Table 15.12 Influence of Smoking on Development of Simian Creases

Gender of Infant	Maternal Smoking	Birthweight Interval (lb)			
		<6	6–6.99	7–7.99	≥8
Female	No	2/45	5/156	9/242	11/216
	Yes	4/48	8/107	6/110	3/44
Male	No	5/40	5/109	23/265	18/278
	Yes	10/55	6/84	10/106	6/74

(a) These data can be analyzed by the multidimensional contingency table approach of Chapter 7. However, we can also treat it as a problem in standardization. Describe how indirect standardization can be carried out using the total sample as the reference population, to compare "risk" of Simian creases in smokers and nonsmokers adjusted for birthweight and gender of the infants.

(b) Carry out the indirect standardization procedure and compare the standardized rates for smokers and nonsmokers. State your conclusions.

(c) Carry out the logistic model analysis of Chapter 7.

***15.14** Show that the variance of the standardized mortality ratio, equation (3), is approximately equal to equation (4).

REFERENCES

Annegers, J. F., Elveback, L. R., Labarthe, D. R., and Hauser, W. A. [1976]. Ischemic heart disease in patients with epilepsy. *Epilepsia*, **17**: 11–14.

Bruce, E., Frederick, R., Bruce, R., and Fisher, L. D. [1976]. Comparison of active participants and dropouts in CAPRI cardiopulmonary rehabilitation programs. *American Journal of Cardiology*, **37**: 53–60.

Bunker, J. P., Forest, W. H., Jr., Mosteller, F., and Vandam, L. D. [1969]. *The National Halothane Study: A Study of the Possible Association between Halothane Anesthesia and Postoperative Hepatic Necrosis*. National Institute of Health/National Institute of Several Medical Sciences, Bethesda, MD.

Clark, D. A., Stinson, E. B., Griepp, R. B., Schroeder, J. S., Shumway, N. E., and Harrison, D. C. [1971]. Cardiac transplantation: VI. Prognosis of patients selected for cardiac transplantation. *Annals of Internal Medicine*, **75**: 15–21. Used with permission.

Foy, H. M., Kenny, G. E., Cooney, M. K., Allan, I. D., and van Belle, G. [1983]. Naturally acquired immunity to mycoplasma pneumonia infections. *Journal of Infectious Diseases*, **147**: 967–973. Used with permission from University of Chicago Press.

Friis, R., Nanjundappa, G., Prendergast, J. J., Jr., and Welsh, M. [1981]. Coronary heart disease mortality and risk among hispanics and non-hispanics in Orange County, CA. *Public Health Reports*, **96**: 418–422.

Janerich, D. T., Skalko, R. G., and Porter, I. H. (eds.) [1974]. *Congenital Defects: New Directions in Research*. Academic Press, New York.

Kalbfleisch, J. D., and Prentice, R. L. [2002]. *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, New York.

National Cancer Institute [1975]. *Third National Cancer Survey: Incidence Data*. Monograph 41. DHEW Publication (NIH) 75–787. U.S. Government Printing Office, Washington, DC.

National Center for Health Statistics [1976]. *Selected Vital and Health Statistics in Poverty and Non-poverty Areas of 19 Large Cities: United States, 1969–1971*. Series 21, No. 26. U.S. Government Printing Office, Washington, DC.

Polednak, A. P., Stehney, A. F., and Rowland, R. E. [1978]. Mortality among women first employed before 1930 in the U.S. radium dial-painting industry. *American Journal of Epidemiology*, **107**: 179–195.

CHAPTER 16

Analysis of the Time to an Event: Survival Analysis

16.1 INTRODUCTION

Many biomedical analyses study the time to an event. A cancer study of combination therapy using surgery, radiation, and chemotherapy may examine the time from the onset of therapy until death. A study of coronary artery bypass surgery may analyze the time from surgery until death. In each of these two cases, the event being used is death. Other events are also analyzed. In some cancer studies, the time from successful therapy (i.e., a patient goes into remission) until remission ends is studied. In cardiovascular studies, one may analyze the time to a heart attack or death, whichever event occurs first. A health services project may consider the time from enrollment in a health plan until the first use of the facilities. An analysis of children and their need for dental care may use the time from birth until the first cavity is filled. An assessment of an ointment for contact skin allergies may consider the time from treatment until the rash has cleared up.

In each of the foregoing situations, the data consisted of the time from a fixed or designated initial point until an event occurs. In this chapter we show how to analyze such *event data*. When the event of interest is death, the subject is called *survival analysis*. In medicine and public health this name is often used generically, even when the endpoint or event being studied is not death but something else. In industrial settings the study of the lifetime of a component (until failure) is called *reliability theory*, and social scientists use the term *event history analysis*. For concreteness, we often speak of the event as death and the time as survival time. However, it should always be kept in mind that there are other uses.

In this chapter we consider the presentation of time to event data, estimation of the time to an event, and its statistical variability. We also consider potential predictor or explanatory variables. A third topic is to compare the time to event in several different groups. For example, a study of two alternative modes of cancer therapy may examine which group has the best survival experience.

When the event is not death, there may be multiple occurrences for a given person or multiple types of event. It is usually possible to restrict the analysis to the first event as we did in the situations described above. This restriction trades a considerable gain statistical simplicity for an often modest loss in power. We discuss the analysis of multiple events only briefly.

16.2 SURVIVORSHIP FUNCTION OR SURVIVAL CURVE

In previous chapters we examined means of characterizing the distribution of a variable using, for example, the cumulative distribution function and histograms. One might take survival data

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

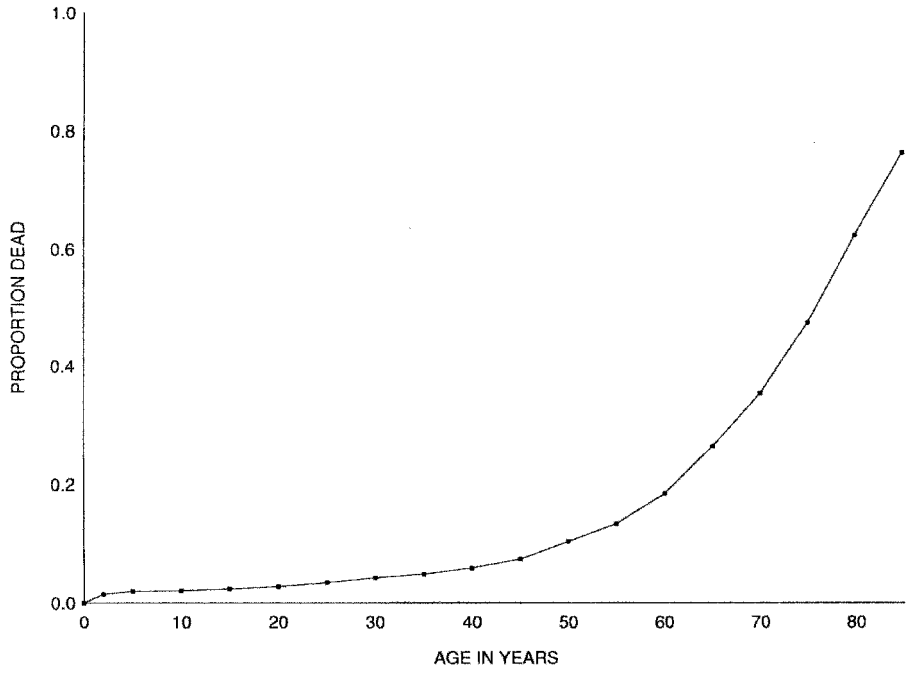


Figure 16.1 Cumulative probability of death, United States, 1974. (From U.S. Department of Health, Education, and Welfare [1976].)

and present the cumulative distribution function. Figure 16.1 shows an estimate for the U.S. population in 1974 of the probability of dying before a fixed age. This is an estimate of the cumulative distribution of survival in the United States in 1974. Note that there is an increase in deaths during the first year; after this the rate levels off but then climbs progressively in the later years. This cumulative probability of death is then an estimate of the probability that a person dies at or before the given time. That is,

$$F(t) = P[\text{person dies at a time} \leq t]$$

If we had observed the entire survival experience of the 1974 population, we would estimate this quantity as we estimated the cumulative distribution function previously. We would estimate it as

$$F(t) = \frac{\text{number of people who die at or before time } t}{\text{total number observed}} \tag{1}$$

Note, however, that we cannot estimate the survival experience of the 1974 population this way because we have not observed all of its members until death. This is a most fortunate circumstance since the population includes all of the authors of this book as well as many of its readers. In the next section, we discuss some methods of estimating survival when one does not observe the true survival of the entire population.

It is depressing to speak of death; it is more pleasant to speak of life. In analyzing survival data, the custom has grown not of using the cumulative probability of death but of using an equivalent function called the *survivorship function* or *survival curve*. This function is merely the percent of people who live to a fixed time or beyond.

Definition 16.1. The *survival curve*, or *survivorship function*, is the proportion or percent of people living to a fixed time t or beyond. The curve is then a function of t:

$$S(t) = \begin{cases} \text{percent of people surviving to time } t \text{ or beyond if} \\ \quad \text{expressed as a percent} \\ \text{proportion of people surviving to time } t \text{ or beyond} \\ \quad \text{if expressed as a proportion} \end{cases} \tag{2}$$

If we have a sample from a population, there is a distinction between the population survival curve and the sample or estimated population survival curve. In practice, there is no distinct notation unless it is necessary to emphasize the difference. The context will usually show which of the two is meant.

The cumulative distribution function of the survival and the survival curve are closely related. If the two curves are continuous, they are related by

$$S(t) = 100[1 - F(t)] \quad \text{or} \quad S(t) = 1 - F(t)$$

(When we look at the sample curves, the curves are equal at all points except for the points where the curves jump. At these points there is a slight technical problem because we have used $\leq$ in one instance and $\geq$ in the other instance. But for all practical purposes, the two curves are related by the equation above.)

Figure 16.2 shows the survival curve for the U.S. population as given in Figure 16.1. As you can see, the survival curve results by "flipping over" the cumulative probability of death and using percentages. As mentioned above, the estimate of the curve in Figure 16.2 is complicated by the fact that many people in the 1974 U.S. population are happily alive. Thus, their true

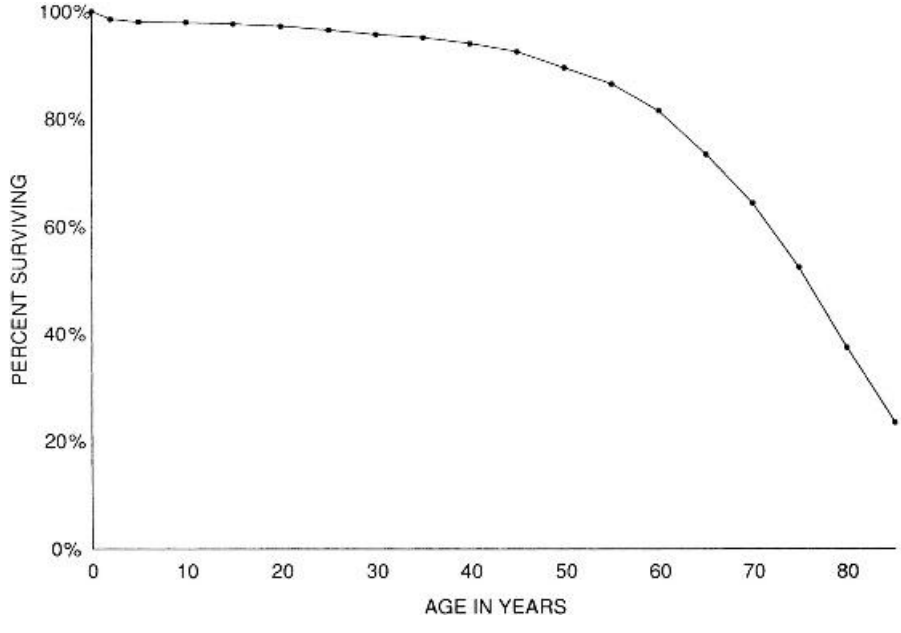


Figure 16.2 Survival curve of the U.S. population, 1974. Same data as used in Figure 16.1.

survival is not yet observed. The survival in the overall population is not yet observed. The survival in the overall population is estimated by the method discussed in the next section.

Sometimes the *proportion* surviving to time t or beyond is used. We will use them interchangeably. The two are simply related; to find the percent, merely multiply the proportion by 100.

If we observe the survival of all persons, it is easy to estimate the survival curve. In analogy with the estimate of the cumulative distribution function, the estimate of the survival curve at a fixed t is merely the percent of people whose survival was equal to the value t or greater. That is,

$$S(t) = 100\left(\frac{\text{number of people who survive to or beyond } t}{\text{total number observed}}\right) \tag{3}$$

In many instances, we are not able to observe everyone until they reach the event of interest. This makes the estimation problem more challenging. We discuss the estimates in the next section.

16.3 ESTIMATION OF THE SURVIVAL CURVE: ACTUARIAL OR LIFE TABLE METHOD

Consider a clinical study of a procedure with a high initial mortality rate: for example, very delicate high-risk surgery during its development period. Suppose that we design a study to follow a group of such people for two years. Because most of the mortality is expected during the first year, it is decided to concentrate the effort on the first year. Two thousand people are to be entered in the study; half of them will be followed for two years, while one-half will be followed only for the critical first year. The people are randomized into two groups, group 1 to be followed for one year and group 2 to be followed for both years. Suppose that the data are as follows:

	Group 1		Group 2	
Year	Number Observed	Number Who Died	Number Observed	Number Who Died
1	1000	240	1000	200
2	—	—	800	16

We wish to estimate one- and two-year survival. We consider three methods of estimation. The first two methods will not be appropriate but are used to motivate the correct life table method to follow.

One way of estimating survival might be to estimate separately the one- and two-year survival. Since it is wasteful to "throw away" data and the reason that 2000 people were observed for one year was because that year was considered crucial, it is natural to estimate the percent surviving for one year by the total population. This percentage is as follows:

$$\text{percent of one-year survival} = 100\left(\frac{2000 - 240 - 200}{2000}\right) = 78.0\%$$

To estimate two-year survival, we did not observe what happened to the subjects in group 1 during the second year. Thus, we might estimate the survival using only those in group 2. This

estimate is

$$\text{percent of two-year survival} = 100\left(\frac{1000-200-16}{1000}\right) = 78.4\%$$

There are two problems with this estimation method. The first is that we need to know the potential follow-up time (one year or two years) for everyone. In a clinical trial this is reasonable, but in a cohort study we may not know whether someone who in fact died after six months would have been followed up for one year or two years if he or she had not died. Nor is it reasonable that our estimate of the survival should depend on this unobservable potential follow-up.

More importantly, we have a problem in that the estimated percent surviving one year is less than the percent surviving two years! Clearly, as time increases, the percent surviving must decrease, but the sampling variability in the estimate has led to the second-year estimate being larger than the first-year estimate. Although this method is approximately unbiased and uses all the available data, it is not a desirable way to estimate our survival curve.

One way to get around this problem is to use only the subjects from group 2 who are observed for two years. Then we have a straightforward estimate of survival at each time period. The percent surviving one year or more is 80%, while the percent surviving two or more years is, as before, 78.4%. This gives a consistent pattern of survival but seems quite wasteful; we deliberately designed the study to allow us to observe more subjects in the first year, when the mortality was expected to be high. It does not seem appropriate to throw away the 1000 subjects who were only observed for one year. If we need to do this, we had an extremely poor experimental design.

The solution to our problem is to note that we can efficiently estimate the probability of one-year survival using both groups of people. Further, using the second group, we can estimate the probability of surviving the second year *conditionally upon having survived the first year*. The two estimates as percentages are

$$\text{percent of one-year survival} = 78.0\%$$

$$\text{percent surviving year 2} = 100\left(\frac{800-16}{800}\right) = 98.0\%$$

We can then combine these to get an estimate of the probability of surviving in the first year and the second year by using the concept of conditional probability. We see that the probability of two-year survival is the probability of one-year survival times the probability of two-year survival given one-year survival, and so cannot be larger than the probability of one-year survival. The probability of two-year survival is as follows:

$$P[A \text{ and } B] = P[A]P[B|A]$$

Let A be the survival of one year and B the survival of two years. Then

$$\begin{aligned} P[\text{one-year survival}] &= P[\text{one-year survival}] \\ &\quad \times P[\text{two-year survival}|\text{one-year survival}] \\ &= 0.78 \times 0.98 = 0.7644 \end{aligned}$$

For these probability calculations, note that it is more convenient to have probabilities than percents because the probabilities multiply. If we had percents, the formula would have an extra factor of 100. For this reason the calculations on the survival curves are usually done as probabilities and then switched to percentages for graphical presentation. We will adhere to this.

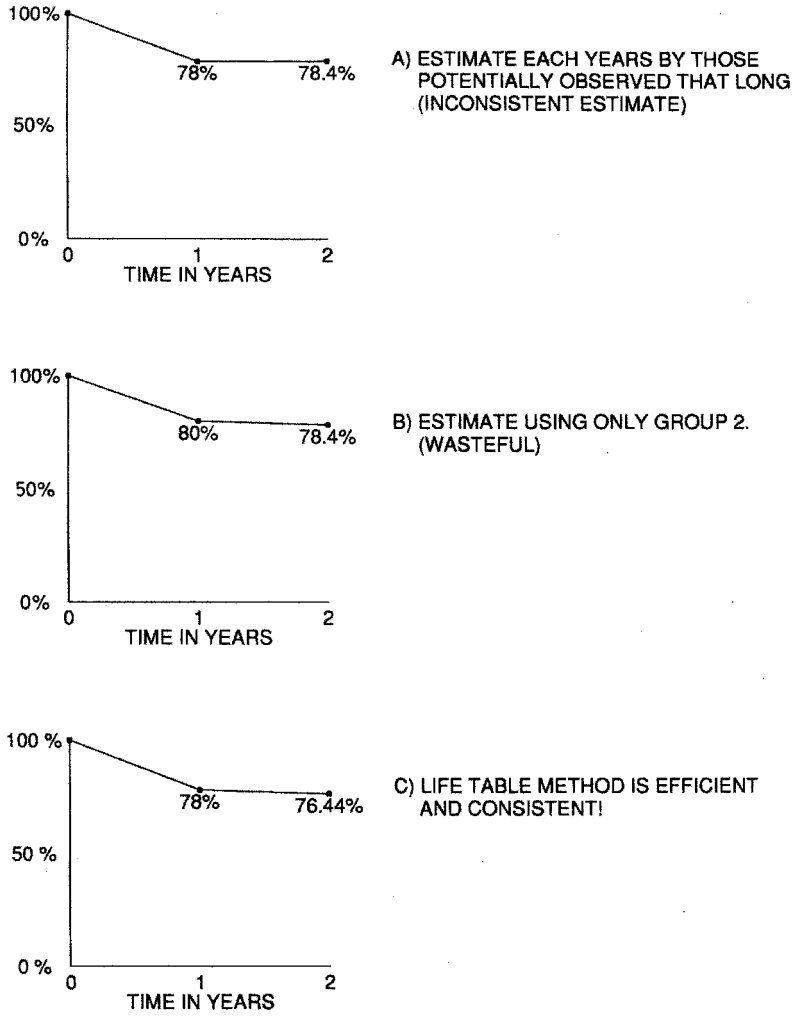


Figure 16.3 Three methods of estimating survival.

Figure 16.3 presents the three estimates; for these data they are all close. The third estimate gives a self-consistent estimate of the curve (i.e., the curve will never increase) and the estimate is efficient (because it uses all the data); it is the correct method for estimating survival. This idea can easily be generalized to more than two intervals.

When the data are grouped into time intervals, we can estimate the survival in each interval. Let x denote the lower endpoint of each interval. [x rather than t is used here to conform to standard notation in the actuarial field. When it is necessary to index the intervals, we will use $i(x)$ to denote the inverse relationship.] Let $\prod_i$ denote the probability of surviving to $x(i)$, where $x(i)$ is the lower endpoint of the ith interval; that is,

$$\prod_i = S(x(i))$$

where S is the survival curve (expressed here as the proportion surviving). Further, let π_i be the probability of living through the interval, with lower endpoint $x(i)$, conditionally upon the event

of being alive at the beginning of the interval. Using the definition of a conditional probability,

$$\pi_i = \frac{\prod_{i+1}}{\prod_i} = \frac{P[\text{survive to the end of the } i\text{th interval}]}{P[\text{survive to the end of the } (i-1)\text{st interval}]} \tag{4}$$

From this,

$$\prod_{i+1} = \pi_i \prod_i$$

and

$$\prod_{i+1} = \pi_1 \pi_2 \cdots \pi_i \qquad \text{where} \quad \prod_1 = 1 \tag{5}$$

In presenting group data graphically, one plots points corresponding to the time of the lower endpoint of the interval and the corresponding $\prod_i$ value. The plotted points are then joined by straight-line segments, as in Figure 16.4.

There is one further complication before we present the life table estimates. If we are following people periodically (e.g., every six months or every year), it will occasionally happen that people cannot be located. Such subjects are called *lost to follow-up* in the study. Further, subjects may be withdrawn from the study for a variety of reasons. In clinical studies in the United States, all subjects have the right to withdraw from participation at any time. Or we might be trying to examine a medical survival in patients who could potentially be treated with surgery. Some of them may subsequently receive surgery; we could withdraw such patients from the analysis at the time they received surgery. The rationale for this would be that after they received surgery, their survival experience is potentially altered. Whatever the reason for a person being lost to follow-up or withdrawn, this fact must be considered in the life table analysis.

To estimate the survival curve from data, the method is to estimate the π_i and $\prod_i$ by the product of the estimates of the π_i according to equation (5). The data are usually presented in the form of Table 16.1. How might one estimate the probability of dying in the interval whose

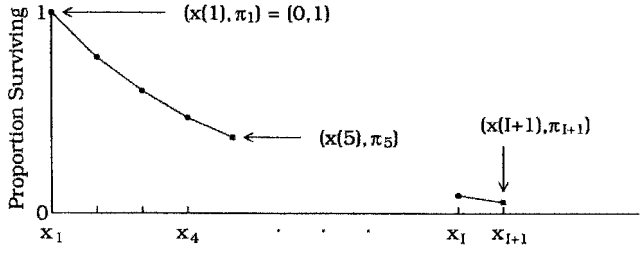


Figure 16.4 Form of the presentation of the survival curve for grouped survival data.

Table 16.1 Presentation of Life Table Data

Interval	Number of Subjects Observed Alive at Beginning of Interval	Died during Interval	Lost to Follow-up during Interval	Withdrawn Alive during Interval
x to $x + \Delta x$	l_x	d_x	u_x	w_x
$x(1) - x(2)$	$l_{x(1)}$	$d_{x(1)}$	$u_{x(1)}$	$w_{x(1)}$
$x(2) - x(3)$	$l_{x(2)}$	$d_{x(2)}$	$u_{x(2)}$	$w_{x(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x(I) - x(I+1)$	$l_{x(I)}$	$d_{x(I)}$	$u_{x(I)}$	$w_{x(I)}$

lower endpoint is x conditionally upon being alive at the beginning of the interval? At first glance one might reason that there were l_x subjects, of whom (a binomial) d_x died, so that the estimate should be d_x/l_x. The problem is that those who were lost to follow-up or withdrew during the interval might have died during the interval *after* withdrawing, and this would not be counted. If such persons were equally likely to withdraw at any time during the interval, on the average they would be observed only one-half of the time. Thus, they really represent only one-half a person at risk. Thus the effective number of persons at risk, l'_x, is

$$\begin{aligned} l'_x &= \underbrace{l_x - (u_x + w_x)}_{\substack{\text{number observed} \\ \text{over entire interval}}} + \underbrace{\tfrac{1}{2}(u_x + w_x)}_{\substack{\text{number observed} \\ \text{over } \frac{1}{2} \text{ interval}}} \\ &= l_x - \tfrac{1}{2}(u_x + w_x) \end{aligned} \tag{6}$$

where the u_x is the number lost to follow-up and w_x is the number withdrawing. The estimate of the proportion dying, q_x, is thus

$$q_x = \frac{d_x}{l'_x}$$

The estimate of π_i, the probability of surviving the interval $x(i)$ to $x(i+1)$, is

$$p_{x(i)} = 1 - q_{x(i)}$$

Finally, the estimate of $\prod_i = \pi_1 \pi_2 \cdots \pi_{i-1}$, $\prod_1 = 1$ is

$$P_{x(i)} = p_{x(1)} p_{x(2)} \cdots p_{x(i-1)}, \quad P_{x(0)} = 1 \tag{7}$$

Note that those who are lost to follow-up and those who are withdrawn alive are treated together; that is, in the estimates, only the sum of the two is used. In many presentations such people are lumped together as *withdrawn* or *censored*.

Before presenting the estimates, it is also clear that an estimate of the survival curve will be more useful if some idea of its variability is given.

An estimate of the standard error of the P_x is given by Greenwood's formula [Greenwood, 1926]:

$$\begin{aligned} \mathrm{SE}(P_{x(i)}) &= P_{x(i)} \sqrt{\sum_{j=1}^{i-1} \frac{q_{x(j)}}{l'_{x(j)} - d_{x(j)}}} \\ &= P_{x(i)} \sqrt{\sum_{j=1}^{i-1} \frac{q_{x(j)}}{l'_{x(j)} p_{x(j)}}} \end{aligned} \tag{8}$$

Confidence intervals constructed using ± 1.96 times this standard error are valid only in relatively large samples. For example, it is easy to see that these confidence intervals could extend outside the interval [0, 1], where the probability must lie. Better confidence intervals in small samples can be obtained by transforming $P(t)$; they are discussed in the Notes to this chapter.

Example 16.1. The method is illustrated by data of Parker et al. [1946], as discussed in Gehan [1969]. Those data are from 2418 males with a diagnosis of angina pectoris (chest pain thought to be of cardiac origin) at the Mayo Clinic between January 1, 1927 and December 31, 1936. The life table of survival time from diagnosis (in yearly intervals) is shown in Table 16.2.

Table 16.2 Life Table Analysis of 2418 Males with Angina Pectoris

x to $x + \Delta x$ (yr)	l_x	d_x	u_x	w_x	l'_x	q_x	p_x	P_x	SE(P_x)
0–1	2418	456	0	0	2418	0.1886	0.8114	1.0000	—
1–2	1962	226	39	0	1942.5	0.1163	0.8837	0.8114	0.0080
2–3	1697	152	22	0	1686.0	0.0902	0.9098	0.7170	0.0092
3–4	1523	171	23	0	1511.5	0.1131	0.8869	0.6524	0.0097
4–5	1329	135	24	0	1317.0	0.1025	0.8975	0.5786	0.0101
5–6	1170	125	107	0	1116.5	0.1120	0.8880	0.5139	0.0103
6–7	938	83	133	0	871.5	0.0952	0.9048	0.4611	0.0104
7–8	722	74	102	0	671.0	0.1103	0.8897	0.4172	0.0105
8–9	546	51	68	0	512.0	0.0996	0.9004	0.3712	0.0106
9–10	427	42	64	0	395.0	0.1063	0.8937	0.3342	0.0107
10–11	321	43	45	0	298.5	0.1441	0.8559	0.2987	0.0109
11–12	233	34	53	0	206.5	0.1646	0.8354	0.2557	0.0111
12–13	146	18	33	0	129.5	0.1390	0.8610	0.2136	0.0114
13–14	95	9	27	0	81.5	0.1104	0.8896	0.1839	0.0118
14–15	59	6	23	0	47.5	0.1263	0.8737	0.1636	0.0123

Source: Data from Gehan [1969].

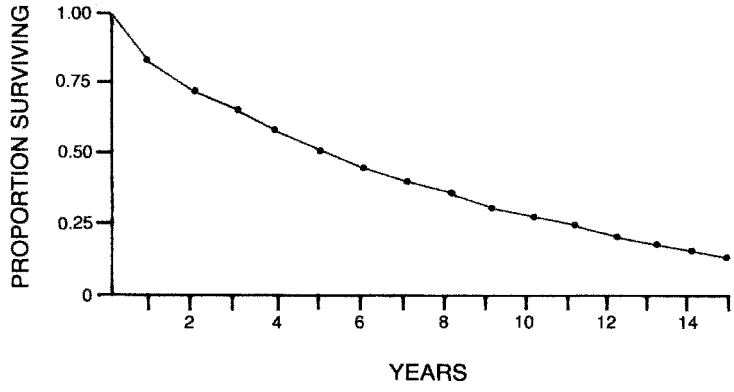


Figure 16.5 Survivorship function. (Data from Gehan [1969]; see Table 16.2.)

The survival data are given graphically in Figure 16.5. Note that in this case the proportion rather than the percent is presented.

As a second example, we consider patients with the same diagnosis, angina pectoris; these data are more recent.

Example 16.2. Passamani et al. [1982] studied patients with chest pain who were studied for possible coronary artery disease. Chest pain upon exertion is often associated with coronary artery disease. The chest pain was evaluated by a physician as definitely angina, probably angina, probably not angina, and definitely not angina. The definitions of these four classes were:

- *Definitely angina:* a substantial discomfort that is precipitated by exertion, relieved by rest and/or nitroglycerin in less than 10 minutes, and has a typical radiation to either shoulder, jaw, or the inner aspect of the arm. At times, definite angina may be isolated to the shoulder, jaw, arm, or upper abdomen.

- *Probably angina:* has most of the features of definite angina but may not be entirely typical in some aspects.
- *Probably not angina:* an atypical overall pattern of chest pain symptoms which does not fit the description of definite angina.
- *Definitely not angina:* a pattern of chest pain symptoms that are unrelated to activity, unrelieved by nitroglycerin and/or rest, and appear clearly noncardiac in origin.

The data are plotted in Figure 16.6. Note how much improved the survival of the angina patients (definite and probable) is compared with the Mayo data of Figure 16.5. Those data had a 52% five-year survival. These data have 91% and 85% five-year survival! This indicates the great difficulty of using historical control data. A statistic and p-value for testing differences among the four groups is discussed in Section 16.6.

Table 16.3 gives the calculation using 91-day intervals and four intervals to approximate a year for one of the four groups, the definite angina patients. As a sample calculation, consider the interval from 637 to 728 days. We see that

$$
\begin{aligned}
l_x &= 2704, \qquad u_x + d_x = 281 \\
l'_x &= 2704 - \frac{281}{2} = 2563.5 \\
q_x &= \frac{12}{2563.5} = 0.0047 \\
p_x &= 1 - 0.0047 = 0.9953 \\
P_x &= 0.9350 \times 0.9953 = 0.9306
\end{aligned}
$$

Note that the definite angina cases have the worst survival, followed by the probable angina cases (91%). The other two categories are almost indistinguishable.

As we have seen, in the life table method we have some data for which the event in question is not observed, often because at the time of the end of data collection and analysis, patients are still alive. One term used for such data is *censoring*, a term that brings to mind a powerful, possibly sinister figure throwing away data to mislead one in the data analysis. In this context

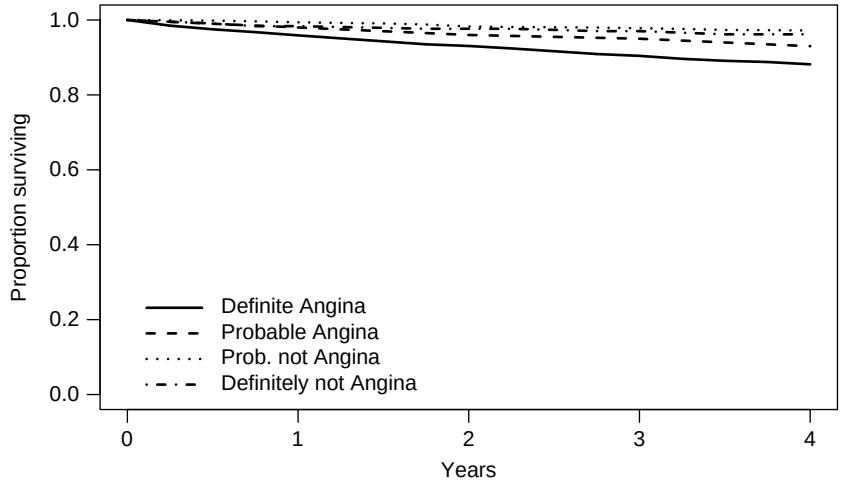


Figure 16.6 Survival by classification of chest pain. (Data from Passamani et al. [1982].)

Table 16.3 Life Table for Definite Angina Patients. Time in Days

$t(i)$	Enter	At Risk	Dead	Withdrawn Alive	Proportion Dead	Cumulative Survival of the End of Interval	SE	Effective Sample Size
0.0–90.9	2894	2894.0	44	0	0.0152	0.9848	0.002	2893.99
91.0–181.9	2850	2850.0	28	0	0.0098	0.9751	0.003	2893.99
182.0–272.9	2822	2822.0	22	0	0.0078	0.9675	0.003	2894.00
273.0–363.9	2800	2799.0	25	2	0.0089	0.9589	0.004	2893.77
364.0–454.9	2773	2773.0	23	0	0.0083	0.9509	0.004	2893.46
455.0–545.9	2750	2750.0	23	0	0.0084	0.9430	0.004	2893.23
546.0–636.9	2727	2727.0	23	0	0.0084	0.9350	0.005	2893.06
637.0–727.9	2704	2563.5	12	281	0.0047	0.9306	0.005	2882.32
728.0–818.9	2411	2394.0	17	34	0.0071	0.9240	0.005	2850.22
819.0–909.9	2360	2359.0	19	2	0.0081	0.9166	0.005	2818.52
910.0–1000.9	2339	2336.5	19	5	0.0081	0.9091	0.005	2792.12
1001.0–1091.9	2315	2035.5	11	559	0.0054	0.9042	0.006	2753.73
1092.0–1182.9	1745	1722.5	15	45	0.0087	0.8963	0.006	2654.36
1183.0–1273.9	1685	1685.0	19	0	0.0059	0.8910	0.006	2596.11
1274.0–1364.9	1675	1670.5	6	9	0.0036	0.8878	0.006	2564.52
1365.0–1455.9	1660	1274.5	9	771	0.0071	0.8816	0.007	2449.65

it refers to the fact that although one is interested in survival times, the actual survival times are not observed for all the subjects. We have seen several sources of censored data. Subjects may be alive at the time of analysis; (subjects) may be lost to follow-up; (subjects) may refuse to participate further in research; or (subjects) may undergo a different therapy which removes them from estimates of the survival in a particular therapeutic group.

The *life table* or *actuarial method* that we have used above has the strength of allowing censored data and also uses the data with maximum efficiency. There is an important underlying assumption if we are to get unbiased estimates of the survival in a population from which such subjects may be considered to come. *It is necessary that the withdrawal or censoring not be associated with the endpoint*. Obviously, if everyone is withdrawn because their situation deteriorates, one would expect a bias in the estimation of death. Let us emphasize this again. The life table estimate gives *biased* estimates if subjects who are censored at a given time have higher or lower chance of failure than those not censored at that time. The assumption we need is technically called *noninformative censoring*; the term *independent censoring* is also used.

We return later in the chapter to the related but distinct problem of competing causes of death, for example, examining the differences in death from cardiovascular causes in an elderly population where many people die of cancer or infectious disease during the study.

16.4 HAZARD FUNCTION OR FORCE OF MORTALITY

In the analysis of survival data, one is often interested in examining which periods have the highest or lowest risk of death. By risk of death, one has in mind the risk or probability among those alive at that time. For example, in very old age there is a high risk of dying each year *among* those reaching that age. The probability of any person dying, say, in the 100th year is small because so few people live to be 100 years old.

This concept is made rigorous by the idea of the hazard function or *hazard rate*. (A very precise definition of the hazard function requires ideas beyond the scope of this book and is discussed briefly in the Notes at the end of this chapter.) The hazard function is also called the *force of mortality*, *age-specific death rate, conditional failure rate*, and *instantaneous death rate*.

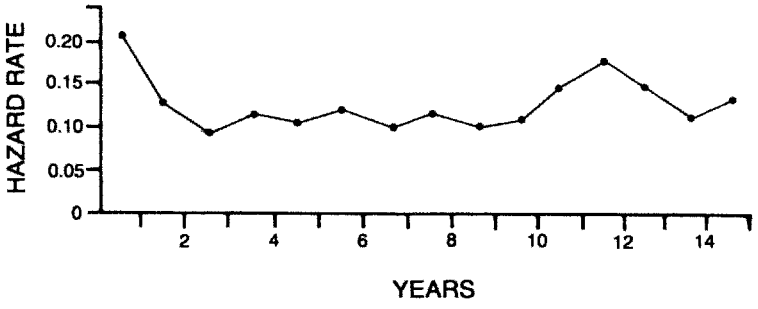


Figure 16.7 Hazard function for Example 16.1. (Data from Parker et al. [1946].)

Definition 16.2. In a life table situation, the (*interval or actuarial*) *hazard rate* is the expected number dying in the interval, divided by the product of the average number exposed in the interval and the interval width.

In other words, the hazard rate, λ, is the probability of dying per unit time given survival to the time point in question. The estimate h of the hazard function is given by

$$h_x = \frac{d_x}{l'_x - d_x/2} \frac{1}{\Delta x} \tag{9}$$

where $\Delta x(1) = x(i+1) - x(i)$, the interval width. This is an estimate of the form

$$\frac{\text{number dying}}{\text{total exposure time}}$$

l'_x is an estimate of the number at risk of death. Note that this estimate is analogous to the definition in Section 15.4. Those who die will on average have been exposed for approximately one-half of the time interval, so the number of intervals of observed time is approximately $(l'_x - d_x/2)\Delta x$. Thus, the hazard rate is a death rate; its units are proportion per unit time (e.g., percent per year). If the hazard rate has a constant value λ over time, the survival is exponential, that is, $S(t) = 100e^{-\lambda t}$, a point returned to later. The estimated hazard rate for Parker's data of Example 16.1 is given in Figure 16.7.

A large-sample approximation from Gehan [1969] for the SE of h is

$$\text{SE}(h_x) = \left\{ \frac{h_x^3}{l_x q_x} \left[1 - \left(\frac{h_x \Delta x}{2} \right)^2 \right] \right\}^{1/2} \tag{10}$$

For the data of Example 16.1, we compute the hazard function for the second interval. We find that

$$h_1 = \left(\frac{226}{1942.5 - 226/2} \right) \left(\frac{1}{1} \right) = 0.124$$

16.5 PRODUCT LIMIT OR KAPLAN–MEIER ESTIMATE OF THE SURVIVAL CURVE

If survival data are recorded in great detail, accuracy is preserved by placing the data into smaller rather than larger intervals. Obviously, if data are grouped, for example, into five-year

intervals while the time of death is recorded to the nearest day, considerable detail is lost. The *product limit* or *Kaplan–Meier estimate* is based on the idea of taking more and more intervals. In the limit, the intervals become arbitrarily small.

Suppose in the following that the time at which data are censored (lost to follow-up or withdrawn from the study) and the time of death (when observed) are measured to a high degree of accuracy. The product limit *or* Kaplan–Meier (see Kaplan and Meier [1958]) estimate (KM estimate) results from the actuarial or life table method of Section 16.4 as the number of intervals increases in such a way that the maximum interval width approaches zero. In this case it can be seen that the estimated survival curve is constant except for jumps at the observed times of death. The values of the survival probability before a time of death(s) is multiplied by the estimated probability of surviving past the time of death to find the new value of the survival curve.

To be more precise, suppose that n persons are observed. Further, suppose that the time of death is observed in l of the subjects at k distinct times $t_1 < t_2 < \cdots < t_k$. Let m_i be the number of deaths at time t_i. The other $n - l$ subjects are censored observations. If a censoring time and a death occur at the same time, it is assumed that the true time of death for the censored subject is greater than the censoring time observed. Let n_i be the number of subjects at risk of dying at time t_i. That is, $n_i = n$ minus the number of deaths prior to t_i and minus the number of subjects whose observations were censored prior to time t_i. The product limit estimate of the survival curve expressed as a proportion is

$$S(t) = \begin{cases} 1 & \text{for } t < t_1 \\ \prod_{j=1}^{i} \dfrac{n_i - m_i}{n_i}, & t_i \le t < t_{i+1} (i < k) \\ 0 & \text{for } t_k \le t \quad \text{if } m_k = n_k \text{ (i.e., no one survives past time } t_k) \\ \prod_{j=1}^{k} \dfrac{n_i - m_i}{n_i} & \text{for } t_k \le t \le \text{ largest observed censored observation} \end{cases} \tag{11}$$

If $m_k < n_k$, then $S(t)$ is undefined for $t >$ largest observed censored observation. Some software will report either $S(t) = 0$ or $S(t) = S(t_k)$ for times after the last censored observation, but this should not be encouraged.

We illustrate the method with an example.

Example 16.3. We again use the Stanford heart transplant data discussed in Section 15.4. Suppose that we wished to estimate the survival of these patients given medical treatment only. A complication is that when a donor heart becomes available, the patient has a heart transplant; we can no longer observe what the survival without a transplant would have been. One *incorrect* way to analyze such data would be the following. Since we are interested in medical survival, we should not worry about patients who have had surgery. We should go through the records and look at the survival curves only for patients who did not have surgery. Since by definition such people died awaiting the donor heart, their early survival experience would be quite poor.

At the time of the Stanford study, waiting lists were short and a donor heart was transplanted to the best-matching recipient on the waiting list [Crowley and Hu, 1977]. Thus, we may use surgery for heart transplantation as a source of censoring for medical survival: The availability of a heart should not be related to the severity of illness of the recipient. Current practice is quite different; more seriously ill patients are more likely to receive a transplant (*http://www.optn.org/*), so the censoring by surgery would be *informative* (biased) in a modern study.

Table 16.4 presents the medical survival data using surgery as the source of censoring for the Stanford heart transplant patients. The computations as described above are given. The product limit estimate of the correct survival curve is shown by solid lines in Figure 16.8. Lines with x's is the incorrect curve if one ignores the effect of surgery as censoring and totally eliminates

Table 16.4 Survival Data for Heart Transplant Patients

t (days)	Death (*)	n_i	$(n_i - m_i)/n_i$	$S(t),\ t_i \le t < t_{i+1}$
1	*	34	33/34	0.971
1		33		
2		32		
5	*	31	30/31	0.939
7	*	30	29/30	0.908
7		29		
11		28		
11		27		
12	*	26	25/26	0.873
15	*	25	24/25	0.838
15		24		
16		23		
17	*	22	21/22	0.800
17		21		
17		20		
19		19		
22		18		
24		17		
24		16		
26		15		
34	*	14	13/14	0.743
34		13		
35	*	12	11/12	0.681
36	*	11	10/11	0.619
36		10		
40	*	9	8/9	0.550
49	*	8	7/8	0.481
49		7		
50		6		
69		5		
81		4		
84	*	3	2/3	0.321
111	*	2	1/2	0.160
480		1		

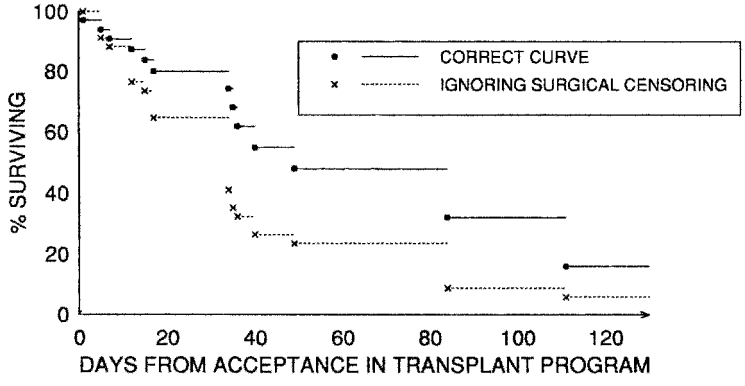


Figure 16.8 Days from acceptance in transplant program. Kaplan–Meier survival curve.

such subjects from the analysis. Finally, note that there was one patient who spontaneously improved under medical treatment and was reported alive at 16 months. The data of that subject are reported in the medical survival data as a 480-day survivor. As before, an asymptotic formula for the standard error of the estimate may be given. Greenwood's formula for the approximate standard error of the estimate also holds in this case. The form it takes is

$$\mathrm{SE}(S(t)) \doteq S(t)\sqrt{\sum_{j=1}^{i}\frac{m_j}{n_j(n_j-m_j)}} \qquad \text{for } t_i \le t < t_{i+1} \tag{12}$$

16.6 COMPARISON OF DIFFERENT SURVIVAL CURVES: LOG-RANK TEST

In this section we consider a test statistic for comparing two or more survival curves for different groups of subjects. This statistic is based on the following idea. Take a particular interval in which deaths occur, or in the case of the product limit curve, a time when one or more deaths occur. Suppose that the first group considered has one-third of the subjects being observed. How many deaths would we expect in the first group if, in fact, the survival experience is the same for all the groups? We expect the number of deaths to be proportional to the fraction of the people at risk of dying in the group. That is, for the first group the expected number of deaths would be the observed number of deaths at that time divided by 3. The log-rank test uses this simple fact. At each interval or time of death we take the observed number of deaths and calculate the expected number of deaths that would occur in each of the groups if all had the same risk of dying. For each group, the expected number of deaths is summed over all intervals and then compared to the observed number of deaths. Using this comparison, we get a statistic, the *log-rank statistic*, which has approximately a chi-square distribution with $k-1$ degrees of freedom when k groups are observed. We formalize this.

Suppose that one is interested in comparing the survival experience of k populations. Suppose that there are M different times at which deaths appear. For the life table method, this will usually be each interval. In the product limit approach, each death observed will be associated with a unique time. At the mth time, let d_{im} be the number of deaths observed in the ith population and l_{im} be the number at risk of dying. (For the life table approach with withdrawals, l_{ij} is the appropriate l'_x.) The data may be presented in M $2 \times k$ contingency tables with totals:

1	2	$\cdots$	k		
d_{1m}	d_{2m}	$\cdots$	d_{km}	D_m	dying
$l_{1m}-d_{1m}$	$l_{2m}-d_{2m}$	$\ldots$	$l_{km}-d_{km}$	A_m	alive
l_{1m}	l_{2m}	$\cdots$	l_{km}	T_m	total

$$m = 1, 2, \ldots, M$$

If all of the k populations are at equal risk of death, the probability of death will be the same in each population, and conditionally upon the row and column totals,

$$E(d_{im}) = \frac{l_{im}D_m}{T_m} \tag{13}$$

as in the chi-square test for contingency tables.

In the ith population, the total number of deaths observed is

$$O_i = \sum_{m=1}^{M} d_{im} \tag{14}$$

Examining all of the times of death, the expected number of deaths in the ith population is

$$E_i = \sum_{m=1}^{M} E(d_{im}) = \sum_{m=1}^{M} \frac{l_{im} D_m}{T_m} \tag{15}$$

The test statistic is then computed from the observed minus expected values. A simple approximate statistic suitable for hand calculation is

$$X^2 = \sum_{i=1}^{k} (O_i - E_i)^2 \Big/ E_i \tag{16}$$

The statistic is written in the familiar form of the chi-square test for comparing observed and expected values. [If any $E_i = 0$, define $(O_i - E_i)^2/E_i = 0$.] Under the null hypothesis of equal survival curves in the k groups this statistic will have approximately a chi-square distribution with $k-1$ degrees of freedom. The approximation is good when the subjects at risk are distributed over the k groups in roughly the same proportions at all times. The complete formulas for the log-rank test, which is implemented in most major statistics packages, are given in Note 16.3.

The log-rank test is illustrated by using the data of the Stanford transplant patients (Table 16.4) and comparing them with the data of Houston heart transplant patients, as reported in Messmer et al. [1969]. The time of survival for 15 Houston patients is read from Figure 16.9 and therefore has some inaccuracy.

Ordering both the Stanford and Houston transplant patients by their survival time after transplantation and status (dead or alive) gives Table 16.5. The dashes for the d_{im} values indicate where withdrawals occur, and those lines could have been omitted in the calculation. One stops when there are no future deaths at a time when members of both populations are present.

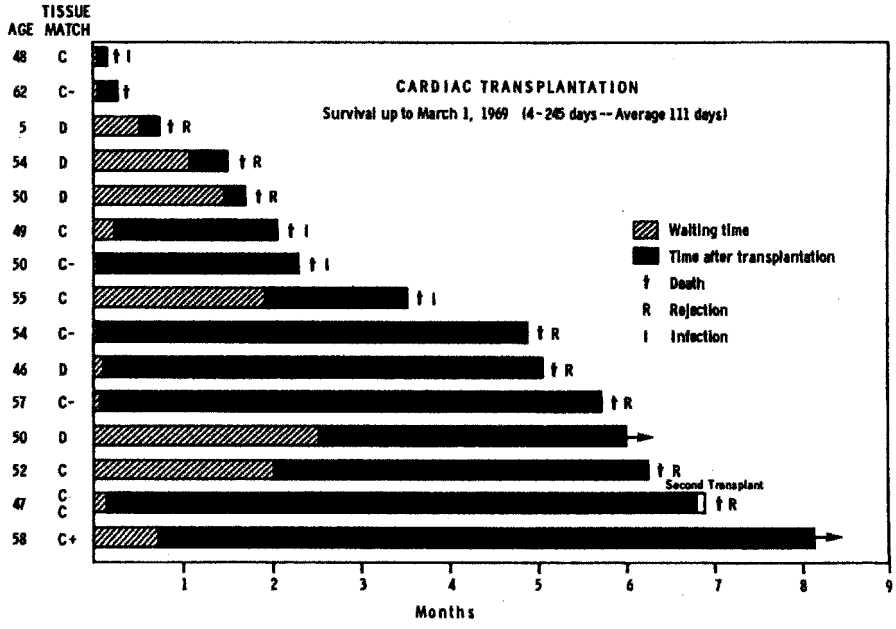


Figure 16.9 Survival of 15 patients given a cardiac allograft. Arrows indicate patients still alive on March 1, 1969. (Data from Messmer et al. [1969].)

Table 16.5 Stanford and Houston Survival Data

	Stanford		Houston			
Day	l_{1m}	d_{1m}	l_{2m}	d_{2m}	$E(d_{1m})$	$E(d_{2m})$
1	20	1	15	0	0.571	0.429
3	19	1	15	0	0.559	0.441
4	18	0	15	1	0.545	0.455
6	18	0	14	2	1.125	0.875
7	18	0	12	1	0.600	0.400
10	18	1	11	0	0.621	0.379
12	17	0	11	1	0.607	0.393
15	17	1	10	0	0.630	0.370
24	16	1	10	0	0.615	0.385
39	15	1	10	0	0.600	0.400
46	14	1	10	0	0.583	0.417
48	13	0	10	1	0.565	0.435
54	13	0	9	1	0.591	0.409
60	13	1	8	0	0.619	0.381
61	12	1	8	1	1.200	0.800
102	11	0	7	0	—	—
104	10	0	6	0	—	—
110	10	0	6	1	0.625	0.375
118	10	0	5	0	—	—
127	9	1	5	0	0.643	0.357
136	8	1	5	0	0.615	0.385
146	7	0	5	1	0.583	0.417
148	7	0	4	1	0.636	0.364
169	7	0	3	1	0.700	0.300
200	7	0	2	1	0.778	0.222

Summing the appropriate columns, one finds that

$$O_1 = \sum_m d_{1m} = 11$$

$$E_1 = \sum_m E(d_{1m}) = 14.611$$

$$O_2 = \sum_m d_{2m} = 13$$

$$E_2 = \sum_m E(d_{2m}) = 9.389$$

The log-rank statistic is 2.32. The simple, less powerful approximation is $X^2 = (11 - 14.611)^2/14.611 + (13 - 9.389)^2/9.389 = 2.28$. Looking at the critical values of the chi-square distribution with one degree of freedom, there is not a statistically significant difference in the survival experience of the two populations.

Another approach is to look at the difference between survival curves at a fixed time point. Using either the life table or Kaplan–Meier product limit estimate at a fixed time T_o, one can estimate the probability of survival to T_o, say, $S(T_o)$ and the standard error of $S(T_o)$, $\mathrm{SE}(S(T_o))$, as described in the sections above. Suppose that a subscript is used on S to denote estimates for different populations. To compare the survival experience of two populations with regard to

surviving to T_o, the following statistic is $N(0, 1)$, as the sample sizes become large [when the null hypothesis of $S_1(T_o) = S_2(T_o)$ is valid]:

$$Z = \frac{S_1(T_o) - S_2(T_o)}{\sqrt{\text{SE}(S_1(T_o))^2 + \text{SE}(S_2(T_o))^2}} \tag{17}$$

A one- or two-sided test may be performed, depending on the alternative hypothesis of interest. For k groups, to compare the probability of survival to time T_o, the estimated values may be compared by constructing multiple comparison confidence intervals.

16.7 ADJUSTMENT FOR CONFOUNDING FACTORS BY STRATIFICATION

In Example 16.2, in the Coronary Artery Surgery Study (Passamani et al., 1982), the degree of impairment due to chest pain pattern was related to survival. Patients with pain definitely not angina had a better survival pattern than patients with definite angina. The chest pain status is predictive of survival. These patients were studied by coronary angiography; the amount of disease in their coronary arteries as well as their left ventricular performance (the performance of the pumping part of the heart) were also evaluated. One might argue that the amount of disease is a more fundamental predictor than type of chest pain. If the pain results from coronary artery disease that affects the arteries and ventricle, the latter affects survival more fundamentally. We might ask the question: Is there additional prognostic information in the type of chest pain if one takes into account, or adjusts for, the angiographic findings?

We have used various methods of adjusting for variables. As discussed in Chapter 2, twin studies adjust for genetic variation by matching people with the same genetic pattern. Analogously, matched-pairs studies match people to be (effectively) twins in the pertinent variables; this adjusts for covariates. One step up from this is *stratified analysis*. In this case, the strata are to be quite homogeneous. People in the same strata are (to a good approximation) the same with respect to the variable or variables used to define the strata. One example of stratified analysis occurred with the Mantel–Haenszel procedure for summing 2×2 tables. The point of the stratification was to adjust for the variable or variables defining the strata. In this section we consider the same approach to the analysis of the life table or actuarial method of comparing survival curves from different groups.

16.7.1 Stratification of Life Table Analyses: Log-Rank Test

To extend the life table approach to stratification is straightforward. The first step is to perform the life table survival analysis *within each stratum*. If we do this for the four chest pain classes as discussed in Example 16.2 to adjust for angiographic data, we would use strata that depend on the angiographic findings. This is done below. Within each of the strata, we will be comparing persons with the same angiographic findings but different chest pain status. The log-rank statistic may be computed *separately* for each of the strata, giving us an observed and expected number of deaths for each group being studied. Somehow we want to combine the information across all the strata. This was done, for example, in the Mantel–Haenszel approach to 2×2 tables. We do this by summing the values for each group of the observed and expected numbers of deaths for the different strata. These observed and expected numbers are then combined into a final log-rank statistic. Note 16.3 gives the details of the computation of the statistic. Because it is based on many more subjects, the final statistic will be much more powerful then the log-rank statistic for any one stratum, *provided* that there is a consistent trend in the same direction within strata. We illustrate this by example.

Example 16.2. (*continued*) We continue with our study of chest pain groups. We would like to adjust for angiographic variables. A study of the angiographic variables showed that most of the prognostic information is contained within these variables:

1. The number of vessels diseased of the three major coronary vessels
2. The number of proximal vessels diseased (i.e., the number of diseased vessels where the disease is near the point where the blood pumps into the heart)
3. The left ventricular function, measured by a variable called LVSCORE

Various combinations of these three variables were used to define 30 different strata. Table 16.6 gives the values of the variables and the strata. Separate survival curves result in the differing strata. Figures 16.10 and 16.11 present the survival curves for two of the different strata used.

Note that the overall p-value is 0.69, a result that is not statistically significant. Thus although the survival patterns differ among chest pain categories, the differences may be explained by different amounts of underlying coronary artery disease. In other words, adjustment for the arteriographic and ventriculographic findings removed the group differences.

Note that of 30 strata, one p-value, that of stratum 25, is less than 0.05. Because of the multiple comparison problem, this is not a worry. Further, in this stratum, the definite angina cases have one observed and 0.03 expected deaths. As the log-rank statistic has an *asymptotic* chi-square distribution, the small expected number of deaths make the asymptotic distribution inappropriate in this stratum.

16.8 COX PROPORTIONAL HAZARD REGRESSION MODEL

In earlier work on the life table method, we observed various ways of dealing with factors that were related to survival. One method is to plot data for different groups, where the groups were defined by different values on the factor(s) being analyzed. When we wanted to adjust for covariates, we examined stratified life table analyses. These approaches are limited, however, by the numbers involved. If we want to divide the data into strata on 10 variables simultaneously, there will be so many strata that most strata will contain no one or at most one person. This makes comparisons impossible. One way of getting around the number problem is to have an appropriate mathematical model with covariates. In this section we consider the *Cox proportional hazards regression model*. This model is a mathematical model of survival that allows covariate values to be taken into account. Use of the model in survival analysis is quite similar to the multiple regression analysis of Chapter 11. We first turn to examination of the model itself.

16.8.1 Cox Proportional Hazard Model

Suppose that we want to examine the survival pattern of two people, one of whom initially is at higher risk than the other. A natural way to quantify the idea of risk is the hazard function discussed previously. We may think of the hazard function as the instantaneous probability of dying given that a person has survived to a particular time. The person with the higher risk will have a higher value for the hazard function than a person who has lower risk at the particular time. The Cox proportional hazard model works with covariates; the model expresses the hazard as a function of the covariate values. The major assumption of the model is that if the first person has a risk of death at the initial time point that is, say, twice as high as that of a second person, the risk of death at later times is also twice as large. We now express this mathematically.

Suppose that at the average value of all of our covariates in the population, the hazard at time t, is denoted by $h_0(t)$. Any other person whose values on the variables being considered are not equal to the mean values will have a hazard function proportional to $h_0(t)$. This proportionality constant varies from person to person depending on the values of the variables. We develop this

Table 16.6 Stratified Analysis of Survival by Chest Pain Classification

Stratification Variables				Deaths								
				Definite Angina		Probable Angina		Probably Not Angina		Definitely Not Angina		Log-Rank Statistic
Stratum Number	Number of Vessels	Number of Prox. Vessels	Left Ventricular Score	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	*p*-Value
1	0	0	5–11	9	10.07	42	38.33	39	43.35	9	7.25	0.74
2	0	0	12–16	0	0.79	2	1.25	1	0.87	0	0.09	0.73
3	0	0	17–30	0	0.00	0	0.00	0	0.00	0	0.00	1.00
4	1	0	5–11	19	18.88	26	23.84	5	6.71	0	0.56	0.85
5	1	0	12–16	3	3.46	5	3.25	0	1.06	0	0.23	0.52
6	1	0	17–30	1	0.31	0	0.62	0	0.08	—	—	0.43
7	1	1	5–11	14	13.36	13	13.19	2	2.00	0	0.45	0.96
8	1	1	12–16	1	2.53	3	2.05	0	0.27	1	0.15	0.15
9	1	1	17–30	4	3.49	2	2.22	0	0.30	—	—	0.93
10	2	0	5–11	17	18.54	16	14.62	2	2.29	1	0.55	0.93
11	2	0	12–16	7	6.81	2	3.90	3	1.11	0	0.18	0.20
12	2	0	17–30	5	3.49	3	3.99	1	0.98	0	0.53	0.72
13	2	1	5–11	18	15.50	10	14.91	1	1.07	2	0.24	0.11
14	2	1	12–16	9	9.06	6	4.99	0	0.80	0	0.14	0.77
15	2	1	17–30	3	3.40	3	2.38	0	0.22	—	—	0.93
16	2	2	5–11	18	17.36	13	13.56	1	0.92	0	0.16	0.99
17	2	2	12–16	19	6.70	4	5.98	0	0.32	—	—	0.59
18	2	2	17–30	3	4.67	4	2.33	—	—	—	—	0.62
19	3	0	5–11	11	11.75	9	7.44	0	0.72	0	0.10	0.76
20	3	0	12–16	8	7.49	7	6.69	—	—	0	0.83	0.83
21	3	0	17–30	4	4.31	1	0.69	—	—	—	—	0.98
22	3	1	5–11	28	23.67	15	17.78	0	1.54	—	—	0.37
23	3	1	12–16	17	16.66	6	6.34	—	—	—	—	1.00
24	3	1	17–30	9	7.32	5	6.15	0	0.53	—	—	0.72
25	3	2	5–11	36	32.08	11	17.55	2	0.34	1	0.03	0.01
26	3	2	12–16	20	16.48	6	8.45	0	1.07	—	—	0.42
27	3	2	17–30	8	9.34	7	5.17	—	—	0	0.49	0.72
28	3	3	5–11	17	22.42	19	14.36	1	0.22	—	—	0.11
29	3	3	12–16	16	14.62	6	8.24	1	0.14	—	—	0.09
30	3	3	17–30	11	12.93	4	2.07	—	—	—	—	0.56
Total				325	317.49	250	251.63	59	66.91	14	11.97	0.69[a]

[a] A dash indicates no individuals in the group in the given stratum. Obs., observed; Exp., expected; log-rank statistic = 1.47 with 3 degrees of freedom.

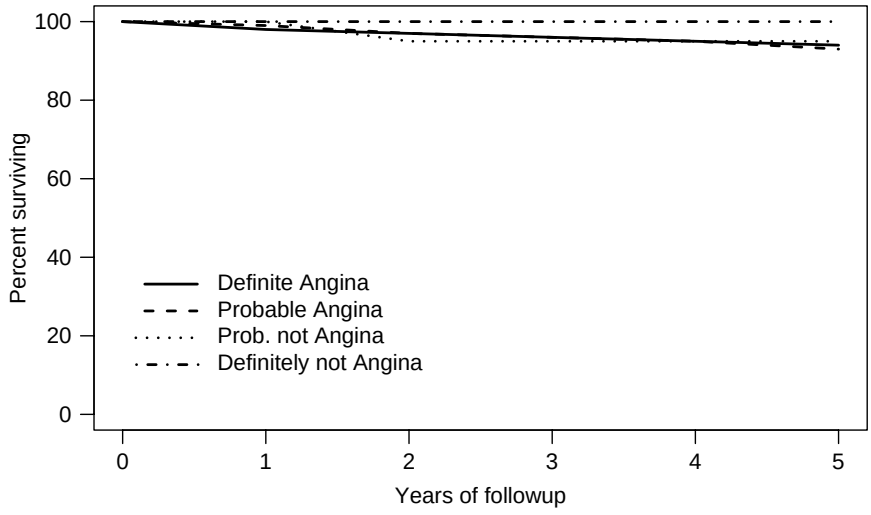


Figure 16.10 Example 16.4: survival curves for stratum 7. Cases have one proximal vessels diseased with good ventricular function (LVSCORE of 5–11).

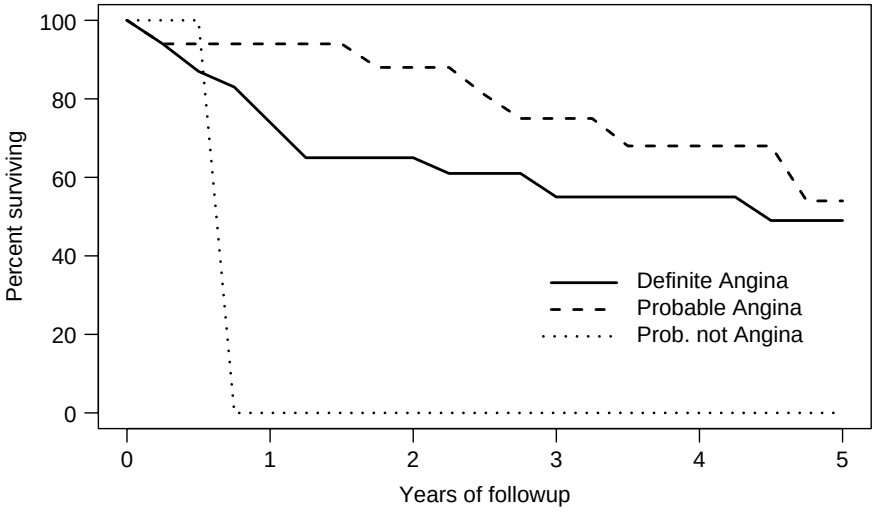


Figure 16.11 Example 16.4: survival curves for stratum 29. Cases have three proximal vessels diseased with impaired ventricular function (LVSCORE of 12–17).

algebraically. There are variables $X_1, \ldots, X_p$ to be considered. Let $\mathbf{X}$ denote the values of all the X_i, that is, $\mathbf{X} = (X_1, \ldots, X_p)$.

1. If a person has $\mathbf{X} = \overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_p)$, the hazard function is $h_0(t)$.
2. If a person has different values for $\mathbf{X}$, the hazard function is $h_0(t)C$, where C is a constant that depends on the values of $\mathbf{X}$. If we think of the hazard as depending on $\mathbf{X}$, as well as t, the hazard is

$$h_0(t)C(\mathbf{X})$$

3. For any two people with values of $\mathbf{X} = \mathbf{X}(1)$ and $\mathbf{X} = \mathbf{X}(2)$, respectively, the ratio of their two hazard functions is

$$\frac{h_0(t)C(\mathbf{X}(1))}{h_0(t)C(\mathbf{X}(2))} = \frac{C(\mathbf{X}(1))}{C(\mathbf{X}(2))} \tag{18}$$

The hazard functions are *proportional*; the ratio does not depend on t.

Let us reiterate this last point. Given two people, if one has one-half as much risk initially as a second person, then at all time points, risk is one-half that of the second person. Thus, the two hazard functions are proportional, and such models are called *proportional hazard models*.

Note that proportionality of the hazard function is an assumption that does not necessarily hold. For example, if two people were such that one is to be treated medically and the second surgically by open heart surgery, the person being treated surgically may be at higher risk initially because of the possibility of operative mortality; later, however, the risk may be the same or even less than that of the equivalent person being treated medically. In this case, if one of the covariate values indicates whether a person is treated medically or surgically, the proportional hazards model will not hold. In a given situation you need to examine the plausibility of the assumption. The model has been shown empirically to hold reasonably well for many populations over moderately long periods, say five to 10 years. Still, proportional hazards is an assumption.

As currently used, one particular parametric form has been chosen for the proportionality constant $C(\mathbf{X})$. Since it multiplies a hazard function, this constant must always be positive because the resulting hazard function is an instantaneous probability of an endpoint and consequently must be nonnegative. A convenient functional form that reasonably fits many data sets is

$$C(\mathbf{X}) = e^{\alpha+\beta_1 X_1+\cdots+\beta_p X_p}, \qquad \text{where} \quad \alpha = -\beta_1 \overline{X}_1 - \cdots - \beta_p \overline{X}_p \tag{19}$$

In this parameterization, the unknown population parameters β_i are to be estimated from a data set at hand.

With hazard $h_0(t)$, let $S_{0,\text{pop}}(t)$ be the corresponding survival curve. For a person with covariate values $\mathbf{X} = (X_1, \ldots, X_p)$, let the survival be $S(t|\mathbf{X})$. Using the previous equations, the survival curve is

$$S(t|\mathbf{X}) = (S_{0,\text{pop}}(t))^{\exp(\alpha+\beta_1 X_1+\cdots+\beta_p X_p)} \tag{20}$$

That is, the survival curve for any person is obtained by raising a standard survival curve $[S_{0,\text{pop}}(t)]$ to an appropriate power. To estimate this quantity, the following steps are performed:

1. Estimate $S_{0,\text{pop}}$ and $\alpha, \beta_1, \ldots, \beta_p$ by $S_0(t), a, b_1, \ldots, b_p$. This is done by a computer program. The estimation is too complex to do by hand.
2. Compute $Y = a + b_1 X_1 + \cdots + b_p X_p$ [where $\mathbf{X} = (X_1, \ldots, X_p)$].
3. Compute $k = e^Y$.
4. Finally, compute $S_0(t)^k$.

The estimated survival curve is the population curve (the curve for the mean covariate values) raised to a power. If the power k is equal to 1, corresponding to e^0, the underlying curve for S_0 results. If k is greater than 1, the curve lies below S_0, and if k is less than 1, the curve lies above S_0. This is presented graphically in Figure 16.12.

Note several factors about these curves:

1. The curves do not cross each other. This means that a procedure having a high initial mortality, such as a high dose of radiation in cancer therapy, but better long-term survival,

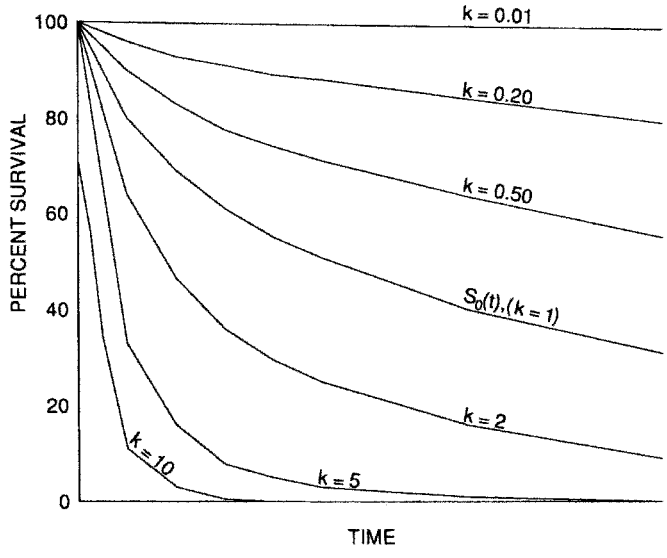


Figure 16.12 Proportional hazard survival curves as a function of $k = e^{a+b_1x_1+\cdots+b_px_p}$.

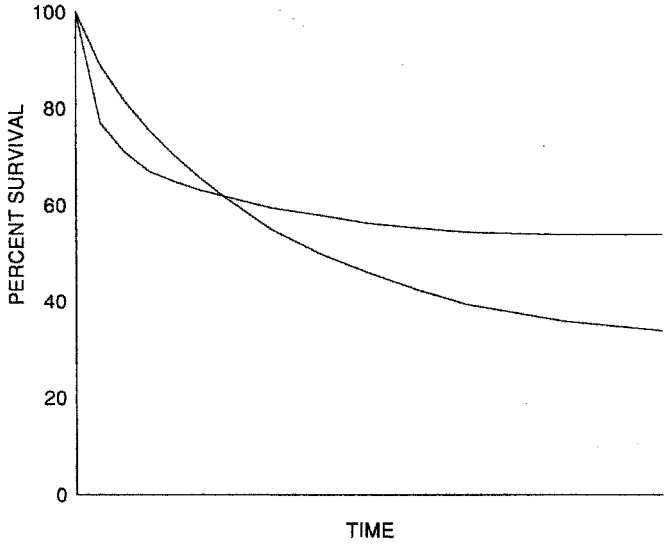


Figure 16.13 Two survival curves without proportional hazards.

as in Figure 16.13, could not be modeled by the proportional hazard model with one of the variables, say X_1, equal to 1 if the therapy were radiation and 0 if an alternative therapy were used.

2. The proportionality constant in the proportional hazard model,

$$e^{\alpha+\beta_1x_1+\cdots+\beta_px_p}$$

is parametric. We have not specified the form of the underlying survival S_0. This curve is not estimated by a parametric model but by other means.

3. Where there is a plateau in one curve, the other curve has a plateau at the same time points. The proportional hazards assumption implies that covariates do not affect the timing of plateaus or other distinctive features of the curves, only their height.

16.8.2 Example of the Cox Proportional Hazard Regression Model

The *Cox proportional hazard model* is also called the *Cox proportional regression model* or the *Cox regression model*. The reason for calling this model a regression model is that the dependent variable of interest, survival, is modeled upon or "regressed upon" the values of the covariates or independent variables. The analogies between multiple regression and the Cox regression are quite good, although there is not a one-to-one correspondence between the techniques. Computer software for Cox regression typically produces at least the quantities shown in Table 16.7.

The following example illustrates the use of the Cox proportional hazards model.

Example 16.4. The left main coronary artery is a short segment of the arteries delivering blood to the heart. Two of the three major arterial systems branch off the left main coronary artery. If this artery should close, death is almost certain. Two randomized clinical trials (Veterans' Administration Study Group, Takaro et al. [1976] and the European Coronary Surgery Study Group [1980]) reported superior survival in patients undergoing coronary artery bypass surgery. Chaitman et al. [1981] examined the observational data of the Coronary Artery Surgery Study (CASS), registry. Patients were analyzed as being in the medical group until censored at the time of surgery. They were then entered into the surgical survival experience at the day of surgery.

A Cox model using a therapy indicator variable was used to examine the effect of therapy. Eight variables were used in this model:

- *CHFSCR:* a score for congestive heart failure (CHF). The score ranged from 0 to 4; 0 indicated no CHF symptoms. A score of 4 was indicative of severe, treated CHF.
- *LMCA:* the percent of diameter narrowing of the left main coronary artery due to atherosclerotic heart disease. By selection, all cases had at least 50% narrowing of the left main coronary artery (LMCA).
- *LVSCR:* a measure of ventricular function, the pumping action of the heart. The score ranged from 5 (normal) to a potential maximum of 30 (not attained). The higher the score, the worse the ventricular function.
- *DOM:* the dominance of the heart shows whether the right coronary artery carries the usual amount of blood; there is great biological variability. Patients are classed as right or balanced dominance (DOM = 0). A left-dominant subject has a higher proportion of blood flow through the LMCA, making left main disease even more important (DOM = 1).
- *AGE:* the patient's age in years.
- *HYPTEN:* Is there a history of hypertension? HYPTEN = 1 for yes and HYPTEN = 0 for no.
- *THRPY:* This is 1 for medical therapy and 2 for surgical therapy.
- *RCA:* This variable is 1 if the right coronary artery has $\geq$ 70% stenosis and is zero otherwise.

The Cox model produces the results shown in Table 16.8. The chi-square value for CHFSCR is found by the square of β divided by the standard error. For example, $(0.2985/0.0667)^2 = 20.03$, which is the chi-square value to within the numerical accuracy. The underlying survival curve (at the mean covariate values) has probabilities 0.944 and 0.910 of one- and two-year survival, respectively. The first case in the file has values CHFSCR = 3, LMCA = 90, LVSCR = 18, DOM = 0, AGE = 49, HYPTEN = 1, THRPY = 1, and RCA = 1. What is the estimated

Table 16.7 Computer Output for Cox Regression

Output	Description	Use of Output
b_i	Estimate of the regression coefficient β_i	**1.** The b_i give an estimate of the increase in risk (the hazard function) for different values of $X_1, \ldots, X_p$. **2.** The regression coefficients allow estimation of $e^{\alpha+\beta_1 X_1+\cdots+\beta_p X_p}$ by $e^{a+b_1 x_1+\cdots+b_p x_p}$. By using this and the estimate of $S_0(t)$, we can estimate survival for any person in terms of the values of $X_1, \ldots, X_p$ for each time t.
$\text{SE}(b_i)$	Estimated standard error of b_i	**1.** The distribution of b_i is approximately $N(\beta_i, \text{SE}(b_i)^2)$ for large sample sizes. We can obtain $100(1-\alpha)\%$ confidence intervals for β_i as $(b_i - z_{1-\alpha/2}\text{SE}(b_i), b_i + z_{1-\alpha/2}\text{SE}(b_i))$. **2.** We test for statistical significance of β_i (in a model with the other X_j's) by rejecting $\beta_i = 0$ if $b_i^2/[\text{SE}(b_i)]^2 \geq \chi^2_{1,1-\alpha}$. $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ percentile of the χ^2 distribution with one degree of freedom. This χ^2 test or the equivalent z test is also given by most software.
Model chi-square	Chi-square value for the entire model with p degrees of freedom	**1.** For nested models the chi-square values may be subtracted (as are the degrees of freedom) to give a chi-square test. **2.** For a single model this chi-square statistic tests for *any* relationships among the $X_1, \ldots, X_p$ and the survival experience. The null hypothesis tested is $\beta_1 = \cdots = \beta_p = 0$, which is only occasionally an interesting null hypothesis. This is analogous to testing for zero multiple correlation between survival and $(X_1, \ldots, X_p)$ in a multiple regression setting.
$S_0(t)$ and α, or $S_0(t)^\alpha$	Estimate of the survival function for a person with covariate values equal to the mean of each variable, or for a person with zero values of the covariate	**1.** With $S_0(t)$ and a, or $S_0(t)^\alpha$ and the b_i, we may plot the estimated survival experience of the population for any fixed value of the covariates. **2.** For a fixed time, say t_0, by varying the values of the covariates $\mathbf{X}$, we may present the effect of combinations of the covariate values (see Example 16.5).

probability of one- and two-year survival for this person?

$$\begin{aligned} a + b_1 X_1 + \cdots + b_n X_n &= -2.8968 + (0.2985 \times 3) + (0.0178 \times 90) \\ &\quad + (0.1126 \times 18) + (1.2331 \times 0) + (0.0423 \times 49) \\ &\quad + (-0.5428 \times 1) + (-1.0777 \times 1) \\ &\quad + (0.5285 \times 1) \\ &= 2.6622 \end{aligned}$$

Table 16.8 Results of Cox Model Fitting

Variable	Beta	Standard Error	Chi-Square	Probability
CHFSCR	0.2985	0.0667	20.01	0.0000
LMCA	0.0178	0.0049	13.53	0.0002
LVSCR	0.1126	0.0182	38.41	0.0000
DOM	1.2331	0.3564	11.97	0.0006
AGE	0.0423	0.0098	18.75	0.0000
HYPTEN	−0.5428	0.1547	12.31	0.0005
THRPY	−1.0777	0.1668	41.77	0.0000
RCA	0.5285	0.2923	3.27	0.0706
Constant	−2.8968			

$$\begin{aligned}\text{estimated probability of one-year survival} &= 0.944^{e^{2.6622}}\\ &= 0.944^{14.328}\\ &= 0.438\\ \text{estimated probability of two-year survival} &= 0.910^{14.328}\\ &= 0.259\end{aligned}$$

The estimated probability of survival under medical therapy is 44% for one year and 26% for two years. This bad prognosis is due largely to heart failure (CHFSCR) and very poor ventricular function (LVSCR).

16.8.3 Interpretation of the Regression Coefficients β_i

In the multiple regression setting, the regression coefficients may be interpreted as the average difference in the response variables between cases where the predictor variable differs by one unit, with everything else the same. In this section we look at the interpretation of the β_i for the Cox proportional hazard model. Recall that the hazard function is proportional to the probability of failure in a short time interval. Suppose that we have two patients whose covariate values are the same on all the p regression variables for the Cox model with the exception of the ith variable. If we take the ratio of the hazard functions for the two people at some time t, we have the ratio of the probability of an event in a short interval after time t. The ratio of these two probabilities is the relative risk of an event during this time period. This is also called the *instantaneous relative risk*. For the Cox proportional hazards model, we find that

$$\begin{aligned}\text{instantaneous relative risk (RR)} &= \frac{h_0(t)e^{\alpha+\beta_1 X_1+\cdots+\beta_i X_i^{(1)}+\cdots+\beta_p X_p}}{h_0(t)e^{\alpha+\beta_1 X_1+\cdots+\beta_i X_i^{(2)}+\cdots+\beta_p X_p}}\\ &= e^{\beta_i(X_i^{(1)}-X_i^{(2)})}\end{aligned} \tag{21}$$

An equivalent formulation is to take the logarithm of the instantaneous relative risk (RR). The logarithm is given by

$$\ln(\text{RR}) = \beta_i(X_i^{(1)} - X_i^{(2)}) \tag{22}$$

In words, the regression coefficients β of the Cox proportional hazard model are equal to the logarithm of the relative risk if the variable X is increased by one unit.

16.8.4 Evaluating the Proportional Hazards Assumption

One graphical assessment of the proportional hazards assumption for binary (or categorical) variables plots the cumulative hazard in each group on a logarithmic scale. Under the proportional hazards assumption, the resulting curves should be parallel, that is, separated by a constant vertical difference (the reason is given in Section 16.8.3). Although popular, these *log-log plots* are not particularly useful. Judging whether two curves (as opposed to straight lines) are parallel is difficult, and the problem is compounded by the fact that the uncertainty in the estimated log hazard varies substantially along the curves.

An better approach to judging proportional hazards involves smoothed plots of the *scaled Schoenfeld residuals*, proposed by Therneau and Grambsch [2000]. These plots, available in Stata and S, estimate how a coefficient β_i varies over time. In addition to an easier visual interpretation, the Schoenfeld residual methods provide a formal test of the proportional hazards assumption and are valid for continuous as well as categorical variables.

The technical details of the Schoenfeld residual methods are complex, but there is a simple underlying heuristic. Suppose that the hazard ratio for, say, hypertension is greater than unity. Hypertensive persons will be overrepresented among the deaths in any given period. If, in addition, the hazard ratio increases with time, overrepresentation of hypertensives among the deaths will increase with time. By calculating the proportion of hypertensives among the deaths and the population at risk in each short interval of time, we should be able to detect the increasing hazard ratio.

If there is substantial nonproportionality of hazards, it may be desirable to stratify the model (see Section 16.10.2) on the variable in question, or to define a time-dependent variable as in Example 16.7 in Section 16.10.2.

Example 16.5. Primary biliary cirrhosis is a rare, autoimmune disease of the liver. Until the advent of liver transplantation, it was untreatable and eventually fatal. The Mayo Clinic performed a randomized trial of one proposed treatment, D–penicillamine, in 312 patients. The treatment was not effective, but the data from the trial have been used to develop a widely used prognostic model for survival of this disease. The data for this model have been made available on the Web by Terry Therneau of the Mayo Clinic and are linked in the Web appendix.

The Mayo model includes five covariates:

- *BILI:* logarithm of serum bilirubin concentration. Bilirubin is excreted in the bile and accumulates in liver disease.
- *PROTIME:* logarithm of the prothrombin time, a measure of blood clotting. Prothrombin time is increased when the liver fails to produce certain clotting factors.
- *ALBUMIN:* logarithm of serum albumin concentration. The liver produces albumin to prevent blood plasma from leaking out of capillaries.
- *EDTRT:* edema (fluid retention), coded as 0 for no edema, $\frac{1}{2}$ for untreated edema or edema resolved by treatment, 1 for edema present despite treatment.
- *AGE:* in tens of years. Age affects the risk for almost any cause of death.

Figure 16.14 shows scatter plots for two of these covariates against survival time. The censored observations are indicated by open triangles, the deaths by filled circles. There is clearly a relationship with both variables. It is also interesting to note that according to Fleming and Harrington [1991, Chap. 5], the outlying value of 18 for prothrombin time was a data-entry error; it should be 11.

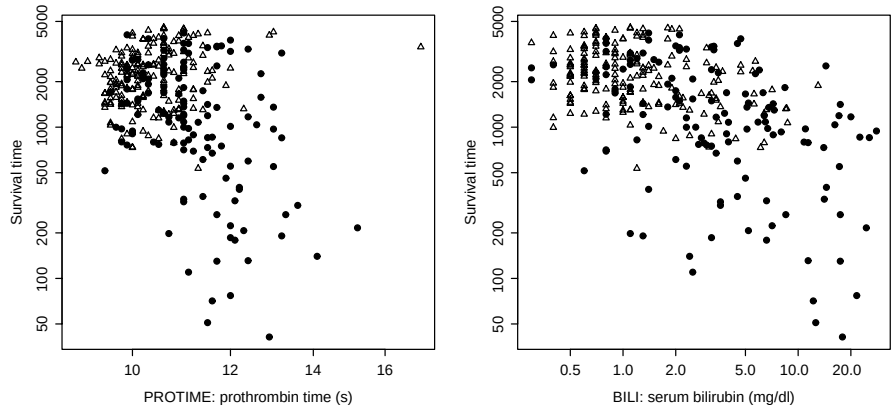


Figure 16.14 Scatter plots of survival time vs. PROTIME and BILI in the Mayo PBC data. Triangles indicate censored times.

The Mayo model has the following coefficients:

Variable	b	SE(b)
BILI	0.88	0.10
EDTRT	0.79	0.30
ALBUMIN	−3.06	0.72
PROTIME	3.01	1.02
AGE	0.33	0.08

The survival function for someone with no edema, albumin of 3.5 mg/dL, prothrombin time of 10 seconds, bilirubin of 1.75 mg/dL, and age 50 is:

t (yr)	$S(t)$ (%)	t (yr)	$S(t)$ (%)
1	98	6	80
2	97	7	74
3	92	8	68
4	88	9	61
5	84	10	51

Figure 16.15 shows a scaled Schoenfeld residual plot for PROTIME. The smooth curve that estimates $\beta(t)$ shows that the logarithm of the hazard ratio for elevated prothrombin time is very high initially and then decreases to near zero over the first three to four years. That is, a patient with high prothrombin time is at greatly increased risk of death, but a patient who had a high prothrombin time four years ago and is still alive is not at particularly high risk. The p-value for nonproportionality for PROTIME is 0.055, so there is moderately strong evidence that the pattern we see in Figure 16.15 is real.

16.8.5 Use of the Cox Model as a Method of Adjustment

In Section 16.7 we considered stratified life table analyses to adjust for confounding factors or covariates. The Cox model may be used for the same purpose. As in the multiple linear regression model, there are two ways in which we may adjust. One is to consider a variable whose effect we want to study in relationship to survival. Suppose that we want adjust for

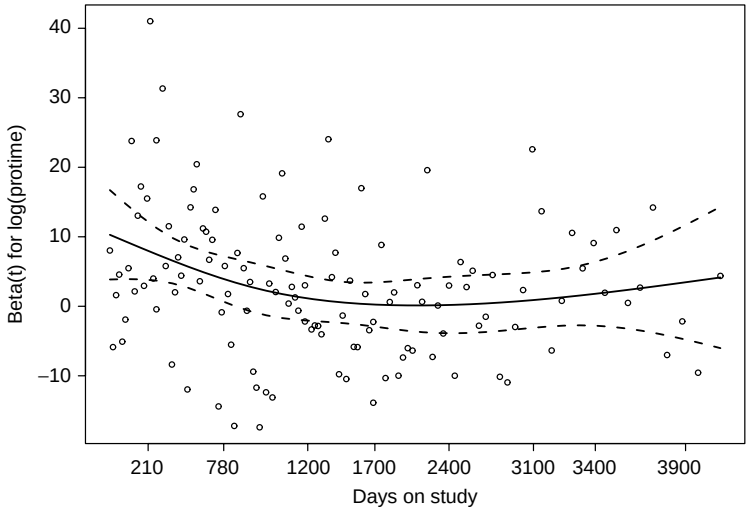


Figure 16.15 Assessing proportional hazards for PROTIME with scaled Schoenfeld residuals.

variables $X_1, \ldots, X_k$. We run the Cox proportional hazards regression model with the variable of interest and the adjustment covariates in the model. The statistical significance of the variable of interest may be tested by taking its estimated regression coefficient, dividing by its standard error and using a normal probability critical value. An equivalent approach, similar to nested hypotheses in the multiple linear regression model, is to run the Cox proportional hazards model with only the adjusting covariates. This will result in a chi-square statistic for the entire model. A second Cox proportional hazards model may be run with the variable of interest in the model in addition to the adjustment covariates. This will result in a second chi-square statistic for the model. The chi-square statistic for the second model minus the chi-square statistic for the first model will have approximately a chi-square distribution with one degree of freedom if the variable of interest has no effect on the survival after adjustment for the covariates $X_1, \ldots, X_p$.

Example 16.5. (*continued*) Of the 418 patients in the Mayo Clinic PBC data set, 312 agreed to participate in the randomized trial and 106 refused. As the data from the randomized trial were used to develop a predictive model for survival, it is important to know whether the randomized and nonrandomized patients differ in important ways.

A simple comparison of survival times in these two groups does not answer quite the right question. Suppose that patients agreeing to be randomized had longer survival times but also had lower levels of bilirubin that were sufficient to explain their improved survival. This discrepancy in survival times does not invalidate the model. Conversely, if the two groups had very similar survival times despite a difference in average bilirubin levels, this would be evidence against the model.

We can estimate the adjusted difference between the randomized and nonrandomized patients by fitting a Cox model that has the five Mayo model predictors and an additional variable indicating which group the patient is in. The estimated hazard ratio for nonrandomized patients is 0.97, with a 95% confidence interval from 0.66 to 1.41. We would not typically report coefficients and confidence intervals for the other adjusting covariates; their associations with survival are not of direct interest in this analysis.

There is no evidence of any difference in survival between randomized and nonrandomized patients in this study, but the confidence intervals are quite wide, so these differences have not been ruled out.

Other examples of estimating adjusted contrasts using the Cox model appear in Section 16.10.

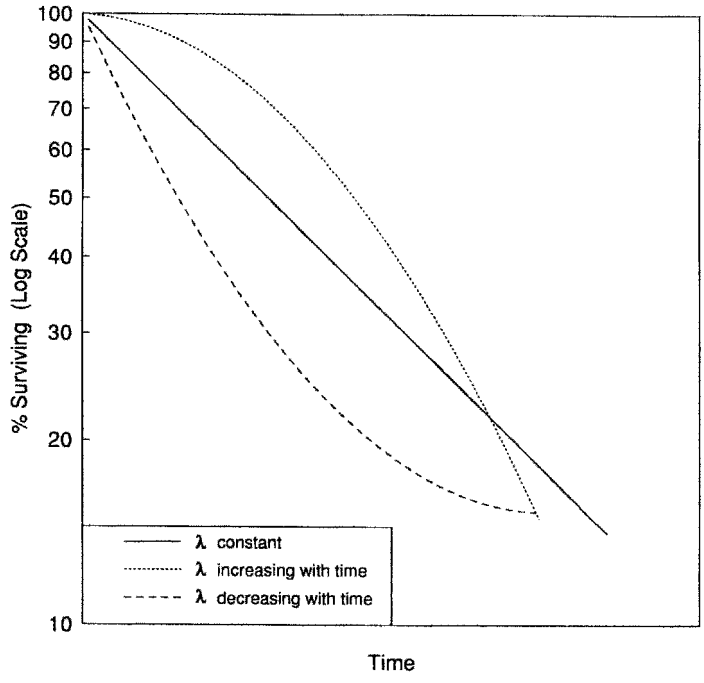


Figure 16.16 Log plot for exponential survival.

16.9 PARAMETRIC MODELS

16.9.1 Exponential Model; Rates

Suppose that at each instant of time, the instantaneous probability of death is the same. That is, suppose that the hazard rate or force of mortality is constant. Although in human populations this is not a useful assumption over a wide time interval, it may be a valid assumption over a five- or 10-year interval, say.

If the constant hazard rate is λ, the survival curve is $S(t) = e^{-\lambda t}$. From this expression the term *exponential survival* arises. The expected length of survival is $1/\lambda$. If the exponential situation holds, the parameter λ is estimated by the number of events divided by total exposure time. The methods and interpretation of rates are then appropriate. If $S(t)$ is exponential, $\log S(t) = -\lambda t$ is a straight line with slope $-\lambda$. Plotting an estimate of $S(t)$ on a logarithmic scale is one way of visually examining the appropriateness of assuming an exponential model. Figure 16.16 shows some of the patterns that one might observe.

To illustrate this we return to the Mayo primary biliary cirrhosis data set but now consider an analysis of time until loss to follow-up, that is, a survival analysis where the event is loss to follow-up. To avoid confusing patients lost to follow-up with those alive and under observation at the end of the study, we look at just the first eight years of the study. From the plot one sees that the data do *not* look exponential (Figure 16.17). Rather, it appears that the hazard of dropping out is initially very low and increases progressively.

16.9.2 Two Other Parametric Models for Survival Analysis

There are a variety of parametric models for survival distributions. In this section, two are mentioned. For details of the distributions and parameter estimates, the reader is referred to

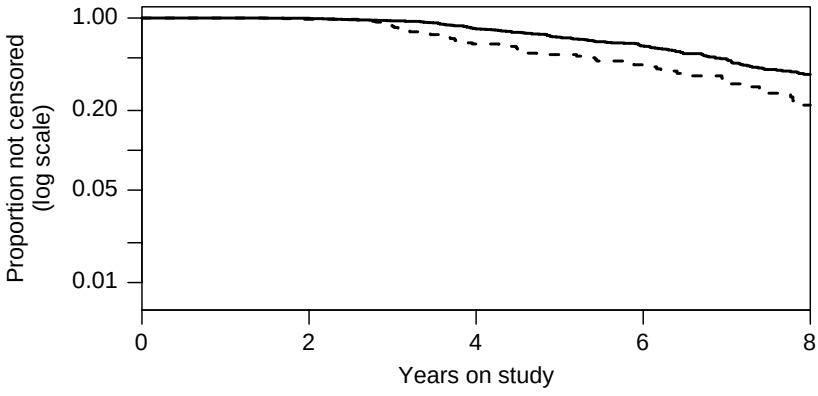


Figure 16.17 Loss to follow-up of 312 randomized and 106 nonrandomized patients with primary biliary cirrhosis.

texts by Mann et al. [1974] and Gross and Clark [1975]. These books also present a variety of models not touched on here.

The two-parameter *Weibull distribution* has a survival curve of the form

$$S(t) = e^{-\alpha t^{\beta}} \qquad \text{for } t > 0\,(\alpha > 0, \beta > 0) \tag{23}$$

If $\beta = 1$, the Weibull distribution is the exponential model with constant hazard rate. The hazard rate decreases with time if $\beta < 1$ and increases with time if $\beta > 1$. Often, if the time of survival is measured from diagnosis of a disease, a Weibull with $\beta > 1$ will reasonably model the situation. Estimates are made by computer.

Another distribution, the *lognormal distribution*, assumes that the logarithm of the survival time is normally distributed. If there is no censoring of data, one may work with the logarithm of the survival times and use methods appropriate for the normal distribution.

Regression versions of the exponential, lognormal, Weibull, and other parametric survival models are also available in many statistical packages. The exponential and Weibull models are special cases of the Cox proportional hazards model and have little advantage over the Cox model. The lognormal model is not related to the Cox model.

16.10 EXTENSIONS

16.10.1 Cox Model with Time-Dependent Covariates

If two groups are defined by some baseline measurement, such as smokers and nonsmokers, their hazard ratio would be expected to change over time simply because some of the smokers will stop smoking and lower their risk of death. For this reason it may be desirable to base the hazard ratio at time t on the most recent available values of covariates rather than on the values at the start of follow-up. The Cox model is then most naturally written in terms of the hazard rather than the survival:

$$\text{hazard at time } t = h_0(t) \exp\left[\alpha + \beta_1 X_1(t) + \beta_2 X_2(t) + \cdots + \beta_p X_p(t)\right]$$

and we write $X_1(t)$ for the value of X_1 at time t.

The hazard ratio between two subjects with covariates $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is then

$$\begin{aligned}\frac{h(t;\mathbf{X}^{(1)})}{h(t;\mathbf{X}^{(2)})} &= \frac{h_0(t)\exp[\alpha+\beta_1 X_1(t)^{(1)}+\beta_2 X_2(t)^{(1)}+\cdots+\beta_p X_p(t)^{(1)}]}{h_0(t)\exp[\alpha+\beta_1 X_1(t)^{(2)}+\beta_2 X_2(t)^{(2)}+\cdots+\beta_p X_p(t)^{(2)}]} \\ &= \frac{\exp[\beta_1 X_1(t)^{(1)}+\beta_2 X_2(t)^{(1)}+\cdots+\beta_p X_p(t)^{(1)}]}{\exp[\beta_1 X_1(t)^{(1)}+\beta_2 X_2(t)^{(1)}+\cdots+\beta_p X_p(t)^{(1)}]} \\ &= \exp\left\{\beta_1\left[X_1(t)^{(1)}-(X_1(t)^{(2)}\right]+\beta_2\left[X_2(t)^{(1)}-(X_2(t)^{(2)}\right]\right. \\ &\quad \left.+\cdots+\beta_p\left[X_p(t)^{(1)}-(X_p(t)^{(2)}\right]\right\}\end{aligned}$$

In the constant-covariate situation, the proportional hazards assumption means that the hazard ratio does not change over time; in the time-dependent situation, it means that the hazard ratio changes only due to changes in the covariates over time.

Example 16.6. An example of time-dependent covariates comes from a study by Holt and colleagues [2002] that examined the effects of court protective orders on abuse of women by their domestic partners. In this study the time-dependent covariates were the presence (1) or absence (0) of temporary restraining orders and permanent restraining orders. At the start of the study, after the first police report of abuse, both variables would be zero. Most of the women in the study (2366) never obtained a protection order, so the variable remained at zero. Of those who obtained a two-week temporary order (325), about half (185) later obtained a permanent order. The time-dependent Cox model compares the risk of abuse in women who do and do not have each type of protective order *at the same time after their initial incident*. Cox models thus reduce the potential for confounding by time since the initial incident: Since permanent protective orders tend to happen later in time, when risks are already lower, they might appear protective even if they actually had no effect.

Temporary restraining orders were associated with an increase in the hazard of psychological abuse (hazard ratio 4.9, 95% confidence interval 2.6 to 8.6) and no change in the hazard of physical abuse (hazard ratio 1.6, 95% CI 0.6 to 4.4). Permanent restraining orders appeared to reduce physical abuse (hazard ratio 0.2, 95% CI 0.1 to 0.8) and have no effect on psychological abuse (hazard ratio 0.9, 95% CI 0.5 to 1.7).

In some settings it may be more appropriate to use values of covariates for some short or long period in the past rather than the instantaneously updated values. These time-dependent variables reflect the history of exposure rather than just the current status.

Example 16.7. Heckbert et al. [2001] studied how the risk of a recurrent heart attack changed over time in women who had already had one heart attack and were taking hormone replacement therapy (HRT). Estrogen, the active ingredient of HRT, is known to improve cholesterol levels but also to increase blood clotting, and so might have positive or negative effects on heart disease. A recent randomized trial, HERS [Hulley et al., 1998], suggested that the balance of risk and benefit might change over time.

The researchers hypothesized that having recently started hormone replacement therapy would increase the risk of heart attack, but that long-term therapy might not increase the risk. They defined three time-dependent exposure variables:

- *STARTING:* 1 for women taking HRT who started less than 60 days ago, 0 otherwise
- *RECENT:* 1 for women taking HRT who started between 60 and 365 days previously, 0 otherwise
- *LONGTERM:* 1 for women taking HRT who started more than a year ago, 0 otherwise

The hypothesis was that the coefficients for STARTING would be positive (increased risk), but that coefficients for RECENT and LONGTERM would be lower, and possibly negative. They found that the hazard ratio e^b for STARTING was 2.16, with a 95% confidence interval, 0.94 to 4.95, not quite excluding 1. The hazard ratio for LONGTERM was 0.76, a with 95% confidence interval 0.42 to 1.36.

Time-dependent covariates are not always appropriate. In particular, they do not result in useful predictive models: In order to estimate the chance of surviving for the next five years, it is necessary to have covariate values for the next five years to plug into the model.

Even when time-dependent models are appropriate, they involve significantly more complex computation, however, good facilities for time-dependent Cox models are now available in many major statistics packages. Computational details vary between packages, and between versions of the same package, but the basic approach is to break each person's data into many short time intervals on which their covariates are constant. These time intervals are treated as if they came from separate people, which is valid as long as each person can have only one event.

Time-dependent covariates are discussed in many of the recent textbooks on survival analysis, including Therneau and Grambsch [2000], Klein and Moeschberger [1997], and Kleinbaum [1996] and in older references such as Kalbfleisch and Prentice [1980] and Breslow and Day [1987].

16.10.2 Stratification in the Cox Model

The Cox model, which assumes that hazards are proportional over time, can be extended to a stratified model in which hazards need only be proportional within the same stratum and can differ arbitrarily between strata. Stratification can be useful when a small number of important variables do not satisfy the proportional hazards assumption. In addition to the usual difficulties that occur with stratifying on too many variables, the stratified model also suffers from the fact that it is not possible to test the effects of the stratifying variables.

For example, Lumley et al. [2002] constructed a predictive model for the risk of stroke in elderly people. The rates of stroke were not proportional between men and women, so a model stratified by gender was used. Instead of a single underlying survival curve $S_o(t)$, the model has curves $S_m(t)$ for men and $S_w(t)$ for women. The hazard ratio for other covariates, such as diabetes or smoking, is assumed to be constant over time within each stratum. The hazard ratio may be constrained to be the same for women and men or allowed to differ. As Table 16.9 shows, the stroke prediction model used a common hazard ratio for diabetes in men and women, but the hazard ratio for history of heart disease was allowed to differ between men and women. A Java applet showing this model is linked from the Web appendix.

Table 16.9 Stratified Cox Model for Risk of Stroke

	Mean		Coefficient	
	2495 Men	3393 Women	Men	Women
Left ventricular hypertrophy by ECG (%)	5.1	4.9	0.501	
Diabetes (%)	14.9	12.5	0.521	
Elevated fasting glucose (%)	19.0	14.4	0.347	
Creatinine >1.25 mg/dL (%)	39.6	8.1	0.141	
Time to walk 15 ft (s)	5.5	6.0	0.099	
Systolic blood pressure (mmHg)	143	144	172/10	
History of heart disease (%)	26.5	16.1	0.445	0.073
Atrial fibrillation by ECG (%)	3.5	2.1	0.4097	1.346
Age (yr)	73	73	0.382/10	0.613/10

16.10.3 Left Truncation

In the examples discussed so far, the survival time has been measured from the beginning of the study, so that all subjects are under observation from time 0 under they die or are censored. There are situations where this is not feasible. Consider a study of occupational exposure to a potential carcinogen, where workers at a factory are interviewed about their past exposure and other risk factors such as cancer, and then followed up.

It would be desirable to set time zero to be when each worker was first employed at the factory rather than the date when the study was performed. This would more accurately approximate the ideal study that recruited everyone as they entered employment and followed them for the rest of their lives. There is a serious complication, however. Workers who died before the study started will not be included, making the sample biased. This phenomenon is called *left truncation*. Truncation is not quite the same as censoring, although both involve incomplete information. With censoring, we have information on only part of a person's life. With truncation, we have no information on some people and complete information on others.

The solution to left truncation is similar to the solution to right censoring. If we break time up into short intervals, each person contributes information about the probability of surviving through an interval given that one is alive at the start of the interval. These probabilities can be multiplied to give an overall survival probability. Most statistical software will allow you to specify an *entry* time as well as a survival or censoring time, and will fit Cox regression models to data specified in this way.

In the occupational exposure example, consider a worker who started at the factory in 1955, who entered the study in 1985, and who died in 1995. We want to take time to be 0 in 1955, so the *entry* time is $1985 - 1955$, or 30 years, and the survival time is $1995 - 1955$, or 40 years. Another worker might have started at the factory in 1975, been recruited in 1985, and still be alive at the end of the study in 2000. This would give an *entry* time of $1985 - 1975$, or 10 years, and a censoring time of $2000 - 1975$, or 25 years.

Breslow and Day [1987] discuss an example of this sort in some detail, comparing the effects of placing time zero at different events in analyzing the cancer risks of workers at a nickel refinery.

16.10.4 Other References Dealing with Survival Analysis and Heart Transplant Data

The first heart transplant data has been used extensively as an illustration in the development of survival techniques. Further references are Mantel and Byar [1974], Turnbull et al. [1974], and Crowley and Hu [1977].

NOTES

16.1 Recurrent Events

Some events can occur more than once for the same person. Although it is usually possible to study just the time until the first event, it may be useful to incorporate subsequent events to increase the information available. The hazard formulation of survival analysis extends naturally to recurrent events. The hazard (now often called the *intensity*) is still defined in terms of the probability of having an event in a small interval of time, conditional on being alive and under observation. The difference is that now a person can still be alive and under observation after an event occurs. Although computation for recurrent event models is fairly straightforward, there are a number of important methodologic issues that need to be considered. In particular, there is no really satisfactory way to handle recurrent events and deaths in the same analysis. Volume 16, No. 18 of *Statistics in Medicine* (April 30, 1997) has a number of papers discussing these issues. The Web appendix to this chapter includes some examples of analyses of recurrent infections in children with chronic granulomatous disease, a genetic immune deficiency.

16.2 More on the Hazard Rate and Proportional Hazards

Many of the concepts presented in this chapter are analogs of continuous quantities that are best defined in terms of calculus. If the survival function is $S(t)$, its probability density function is

$$f(t) = -\frac{dS(t)}{dt}$$

The hazard rate is then

$$h(t) = \frac{f(t)}{S(t)}$$

From this it follows that the survival is found from the hazard rate by the equation

$$S(t) = e^{-\int_0^t h(x)\,dx}$$

The quantity

$$H(t) = \int_0^t h(x)\ dx = -\log S(t)$$

is called the *cumulative hazard*. Under the proportional hazards assumption, the cumulative hazards H_1 and H_2 for two groups of cases are related by

$$H_1(t) = \lambda \times H_2(t)$$

so

$$\log H_1(t) = \log \lambda + \log H_2(t)$$

16.3 Log-Rank Statistic and Log-Rank Statistic for Stratified Data

We present the statistic using some matrix ideas. The notation is that of Section 16.6 on the log-rank test. For the ith group at the mth time of a death (or deaths), there were d_{im} deaths and l_{im} persons at risk. Suppose that we have k groups and M times of death. For $i, j = 1, \ldots, k$, let

$$V_{ij} = \begin{cases} \sum_{m=1}^{M} \dfrac{l_{im}(T_m - l_{im})D_m(T_m - D_m)}{T_m^2(T_m - 1)}, & i = j \\ \sum_{m=1}^{M} \dfrac{-l_{im}l_{jm}D_m(T_m - D_m)}{T_m^2(T_m - 1)}, & i \neq j \end{cases}$$

Define the $(k-1) \times (k-1)$ matrix V by

$$V = \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1,k-1} \\ V_{21} & & & \vdots \\ \vdots & & & \vdots \\ V_{k-1,1} & \cdots & \cdots & V_{k-1,k-1} \end{pmatrix}$$

Define vectors of observed and expected number of deaths in groups $1, 2, \ldots, k-1$ by

$$\mathbf{O} = \begin{pmatrix} O_1 \\ \vdots \\ O_{k-1} \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} E_1 \\ \vdots \\ E_{k-1} \end{pmatrix}$$

The log-rank statistic is

$$(\mathbf{O} - \mathbf{E})' V^{-1} (\mathbf{O} - \mathbf{E})$$

where $'$ denotes a transpose and -1 a matrix inverse. If there are $s = 1, \ldots, S$ strata, for each stratum we have **O, E,** and V. Let these values be indexed by s to denote the strata. The log-rank statistic is

$$\left[\sum_{s=1}^{S} (\mathbf{O}_s - \mathbf{E}_s)\right]' \left(\sum_{s=1}^{S} V_s\right)^{-1} \left[\sum_{s=1}^{S} (\mathbf{O}_s - \mathbf{E}_s)\right]$$

16.4 Estimating the Probability Density Function in Life Table Methods

The density function in the interval from $x(i)$ to $x(i+1)$ for the life table is estimated by

$$f_i = \frac{P_i - P_{i+1}}{x(i+1) - x(i)}$$

The standard error of f_i is estimated by

$$\frac{p_i q_i}{\sqrt{x(i+1) - x(i)}} \left(\sum_{j=1}^{i-1} \frac{q_j}{l'_j p_j} + \frac{p_i}{l'_i q_i}\right)^{1/2}$$

16.5 Other Confidence Intervals for the Survival Function

Direct use of Greenwood's formula to construct confidence intervals in small samples can lead to confidence intervals that cross 0% or 100% survival. Even when this does not occur, the confidence intervals do not perform very well. Better confidence intervals are obtained by multiplying, rather than adding, the same quantity above and below the estimated survival function. That is, the confidence interval is given by

$$\left[\hat{S}(t) \times \exp\left(-z_{\alpha/2} \frac{\text{SE}(\hat{S}(t))}{\hat{S}(t)}\right), \hat{S}(t) \times \exp\left(z_{\alpha/2} \frac{\text{SE}(\hat{S}(t))}{\hat{S}(t)}\right)\right]$$

Bie et al. [1987] studied this interval and a more complicated one based on transforming $S(t)$ to $\arcsin\{\exp[-S(t)/2]\}$ and found that both performed well even with only 25 observations, half of which were censored.

16.6 Group Expected Survival

The baseline survival curve $S_0(t)$ estimates the survival probability at time t for a person whose covariates equal the average of the population. This is not the same as the survival curve expected for the population $S(t)$ as estimated by the Kaplan–Meier method. The population curve $S(t)$

decreases faster than $S_0(t)$ initially, as those with worse-than-average covariates die and then flattens out relative to $S_0(t)$, as the remaining sample has better-than-average covariates. The difference between $S(t)$ and $S_0(t)$ is more pronounced when covariate effects are strong and when there is little censoring.

The relationship between the curves is that the population curve is the average of all the predicted individual survival curves:

$$S(t) = \sum_i S_0(t)^{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

This relationship can be used to predict the population curve for a new population and compare it to the expected population, an extension of the direct standardization of rates in Chapter 15. For example, the predictions of a Cox model can be validated in a new population by dividing the new population into groups and comparing the expected $S(t)$ for each group with the observed survival curve calculated by the Kaplan–Meier method.

Example 16.5. (*continued*) Figure 16.18 compares the expected and observed survival rates for the 106 nonrandomized patients from the Mayo Clinic PBC data. These patients were divided into three equal groups based on the risk predicted by the Mayo model. The Kaplan–Meier survival curve and the group expected survival curve were calculated for each of the three groups. The relatively smooth lines are the expected survival; the stepped lines are the Kaplan–Meier estimates. There is no suggestion that the expected and observed curves differ importantly.

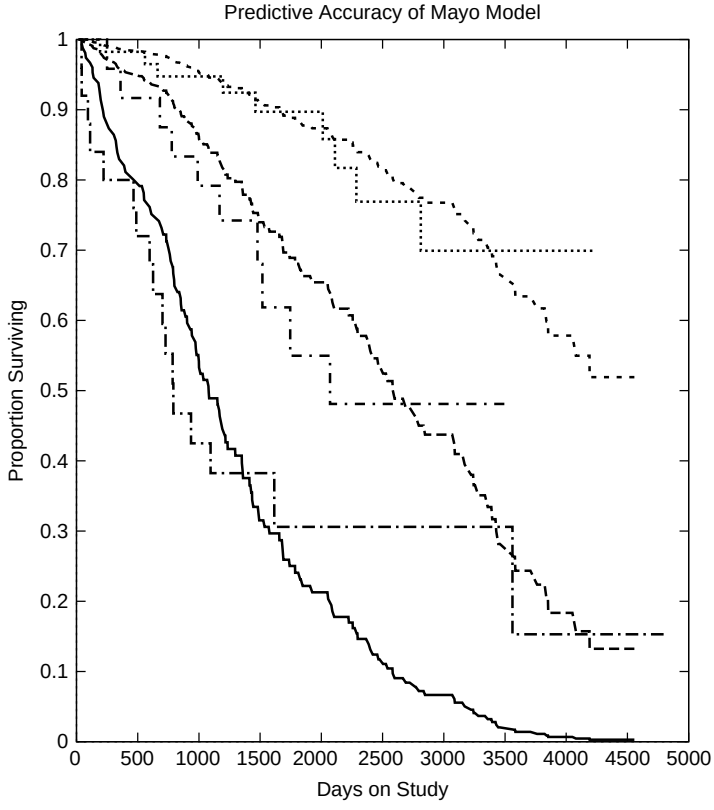


Figure 16.18 Expected and observed survival curves for three groups of nonrandomized patients.

For a stratified life table analysis, the same calculation of expected survival can be done more easily. In this context it is called the method of *direct adjustment*. Suppose that we want to compare survival in treatment groups $j = 1, 2$ and we have strata $i = 1, 2, \ldots, m$. We calculate the survival curve for each treatment group in each stratum $S_{ij}(t)$ and then add up over strata

$$S_j(t) = \sum_{i=1}^{m} S_{ij}(t) r_i$$

where r_i is the proportion of subjects in stratum i.

16.7 Competing Risks

In certain situations one is only interested in certain causes of death that may be linked to the disease in question. For example, in a study of heart disease a death in a plane crash might be considered an unreasonable endpoint to attribute to the disease. It is tempting to censor people who die of genuinely unrelated causes. This cannot be true *noninformative censoring*, as someone who dies in a plane crash certainly has a reduced (zero) risk of heart disease in the future. On the other hand, there seems to be no way that these deaths would bias the remaining sample. It turns out that conclusions from Cox regression in this case are basically valid but that estimated survival curves need to be rethought. Such endpoints are called *competing risks*.

In a more complicated version of the problem, there is often interest in the effects of a treatment on more than one type of event. Lowering blood pressure reduces the risk of death from stroke, heart attack, cardiac arrest, and congestive heart failure, but different drugs may affect these events differently. Inference for these *dependent* competing risks is much more difficult and is complicated further by the fact that it is theoretically impossible to determine whether competing risks are dependent or independent. When all the events are rare, as in primary prevention of cardiovascular disease, ignoring the competing-risks problem may be a satisfactory practical approach. With more common events, this is not possible.

In some cases it is appropriate to treat deaths from other causes as indicating indefinitely long "survival" for the cause of interest. For example, consider a study of time to stroke in elderly people (e.g., Section 16.10.2). If a subject dies from breast cancer at 3.5 years follow-up, her chance of ever having a stroke is known exactly: She never will. This can be represented by censoring her observation time not at the time of death but at a time after the end of the study. The resulting survival curve will estimate the proportion of people who have not had strokes, which will not decrease to zero as follow-up time increases. In other cases this approach is undesirable because decreases in stroke risk and increases in other risks have the same impact—in a clinical trial of stroke prevention one would not want to declare the treatment successful just because it made people die of other causes.

Kalbfleisch and Prentice [2003], Gross and Clark [1975], and Prentice et al. [1978] discuss such issues. Pepe and Mori [1993] discuss alternatives to estimating the cause-specific survival function. Misuse of the cause-specific survival function has been an important issue in radiation oncology and is discussed by Gelman et al. [1990]. The impossibility of testing for dependent competing risks was shown by Tsiatis [1978]. The proof is highly technical but the result should be intuitively plausible: No data are available after censoring, so there should be no way to tell if survival is the same as for noncensored people.

A related issue is multivariate failure time, where events of different types can be observed for the same person. These could be ordered events, such as cancer recurrence and death; multiple versions of the same event, such as time to vision impairment in left and right eyes; or separate events, such as time to marriage and time to having children. Therneau and Grambsch [2000] discuss multivariate failure times, as does Lin [1994]. Somewhat surprisingly, this is a more tractable problem than competing risks.

16.8 Counting Process Notation

Many modern books on survival analysis and most recent statistical papers on the subject use a different mathematical notation from ours, the *counting process notation*. We have described each person's data by a covariate vector $\mathbf{X}_i$, an observation time T_i, and a censoring indicator Δ_i. The counting process notation replaces the time and censoring indicator with two functions of time: $N_i(t)$, which counts the number of times the person has been observed to "die" by time t, and $Y_i(t)$, which is 1 when the person is under observation and 0 otherwise. The covariate vector is usually called $\mathbf{Z}_i(t)$ rather than $\mathbf{X}_i$.

For ordinary survival data this means $N_i(t) = 0$ and $Y_i(t) = 1$ for $t < T_i$, $N_i(t) = \Delta_i$ and $Y_i(t) = 1$ for $t = T_i$, and $N_i(t) = \Delta_i$ and $Y_i(t) = 0$ for $t > T_i$. The notation $dN_i(t)$ means the jump in N_i at time t. This is zero except at the time of a death, when it is 1.

As a final complication, integral notation is used to indicate sums over a time point. For example, the notation $\int Z_i(t)\, dNi(t)$ means the sum of $Z_i(t) \times dN_i(t)$ over all time points. As $dNi(t) = 0$ except at the time of death, this is 0 if the person is censored and is $Z_i(T_i)$ if the person dies at time T_i.

This apparently cumbersome notation was introduced initially for purely mathematical reasons. It becomes more obviously useful when handling recurrent events [when $N_i(t)$ counts the number of events that have occurred], or left-truncation, when $Y_i(t) = 0$ before entry into the study to indicate that a death at that time would not have been observed. Klein and Moeschberger [1997] provide a reasonably accessible treatment of survival analysis using counting process notation.

PROBLEMS

The first four problems deal with the life table or actuarial method of estimating the survival curve. In each case, fill in the question marks from the other numbers given in the table.

16.1 Example 16.2 deals with chest pain in groups in the Coronary Artery Surgery Study; all times are in days. The life table for the individuals with chest pain thought probably not to be angina is given in Table 16.10.

16.2 From Example 16.2 for patients with chest pain thought definitely to be angina the life table is as given in Table 16.11.

16.3 Patients from Example 16.4 on a beta-blocking drug are used here and those not on a beta-blocking drug in Problem 16.4. The life table for those using such drugs at enrollment is given in Table 16.12.

16.4 Those not using beta-blocking drugs have the survival experience shown in Table 16.13.

16.5 Take the Stanford heart transplant data of Example 16.3. Place the data in a life table analysis using 50-day intervals. Plot the data over the interval from zero to 300 days. (Do not compute the Greenwood standard errors.)

16.6 For Problem 16.1, compute the hazard function (in probability of dying/day) for intervals:

- **(a)** 546–637
- **(b)** 1092–1183
- **(c)** 1456–1547

16.7 For the data of Problem 16.2, compute the hazard rate for the patients:

- **(a)** 0–91
- **(b)** 91–182
- **(c)** 819–910

Table 16.10 Life Table for Patients with Chest Pain Probably Not Angina

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	2404	2404.0	2	0	0.0008	0.9992	?
91.0–181.9	2402	?	2	0	0.0008	0.9983	?
182.0–272.9	2400	2400.0	?	0	0.0021	0.9963	0.001
273.0–363.9	2395	2395.0	6	0	?	0.9938	0.002
364.0–454.9	?	2388.0	4	2	0.0017	0.9921	0.002
455.0–545.9	2383	2383.0	3	0	0.0013	?	0.002
546.0–636.9	2380	2380.0	7	0	0.0029	0.9879	0.002
637.0–727.9	2373	?	12	300	?	?	0.003
728.0–818.9	2061	2051.5	?	19	0.0015	0.9812	0.003
819.0–909.9	?	2039.0	1	0	0.0005	0.9807	0.003
910.0–1000.9	2038	2037.0	2	?	0.0010	0.9797	0.003
1001.0–1091.9	2034	?	3	517	0.0017	0.9781	0.003
1092.0–1182.9	1514	1494.0	3	40	0.0020	0.9761	0.003
1183.0–1273.9	1471	1471.0	4	0	?	0.9734	0.004
1274.0–1364.9	1467	1466.5	1	1	0.0007	0.9728	0.004
1365.0–1455.9	?	1144.0	1	642	0.0009	0.9719	0.004
1456.0–1546.9	822	777.5	1	?	0.0013	0.9707	0.004
1547.0–1637.9	732	732.0	1	0	0.0014	?	0.004
1638.0–1728.9	731	730.0	2	2	0.0027	0.9667	0.004
1729.0–1819.9	727	449.0	1	?	0.0022	0.9645	0.005

Table 16.11 Life Table for Patients with Definite Angina

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	426	426.0	2	?	0.0047	0.9953	0.003
91.0–181.9	?	424.0	2	0	0.0047	0.9906	?
182.0–272.9	422	?	3	0	?	?	0.006
273.0–363.9	419	419.0	0	0	0.0000	0.9836	0.006
364.0–454.9	419	419.0	1	0	0.0024	0.9812	0.007
455.0–545.9	418	417.5	?	1	0.0024	0.9789	0.007
546.0–636.9	416	416.0	1	0	0.0024	0.9765	0.007
637.0–727.9	415	382.0	0	?	0.0000	0.9765	0.007
728.0–818.9	349	343.0	0	11	0.0000	0.9765	0.007
819.0–909.9	338	338.0	1	0	0.0030	0.9736	0.008
910.0–1000.9	337	336.5	0	1	0.0000	0.9736	0.008
1001.0–1091.9	336	?	1	97	?	?	0.009
1092.0–1182.9	238	232.5	0	11	0.0000	0.9702	0.009
1183.0–1273.9	227	?	1	1	0.0044	0.9660	0.010
1274.0–1364.9	?	224.5	1	1	0.0045	0.9617	0.010
1365.0–1455.9	?	170.0	0	106	0.0000	0.9617	0.010
1456.0–1446.9	117	114.0	?	6	0.0000	0.9617	0.010
1547.0–1637.9	?	?	0	1	0.0000	0.9617	0.010
1638.0–1728.9	110	109.5	0	1	0.0000	0.9617	0.010
1729.0–1819.9	109	65.5	0	87	0.0000	0.9617	0.010

Table 16.12 Life Table for Patients Taking a β-Blocker

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	4942	4942.0	?	0	0.0097	0.9903	0.001
91.0–181.9	4894	4894.0	33	0	0.0067	0.9836	0.002
182.0–272.9	4861	4861.0	?	?	0.0058	0.9779	?
273.0–363.9	4833	4832.5	28	1	0.0058	0.9723	0.002
364.0–454.9	4804	4804.0	17	0	0.0035	?	0.002
455.0–545.9	4787	4786.5	29	1	?	?	0.003
546.0–636.9	4757	4757.0	22	0	0.0046	0.9585	0.003
637.0–727.9	4735	4376.0	25	718	0.0057	0.9530	0.003
728.0–818.9	?	?	?	62	0.0043	0.9489	0.003
819.0–909.9	3913	3912.0	23	2	?	0.9434	0.003
910.0–1000.9	3888	3884.5	19	7	0.0049	0.9388	0.004
1001.0–1091.9	?	?	?	1191	0.0040	0.9350	0.004
1092.0–1182.9	2658	2624.5	14	67	0.0053	0.9300	0.004
1183.0–1273.9	2577	2576.5	11	1	0.0043	0.9261	0.004
1274.0–1364.9	2565	2561.0	15	8	?	0.9206	0.004
1365.0–1455.9	2542	1849.5	12	1385	0.0065	0.9147	0.005
1456.0–1446.9	1145	1075.0	5	?	0.0047	?	0.005
1547.0–1637.9	1000	999.0	4	2	0.0040	0.9068	0.005
1638.0–1728.9	994	989.0	4	10	0.0040	0.9031	0.006
1729.0–1819.9	980	580.0	5	800	0.0086	0.8953	0.006

Table 16.13 Life Table for Patients Not Taking a β-Blocker

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	6453	?	45	0	?	?	?
91.0–181.9	6408	?	28	0	?	?	?
182.0–272.9	6380	?	42	0	?	?	?
273.0–363.9	6338	?	25	2	?	?	?
364.0–454.9	6311	6310.0	24	2	0.0038	0.9746	0.002
455.0–545.9	6285	6285.0	32	0	0.0051	0.9696	0.002
546.0–636.9	6253	6253.0	?	0	0.0048	0.9650	0.002
637.0–727.9	6223	5889.0	23	668	0.0039	0.9612	0.002
728.0–818.9	?	?	23	40	0.0042	0.9572	0.003
819.0–909.9	?	5467.0	17	4	?	0.9542	0.003
910.0–1000.9	5448	5444.5	23	7	0.0042	0.9502	0.003
1001.0–1091.9	5418	4787.4	25	1261	0.0052	0.9452	0.003
1092.0–1182.9	4132	4082.0	?	100	0.0054	0.9401	0.003
1183.0–1273.9	4010	4010.0	23	0	0.0057	0.9347	0.003
1274.0–1364.9	3987	3981.0	18	?	0.0020	0.9329	0.003
1365.0–1455.9	3967	3100.0	13	1734	0.0042	0.9289	0.003
1456.0–1446.9	2220	2104.0	13	?	0.0062	0.9232	0.004
1547.0–1637.9	1975	1974.0	?	2	0.0020	0.9213	0.004
1638.0–1728.9	1969	1961.5	11	15	0.0056	0.9162	0.004
1729.0–1819.9	1943	1212.0	17	7	0.0058	0.9109	0.005

16.8 Data used by Pike [1966] are quoted in Kalbfleisch and Prentice [2003]. Two groups of rats with different pretreatment regimes were exposed to the carcinogen DBMA. The time to mortality from vaginal cancer in the two groups was: (* indicates a censored observation):

- *Group 1:* 143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 216*, 220, 227, 230, 234, 244*, 246, 265, 304
- *Group 2:* 142, 156, 163, 198, 204*, 205, 232, 232, 233, 233, 233, 233, 239, 240, 261, 280, 280, 296, 296, 323, 344*

(a) Compute and graph the two product limit curves of the groups.

(b) Compute the expected number of deaths in each group and the value of the approximation $[\sum(O - E)^2/E]$ to the log-rank test. Are the survival times different in the two groups at the 5% significance level?

(c) How close is the approximate log-rank statistic to the exact value reported by your favorite statistics software?

16.9 The data of Problems 16.3 and 16.4, where stratified into the 30 strata discussed in the text, give the results shown in Table 16.14.

(a) What are the observed and expected numbers in the two groups? (Why do you have to add only three columns?)

(b) Two strata (12 and 17) are significant with $p = 0.02$. If the true survival patterns (in the conceptual underlying populations) are the same, does this surprise you?

(c) What is $\sum(O-E)^2/E$? How does this compare to the more complicated log-rank statistic which can be shown to be 6.510?

16.10 The paper by Chaitman et al. [1981] studied patients with left main coronary artery disease, as discussed in Example 16.4. Separate Cox survival runs were performed for the medical and surgical groups. The data are presented in Table 16.15. The survival, at the mean covariate values, for one, two, and three years are given by $S_0(1)$, $S_0(2)$, and $S_0(3)$, respectively. The zero–one variables are 0 for no and 1 for yes. Consider five patients with the variable values given in Table 16.16.

(a) What is the estimate of the two-year medical survival for patients 1, 2, and 3?

(b) What is the estimate of the three-year surgical survival for patients 4 and 5?

(c) What are the estimated one-year medical and one-year surgical survival rates for patient 1? For patient 3?

(d) What is the logarithm of the instantaneous relative risk for two individuals treated medically who differ by 20 years, but otherwise have the same values for the variables? What is the instantaneous relative risk?

(e) What is the instantaneous relative risk due to diabetes (yes vs. no) for surgical cases?

***(f)** What is the standard error for the LV score coefficient for the surgical group? For the age coefficient for the medical group? Form an approximate 95% confidence interval for the age coefficient in the medical group.

16.11 Alderman et al. [1983] studied the medical and surgical survival of patients with poor left ventricular function; that is, they studied patients whose hearts pumped poorly. Their model (in one analysis) included the following variables:

Table 16.14 Drug Use Data for Problem 16.9

	Drug Use		No Drug Use		
Stratum	Obs.	Exp.	Obs.	Exp.	p-Value
1	45	43.30	71	72.70	0.74
2	2	2.23	4	3.77	0.84
3	0	0.20	1	0.80	0.54
4	27	28.54	37	35.46	0.69
5	6	4.84	5	6.16	0.48
6	2	0.76	1	2.24	0.08
7	20	16.87	20	23.13	0.31
8	4	5.25	10	8.75	0.49
9	3	3.17	5	4.83	0.90
10	18	16.55	21	22.45	0.63
11	5	6.68	9	7.32	0.35
12	8	4.58	1	4.42	0.02
13	21	16.04	13	17.96	0.08
14	6	8.95	16	13.05	0.19
15	2	2.63	5	4.37	0.61
16	16	16.82	20	19.81	0.78
17	5	9.86	15	10.14	0.02
18	4	4.40	5	4.60	0.78
19	7	11.48	16	11.52	0.06
20	10	8.98	8	9.02	0.62
21	4	2.89	2	3.11	0.34
22	21	19.67	24	25.33	0.68
23	13	14.59	20	18.41	0.56
24	5	6.86	11	9.14	0.32
25	35	29.64	21	26.36	0.14
26	18	14.82	13	16.18	0.24
27	7	8.89	8	6.11	0.29
28	22	17.08	18	22.92	0.10
29	11	11.24	15	14.76	0.92
30	8	9.11	8	6.89	0.52

- *Impairment:* impairment due to congestive heart failure (CHF); 0 = never had CHF; 1 = had CHF but have no impairment; 2 = mild CHF impairment; 3 = moderate CHF impairment; and 4 = severe CHF impairment
- *Age:* in years
- *LMCA:* percent of diameter narrowing of the left main coronary artery
- *EF:* ejection fraction, the percent of the blood in the pumping chamber (left ventricle) of the heart pumped out during heartbeat
- *Digitalis:* Does the patient use digitalis? 1 = yes, 2 = no
- *Therapy:* 1 = medical; 2 = surgical
- *Vessel:* number (0 to 3) of vessels diseased with 70% or more stenosis

The β values and their standard errors are given in Table 16.17.

(a) Fill in the chi-square value column where missing.

(b) For which variables is $p < 0.10$? 0.05? 0.01? 0.001?

Table 16.15 Significant Independent Predictors of Mortality in Patients with Greater Than 50% Stenosis of the Left Main Coronary Artery

	Medical Group		Surgical Group	
Variable	X^{2a}	β_i	X^{2a}	β_i
LV score (5–30)	19.12	0.1231	18.54	0.1176
CHF score (0–4)	9.39	0.2815	8.16	0.2964
Age	14.42	0.0526	6.98	0.0402
% LMCA stenosis (50–100)	19.81	0.0293	—	—
Hypertension (0–1)	9.41	0.7067	5.74	0.5455
Left dominance (0–1)	—	—	10.23	1.0101
Smoking (1 = never, 2 = ever, 3 = present)	7.26	0.4389	—	—
MI status (0 = none, 1 = single, 2 = multiple)	4.41	−0.2842	—	—
Diabetes (0–1)	—	—	4.67	0.5934
Total chi-square	90.97	—	67.11	—
Degrees of freedom	7	—	6	—
p	<0.0001	—	<0.0001	—
Constant c		−7.2956		−3.7807
Estimated survival				
$S_0(1)$		0.90		0.97
$S_0(2)$		0.83		0.95
$S_0(3)$		0.76		0.93

[a] Adjusted chi-square (X^2) statistics were computed with all variables considered together. Chi-square >6.63 corresponds to $p < 0.01$, and chi-square >10.83, to $p < 0.001$. β, beta coefficient; CHF, congestive heart failure; LMCA, left main coronary artery; LV, left ventricular; MI, myocardial infarction. Dashes indicate a variable not in the particular model.

Table 16.16 Variable Data for Problem 16.10

	Patient Number				
Variable	1	2	3	4	5
LV score	13	5	7	8	12
CHF score	2	0	1	0	3
Age	71	62	42	55	46
Percent LMCA stenosis	75	90	50	70	95
Hypertension	No	Yes	Yes	No	No
Left dominance	No	No	No	Yes	No
Smoking	Ever	Present	Ever	Ever	Present
MI status	Multiple	None	Single	None	Single
Diabetes	No	No	No	Yes	No

Table 16.17 Data for Problem 16.11

Variable	Beta	Standard Error	Chi-Square
Impairment	0.2677	0.0505	?
Age	0.0430	0.0084	26.02
LMCA	0.0090	0.0024	?
EF	−0.0362	0.0098	?
Digitalis	−0.3802	0.1625	?
Therapy	−0.3418	0.1458	5.49
Vessel	0.2081	0.1012	4.23
Constant	−1.2873		

Table 16.18 Variable Data for Problem 16.11

	Patient Number		
Variable	1	2	3
Impairment	Severe	Mild	Moderate
Age	64	51	59
LMCA	50%	0%	0%
EF	15	32	23
Digitalis	Yes	Yes	Yes
Therapy	Medical	Surgical	Medical
Vessel	3	2	3

(c) What is the instantaneous relative risk of 70% LMCA compared to 0% LMCA?

(d) Consider three patients with the covariate values given in Table 16.18.

At the mean values of the data, the one- and two-year survival were 88.0% and 80.16%, respectively. Find the probability of one- and two-year survival for these three patients.

(e) With this model: (**i**) Can surgery be better for one person and medical treatment for another? Why? What does this say about unthinking application of the model? (**ii**) Under surgical therapy, can the curve cross over the estimated medical survival for some patients? For heavy surgical mortality, would a proportional hazard model always seem appropriate?

16.12 The Clark et al. [1971] heart transplant data were collected as follows. People with failing hearts waited for a donor heart to become available; this usually occurred within 90 days. However, some patients died before a donor heart became available. Figure 16.19 plots the survival curves of (1) those not transplanted (indicated by circles) and (2) the transplant patients from time of surgery (indicated by the triangles).

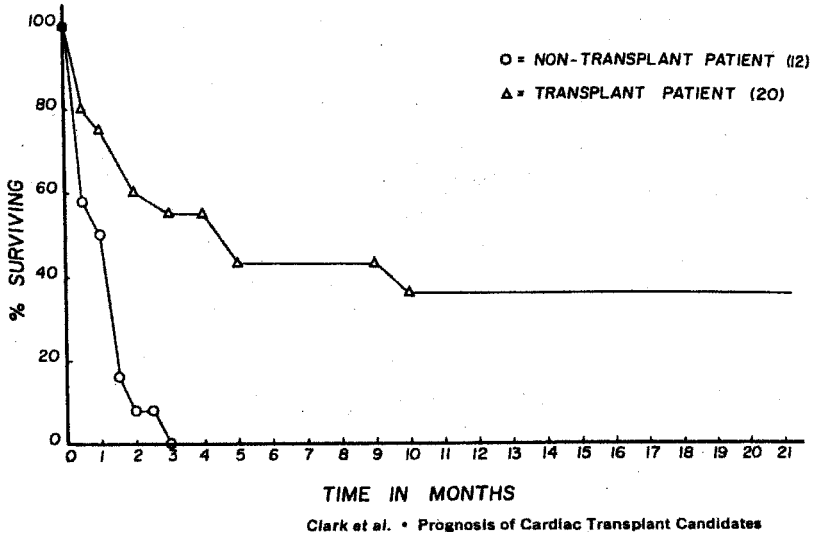


Figure 16.19 Survival calculated by the life table method. Survival for transplanted patients is calculated from the time of operation; survival of nontransplanted patients is calculated from the time of selection for transplantation.

(a) Is the survival of the nontransplanted patients a reasonable estimate of the nonoperative survival of candidates for heart transplant? Why or why not?

(b) Would you be willing to conclude from the figure (assuming a statistically significant result) that 1960s heart transplant surgery prolonged life? Why or why not?

(c) Consider a Cox model fitted with transplantation as a time-dependent covariate:

$$h_i(t) = h_0(t)e^{\exp(\alpha+\beta\times\text{TRANSPLANT}(t))}$$

The estimate of β is 0.13, with a 95% confidence interval $(-0.46, 0.72)$. (Verify this if you have access to suitable software.) What is the interpretation of this estimate? What would you conclude about whether 1960s-style heart transplant surgery prolongs life?

(d) A later, expanded version of the Stanford heart transplant data includes the age of the participant and the year of the transplant (from 1967 to 1973). Adding these variables gives the following coefficients:

Variable	β	SE(β)	p-value
Transplant	−0.030	0.318	0.92
Age	0.027	0.014	0.06
Year	−0.179	0.070	0.01

What would you conclude from these results, and why?

16.13 Simes et al. [2002] analyzed results from the LIPID trial that compared the cholesterol-lowering drug pravastatin to placebo in preventing coronary heart disease events. The outcome defined by the trial was time until fatal coronary heart disease or nonfatal myocardial infarction.

(a) The authors report that Cox model with one variable coded 1 for pravastatin and 0 for placebo gives a reduction in the risk of 24% (95% confidence interval, 15 to 32%). What is the hazard ratio? What is the coefficient for the treatment variable?

(b) A second model had three variables: treatment, HDL (good) cholesterol level after treatment, and total cholesterol level after treatment. The estimated risk reduction for the treatment variable in this model is 9% (95% confidence interval, −7 to 22%). What is the interpretation of the coefficient for treatment in this model?

16.14 In an elderly cohort, the death rate from heart disease was approximately constant at 2% per year, and from other causes was approximately constant at 3% per year.

(a) Suppose that a researcher computed a survival curve for time to heart disease death, treating deaths from other causes as censored. As described in Section 16.9.1, the survival function would be approximately $S(t) = e^{-0.02t}$. Compute this function at 1, 2, 3, . . . ,10 years.

(b) Another researcher computed a survival curve for time to non-heart-disease death, censoring deaths from heart disease. What would the survival function be? Compute it at 1, 2, 3, . . . ,10 years.

(c) What is the true survival function for deaths from all causes? Compare it to the two cause-specific functions and discuss why they appear inconsistent.

REFERENCES

Alderman, E. L., Fisher, L. D., Litwin, P., Kaiser, G. C., Myers, W. O., Maynard, C., Levine, F., and Schloss, M. [1983]. Results of coronary artery surgery in patients with poor left ventricular function (CASS). *Circulation*, **68**: 785–789. Used with permission from the American Heart Society.

Bie, O., Borgan, Ø., and Liestøl, K. [1987]. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, **14**: 221–223.

Breslow, N. E., and Day, N. E. [1987]. *Statistical Methods in Cancer Research*, Vol. II. International Agency for Research on Cancer, Lyon, France.

Chaitman, B. R., Fisher, L. D., Bourassa, M. G., Davis, K., Rogers, W. J., Maynard, C., Tyras, D. H., Berger, R. L., Judkins, M. P., Ringqvist, I., Mock, M. B., and Killip, T. [1981]. Effect of coronary bypass surgery on survival patterns in subsets of patients with left main coronary disease. *American Journal of Cardiology*, **48**: 765–777.

Clark, D. A., Stinson, E. B., Grieppe, R. B., Schroeder, J. S., Shumway, N. E., and Harrison, D. B. [1971]. Cardiac transplantation in man: VI. Prognosis of patients selected for cardiac transplantation. *Annals of Internal Medicine*, **75**: 15–21.

Crowley, J., and Hu, M. [1977]. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72**: 27–36.

European Coronary Surgery Study Group [1980]. Prospective randomized study of coronary artery bypass surgery in stable angina pectoris: second interim report. *Lancet*, Sept. 6, **2**: 491–495.

Fleming, T. R., and Harrington, D. [1991]. *Counting Processes and Survival Analysis.* Wiley, New York.

Gehan, E. A. [1969]. Estimating survival functions from the life table. *Journal of Chronic Diseases*, **21**: 629–644. Copyright © 1969 by Pergamon Press, Inc. Used with permission.

Gelman, R., Gelber, R., Henderson I. C., Coleman, C. N., and Harris, J. R. [1990]. Improved methodology for analyzing local and distant recurrence. *Journal of Clinical Oncology*, **8**(3): 548–555.

Greenwood, M. [1926]. *Reports on Public Health and Medical Subjects*, No. 33, App. I, The errors of sampling of the survivorship tables. H. M. Stationary Office, London.

Gross, A. J. and Clark, V. A. [1975]. *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley, New York.

Heckbert, S. R., Kaplan, R. C., Weiss, N. S., Psaty, B. M., Lin, D., Furberg, C. D., Starr, J. S., Anderson, G. D., and LaCroix, A. Z. [2001]. Risk of recurrent coronary events in relation to use and recent initiation of postmenopausal hormone therapy. *Archives of Internal Medicine*, **161**(14): 1709–1713.

Holt, V. L., Kernic, M. A., Lumley, T., Wolf, M. E., and Rivara, F. P. [2002]. Civil protection orders and risk of subsequent police-reported violence. *Journal of the American Medical Association*, **288**(5): 589–594

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., and Vittinghoff, E. [1998]. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association*, **280**(7): 605–613.

Kalbfleisch, J. D., and Prentice, R. L. [2003]. *The Statistical Analysis of Failure Time Data*. 2nd edition Wiley, New York.

Kaplan, E. L., and Meier, P. [1958]. Nonparametric estimation for incomplete observations. *Journal of the American Statistical Association*, **53**: 457–481.

Klein, J. P., and Moeschberger, M. L. [1997]. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.

Kleinbaum, D. G. [1996]. *Survival Analysis: A Self-Learning Text*. Springer-Verlag, New York.

Lin, D. Y. [1994]. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, **13**: 2233–2247.

Lumley, T., Kronmal, D., Cushman, M., Monolio, T. A. and Goldstein, S. [2002]. Predicting stroke in the elderly: validation and web-based application. *Journal of Clinical Epidemiology*, **55**: 129–136.

Mann, N. R., Schafer, R. C. and Singpurwalla, N. D. [1974]. *Methods for Statistical Analysis of Reliability and Life Data*. Wiley, New York.

Mantel, N., and Byar, D. [1974]. Evaluation of response time 32 data involving transient states: an illustration using heart transplant data. *Journal of the American Statistical Association*, **69**: 81–86.

Messmer, B. J., Nora, J. J., Leachman, R. E., and Cooley, D. A. [1969]. Survival times after cardiac allographs. *Lancet*, May 10, **1**: 954–956.

Miller, R. G. [1981]. *Survival Analysis*. Wiley, New York.

Parker, R. L., Dry, T. J., Willius, F. A., and Gage, R. P. [1946]. Life expectancy in angina pectoris. *Journal of the American Medical Association*, **131**: 95–100.

Passamani, E. R., Fisher, L. D., Davis, K. B., Russel, R. O., Oberman, A., Rogers, W. J., Kennedy, J. W., Alderman, E., and Cohen, L. [1982]. The relationship of symptoms to severity, location and extent of coronary artery disease and mortality. Unpublished study.

Pepe, M. S., and Mori, M. [1993]. Kaplan–Meier, marginal, or conditional probability curves in summarizing competing risks failure time data. *Statistics in Medicine*, **12:** 737–751.

Pike, M. C. [1966]. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, **26**: 579–581.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. L. [1978]. The analysis of failure times in the presence of competing risks. *Biometrics*, **34**: 541–554.

Simes, R. S., Masschner, I. C., Hunt, D., Colquhoun, D., Sullivan, D., Stewart, R. A. H., Hague, W., Kelch, A., Thompson, P., White, H., Shaw, V., and Torkin, A. [2002]. Relationship between lipid levels and clinical outcomes in the long-term intervention with Pravastatin in ischemic disease (LIPID) trial: to what extent is the reduction in coronary events with Pravastatin explained by on-study lipid levels? *Circulation*, **105**: 1162–1169.

Takaro, T., Hultgren, H. N., Lipton, M. J., Detre, K. M., and participants in the study group [1976]. The Veteran's Administration cooperative randomized study of surgery for coronary arterial occlusive disease: II. Subgroup with significant left main lesions. *Circulation Supplement 3*, **54**: III-107 to III-117.

Therneau, T. M., and Grambsch, P. [2000]. *Modelling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tsiatis, A. A. [1978]. An example of non-identifiability in competing risks. *Scandinavian Actuarial Journal*, 235–239.

Turnbull, B., Brown, B., and Hu, M. [1974]. Survivorship analysis of heart transplant data. *Journal of the American Statistical Association*, **69**: 74–80.

U.S. Department of Health, Education, and Welfare [1976]. *Vital Statistics of the United States, 1974*, Vol. II, Sec. 5, Life tables. U.S. Government Printing Office, Washington, DC.

CHAPTER 17

Sample Sizes for Observational Studies

17.1 INTRODUCTION

In this chapter we deal with the problem of calculating sample sizes in various observational settings. There is a very diverse literature on sample size calculations, dealing with many interesting areas. We can only give you a feeling for some approaches and some pointers for further study.

We start the chapter by considering the topic of screening in the context of adverse effects attributable to drug usage, trying to accommodate both the "rare disease" assumption and the multiple comparison problem. Section 17.3 discusses sample-size considerations when costs of observations are not equal, or the variability is unequal; some very simple but elegant relationships are derived. Section 17.4 considers sample size consideration in the context of discriminant analysis. Three questions are considered: (1) how to select variables to be used in discriminating between two populations in the face of multiple comparisons; (2) given that m variables have been selected, what sample size is needed to discriminate between two populations with satisfactory power; and (3) how large a sample size is needed to estimate the probability of correct classification with adequate precision and power. Notes, problems, and references complete the chapter.

17.2 SCREENING STUDIES

A screening study is a scientific fishing expedition: for example, attempting to relate exposure to one of several drugs to the presence or absence of one or more side effects (disease). In such screening studies the number of drug categories is usually very large—500 is not uncommon—and the number of diseases is very large—50 or more is not unusual. Thus, the number of combinations of disease and drug exposure can be very large—25,000 in the example above. In this section we want to consider the determination of sample size in screening studies in terms of the following considerations: many variables are tested and side effects are rare. A cohort of exposed and unexposed subjects is either followed or observed. We have looked at many diseases or exposures, want to "protect" ourselves against a large Type I error, and want to know how many observations are to be taken. We proceed in two steps: First, we derive the formula for the sample size without consideration of the multiple testing aspect, then we incorporate the multiple testing aspect. Let

$$X_1 = \text{number of occurrences of a disease of interest (per 100,000 person-years, say) in the unexposed population}$$

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

X_2 = number of occurrences (per 100,000 person-years) in the exposed population

If X_1 and X_2 are rare events, $X_1 \sim \text{Poisson}(\theta_1)$ and $X_2 \sim \text{Poisson}(\theta_2)$. Let $\theta_2 = R\theta_1$; that is, the risk in the exposed population is R times that in the unexposed population ($0 < R < \infty$). We can approximate the distributions by using the variance stabilizing transformation (discussed in Chapter 10):

$$Y_1 = \sqrt{X_1} \sim N(\sqrt{\theta_1}, \sigma^2 = 0.25)$$
$$Y_2 = \sqrt{X_2} \sim N(\sqrt{\theta_2}, \sigma^2 = 0.25)$$

Assuming independence,

$$Y_2 - Y_1 \sim N\left(\sqrt{\theta_1}(\sqrt{R} - 1), \sigma^2 = 0.5\right) \tag{1}$$

For specified Type I and Type II errors α and β, the *number of events* n_1 and n_2 in the unexposed and exposed groups required to detect a relative risk of R with power $1 - \beta$ are given by the equation

$$n_1 = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{2(\sqrt{R} - 1)^2}, \qquad n_2 = Rn_1 \tag{2}$$

Equation (2) assumes a two-sided, two-sample test with an equal number of subjects observed in each group. It is an approximation, based on the normality of the square root of a Poisson random variable. If the prevalence, π_1, in the unexposed population is known, the number of subjects per group, N, can be calculated by using the relationship

$$N\pi_1 = n_1 \quad \text{or} \quad N = n_1/\pi_1 \tag{3}$$

Example 17.1. In Section 15.4, mortality was compared in active participants in an exercise program and in dropouts. Among the active participants, there were 16 deaths in 593 person-years of active participation; in dropouts there were 34 deaths in 723 person-years. Using an α of 0.05, the results were not significantly different. The relative risk, R, for dropouts is estimated by

$$R = \frac{34/723}{16/593} = 1.74$$

Assuming equal exposure time in the active participants and dropouts, how large should the sample sizes n_1 and n_2 be to declare the relative risk, $R = 1.74$, significant at the 0.05 level with probability 0.95? In this case we use a two-tailed test and $Z_{1-\alpha/2} = 1.960$ and $Z_{1-\beta} = 1.645$, so that

$$n_1 = \frac{(1.960 + 1.645)^2}{2(\sqrt{1.74} - 1)^2} = 63.4 \doteq 64 \quad \text{and} \quad n_2 = (1.74)n_1 = 111$$

for a total number of observed events $= n_1 + n_1 = 64 + 111 = 175$ deaths. We would need approximately $(111/34) \times 723 = 2360$ person-years exposure in the dropouts and the same number of years of exposure among the controls. The exposure years in the observed data are not split equally between the two groups. We discuss this aspect further in Note 17.1.

If there is only one observational group, the group's experience perhaps being compared with that of a known population, the sample size required is $n_1/2$, again illustrating the fact that comparing two groups requires four times more exposure time than comparing one group with a known population.

Table 17.1 Relationship between Overall Significance Level α, Significance Level per Test, Number of Tests, and Associated Z-Values, Using the Bonferroni Inequality

Number of Tests (K)	Overall α Level	Required Level per Test (α)	Z-Values	
			One-Tailed	Two-Tailed
1	0.05	0.05	1.645	1.960
2	0.05	0.025	1.960	2.241
3	0.05	0.01667	2.128	2.394
4	0.05	0.0125	2.241	2.498
5	0.05	0.01	2.326	2.576
10	0.05	0.005	2.576	2.807
100	0.05	0.0005	3.291	3.481
1000	0.05	0.00005	3.891	4.056
10000	0.05	0.000005	4.417	4.565

We now turn to the second aspect of our question. Suppose that the comparison above is one of a multitude of comparisons? To maintain a per experiment significance level of α, we use the Bonferroni inequality to calculate the per comparison error rate. Table 17.1 relates the per comparison critical values to the number of tests performed and the per experiment error rate. It is remarkable that the critical values do not increase too rapidly with the number of tests.

Example 17.2. Suppose that the FDA is screening a large number of drugs, relating 10 kinds of congenital malformations to 100 drugs that could be taken during pregnancy. A particular drug and a particular malformation is now being examined. Equal numbers of exposed and unexposed women are to be selected and a relative risk of $R = 2$ is to be detected with power 0.80 and per experiment one-sided error rate of $\alpha = 0.05$. In this situation $\alpha^* = \alpha/1000$ and $Z_{1-\alpha^*} = Z_{1-\alpha/1000} = Z_{0.99995} = 3.891$. The required number of events in the unexposed group is

$$n_1 = \frac{(3.891 + 0.842)^2}{2(\sqrt{2} - 1)^2} = \frac{22.4013}{0.343146} = 65.3 \doteq 66$$

$$n_2 = 2n_1 = 132$$

In total, $66 + 132 = 198$ malformations must be observed. For a particular malformation, if the congenital malformation rate is on the order of 3/1000 live births, approximately 22,000 unexposed women and 22,000 women exposed to the drug must be examined. This large sample size is not only a result of the multiple testing but also the rarity of the disease. [The comparable number testing only once, $\alpha^* = \alpha = 0.05$, is $n_1 = \frac{1}{2}(1.645 + 0.842)^2/(\sqrt{2} - 1)^2 = 18$, or 3000 women per group.]

17.3 SAMPLE SIZE AS A FUNCTION OF COST AND AVAILABILITY

17.3.1 Equal-Variance Case

Consider the comparison of means from two independent groups with the same variance σ; the standard error of the difference is

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{4}$$

where n_1 and n_2 are the sample sizes in the two groups. As is well known, for fixed N the standard error of the difference is minimized (maximum precision) when

$$n_1 = n_2 = N$$

That is, the sample sizes are equal. Suppose now that there is a differential cost in obtaining the observations in the two groups; then it may pay to choose n_1 and n_2 unequal, subject to the constraint that the standard error of the difference remains the same. For example,

$$\frac{1}{10} + \frac{1}{10} = \frac{1}{6} + \frac{1}{30}$$

Two groups of equal sample size, $n_1 = n_2 = 10$, give the same precision as two groups with $n_1 = 6$ and $n_2 = 30$. Of course, the total number of observations N is larger, 20 vs. 36.

In many instances, sample size calculations are based on additional considerations, such as:

1. Relative cost of the observations in the two groups
2. Unequal hazard or potential hazard of treatment in the two groups
3. The limited number of observations available for one group

In the last category are case–control studies where the number of cases is limited. For example, in studying sudden infant death syndrome (SIDS) by means of a case–control study, the number of cases in a defined population is fairly well fixed, whereas an arbitrary number of (matching) controls can be obtained.

We now formalize the argument. Suppose that there are two groups, G_1 and G_2, with costs per observations c_1 and c_2, respectively. The total cost, C, of the experiment is

$$C = c_1 n_1 + c_2 n_2 \tag{5}$$

where n_1 and n_2 are the number of observations in G_1 and G_2, respectively. The values of n_1 and n_2 are to be chosen to minimize (maximum precision),

$$\frac{1}{n_1} + \frac{1}{n_2}$$

subject to the constraint that the total cost is to be C. It can be shown that under these conditions the required sample sizes are

$$n_1 = \frac{C}{c_1 + \sqrt{c_1 c_2}} \tag{6}$$

and

$$n_2 = \frac{C}{c_2 + \sqrt{c_1 c_2}} \tag{7}$$

The ratio of the two sample sizes is

$$\frac{n_2}{n_1} = \sqrt{\frac{c_1}{c_2}} = h, \qquad \text{say} \tag{8}$$

That is, if costs per observation in groups G_1 and G_2, are c_1 and c_2, respectively, then choose n_1 and n_2 on the basis of the ratio of the square root of the costs. This rule has been termed the *square root rule* by Gail et al. [1976]; the derivation can also be found in Nam [1973] and Cochran [1977].

If the costs are equal, $n_1 = n_2$, as before. Application of this rule can decrease the cost of an experiment, although it will increase the total number of observations. Note that the population means and standard deviation need not be known to determine the ratio of the sample sizes, only the costs. If the desired precision is specified—perhaps on the basis of sample size calculations assuming equal costs—the values of n_1 and n_2 can be determined. Compared with an experiment with equal sample sizes, the ratio ρ of the costs of the two experiments can be shown to be

$$\rho = \frac{1}{2} + \frac{h}{1+h^2} \tag{9}$$

If $h = 1$, then $\rho = 1$, as expected; if h is very close to zero or very large, $\rho = \frac{1}{2}$; thus, no matter what the relative costs of the observations, the savings can be no larger than 50%.

Example 17.3. (After Gail et al. [1976]) A new therapy, G_1, for hypertension is introduced and costs \$400 per subject. The standard therapy, G_2, costs \$16 per subject. On the basis of power calculations, the precision of the experiment is to be equivalent to an experiment using 22 subjects per treatment, so that

$$\frac{1}{22} + \frac{1}{22} = 0.09091$$

The square root rule specifies the ratio of the number of subjects in G_1 and G_2 by

$$\begin{aligned} n_2 &= \sqrt{\frac{400}{16}} n_1 \\ &= 5n_1 \end{aligned}$$

To obtain the same precision, we need to solve

$$\frac{1}{n_1} + \frac{1}{5n_1} = 0.09091$$

or

$$n_1 = 13.2 \quad \text{and} \quad n_2 = 66.0$$

(i.e., $1/13.2 + 1/66.0 = 0.09091$, the same precision). Rounding up, we require 14 observations in G_1 and 66 observations in G_2. The costs can also be compared as in Table 17.2.

A savings of \$3896 has been obtained, yet the precision is the same. The total number of observations is now 80, compared to 44 in the equal-sample-size experiment. The ratio of the savings is

$$\rho = \frac{6656}{9152} = 0.73$$

Table 17.2 Costs Comparisons for Example 17.3

	Equal Sample Size		Sample Size Determined by Cost	
	n	Cost	n	Cost
G_1	22	8800	14	5600
G_2	22	352	66	1056
Total	44	9152	80	6656

The value for ρ calculated from equation (9) is

$$\rho = \frac{1}{2} + \frac{5}{26} = 0.69$$

The reason for the discrepancy is the rounding of sample sizes to integers.

17.3.2 Unequal-Variance Case

Suppose that we want to compare the means from groups with unequal variance. Again, suppose that there are n_1 and n_2 observations in the two groups. Then the standard error of the difference between the two means is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Let the ratio of the variances be $\eta^2 = \sigma_2^2/\sigma_1^2$. Gail et al. [1976] show that the sample size should now be allocated in the ratio

$$\frac{n_2}{n_1} = \sqrt{\frac{\sigma_2^2}{\sigma_1^2}\frac{c_1}{c_2}} = \eta h$$

The calculations can then be carried out as before. In this case, the cost relative to the experiment with equal sample size is

$$\rho^* = \frac{(h+\eta)^2}{(1+h^2)(1+\eta^2)} \tag{10}$$

These calculations also apply when the costs are equal but the variances unequal, as is the case in binomial sampling.

17.3.3 Rule of Diminishing Precision Gain

One of the reasons advanced at the beginning of Section 17.3 for distinguishing between the sample sizes of two groups is that a limited number of observations may be available for one group and a virtually unlimited number in the second group. Case–control studies were cited where the number of cases per population is relatively fixed. Analogous to Gail et al. [1976], we define a rule of diminishing precision gain. Suppose that there are n cases and that an unlimited number of controls are available. Assume that costs and variances are equal. The precision of the difference is then proportional to

$$\sigma\sqrt{\frac{1}{n} + \frac{1}{hn}}$$

where hn is the number of controls selected for the n cases.

We calculate the ratio P_h:

$$P_h = \frac{\sqrt{1/n + 1/hn}}{\sqrt{1/n + 1/n}}$$

$$= \sqrt{\frac{1}{2}\left(1 + \frac{1}{h}\right)}$$

This ratio P_h is a measure of the precision of a case–control study with n and hn cases and controls, respectively, relative to the precision of a study with an equal number, n, of cases and controls. Table 17.3 presents the values of P_h and $100(P_h - P_\infty)/P_\infty$ as a function of h.

Table 17.3 Comparison of Precision of Case Control Study with n and hn Cases and Controls, Respectively

h	P_h	$100[(P_h - P_\infty)/P_\infty]\%$
1	1.00	41
2	0.87	22
3	0.82	15
4	0.79	12
5	0.77	10
10	0.74	5
∞	0.71	0

This table indicates that in the context above, the gain in precision with, say, more than four controls per case is minimal. At $h = 4$, one obtains all but 12% of the precision associated with a study using an infinite number of controls. Hence, in the situation above, there is little merit in obtaining more than four or five times as many controls as cases. Lubin [1980] approaches this from the point of view of the logarithm of the odds ratio and comes to a similar conclusion.

17.4 SAMPLE-SIZE CALCULATIONS IN SELECTING CONTINUOUS VARIABLES TO DISCRIMINATE BETWEEN POPULATIONS

In certain situations, there is interest in examining a large number of continuous variables to explain the difference between two populations. For example, an investigator might be "fishing" for clues explaining the presence (one population) or absence (the other population) of a disease of unknown etiology. Or in a disease where a variety of factors are known to affect prognosis, the investigator may desire to find a good set of variables for predicting which subjects will survive for a fixed number of years. In this section, the determination of sample size for such studies is discussed.

There are a variety of approaches to the data analysis in this situation. With a large, say 50 or more, number of variables, we would hesitate to run stepwise discriminant analysis to select a few important variables, since (1) in typical data sets there are often many dependencies that make the method numerically unstable (i.e., the results coming forth from some computers cannot be relied on); (2) the more complex the mathematical model used, the less faith we have that it is useful in other situations (i.e., the more parameters that are used and estimated, the less confidence we can have that the result is transportable to another population in time or space; here we might be envisioning a discriminant function with a large number of variables); and (3) the multiple-comparison problems inherent in considering the large number of variables at each step in the stepwise procedure make the result of doubtful value.

One approach to the analysis is first to perform a *univariate screen*. This means that variables (used singly, that is, univariately) with the most power to discriminate between the two populations are selected. Second, use these univariate discriminating variables in the discriminant analysis. The sample-size calculations below are based on this method of analysis. There is some danger in this approach, as variables that univariately are not important in discrimination could be important when used in conjunction with other variables. In many practical situations, this is not usually the case. Before discussing the sample-size considerations, we will consider a second approach to the analysis of such data as envisioned here.

Often, the discriminating variables fall naturally in smaller subsets. For example, the subsets for patients may involve data from (1) the history, (2) a physical exam, and (3) some routine tests. In many situations the predictive information of the variables within each subset is roughly

the same. This being the case, a two-step method of selecting the predictive variables is to (1) use stepwise selection within subsets to select a few variables from each subset, and (2) combine the selected variables into a group to be used for another stepwise selection procedure to find the final subset of predictive variables.

After selecting a smaller subset of variables to use in the prediction process, one of two steps is usually taken. (1) The predictive equation is validated (tested) on a new sample to show that it has predictive power. That is, an F-test for the discriminant function is performed. Or, (2) a larger independent sample is used to provide an indication of the accuracy of the prediction. The second approach requires a larger sample size than merely establishing that there is some predictive ability, as in the first approach. In the next three sections we make this general discussion precise.

17.4.1 Univariate Screening of Continuous Variables

To obtain an approximate idea of the sample size needed to screen among k variables, the following is assumed: The variables are normally distributed with the same variance in each population and possibly different means. The power to classify into the two populations depends on δ, the number of standard deviations distance between the two populations means:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

Some idea of the relationship of classificatory power to δ is given in Figure 17.1.

Suppose that we are going to screen k variables and want to be sure, with probability at least $1 - \alpha$, to include all variables with $\delta \geq D$. In this case we must be willing to accept some variables with values close to but less than D. Suppose that at the same time we want probability at least $1 - \alpha$ of not including any variables with $\delta \leq fD$, where $0 < f < 1$. One approach is to look at confidence intervals for the difference in the population means. If the absolute value of the difference is greater than $fD + (1 - f)D/2$, the variable is included. If the

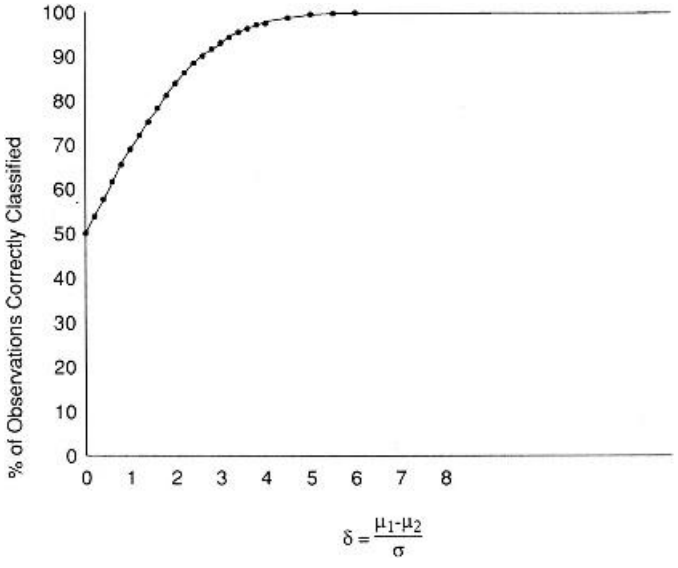


Figure 17.1 Probability of correct classification between $N(0, \sigma^2)$ and $N(\delta\sigma, \sigma^2)$ populations, assuming equal priors and $\delta\sigma/2$ as the cutoff values for classifying into the two populations.

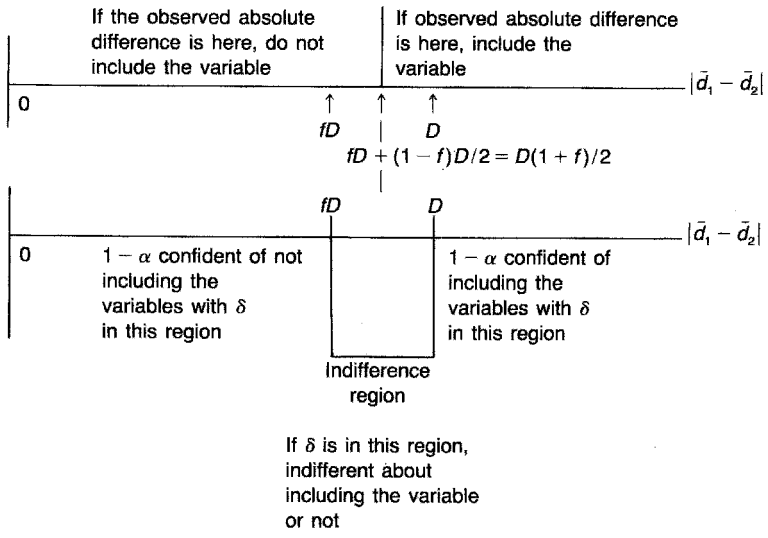


Figure 17.2 Inclusion and exclusion scheme for differences in sample means $|d_1 - d_2|$ from populations G_1 and G_2.

absolute value of the difference is less than this value, the variable is not included. Figure 17.2 presents the situation. To recap, with probability at least $1 - \alpha$, we include for use in prediction all variables with $\delta \geq D$ and do not include those with $\delta \leq fD$. In between, we are willing for either action to take place. The dividing line is placed in the middle.

Let us suppose that the number of observations, n, is large enough so that a normal approximation for confidence intervals will hold. Further, suppose that a fraction p of the data is from the first population and that $1 - p$ is from the second population. If we choose $1 - \alpha^*$ confidence intervals so that the probability is about $1 - \alpha$ that all intervals have half-width $\sigma(1 - f)D/2$, the result will hold.

If n is large, the pooled variance is approximately σ and the half-interval has width (in standard deviation units) of about

$$\sqrt{\frac{1}{Np} + \frac{1}{N(1-p)}} Z_{1-\alpha^*}$$

where $Z_{1-\alpha^*}$ is the $N(0, 1)$ critical value. To make this approximately $(1 - f)D/2$, we need

$$N = \frac{4z^2_{1-\alpha^*}}{p(1-p)D^2(1-f)^2} \tag{11}$$

In Chapter 12 it was shown that $\alpha^* = \alpha/2k$ was an appropriate choice by Bonferroni's inequality. In most practical situations, the observations tend to vary together, and the probability of all the confidence statements holding is greater than $1 - \alpha$. A slight compromise is to use $\alpha^* = [1 - (1 - \alpha)^{1/k}]/2$ as if the tests are independent. This α^* was used in computing Table 17.4.

From the table it is very clear that there is a large price to be paid if the smaller population is a very small fraction of the sample. There is often no way around this if the data need to be collected prospectively before subjects have the population membership determined (by having a heart attack or myocardial infarction, for example).

Table 17.4 Sample Sizes Needed for Univariate Screening When $f = \frac{2}{3}$[a]

	$p = 0.5$			$p = 0.6$			$p = 0.7$			$p = 0.8$			$p = 0.9$		
D	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
$k = 20$	2121	527	132	2210	553	136	2525	629	157	3315	829	204	5891	1471	366
	2478	616	153	2580	642	157	2950	735	183	3872	965	238	6881	1717	429
	3289	825	204	3434	859	213	3923	978	242	5151	1288	319	9159	2287	570
$k = 100$	2920	721	179	3043	761	187	3477	867	217	4565	1139	285	8118	2028	506
	3285	820	204	3421	854	213	3910	978	242	5134	1284	319	9129	2282	570
	4118	1029	255	4288	1071	268	4905	1224	306	6435	1607	400	11445	2860	714
$k = 300$	3477	867	217	3625	905	225	4140	1033	255	5436	1356	336	9665	2414	604
	3846	961	238	4008	999	247	4577	1143	285	6010	1500	374	10685	2669	667
	4684	1169	289	4879	1220	302	5576	1394	349	7323	1828	455	13018	3251	812

[a]For each entry the top, middle, and bottom numbers are for $\alpha = 0.10$, 0.05, and 0.01, respectively.

17.4.2 Sample Size to Determine That a Set of Variables Has Discriminating Power

In this section we find the answer to the following question. Assume that a discriminant analysis is being performed at significance level α with m variables. Assume that one population has a fraction p of the observations and that the other population has a fraction $1 - p$ of the observations. What sample size, n, is needed so that with probability $1 - \beta$, we reject the null hypothesis of no predictive power (i.e., Mahalanobis distance equal to zero) when in fact the Mahalanobis distance is $\Delta > 0$ (where Δ is fixed and known)? (See Chapter 13 for a definition of the Mahalanobis distance.)

The procedure is to use tables for the power functions of the analysis of variance tests as given in the CRC tables [Beyer, 1968 pp. 311–319]. To enter the charts, first find the chart for $v_1 = m$, the number of predictive variables.

The charts are for $\alpha = 0.05$ or 0.01. It is necessary to iterate to find the correct sample size n. The method is as follows:

1. Select an estimate of n.
2. Compute
$$\phi_n = \Delta\sqrt{\frac{p(1-p)}{m+1}} \times \sqrt{n} \tag{12}$$
This quantity indexes the power curves and is a measure of the difference between the two populations, adjusting for p and m.
3. Compute $v_2 = n - 2$.
4. On the horizontal axis, find ϕ and go vertically to the v_2 curve. Follow the intersection horizontally to find $1 - \widetilde{\beta}$.
5. **a.** If $1 - \widetilde{\beta}$ is greater than $1 - \beta$, decrease the estimate of n and go back to step 2.
   b. If $1 - \widetilde{\beta}$ is less than $1 - \beta$, increase the estimate of n and go back to step 2.
   c. If $1 - \widetilde{\beta}$ is approximately equal to $1 - \beta$, stop and use the given value of n as your estimate.

Example 17.4. Working at a significance level 0.05 with five predictive variables, find the total sample size needed to be 90% certain of establishing predictive power when $\cdot\Delta = 1$ and $p = 0.34$. Figure 17.3 is used in the calculation.

We use
$$\phi_n = 1 \times \sqrt{\frac{0.3 \times 0.7}{5+1}}\sqrt{n} = 0.187\sqrt{n}$$

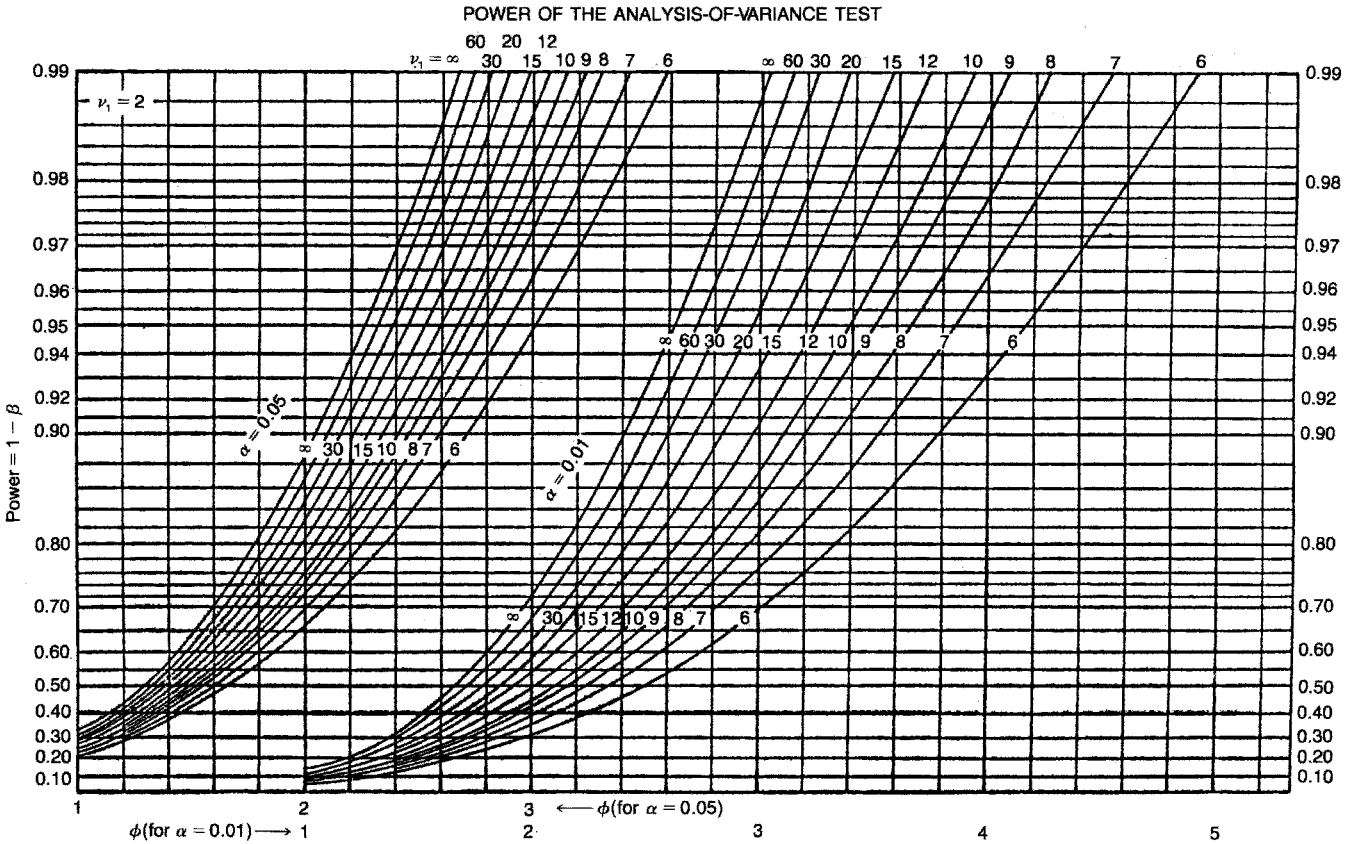


Figure 17.3 Power of the analysis of variance test. (From Beyer [1968].)

The method proceeds as follows:

1. Try $n = 30, \phi = 1.024, v_2 = 28, 1 - \beta \doteq 0.284$.
2. Try $n = 100, \phi = 1.870, v_2 = 98, 1 - \beta \doteq 0.958$.
3. Try $n = 80, \phi = 1.672, v_2 = 78, 1 - \beta \doteq 0.893$.
4. Try $n = 85, \phi = 1.724, v_2 = 83, 1 - \beta \doteq 0.92$.

Use $n = 83$. Note that the method is somewhat approximate, due to the amount of interpolation (rough visual interpretation) needed.

17.4.3 Quantifying the Precision of a Discrimination Method

After developing a method of classification, it is useful to validate the method on a new independent sample from the data used to find the classification algorithm. The approach of Section 17.4.2 is designed to show that there is some classification power. Of more interest is to be able to make a statement on the amount of correct and incorrect classification. Suppose that one is hoping to develop a classification method that classifies correctly $100\pi\%$ of the time.

To estimate with $100(1 - \alpha)\%$ confidence the correct classification percentage to within $100\varepsilon\%$, what number of additional observations are required? The confidence interval (we'll assume n large enough for the normal approximation) will be, letting c equal the number of n trials correctly classified,

$$\frac{c}{n} \pm \sqrt{\frac{1}{n}\frac{c}{n}\left(1 - \frac{c}{n}\right)} z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $N(0, 1)$ critical value. We expect $c/n \doteq \pi$, so it is reasonable to choose n to satisfy $z_{1-\alpha/2} = \varepsilon\sqrt{\pi(1 - \pi)/n}$. This implies that

$$n = z^2_{1-\alpha/2}\pi(1 - \pi)/\varepsilon^2 \tag{13}$$

where $\varepsilon =$ (predicted - actual) probability of misclassification.

Example 17.5. If one plans for $\pi = 90\%$ correct classification and wishes to be 99% confident of estimating the correct classification to within 2%, how many new experimental units must be allowed? From Equation (13) and $z_{0.995} = 2.576$, the answer is

$$n = (2.576)^2 \times \frac{0.9(1 - 0.9)}{(0.02)^2} \doteq 1493$$

17.4.4 Total Sample Size for an Observational Study to Select Classification Variables

In planning an observational study to discriminate between two populations, if the predictive variables are few in number and known, the sample size will be selected in the manner of Section 17.4.2 or 17.4.3. The size depends on whether the desire is to show some predictive power or to have desired accuracy of estimation of the probability of correct classification. In addition, a different sample is needed to estimate the discriminant function. Usually, this is of approximately the same size.

If the predictive variables are to be culled from a large number of choices, an *additional* number of observations must be added for the selection of the predictive variables (e.g., in the manner of Section 17.4.1). Note that the method cannot be validated by application to the observations used to select the variables and to construct the discriminant function: This would lead to an exaggerated idea of the accuracy of the method. As the coefficients and variables were chosen specifically for these data, the method will work better (often considerably better) on these data than on an independent sample chosen as in Section 17.4.2 or 17.4.3.

NOTES

17.1 Sample Sizes for Cohort Studies

Five major journals are sources for papers dealing with sample sizes in cohort and case–control studies: *Statistics in Medicine*, *Biometrics, Controlled Clinical Trials, Journal of Clinical Epidemiology*, and the *American Journal of Epidemiology*. In addition, there are books by Fleiss [1981], Schlesselman [1982], and Schuster [1993].

A cohort study can be thought of as a cross-sectional study; there is no selection on case status or exposure status. The table generated is then the usual 2×2 table. Let the sample proportions be as follows:

	Exposure	No Exposure	
Case	p_{11}	p_{12}	$p_{1\cdot}$
Control	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1

If p_{11}, $p_{1\cdot}$, $p_{2\cdot}$, $p_{\cdot 1}$, and $p_{\cdot 2}$ estimate $\pi_{11}, \pi_{1\cdot}, \pi_{2\cdot}, \pi_{\cdot 1}$, and $\pi_{\cdot 2}$, respectively, then the required total sample size for significance level α, and power $1 - \beta$ is approximately

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \pi_{11}\pi_{1\cdot}\pi_{2\cdot}\pi_{\cdot 1}\pi_{\cdot 2}}{(\pi_{11} - \pi_{1\cdot}\pi_{\cdot 1})^2} \tag{14}$$

Given values of $\pi_{1\cdot}$, $\pi_{\cdot 1}$, and $R = (\pi_{11}/\pi_{\cdot 1})/(\pi_{12}/\pi_{\cdot 2}) =$ the relative risk, the value of π_{11} is determined by

$$\pi_{11} = \frac{R\pi_{\cdot 1}\pi_{1\cdot}}{R\pi_{\cdot 1} + \pi_{\cdot 2}} \tag{15}$$

The formula for the required sample size then becomes

$$n = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \frac{\pi_{\cdot 1}}{1 - \pi_{\cdot 1}} \frac{1 - \pi_{\cdot 1}}{\pi_{\cdot 1}} \left[1 + \frac{1}{\pi_{\cdot 1}(R - 1)}\right]^2 \tag{16}$$

If the events are rare, the Poisson approximation derived in the text can be used. For a discussion of sample sizes in $r \times c$ contingency tables, see Lachin [1977] and Cohen [1988].

17.2 Sample-Size Formulas for Case–Control Studies

There are a variety of sample-size formulas for case–control studies. Let the data be arranged in a table as follows:

	Exposed	Not Exposed	
Case	X_{11}	X_{12}	n
Control	X_{21}	X_{22}	n

and

$$P[\text{exposure}|\text{case}] = \pi_1, \qquad P[\text{exposure}|\text{control}] = \pi_2$$

estimated by $P_1 = X_{11}/n$ and $P_2 = X_{21}/n$ (we assume that $n_1 = n_2 = n$). For a two-sample, two-tailed test with

$$P[\text{Type I error}] = \alpha \quad \text{and} \quad P[\text{Type II error}] = \beta$$

the approximate sample size per group is

$$n = \frac{[Z_{1-\alpha/2}\sqrt{2\overline{\pi}(1-\overline{\pi})} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}]^2}{(\pi_1-\pi_2)^2} \tag{17}$$

where $\overline{\pi} = \frac{1}{2}(\pi_1 + \pi_2)$. The total number of subjects is $2n$, of which n are cases and n are controls. Another formula is

$$n = \frac{[\pi_1(1-\pi) + \pi_2(1-\pi_2)](Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\pi_1-\pi_2)^2} \tag{18}$$

All of these formulas tend to give the same answers, and underestimate the sample sizes required. The choice of formula is primarily a matter of aesthetics.

The formulas for sample sizes for case–control studies are approximations, and several corrections are available to get closer to the exact value. Exact values for equal sample sizes have been tabulated in Haseman [1978]. Adjustment for the approximate sample size have been presented by Casagrande et al. [1978], who give a slightly more complicated and accurate formulation. See also Lachin [1981, 2000] and Ury and Fleiss [1980].

Two other considerations will be mentioned. The first is unequal sample size. Particularly in case–control studies, it may be difficult to recruit more cases. Suppose that we can select n observations from the first population and rn from the second ($0 < r < \infty$). Following Schlesselman [1982], a very good approximation for the exact sample size for the number of cases is

$$n_1 = n\left(\frac{r+1}{2r}\right) \tag{19}$$

and for the number of controls

$$n_2 = n\left(\frac{r+1}{2}\right) \tag{20}$$

where n is determined by equation (17) or (18). The total sample size is then $n((r+1)^2/2r)$. Note that the number of cases can never be reduced to more than $n/2$ no matter what the number of controls. This is closely related to the discussion in Section 17.3. Following Fleiss et al. [1980], a slightly improved estimate can be obtained by using

$$n_1^* = n_1 + \frac{r+1}{r\Delta} = \text{number of cases}$$

and

$$n_2^* = rn_1^* = \text{number of controls}$$

A second consideration is cost. In Section 17.3 we considered sample sizes as a function of cost and related the sample sizes to precision. Now consider a slight reformulation of the problem in the case–control context. Suppose that enrollment of a case costs c_1 and enrollment of a control costs c_2. Pike and Casagrande [1979] show that a reasonable sample size approximation is

$$n_1 = n\left(1 + \sqrt{\frac{c_1}{c_0}}\right)$$

$$n_2 = n\left(1 + \sqrt{\frac{c_0}{c_1}}\right)$$

where n is defined by equations (17) or (18).

Finally, frequently case–control study questions are put in terms of odds ratios (or relative risks). Let the odds ratio be $R = \pi_1(1-\pi_2)/\pi_2(1-\pi_1)$, where π_1 and π_2 are as defined at the beginning of this section. If the control group has known exposure rate π_2, that is, $P[\text{exposure}|\text{control}] = \pi_2$, then

$$\pi_1 = \frac{R\pi_2}{1+\pi_2(R-1)}$$

To calculate sample sizes, use equation (17) for specified values of π_2 and R.

Mantel [1983] gives some clever suggestions for making binomial sample-size tables more useful by making use of the fact that sample size is "inversely proportional to the square of the difference being sought, everything else being more or less fixed."

Newman [2001] is a good reference for sample-size questions involving survival data.

17.3 Power as a Function of Sample Size

Frequently, the question is not "How big should my sample size be" but rather, "I have 60 observations available; what kind of power do I have to detect a specified difference, relative risk, or odds ratio?" The charts by Feigl illustrated in Chapter 6 provided one answer. Basically, the question involves inversion of formulas such as given by equations (17) and (18), solving them for $Z_{1-\beta}$, and calculating the associated area under the normal curve. Besides Feigl, several authors have studied this problem or variations of it. Walter [1977] derived formulas for the smallest and largest relative risk, R, that can be detected as a function of sample size, Type I and Type II errors. Brittain and Schlesselman [1982] present estimates of power as a function of possibly unequal sample size and cost.

17.4 Sample Size as a Function of Coefficient of Variation

Sometimes, sample-size questions are asked in the context of percent variability and percent changes in means. With an appropriate, natural interpretation, valid answers can be provided. Specifically, assume that by *percent variability* is meant the coefficient of variation, call it V, and that the second mean differs from the first mean by a factor f.

Let two normal populations have means μ_1 and μ_2 and standard deviations σ_1 and σ_2. The usual sample-size formula for two independent samples needed to detect a difference $\mu_1 - \mu_2$ in means with Type I error α and power $1-\beta$ is given by

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2}$$

where $z_{1-\gamma}$ is the $100(1-\gamma)$th percentile of the standard normal distribution. This is the formula for a two-sided alternative; n is the number of observations per group. Now assume that $\mu_1 = f\mu_2$ and $\sigma_1/\mu_1 = \sigma_2/\mu_2 = V$. Then the formula transforms to

$$n = (z_{1-\alpha/2} + z_{1-\beta})^2 V^2 \left[1 + \frac{2f}{(f-1)^2}\right] \tag{21}$$

The quantity V is the usual coefficient of variation and f is the ratio of means. It does not matter whether the ratio of means is defined in terms of $1/f$ rather than f.

Sometimes the problem is formulated with the variability V as specified but a percentage change between means is given. If this is interpreted as the second mean, μ_2, being a percent change from the first mean, this percentage change is simply $100(f-1)\%$ and the formula again applies. However, sometimes, the relative status of the means cannot be specified, so an

interpretation of *percent change* is needed. If we know only that $\sigma_1 = V\mu_1$ and $\sigma_2 = V\mu_2$, the formula for sample size becomes

$$n = \frac{V^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\left((\mu_1 - \mu_2)/\sqrt{\mu_1\mu_2}\right)^2}$$

The quantity $\left((\mu_1 - \mu_2)/\sqrt{\mu_1\mu_2}\right)$ is the proportional change from μ_1 to μ_2 as a function of their geometric mean. If the questioner, therefore, can only specify a percent change, this interpretation is quite reasonable. Solving equation (21) for $z_{1-\beta}$ allows us to calculate values for power curves:

$$z_{1-\beta} = -z_{1-\alpha/2} + \frac{\sqrt{n}|f-1|}{V\sqrt{f^2+1}} \tag{22}$$

A useful set of curves as a function of n and a common coefficient of variation $V = 1$ can be constructed by noting that for two coefficients of variation V_1 and V_2, the sample sizes $n(V_1)$ and $n(V_2)$, as functions of V_1 and V_2, are related by

$$\frac{n(V_1)}{n(V_2)} = \frac{\sigma_1^2}{\sigma_2^2}$$

for the same power and Type I error. See van Belle and Martin [1993] and van Belle [2001].

PROBLEMS

17.1 **(a)** Verify that the odds ratio and relative risk are virtually equivalent for

$$P[\text{exposure}] = 0.10, \qquad P[\text{disease}] = 0.01$$

in the following two situations:

$$\pi_{11} = P[\text{exposed and disease}] = 0.005$$

and $\pi_{11} = 0.0025$.

(b) Using equation (2), calculate the number of disease occurrences in the exposed and unexposed groups that would have to be observed to detect the relative risks calculated above with $\alpha = 0.05$ (one-tailed) and $\beta = 0.10$.

(c) How many exposed persons would have to be observed (and hence, unexposed persons as well)?

(d) Calculate the sample size needed if this test is one of K tests for $K = 10, 100,$ and 1000.

(e) In part (d), plot the logarithm of the sample size as a function of $\log K$. What kind of relationship is suggested? Can you state a general rule?

17.2 (After N. E. Breslow) Workers at all nuclear reactor facilities will be observed for a period of 10 years to determine whether they are at excess risk for leukemia. The rate in the general population is 7.5 cases per 100,000 person-years of observation. We want to be 80% sure that a doubled risk will be detected at the 0.05 level of significance.

(a) Calculate the number of leukemia cases that must be detected among the nuclear plant workers.

(b) How many workers must be observed? That is, assuming the null hypothesis holds, how many workers must be observed to accrue 9.1 leukemia cases?

(c) Consider this as a binomial sampling problem. Let $\pi_1 = 9.1/$answer in part (b), and let $\pi_2 = 2\pi_1$. Now use equation (17) to calculate $n/2$ as the required sample size. How close is your answer to part (b)?

17.3 (After N. E. Breslow) The rate of lung cancer for men of working age in a certain population is known to be on the order of 60 cases per 100,000 person-years of observation. A cohort study using equal numbers of exposed and unexposed persons is desired so that an increased risk of $R = 1.5$ can be detected with power $1 - \beta = 0.95$ and $\alpha = 0.01$.

(a) How many cases will have to be observed in the unexposed population? The exposed population?

(b) How many person-years of observation at the normal rates will be required for either of the two groups?

(c) How many workers will be needed assuming a 20-year follow-up?

17.4 (After N. E. Breslow) A case–control study is to be designed to detect an odds ratio of 3 for bladder cancer associated with a certain medication that is used by about one person out of 50 in the general population.

(a) For $\alpha = 0.05$, and $\beta = 0.05$, calculate the number of cases and number of controls needed to detect the increased odds ratio.

(b) Use the Poisson approximation procedure to calculate the sample sizes required.

(c) Four controls can be provided for each case. Use equations (19) and (20) to calculate the sample sizes. Compare this result with the total sample size in part (a).

17.5 The sudden infant death syndrome (SIDS) occurs at a rate of approximately three cases per 1000 live births. It is thought that smoking is a risk factor for SIDS, and a case–control study is initiated to check this assumption. Since the major effort was in the selection and recruitment of cases and controls, a questionnaire was developed that contained 99 additional questions.

(a) Calculate the sample size needed for a case–control study using $\alpha = 0.05$, in which we want to be 95% certain of picking up an increased relative risk of 2 associated with smoking. Assume that an equal number of cases and controls are selected.

(b) Considering smoking just one of the 100 risk factors considered, what sample sizes will be needed to maintain an $\alpha = 0.05$ per experiment error rate?

(c) Given the increased value of Z in part (b), suppose that the sample size is not changed. What is the effect on the power? What is the power now?

(d) Suppose in part (c) that the power also remains fixed at 0.95. What is the minimum relative risk that can be detected?

(e) Since smoking was the risk factor that precipitated the study, can an argument be made for not testing it at a reduced α level? Formulate your answer carefully.

***17.6** Derive the square root rule starting with equations (4) and (5).

***17.7** Derive formula (16) from equation (14).

17.8 It has been shown that coronary bypass surgery does not prolong life in selected patients with relatively mild angina (but may relieve the pain). A surgeon has invented a new

bypass procedure that, she claims, will prolong life substantially. A trial is planned with patients randomized to surgical treatment or standard medical therapy. Currently, the five-year survival probability of patients with relatively mild symptoms is 80%. The surgeon claims that the new technique will increase survival to 90%.

(a) Calculate the sample size needed to be 95% certain that this difference will be detected using an $\alpha = 0.05$ significance level.

(b) Suppose that the cost of a coronary bypass operation is approximately \$50,000; the cost of general medical care is about \$10,000. What is the most economical experiment under the conditions specified in part (a)? What are the total costs of the two studies?

(c) The picture is more complicated than described in part (b). Suppose that about 25% of the patients receiving the medical treatment will go on to have a coronary bypass operation in the next five years. Recalculate the sample sizes under the conditions specified in part (a).

***17.9** Derive the sample sizes in Table 17.4 for $D = 0.5$, $p = 0.8$, $\alpha = 0.5$, and $k = 20, 100, 300$.

***17.10** Consider the situation in Example 17.4.

(a) Calculate the sample size as a function of m, the number of variables, by considering $m = 10$ and $m = 20$.

(b) What is the relationship of sample size to variables?

17.11 Two groups of rats, one young and the other old, are to be compared with respect to levels of nerve growth factor (NGF) in the cerebrospinal fluid. It is estimated that the variability in NGF from animal to animal is on the order of 60%. We want to look at a twofold ratio in means between the two groups.

(a) Using the formula in Note 17.4, calculate the sample size per group using a two-sided alternative, $\alpha = 0.05$, and a power of 0.80.

(b) Suppose that the ratio of the means is really 1.6. What is the power of detecting this difference with the sample sizes calculated in part (a)?

REFERENCES

Beyer, W. H. (ed.) [1968]. *CRC Handbook of Tables for Probability and Statistics*, 2nd ed. CRC Press, Cleveland, OH.

Brittain, E., and Schlesselman, J. J. [1982]. Optimal allocation for the comparison of proportions. *Biometrics*, **38**: 1003–1009.

Casagrande, J. T., Pike, M. C., and Smith, P. C. [1978]. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, **34**: 483–486.

Cochran, W. G. [1977]. *Sampling Techniques*, 3rd ed. Wiley, New York.

Cohen, J. [1988]. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Fleiss, J. L., Tytun, A., and Ury, H. K. [1980]. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, **36**: 343–346.

Gail, M., Williams, R., Byar, D. P., and Brown, C. [1976]. How many controls. *Journal of Chronic Diseases*, **29**: 723–731.

Haseman, J. K. [1978]. Exact sample sizes for the use with the Fisher–Irwin test for 2×2 tables. *Biometrics*, **34**: 106–109.

Lachin, J. M. [1977]. Sample size determinations for $r \times c$ comparative trials. *Biometrics*, **33**: 315–324.

Lachin, J. M. [1981]. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, **2**: 93–113.

Lachin, J. M. [2000]. *Biostatistical Methods*. Wiley, New York.

Lubin, J. H. [1980]. Some efficiency comments on group size in study design. *American Journal of Epidemiology*, **111**: 453–457.

Mantel, H. [1983]. Extended use of binomial sample-size tables. *Biometrics*, **39**: 777–779.

Nam, J. M. [1973]. Optimum sample sizes for the comparison of a control and treatment. *Biometrics*, **29**: 101–108.

Newman, S. C. [2001]. *Biostatistical Methods in Epidemiology*. Wiley, New York.

Pike, M. C., and Casagrande, J. T. [1979]. Cost considerations and sample size requirements in cohort and case-control studies. *American Journal of Epidemiology*, **110**: 100–102.

Schlesselman, J. J. [1982]. *Case–Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

Schuster, J. J. [1993]. *Practical Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton, FL.

Ury, H. K., and Fleiss, J. R. [1980]. On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. *Biometrics*, **36**: 347–351.

van Belle, G. [2001]. *Statistical Rules of Thumb*. Wiley, New York.

van Belle, G., and Martin, D. C. [1993]. Sample size as a function of coefficient of variation and ratio of means. *American Statistician*, **47**: 165–167.

Walter, S. D. [1977]. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *American Journal of Epidemiology*, **105**: 387–397.

CHAPTER 18

Longitudinal Data Analysis

18.1 INTRODUCTION

One of the most common medical research designs is a "pre–post" study in which a single baseline health status measurement is obtained, an intervention is administered, and a single follow-up measurement is collected. In this experimental design, the *change* in the outcome measurement can be associated with the *change* in the exposure condition. For example, if some subjects are given placebo while others are given an active drug, the two groups can be compared to see if the change in the outcome is different for those subjects who are actively treated as compared to control subjects. This design can be viewed as the simplest form of a prospective longitudinal study.

Definition 18.1. A *longitudinal study* refers to an investigation where participant outcomes and possibly treatments or exposures are collected at multiple follow-up times.

A longitudinal study generally yields multiple or "repeated" measurements on each subject. For example, HIV patients may be followed over time and monthly measures such as CD4 counts or viral load are collected to characterize immune status and disease burden, respectively. Such repeated-measures data are correlated within subjects and thus require special statistical techniques for valid analysis and inference.

A second important outcome that is commonly measured in a longitudinal study is the time until a key clinical event such as disease recurrence or death. Analysis of event-time endpoints is the focus of *survival analysis*, which is covered in Chapter 16.

Longitudinal studies play a key role in epidemiology, clinical research, and therapeutic evaluation. Longitudinal studies are used to characterize normal growth and aging, to assess the effect of risk factors on human health, and to evaluate the effectiveness of treatments.

Longitudinal studies involve a great deal of effort but offer several benefits, which include:

1. *Incident events recorded.* A prospective longitudinal study measures the new occurrence of disease. The timing of disease onset can be correlated with recent changes in patient exposure and/or with chronic exposure.

2. *Prospective ascertainment of exposure*. In a prospective study, participants can have their exposure status recorded at multiple follow-up visits. This can alleviate recall bias where subjects who subsequently experience disease are more likely to recall their exposure (a form of measurement error). In addition, the temporal order of exposures and outcomes is observed.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

3. *Measurement of individual change in outcomes*. A key strength of a longitudinal study is the ability to measure change in outcomes and/or exposure at the individual level. Longitudinal studies provide the opportunity to observe individual patterns of change.

4. *Separation of time effects: cohort, period, age*. When studying change over time, there are many time scales to consider. The *cohort scale* is the time of birth, such as 1945 or 1963; *period* is the current time, such as 2003; and *age* is (period − cohort), for example, $58 = 2003 - 1945$, and $40 = 2003 - 1963$. A longitudinal study with measurements at times $t_1, t_2, \ldots, t_n$ can simultaneously characterize multiple time scales such as age and cohort effects using covariates derived from the calendar time of visit and the participant's birth year: the age of subject i at time t_j is $\text{age}_{ij} = t_j - \text{birth}_i$; and their cohort is simply $\text{cohort}_{ij} = \text{birth}_i$. Lebowitz [1996] discusses age, period, and cohort effects in the analysis of pulmonary function data.

5. *Control for cohort effects*. In a cross-sectional study the comparison of subgroups of different ages combines the effects of aging and the effects of different cohorts. That is, comparison of outcomes measured in 2003 among 58-year-old subjects and among 40-year-old subjects reflects both the fact that the groups differ by 18 years (aging) and the fact that the subjects were born in different eras. For example, the public health interventions, such as vaccinations available for a child under 10 years of age, may differ in 1945–1955 compared to the preventive interventions experienced in 1963–1973. In a longitudinal study, the cohort under study is fixed, and thus changes in time are not confounded by cohort differences.

An overview of longitudinal data analysis opportunities in respiratory epidemiology is presented in Weiss and Ware [1996].

The benefits of a longitudinal design are not without cost. There are several challenges posed:

1. *Participant follow-up*. There is the risk of bias due to incomplete follow-up, or dropout of study participants. If subjects who are followed to the planned end of a study differ from subjects who discontinue follow-up, a naive analysis may provide summaries that are not representative of the original target population.

2. *Analysis of correlated data*. Statistical analysis of longitudinal data requires methods that can properly account for the intrasubject correlation of response measurements. If such correlation is ignored, inferences such as statistical tests or confidence intervals can be grossly invalid.

3. *Time-varying covariates*. Although longitudinal designs offer the opportunity to associate changes in exposure with changes in the outcome of interest, the direction of causality can be complicated by "feedback" between the outcome and the exposure. For example, in an observational study of the effects of a drug on specific indicators of health, a patient's current health status may influence the drug exposure or dosage received in the future. Although scientific interest lies in the effect of medication on health, this example has reciprocal influence between exposure and outcome and poses analytical difficulty when trying to separate the effect of medication on health from the effect of health on drug exposure.

18.1.1 Example studies

In this section we give some examples of longitudinal studies and focus on the primary scientific motivation in addition to key outcome and covariate measurements.

Child Asthma Management Program

In the Child Asthma Management Program (CAMP) study, children are randomized to different asthma management regimes. CAMP is a multicenter clinical trial whose primary aim is evaluation of the long-term effects of daily inhaled anti-inflammatory medication use on asthma status and lung growth in children with mild to moderate asthma [The Childhood Asthma Management

Program Research group, 2000]. Outcomes include continuous measures of pulmonary function and categorical indicators of asthma symptoms. Secondary analyses have investigated the association between daily measures of ambient pollution and the prevalence of symptoms. Analysis of an environmental exposure requires specification of a lag between the day of exposure and the resulting effect. In the air pollution literature, short lags of 0 to 2 days are commonly used [Samet et al., 2000; Yu et al., 2000]. For both the evaluation of treatment and exposure to environmental pollution, the scientific questions focus on the association between an exposure (treatment, pollution) and health measures. The within-subject correlation of outcomes is of secondary interest, but must be acknowledged to obtain valid statistical inference.

Cystic Fibrosis Foundation Registry

The Cystic Fibrosis Foundation maintains a registry of longitudinal data for subjects with cystic fibrosis. Pulmonary function measures, such as the 1-second forced expiratory volume (FEV1), and patient health indicators, such as infection with *Pseudomonas aeruginosa*, have been recorded annually since 1966. One scientific objective is to characterize the natural course of the disease and to estimate the average rate of decline in pulmonary function. Risk factor analysis seeks to determine whether measured patient characteristics such as gender and genotype correlate with disease progression or with an increased rate of decline in FEV1. The registry data represent a typical observational design where the longitudinal nature of the data are important for determining individual patterns of change in health outcomes such as lung function.

Multicenter AIDS Cohort Study

The Multicenter AIDS Cohort Study (MACS) enrolled more than 3000 men who were at risk for acquisition of HIV1 [Kaslow et al., 1987]. This prospective cohort study observed $N = 479$ incident HIV1 infections and has been used to characterize the biological changes associated with disease onset. In particular, this study has demonstrated the effect of HIV1 infection on indicators of immunologic function such as CD4 cell counts. One scientific question is whether baseline characteristics such as viral load measured immediately after seroconversion are associated with a poor patient prognosis as indicated by a greater rate of decline in CD4 cell counts. We use these data to illustrate analysis approaches for continuous longitudinal response data.

HIVNET Informed Consent Substudy

Numerous reports suggest that the process of obtaining informed consent in order to participate in research studies is often inadequate. Therefore, for preventive HIV vaccine trials a prototype informed consent process was evaluated among $N = 4892$ subjects participating in the Vaccine Preparedness Study (VPS). Approximately 20% of subjects were selected at random and asked to participate in a mock informed consent process [Coletti et al., 2003]. Participant knowledge of key vaccine trial concepts was evaluated at baseline prior to the informed consent visit, which occurred during a special three-month follow-up visit for the intervention subjects. Vaccine trial knowledge was then assessed for all participants at the scheduled six-, 12-, and 18-month visits. This study design is a basic longitudinal extension of a pre–post design. The primary outcomes include individual knowledge items and a total score that calculates the number of correct responses minus the number of incorrect responses. We use data on a subset of men and women VPS participants. We focus on subjects who were considered at high risk of HIV acquisition, due to injection drug use.

18.1.2 Notation

In this chapter we use Y_{ij} to denote the outcome measured on subject i at time t_{ij}. The index $i = 1, 2, \ldots, N$ is for subject, and the index $j = 1, 2, \ldots, n$ is for observations within a subject. In a designed longitudinal study the measurement times will follow a protocol with

a common set of follow-up times, $t_{ij} = t_j$. For example, in the HIVNET Informed Consent Study, subjects were measured at baseline, $t_1 = 0$, at six months after enrollment, $t_2 =$ six months, and at 12 and 18 months, $t_3 = 12$ months, $t_4 = 18$ months. We let X_{ij} denote covariates associated with observation Y_{ij}. Common covariates in a longitudinal study include the time, t_{ij}, and person-level characteristics such as treatment assignment or demographic characteristics.

Although scientific interest often focuses on the mean response as a function of covariates such as treatment and time, proper statistical inference must account for the within-person correlation of observations. Define $\rho_{jk} = \text{corr}(Y_{ij}, Y_{ik})$, the within-subject correlation between observations at times t_j and t_k. In the following section we discuss methods for exploring the structure of within-subject correlation, and in Section 18.5 we discuss estimation methods that model correlation patterns.

18.2 EXPLORATORY DATA ANALYSIS

Exploratory analysis of longitudinal data seeks to discover patterns of systematic variation across groups of patients, as well as aspects of random variation that distinguish individual patients.

18.2.1 Group Means over Time

When scientific interest is in the average response over time, summary statistics such as means and standard deviations can reveal whether different groups are changing in a similar or different fashion.

Example 18.1. Figure 18.1 shows the mean knowledge score for the informed consent subgroups in the HIVNET Informed Consent Substudy. At baseline the intervention and control groups have very similar mean scores. This is expected since the group assignment is determined by randomization that occurs after enrollment. At an interim three-month visit the intervention subjects are given a mock informed consent for participation in a hypothetical phase III vaccine efficacy trial. The impact of the intervention can be seen by the mean scores at the six-month visit. In the control group the mean at six months is 1.49(SE = 0.11), up slightly from the baseline mean of 1.16(SE = 0.11). In contrast, the intervention group has a six-month mean score of 3.43(SE = 0.24), a large increase from the baseline mean of 1.09(SE = 0.24). The intervention and control groups are significantly different at six months based on a two-sample t-test. At later follow-up times, further change is observed. The control group has a mean that increases to 1.98 at the 12-month visit and to 2.47 at the 18-month visit. The intervention group fluctuates slightly with means of 3.25(SE = 0.27) at month 12 and 3.76(SE = 0.25) at 18 months. These summaries suggest that the intervention has a significant effect on knowledge, and that a small improvement is seen over time in the control group.

Example 18.2. In the MACS study we compare different groups of subjects formed on the basis of their initial viral load measurement. Low viral load is defined by a baseline value less than 15×10^3, medium as 15×10^3 to 46×10^3, and high viral load is classified for subjects with a baseline measurement greater than 46×10^3. Table 18.1 gives the average CD4 count for each year of follow-up. The mean CD4 declines over time for each of the viral load groups. The subjects with the lowest baseline viral load have a mean of 744.8 for the first year after seroconversion and then decline to a mean count of 604.8 during the fourth year. The $744.8 - 604.8 = 140.0$-unit reduction is smaller than the decline observed for the medium-viral-load group, $638.9 - 470.0 = 168.9$, and the high-viral-load group, $600.3 - 353.9 = 246.4$. Therefore, these summaries suggest that higher baseline viral-load measurements are associated with greater subsequent reduction in mean CD4 counts.

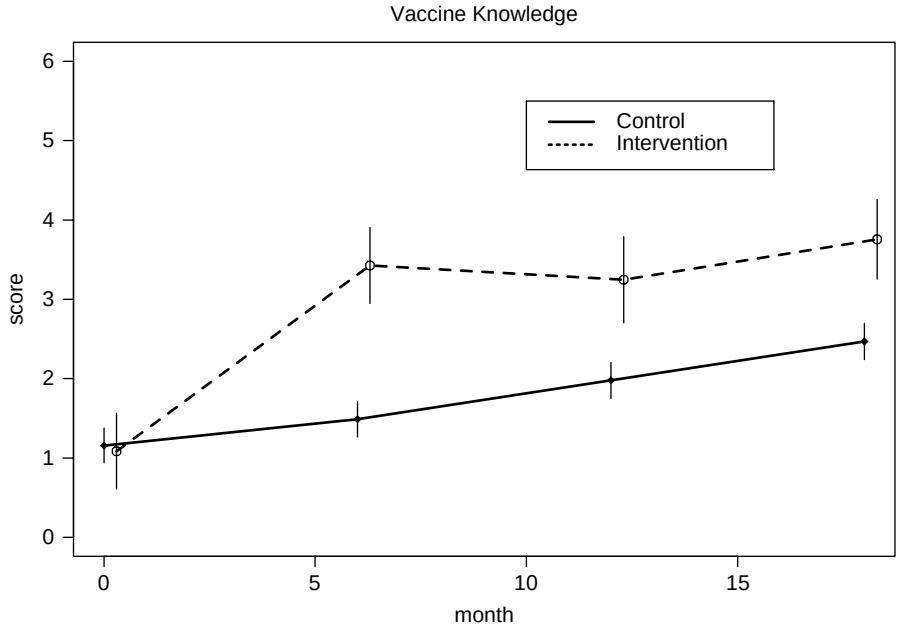


Figure 18.1 Mean knowledge scores over time by treatment group, HIVNET informed consent substudy.

Table 18.1 Mean CD4 Count and Standard Error over Time[a]

	Baseline Viral Load					
	Low		Medium		High	
Year	Mean	SE	Mean	SE	Mean	SE
0–1	744.8	35.8	638.9	27.3	600.3	30.4
1–2	721.2	36.4	588.1	25.7	511.8	22.5
2–3	645.5	37.7	512.8	28.5	474.6	34.2
3–4	604.8	46.8	470.0	28.7	353.9	28.1

[a]Separate summaries are given for groups defined by baseline viral load level.

Example 18.1. (*continued*) In the HIVNET informed consent substudy we saw a substantial improvement in the knowledge score. It is also relevant to consider key individual items that comprise the total score, such as the "safety item" or "nurse item." Regarding safety, participants were asked whether it was true or false that "Once a large-scale HIV vaccine study begins, we can be sure the vaccine is completely safe." Table 18.2 shows the number of responding subjects at each visit and the percent of subjects who correctly answered that the safety statement is false. These data show that the control and intervention groups have a comparable understanding of the safety item at baseline with 40.9% answering correctly among controls, and 39.2% answering correctly among the intervention subjects. A mock informed consent was administered at a three-month visit for the intervention subjects only. The impact of the intervention appears modest, with only 50.3% of intervention subjects correctly responding at six months. This represents a 10.9% increase in the proportion answering correctly, but a two-sample comparison of intervention and control proportions at six months (e.g., 50.3% vs. 42.7%) is not significant

Table 18.2 Number of Subjects and Percent Answering Correctly for the Safety Item from the HIVNET Informed Consent Substudy

	Control Group		Intervention Group	
Visit	*N*	% Correct	*N*	% Correct
Baseline	946	40.9	176	39.2
six-month	838	42.7	171	50.3
12-month	809	41.5	163	43.6
18-month	782	43.5	153	43.1

Table 18.3 Number of Subjects and Percent Answering Correctly for the Nurse Item from the HIVNET Informed Consent Substudy

	Control Group		Intervention Group	
Visit	*n*	% Correct	*n*	% Correct
Baseline	945	54.1	176	50.3
six-month	838	44.7	171	72.1
12-month	808	46.3	163	60.1
18-month	782	48.2	153	66.0

statistically. Finally, the modest intervention impact does not appear to be retained, as the fraction correctly answering this item declines to 43.6% at 12 months and 43.1% at 18 months. Therefore, these data suggest a small but fleeting improvement in participant understanding that a vaccine studied in a phase III trial cannot be guaranteed to be safe.

Other items show different longitudinal trends. Subjects were also asked whether it was true or false that "The study nurse will decide who gets the real vaccine and who gets the placebo." Table 18.3 shows that the groups are again comparable at baseline, but for the nurse item we see a large increase in the fraction answering correctly among intervention subjects at six months with 72.1% answering correctly that the statement is false. A cross-sectional analysis indicates a statistically significant difference in the proportion answering correctly at six months with a confidence interval for the difference in proportions of $(0.199, 0.349)$. Although the magnitude of the separation between groups decreases from 27.4% at six months to 17.8% at 18 months, the confidence interval for the difference in proportions at 18 months is $(0.096, 0.260)$ and excludes the null comparison, $p_1 - p_0 = 0$. Therefore, these data suggest that the intervention has a substantial and lasting impact on understanding that research nurses do not determine allocation to real vaccine or placebo.

18.2.2 Variation among Subjects

With independent observations we can summarize the uncertainty or variability in a response measurement using a single variance parameter. One interpretation of the variance is given as one-half the expected squared distance between any two randomly selected measurements, $\sigma^2 = \frac{1}{2}E[(Y_i - Y_j)^2]$. However, with longitudinal data the "distance" between measurements on different subjects is usually expected to be greater than the distance between repeated measurements taken on the same subject. Thus, although the total variance may be obtained with outcomes from subjects i and i' observed at time t_j, $\sigma^2 = \frac{1}{2}E[(Y_{ij} - Y_{i'j})^2]$ [assuming that $E(Y_{ij}) = E(Y_{i'j}) = \mu$], the expected variation for two measurements taken on the same person

(subject i) but at times t_j and t_k may not equal the total variation σ^2 since the measurements are correlated: $\sigma^2(1-\rho_{jk}) = \frac{1}{2}E[(Y_{ij}-Y_{ik})^2]$ [assuming that $E(Y_{ij}) = E(Y_{ik}) = \mu$]. When $\rho_{jk} > 0$, this shows that *between-subject variation* is greater than *within-subject variation*. In the extreme, $\rho_{jk} = 1$ and $Y_{ij} = Y_{ik}$, implying no variation for repeated observations taken on the same subject.

Graphical methods can be used to explore the magnitude of person-to-person variability in outcomes over time. One approach is to create a panel of individual line plots for each study participant. These plots can then be inspected for both the amount of variation from subject to subject in the overall "level" of the response and the magnitude of variation in the "trend" over time in the response. Such exploratory data analysis can be useful for determining the types of correlated data regression models that would be appropriate. In Section 18.5 we discuss random effects regression models for longitudinal data. In addition to plotting individual series, it is also useful to plot multiple series on a single plot, stratifying on the value of key covariates. Such a plot allows determination of whether the type and magnitude of intersubject variation appears to differ across the covariate subgroups.

Example 18.2. (*continued*) In Figure 18.2 we plot an array of individual series from the MACS data. In each panel the observed CD4 count for a single subject is plotted against the times that measurements were obtained. Such plots allow inspection of the individual response patterns and whether there is strong heterogeneity in the trajectories. Figure 18.2 shows that there can be large variation in the "level" of CD4 for subjects. Subject ID = 1120 in the upper right corner has CD4 counts greater than 1000 for all times, while ID = 1235 in the lower left corner has all measurements below 500. In addition, individuals plots can be evaluated for the change over time. Figure 18.2 indicates that most subjects are either relatively stable in their measurements over time, or tend to be decreasing.

In the common situation where we are interested in correlating the outcome to measured factors such as treatment group or exposure, it will also be useful to plot individual series stratified by covariate group. Figure 18.3 takes a sample of the MACS data and plots lines for each subject stratified by the level of baseline viral load. This figure suggests that the highest viral load group has the lowest mean CD4 count and suggests that variation among measurements may also be lower for the high baseline viral-load group compared to the medium- and low-viral-load groups. Figure 18.3 can also be used to identify those who exhibit time trends that differ markedly from the profiles of others. In the high-viral-load group there is a person who appears to improve dramatically over time, and there is a single unusual measurement where the CD4 count exceeds 2000. Plotting individual series is a useful exploratory prelude to more careful confirmatory statistical analysis.

18.2.3 Characterizing Correlation and Covariance

With correlated outcomes it is useful to understand the strength of correlation and the pattern of correlations across time. Characterizing correlation is useful for understanding components of variation and for identifying a variance or correlation model for regression methods such as mixed-effects models or *generalized estimating equations* (GEEs), discussed in Section 18.5.2. One summary that is used is an estimate of the *covariance matrix*, which is defined as

$$\begin{bmatrix} E[(Y_{i1}-\mu_{i1})^2] & E[(Y_{i1}-\mu_{i1})(Y_{i2}-\mu_{i2})] & \cdots & E[(Y_{i1}-\mu_{i1})(Y_{in}-\mu_{in})] \\ E[(Y_{i2}-\mu_{i2})(Y_{i1}-\mu_{i1})] & E[(Y_{i2}-\mu_{i2})^2] & \cdots & E[(Y_{i2}-\mu_{i2})(Y_{in}-\mu_{in})] \\ \vdots & & \ddots & \cdots \\ E[(Y_{in}-\mu_{in})(Y_{i1}-\mu_{i1})] & E[(Y_{in}-\mu_{in})(Y_{i2}-\mu_{i2})] & \cdots & E[(Y_{in}-\mu_{in})^2] \end{bmatrix}$$

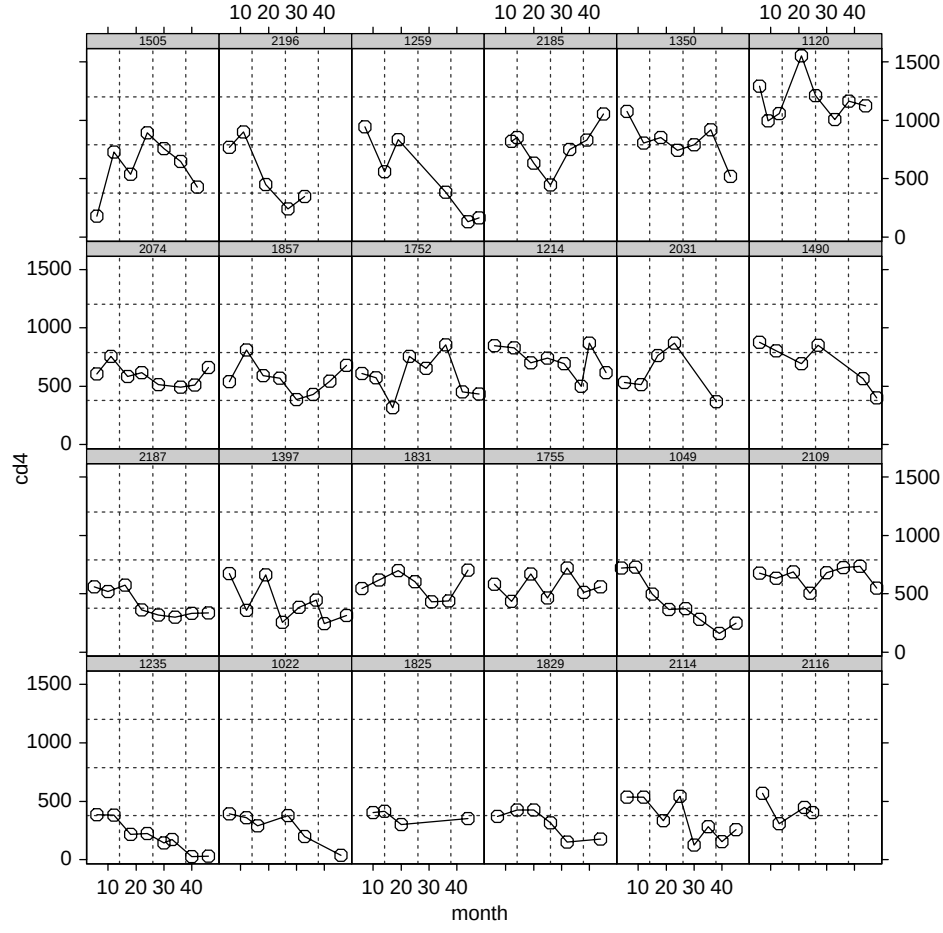


Figure 18.2 A sample of individual CD4 trajectories from the MACS data.

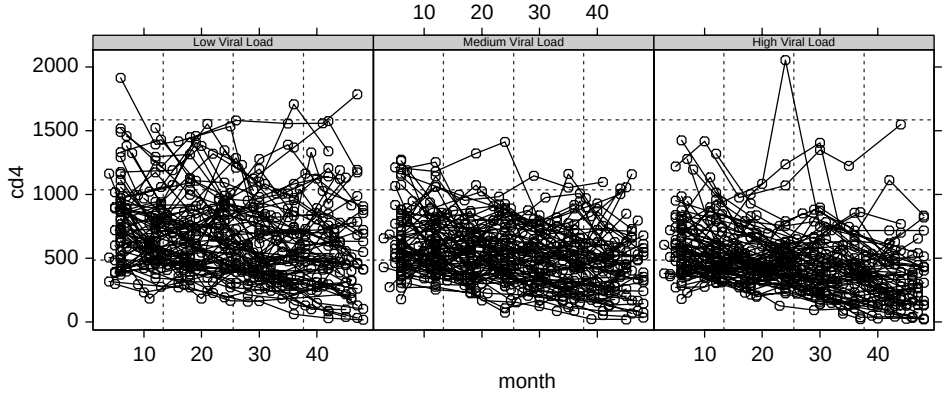


Figure 18.3 Individual CD4 trajectories from the MACS data by tertile of viral load.

The covariance can also be written in terms of the variances σ_j^2 and the correlations ρ_{jk}:

$$\text{cov}(Y_i) = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_n\rho_{1n} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \cdots & \sigma_2\sigma_n\rho_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_n\sigma_1\rho_{n1} & \sigma_n\sigma_2\rho_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Finally, the *correlation matrix* is given as

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \vdots & & \ddots & \vdots \\ \rho_{n1} & \rho_{2n} & \ldots & 1 \end{bmatrix}$$

which is useful for comparing the strength of association between pairs of outcomes, particularly when the variances σ_j^2 are not constant. Sample estimates of the correlations can be obtained using

$$\widehat{\rho}_{jk} = \frac{1}{N-1} \sum_i \frac{(Y_{ij} - \overline{Y}_{\cdot j})}{\widehat{\sigma}_j} \frac{(Y_{ik} - \overline{Y}_{\cdot k})}{\widehat{\sigma}_k}$$

where $\widehat{\sigma}_j^2$ and $\widehat{\sigma}_k^2$ are the sample variances of Y_{ij} and Y_{ik}, respectively (i.e., across subjects for times t_j and t_k).

Graphically, the correlation can be viewed using plots of Y_{ij} vs. Y_{ik} for all possible pairs of times t_j and t_k. These plots can be arranged in an array that corresponds to the covariance matrix and patterns of association across rows or columns can reveal changes in the correlation as a function of increasing time separation between measurements.

Example 18.1. (*continued*) For the HIVNET informed consent data, we focus on correlation analysis of outcomes from the control group. Parallel summaries would usefully characterize the similarity or difference in correlation structures for the control and intervention groups. The correlation matrix is estimated as follows:

	Month 0	**Month 6**	**Month 12**	**Month 18**
Month 0	1.00	0.471	0.394	0.313
Month 6	0.471	1.00	0.444	0.407
Month 12	0.394	0.444	1.00	0.508
Month 18	0.313	0.407	0.508	1.00

The matrix suggests that the correlation in outcomes from the same person is slightly decreasing as the time between the measurements increases. For example, the correlation between knowledge scores from baseline and month 6 is 0.471, while the correlation between baseline and month 12 decreases to 0.394, and decreases further to 0.313 for baseline and month 18. Correlation that decreases as a function of time separation is common among biomedical measurements and often reflects slowly varying underlying processes.

Example 18.2. (*continued*) For the MACS data the timing of measurement is only approximately regular. The following displays both the correlation matrix and the covariance matrix:

	Year 1	Year 2	Year 3	Year 4
Year 1	92,280.4	[0.734]	[0.585]	[0.574]
Year 2	63,589.4	81,370.0	[0.733]	[0.695]
Year 3	48,798.2	57,457.5	75,454.5	[0.806]
Year 4	55,501.2	63,149.9	70,510.1	101,418.2

The correlations are shown in brackets above. The variances are shown on a diagonal below the correlations. For example, the standard deviation among year 1 CD4 counts is $\sqrt{92{,}280.4} = 303.8$, while the standard deviations for years 2 through 4 are $\sqrt{81{,}370.0} = 285.3$, $\sqrt{75{,}454.5} = 274.7$, and $\sqrt{101{,}418.2} = 318.5$, respectively. Below the diagonal are the covariances, which together with the standard deviations determine the correlations. These data have a correlation for measurements that are one year apart of 0.734, 0.733, and 0.806. For measurements two years apart, the correlation decreases slightly to 0.585 and 0.695. Finally, measurements that are three years apart have a correlation of 0.574. Thus, the CD4 counts have a within-person correlation that is high for observations close together in time, but the correlation tends to decrease with increasing time separation between the measurement times.

An alternative method for exploring the correlation structure is through an array of scatter plots showing CD4 measured at year j versus CD4 measured at year k. Figure 18.4 displays these scatter plots. It appears that the correlation in the plot of year 1 vs. year 2 is stronger than for year 1 vs. year 3, or for year 1 vs. year 4. The sample correlations $\widehat{\rho}_{12} = 0.734$, $\widehat{\rho}_{13} = 0.585$, and $\widehat{\rho}_{14} = 0.574$ summarize the linear association presented in these plots.

18.3 DERIVED VARIABLE ANALYSIS

Formal statistical inference with longitudinal data requires either that a univariate summary be created for each subject or that methods for correlated data are used. In this section we review and critique common analytic approaches based on creation of summary measures.

A *derived variable analysis* is a method that takes a collection of measurements and collapses them into a single meaningful summary feature. In classical multivariate methods principal component analysis is one approach for creating a single major factor. With longitudinal data the most common summaries are the average response and the time slope. A second approach is a pre–post analysis which analyzes a single follow-up response in conjunction with a baseline measurement. In Section 18.3.1 we first review average or slope analyses, and then in Section 18.3.2 we discuss general approaches to pre–post analysis.

18.3.1 Average or Slope Analysis

In any longitudinal analysis the substantive aims determine which aspects of the response trajectory are most important. For some applications the repeated measures over time may be averaged, or if the timing of measurement is irregular, an area under the curve (AUC) summary can be the primary feature of interest. In these situations statistical analysis will focus on $\overline{Y}_i = 1/n \sum_{j=1}^{n} Y_{ij}$. A key motivation for computing an individual average and then focusing analysis on the derived averages is that standard methods can be used for inference such as a two-sample t-test. However, if there are any incomplete data, the advantage is lost since either subjects with partial data will need to be excluded, or alternative methods need to be invoked to handle the missingness. Attrition in longitudinal studies is unfortunately quite common, and thus derived variable methods are often more difficult to apply validly than they may first appear.

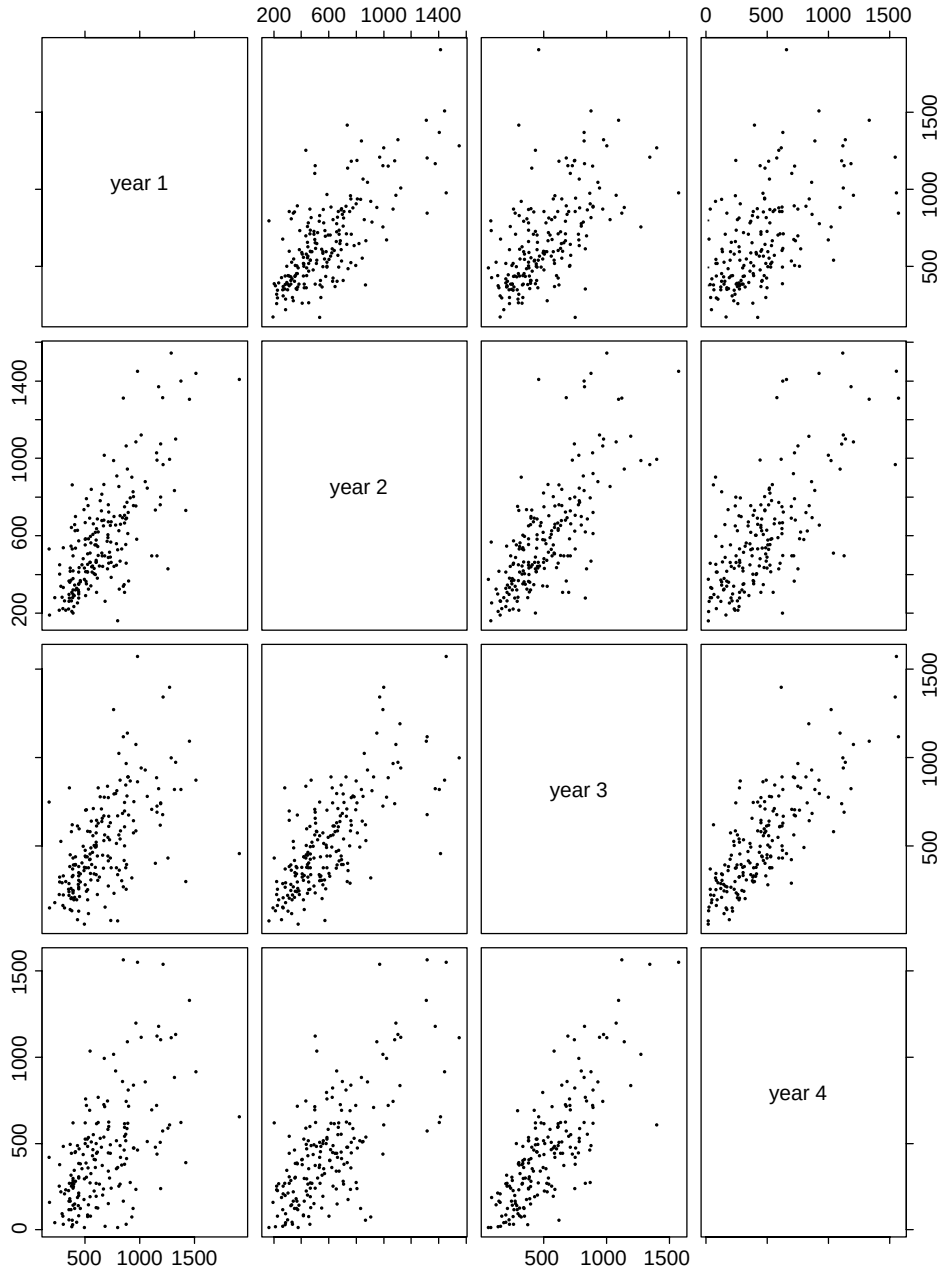


Figure 18.4 Scatter plots of CD4 measurements (counts/mL) taken at years 1 to 4 after seroconversion.

Example 18.1. (*continued*) In the HIVNET informed consent study, the goal is to improve participant knowledge. A derived variable analysis to evaluate evidence for an effect due to the mock informed consent process can be conducted using $\overline{Y}_i = (Y_{i1} + Y_{i2} + Y_{i3})/3$ for the post-baseline times $t_1 =$ six months, $t_2 = 12$ months, and $t_3 = 18$ months. The following table summarizes the data for subjects who have all three post-baseline measurements:

Group	Baseline *N*	Final *N*	Mean	SE	95% CI
Control	947	714	2.038	0.095	
Intervention	177	147	3.444	0.223	
Difference			1.406	0.243	[0.928, 1.885]

First, notice that only $714/947 = 75.4\%$ of control subjects, and $147/177 = 83.1\%$ of intervention subjects have complete data and are therefore included in the analysis. This highlights one major limitation to derived variable analysis: There may be selection bias due to exclusion of subjects with missing data. We discuss missing data issues in Section 18.6. Based on the data above, we would conclude that there is a statistically significant difference between the mean knowledge for the intervention and control groups with a two-sample t-test of $t = 5.796$, $p < 0.001$. Analysis of the single summary for each subject allows the repeated outcome variables to be analyzed using standard independent sample methods.

In other applications, scientific interest centers on the rate of change over time and therefore an individual's slope may be considered as the primary outcome. Typically, each subject in a longitudinal study has only a small number of outcomes collected at the discrete times specified in the protocol. For example, in the MACS data, each subject was to complete a study visit every 6 months and with complete data would have nine measurements between baseline and 48 months. If each subject has complete data, an individual summary statistic can be computed as the regression of outcomes Y_{ij} on times t_j: $Y_{ij} = \beta_{i,0} + \beta_{i,1} t_j + \epsilon_{ij}$; and $\widehat{\beta}_i$ is the ordinary least squares estimate based on data from subject i only. In the case where all subjects have the same collection of measurement times and have complete data, the variation in the estimated slope, $\widehat{\beta}_{i,1}$, will be equal across subjects provided that the variance of ϵ_{ij} is also constant across subjects. Therefore, if

1. The measurement times are common to all subjects: $t_1, t_2, \ldots, t_n$,
2. Each subject has a complete collection of measurements: $Y_{i1}, Y_{i2}, \ldots, Y_{in}$,
3. The within-subject variation $\sigma_i^2 = \text{var}(\epsilon_{ij})$ is constant across subjects: $\sigma_i^2 \equiv \sigma^2$,

then the summaries $\widehat{\beta}_{i,1}$ will have equal variances attributable to using simple linear regression to estimate individual slopes. If any of points 1 to 3 above do not hold, the variance of individual summaries may vary across subjects. This will be the case when each subject has a variable number of outcomes, due to missing data.

When points 1 to 3 are satisfied, simple inference on the derived outcomes $\widehat{\beta}_{i,1}$ can be performed using standard two-sample methods or regression methods. This allows inference regarding factors that are associated with the rate of change over time. If any of points 1 to 3 do not hold, mixed model regression methods (Section 18.5) may be preferable to simple derived variable methods. See Frison and Pocock [1992, 1997] for further discussion of derived variable methods.

Example 18.2. (*continued*) For the MACS data, we are interested in determining whether the rate of decline in CD4 is correlated with the baseline viral load measurement. In Section 18.2 we looked at descriptive statistics comparing the mean CD4 count over time for categories of viral load. We now explore the association between the rate of decline and baseline viral load by obtaining a summary statistic, using the individual time slope $\widehat{\beta}_i$ obtained from a regression of the CD4 count Y_{ij} on measurement time t_{ij}. Figure 18.5 shows a scatter plot of the individual slope estimates plotted against the log of baseline viral load. First notice that plotting symbols of different sizes are used to reflect the fact that the number of measurements per subject, n_i,

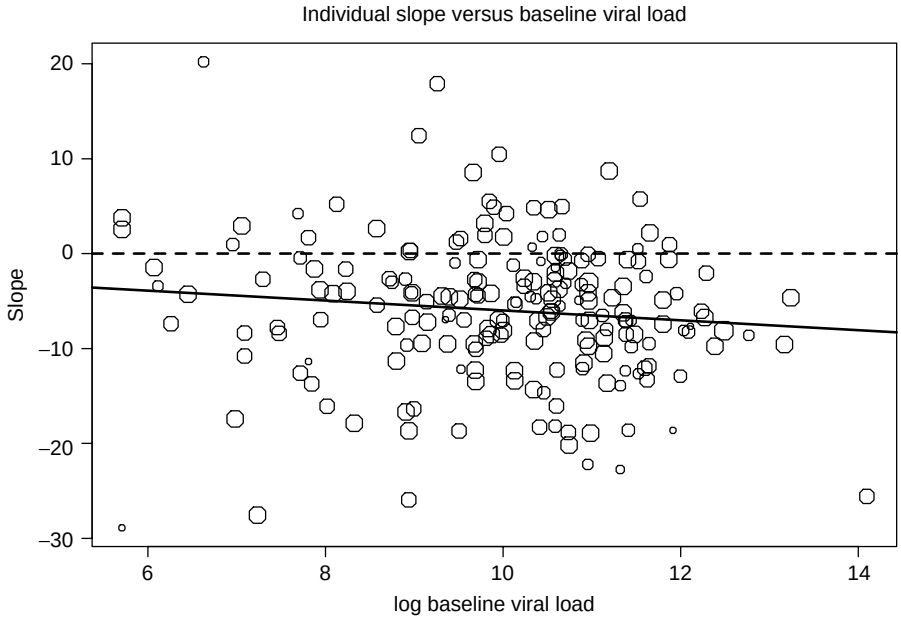


Figure 18.5 Individual CD4 slopes (count/month) vs. log of baseline viral load, MACS data.

is not constant. The plotting symbol size is proportional to n_i. For the MACS data we have the following distribution for the number of observations per subjects over the first four years:

	Number of Observations, n_i								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Number of subjects	5	13	8	10	25	44	82	117	3

For Figure 18.5 the $(5 + 13) = 18$ subjects with either one or two measurements were excluded as a summary slope is either unestimable ($n_i = 1$) or highly variable ($n_i = 2$). Figure 18.5 suggests that there is a pattern of decreasing slope with increasing log baseline viral load. However, there is also a great deal of subject-to-subject variation in the slopes, with some subjects having $\widehat{\beta}_{i,1} > 0$ count/month, indicating a stable or increasing trend, and some subjects having $\widehat{\beta}_{i,1} < 15$ count/month, suggesting a steep decline in their CD4. A linear regression using the individual slope as the response and log baseline viral load as the predictor yields a p-value of 0.124, implying a nonsignificant linear association between the summary statistic $\widehat{\beta}_{i,1}$ and log baseline viral load.

A categorical analysis using tertiles of baseline viral load parallels the descriptive statistics presented in Table 18.1. The average rate of decline in CD4 can be estimated as the mean of the individual slope estimates:

	***N* Subjects**	**Average Slope**	**SE**
Low viral load	66	−5.715	1.103
Medium viral load	69	−4.697	0.802
High viral load	65	−7.627	0.789

We find similar average rates of decline for the medium- and low-viral-load groups and find a greater rate of decline for the high-viral-load group. Using ANOVA, we obtain an F-statistic of 2.68 on 2 and 197 degrees of freedom, with a p-value of 0.071, indicating that we would not reject equality of average rates of decline using the nominal 5% significance level.

Note that neither simple linear regression nor ANOVA accounts for the fact that response variables $\widehat{\beta}_{i,1}$ may have unequal variance due to differing n_i. In addition, a small number of subjects were excluded from the analysis since a slope summary was unavailable. In Section 18.5 we discuss regression methods for correlated data that can efficiently use all of the available data to make inferences with longitudinal data.

18.3.2 Pre–Post Analysis

In this section we discuss analytic methods appropriate when a single baseline and a single follow-up measurement are available. We focus on the situation where interest is in the comparison of two groups: $X_i = 0$ denotes membership in a reference or control group; and $X_i = 1$ denotes membership in an exposure or intervention group. Assume for each subject i that we have a baseline measurement denoted as Y_{i0} and a follow-up measurement denoted as Y_{i1}. The following table summarizes three main analysis options using regression methods to characterize the two-group comparison:

Follow-up only:	$Y_{i1} = \beta_0 + \beta_1 X_i + \epsilon_i$
Change analysis:	$Y_{i1} - Y_{i0} = \beta_0^* + \beta_1^* X_i + \epsilon_i^*$
ANCOVA:	$Y_{i1} = \beta_0^{**} + \beta_1^{**} X_i + \beta_2^{**} Y_{i0} + \epsilon_i^{**}$

Since X_i is a binary response variable we can interpret the coefficients β_1, β_1^*, and β_1^{**} as differences in means comparing $X_i = 1$ to $X_i = 0$. Specifically, for the follow-up only analysis the coefficient β_1 represents the difference in the *mean response at follow-up* comparing $X_i = 1$ to $X_i = 0$. If the assignment to $X_i = 0/1$ was randomized, the simple follow-up comparison is a valid causal analysis of the effect of the treatment. For change analysis the coefficient β_1^* is interpreted as the difference between the *average change* for $X_i = 1$ as compared to the average change for $X_i = 0$. Finally, using ANCOVA estimates β_1^{**}, which represents the difference in the mean follow-up outcome comparing exposed ($X_i = 1$) to unexposed ($X_i = 0$) subjects who are *equal in their baseline response*. Equivalently, we interpret β_1^{**} as the comparison of treated versus control subjects after adjusting for baseline.

It is important to recognize that each of these regression models provides parameters with different interpretations. In situations where the selection of treatment or exposure is not randomized, the ANCOVA analysis can control for "confounding due to indication," or where the baseline value Y_{i0} is associated with a greater or lesser likelihood of receiving the treatment $X_i = 1$. When treatment is randomized, Frison and Pocock [1992] show that $\beta_1 = \beta_1^* = \beta_1^{**}$. This result implies that for a randomized exposure each approach can provide a valid estimate of the average causal effect of treatment. However, Frison and Pocock [1992] also show that the most *precise* estimate of β_1 is obtained using ANCOVA, and that final measurement analysis is more precise than the change analysis when the correlation between baseline and follow-up measurements is less than 0.50. This results from $\text{var}(Y_{i1} - Y_{i0}) = 2\sigma^2(1 - \rho)$, which is less than σ^2 only when $\rho > \frac{1}{2}$.

Example 18.1. (*continued*) To evaluate the effect of the HIVNET mock informed consent, we focus analysis on the baseline and six-month knowledge scores. The following tables give

inference for the follow-up, Y_{i1}:

Group	N	6-month Mean	SE	95% CI
Control	834	1.494	0.111	
Intervention	169	3.391	0.240	
Difference		1.900	0.264	[1.375, 2.418]

and for the change in knowledge score, $Y_{i1} - Y_{i0}$, for the 834/947 control subjects and 169/177 intervention subjects who have both baseline and six-month outcomes:

Group	N	Mean Change	SE	95% CI
Control	834	0.243	0.118	
Intervention	169	2.373	0.263	
Difference		2.130	0.288	[1.562, 2.697]

The correlation between baseline and month 6 knowledge score is 0.462 among controls and 0.411 among intervention subjects. Since $\rho < 0.5$, we expect an analysis of the change in knowledge score to lead to a larger standard error for the treatment effect than a simple cross-sectional analysis of scores at the six-month visit.

Alternatively, we can regress the follow-up on baseline and treatment:

Coefficients	Estimate	SE	Z-value
(Intercept)	0.946	0.105	9.05
Treatment	1.999	0.241	8.30
Baseline (Y_{i0})	0.438	0.027	16.10

In this analysis the estimate of the treatment effect is 1.999, with a standard error of 0.241. The estimate of β_1 is similar to that obtained from a cross-sectional analysis using six-month data only, and to the analysis of the change in knowledge score. However, as predicted, the standard error is smaller that the standard error for each alternative analysis approach. Finally, in Figure 18.6, the six-month knowledge score is plotted against the baseline knowledge score. Separate regression lines are fit and plotted for the intervention and control groups. We see that the fitted lines are nearly parallel, indicating that the ANCOVA assumption is satisfied for these data.

For discrete outcomes, different pre–post analysis options can be considered. For example, with a binary baseline, $Y_{i0} = 0/1$, and a binary follow-up, $Y_{i1} = 0/1$, the difference, $Y_{i1} - Y_{i0}$, takes the values $-1, 0, +1$. A value of -1 means that a subject has changed from $Y_{i0} = 1$ to $Y_{i1} = 0$, while $+1$ means that a subject has changed from $Y_{i0} = 0$ to $Y_{i1} = 1$. A difference of 0 means that a subject had the same response at baseline and follow-up and does not distinguish between $Y_{i0} = Y_{i1} = 0$ and $Y_{i0} = Y_{i1} = 1$. Rather than focus on the difference, it is useful to consider an analysis of change by subsetting on the baseline value. For example, in a comparative study we can subset on subjects with baseline value $Y_{i0} = 1$ and then assess the difference between intervention and control groups with respect to the percent that respond $Y_{i1} = 1$ at follow-up. This analysis allows inference regarding differential change from 0 to 1 comparing

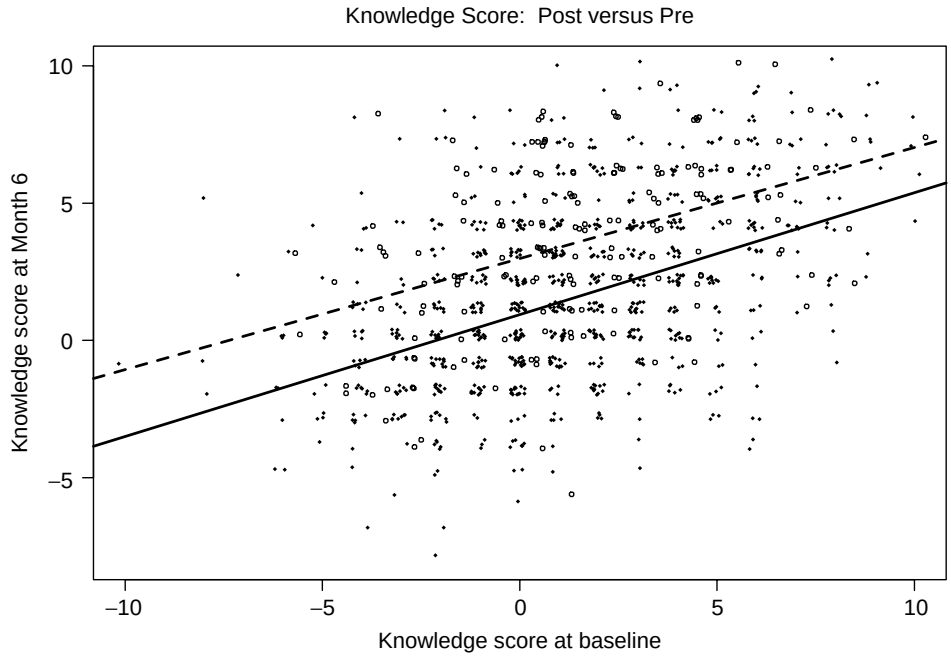


Figure 18.6 Month 6 knowledge score vs. baseline knowledge score (jittered), HIVNET informed consent substudy. Open points and dashed line represent intervention; solid points and line represent control.

the two groups. When a response value of 1 indicates a positive outcome, this analysis provides information about the "corrective" potential for intervention and control groups. An analysis that restricts to subjects with baseline $Y_{i0} = 1$ and then comparing treatment and control subjects at follow-up will focus on a second aspect of change. In this case we are summarizing the fraction of subjects that start with $Y_{i0} = 1$ and then remain with $Y_{i1} = 1$ and thus do not change their outcome but rather, maintain the outcome. When the outcome $Y_{ij} = 1$ indicates a favorable status, this analysis summarizes the relative ability of intervention and control groups to "maintain" the favorable status. Statistical inference can be based on standard two-sample methods for binary data (see Chapter 6). An analysis that summarizes current status at follow-up stratifying on the baseline, or previous outcome, is a special case of a transition model (see Diggle et al. [2002, Chap. 10]).

Example 18.1. (*continued*) The HIVNET informed consent substudy was designed to evaluate whether an informed consent procedure could correct misunderstanding regarding vaccine trial conduct and to reinforce understanding that may be tentative. In Section 18.2 we saw that for the safety item assessment at six months the intervention group had 50% of subjects answer correctly as compared to only 43% of control subjects. For the nurse item the fractions answering correctly at six months were 72% and 45% for intervention and control groups, respectively. By analyzing the six-month outcome separately for subjects that answered incorrectly at baseline, $Y_{i0} = 0$, and for subjects that answered correctly at baseline, $Y_{i0} = 1$, we can assess the mechanisms that lead to the group differences at six months: Does the intervention experience lead to greater rates of "correction" where answers go from $0 \rightarrow 1$ for baseline and six-month assessments; and does intervention appear to help "maintain" or reinforce correct knowledge by leading to increased rates of $1 \rightarrow 1$ for baseline and six-month responses?

The following table stratifies the month 6 safety knowledge item by the baseline response:

Safety Item

"Correction" : $Y_{i0} = 0$		Percent Correct	"Maintain" : $Y_{i0} = 1$		Percent Correct
	N	$Y_{i1} = 1$		N	$Y_{i1} = 1$
Control	488	160/488 = 33%	Control	349	198/349 = 57%
Intervention	105	43/105 = 41%	Intervention	65	42/65 = 65%

This table shows that of the 105 intervention subjects that answered the safety item at baseline incorrectly, a total of 43, or 41%, subsequently answered the item correctly at the 6-month follow-up visit. In the control group only 160/488 = 33% answered this item correctly at six months after they had answered incorrectly at baseline. A two-sample test of proportions yields a p-value of 0.118, indicating a nonsignificant difference between the intervention and control groups in their rates of correcting knowledge of this item. For subjects that answered this item correctly at baseline, 42/65 = 65% of intervention subjects and 198/349 = 57% of control subjects continued to respond correctly. A two-sample test of proportions yields a p-value of 0.230, indicating a nonsignificant difference between the intervention and control groups in their rates of maintaining correct knowledge of the safety item. Therefore, although the intervention group has slightly higher proportions of subjects that switch from incorrect to correct, and that stay correct, these differences are not statistically significant.

For the nurse item we saw that the informed consent led to a large fraction of subjects who answered the item correctly. At six months the intervention group had 72% of subjects answer correctly, while the control group had 45% answer correctly. Focusing on the mechanisms for this difference we find:

Nurse Item

"Correction" : $Y_{i0} = 0$		Percent Correct	"Maintain" : $Y_{i0} = 1$		Percent Correct
	N	$Y_{i1} = 1$		N	$Y_{i1} = 1$
Control	382	122/382 = 32%	Control	455	252/455 = 55%
Intervention	87	59/87 = 68%	Intervention	85	65/85 = 76%

Thus intervention led to a correction for 68% of subjects with an incorrect baseline response compared to 32% among controls. A two-sample test of proportions yields a p-value of <0.001, and a confidence interval for the difference in proportions of (0.250, 0.468). Therefore, the intervention has led to a significantly different rate of correction for the nurse item. Among subjects who correctly answered the nurse item at baseline, only 55% of control subjects answered correctly again at month 6, while 76% of intervention subjects maintained a correct answer at six months. Comparison of the proportion that maintain correct answers yields a p-value of <0.001 and a 95% confidence interval for the difference in probability of a repeat correct answer of (0.113, 0.339). Therefore, the informed consent intervention led to significantly different rates of both correction and maintenance for the safety item.

These categorical longitudinal data could also be considered as multiway contingency tables and analyzed by the methods discussed in Chapter 7.

18.4 IMPACT OF CORRELATION ON INFERENCE

For proper analysis of longitudinal data the within-subject correlation needs to be addressed. In Section 18.3.1 we discussed one method that avoids considering correlation among repeated measures by reducing the multiple measurements to a single summary statistic. In situations where there are variable numbers of observations per subject, alternative approaches are preferable. However, to analyze longitudinal outcomes, either a model for the correlation needs to be adopted or the standard error for statistical summaries needs to be adjusted. In this section we discuss some common correlation models and discuss the impact of the correlation on the standard errors and sample size.

18.4.1 Common Types of Within-Subject Correlation

The simplest correlation structure is the *exchangeable* or *compound symmetric* model, where

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & & & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

In this case the correlation between any two measurements on a given subject is assumed to be equal, $\text{corr}(Y_{ij}, Y_{ik}) = \rho_{jk} \equiv \rho$. The longitudinal outcomes form a simple "cluster" of responses, and the time ordering is not considered when characterizing correlation.

In other models the measurement time or measurement order is used to model correlation. For example, a *banded* correlation is

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-4} \\ \vdots & & & \ddots & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \rho_{n-4} & \cdots & 1 \end{bmatrix}$$

and an *autoregressive* structure is

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \cdots & \rho^{|t_1-t_n|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_2-t_3|} & \cdots & \rho^{|t_2-t_n|} \\ \rho^{|t_3-t_1|} & \rho^{|t_3-t_2|} & 1 & \cdots & \rho^{|t_3-t_n|} \\ \vdots & & & \ddots & \vdots \\ \rho^{|t_n-t_1|} & \rho^{|t_n-t_2|} & \rho^{|t_n-t_3|} & \cdots & 1 \end{bmatrix}$$

Each of these models is a special case of a serial correlation model where the distance between observations determines the correlation. In a banded model correlation between observations is determined by their order. All observations that are adjacent in time are assumed to have an equal correlation: $\text{corr}(Y_{i1}, Y_{i2}) = \text{corr}(Y_{i2}, Y_{i3}) = \cdots = \text{corr}(Y_{in-1}, Y_{in}) = \rho_1$. Similarly, all observations that are two visits apart have correlation ρ_2, and in general all pairs of observations that are k visits apart have correlation ρ_k. A banded correlation matrix will have a total of $n-1$ correlation parameters. The autoregressive correlation model uses a single correlation parameter and assumes that the time separation between measurements determines their correlation through the model

$\text{corr}(Y_{ij}, Y_{ik}) = \rho^{|t_j - t_k|}$. Thus, if $\rho = 0.8$ and observations are 1 unit apart in time, their correlation will be $0.8^1 = 0.8$, while if they are 2 units apart, their correlation will be $0.8^2 = 0.64$. In an autoregressive model the correlation will decay as the distance between observations increases.

There are a large number of correlation models beyond the simple exchangeable and serial models given above. See Verbeke and Molenberghs [2000] and Diggle et al. [2002] for further examples.

18.4.2 Variance Inflation Factor

The impact of correlated observations on summaries such as the mean of all observations taken over time and across all subjects will depend on the specific form of the within-subject correlation. For example,

$$\overline{Y} = \frac{1}{\sum_i n_i} \sum_{i=1}^{N} \sum_{j=1}^{n_i} Y_{ij}$$

$$\text{var}(\overline{Y}) = \frac{1}{(\sum_i n_i)^2} \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \text{var}(Y_{ij}) \; + \; \sum_{j=1}^{n_i - 1} \sum_{k=(j+1)}^{n_i} 2 \times \text{cov}(Y_{ij}, Y_{ik}) \right]$$

If the variance is constant, $\text{var}(Y_{ij}) = \sigma^2$, we obtain

$$\text{var}(\overline{Y}) = \frac{\sigma^2}{(\sum_i n_i)^2} \sum_{i=1}^{N} \left[n_i \; + \; \sum_{j=1}^{n_i - 1} \sum_{k=(j+1)}^{n_i} 2 \times \text{corr}(Y_{ij}, Y_{ik}) \right]$$

Finally, if all subjects have the same number of observations, $n_i \equiv n$, and the correlation is exchangeable, $\rho_{jk} \equiv \rho$, the variance of the mean is

$$\text{var}(\overline{Y}) = \frac{\sigma^2}{Nn} [1 \; + \; (n-1)\rho]$$

The factor $[1 + (n-1) \cdot \rho]$ is referred to as the *variance inflation factor*, since this measures the increase (when $\rho > 0$) in the variance of the mean (calculated using $N \cdot n$ observations) that is due to the within-subject correlation of measurements.

To demonstrate the impact of correlation on the variance of the mean, we calculate the variance inflation factor, $1 + (n-1)\rho$, for various values of cluster size, n, and correlation, ρ, in Table 18.4. This shows that even very small within-cluster correlations can have an important impact on standard errors if clusters are large. For example, a variance inflation factor of 2.0 arises with ($\rho = 0.001, n = 1001$), ($\rho = 0.01, n = 101$), or ($\rho = 0.10, n = 11$). The variance

Table 18.4 Variance Inflation Factors

n Cluster Size	ρ				
	0.001	0.01	0.02	0.05	0.1
2	1.001	1.01	1.02	1.05	1.10
5	1.004	1.04	1.08	1.20	1.40
10	1.009	1.09	1.18	1.45	1.90
100	1.099	1.99	2.98	5.95	10.90
1000	1.999	10.99	20.98	50.95	100.90

inflation factor becomes important when planning a study. In particular, when treatment is given to groups of subjects (e.g., a cluster randomized study), the variance inflation factor needs to be estimated to power the study properly. See Koepsell et al. [1991] or Donner and Klar [1994, 1997] for a discussion of design and analysis issues in cluster randomized studies. For longitudinal data each subject is a "cluster," with individual measurements taken within each subject.

18.5 REGRESSION METHODS

Regression methods permit inference regarding the average response trajectory over time and how this evolution varies with patient characteristics such as treatment assignment or other demographic factors. However, standard regression methods assume that all observations are independent and if applied to longitudinal outcomes may produce invalid standard errors. There are two main approaches to obtaining valid inference: A complete model that includes specific assumptions regarding the correlation of observations within a subject can be adopted and used to estimate the standard error of regression parameter estimates; general regression methods can be used and the standard errors can be corrected to account for the correlated outcomes. In the following section we review a regression method for continuous outcomes that models longitudinal data by assuming random errors within a subject and random variation in the trajectory among subjects.

18.5.1 Mixed Models

Figure 18.7 presents hypothetical longitudinal data for two subjects. In the figure monthly observations are recorded for up to one year, but one person drops out prior to the eight-month visit, and thus the observations for months 8 through 12 are not recorded. Notice that each subject

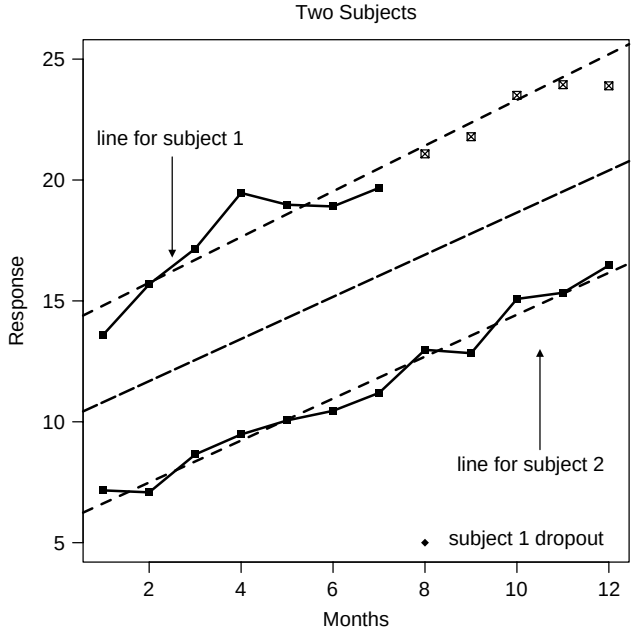


Figure 18.7 Hypothetical longitudinal data for two subjects. Each subject has an individual linear trajectory, and one subject has incomplete data due to dropout.

appears to be tracking his or her own linear trajectory but with small fluctuations about the line. The deviations from the individual observations to the individual's line are referred to as the within-subject variation in the outcomes. If we only had data for a single subject, these would be the typical error terms in a regression equation. In most situations the subjects in a study represent a random sample from a well-defined target population. In this case the specific individual line that a subject happens to follow is not of primary interest, but rather the *typical* linear trajectory and perhaps the magnitude of subject-to-subject variation in the longitudinal process. A dashed line in the center of Figure 18.7 shows the average of individual linear-time trajectories. This average curve characterizes the average for the population as a function of time. For example, the value of the dashed line at month 2 denotes the cross-sectional mean response if the two-month observation for all subjects was averaged. Similarly, the fitted value for the dashed line at 10 months represents the average in the population for the 10-month measurement. Therefore, the average line in Figure 18.7 represents both the typical trajectory and the population average as a function of time.

Linear mixed models make specific assumptions about the variation in observations attributable to variation within a subject and to variation among subjects. The within-subject variation is seen in Figure 18.7 as the deviation between individual observations, Y_{ij}, and the individual linear trajectory. Let $\beta_{i,0} + \beta_{i,1}X_{ij}$ denote the line that characterizes the observation path for subject i. In this example X_{ij} denotes the time of measurement j on subject i. Note that each subject has an individual-specific intercept and slope. Within-subject variation is seen in the magnitude of variation in the deviation between the observations and the individual trajectory, $Y_{ij} - (\beta_{i,0} + \beta_{i,1}X_{ij})$. The between-subject variation is represented by the variation among the intercepts, $\text{var}(\beta_{i,0})$, and the variation among subjects in the slopes, $\text{var}(\beta_{i,1})$.

If parametric assumptions are made regarding the within- and between-subject components of variation, maximum likelihood methods can be used to estimate the regression parameters which characterize the population average, and the variance components which characterize the magnitude of within- and between-subject heterogeneity. For continuous outcomes it is convenient to assume that within-subject errors are normally distributed and to assume that intercepts and slopes are normally distributed among subjects. Formally, these assumptions are written as:

$$\begin{aligned}
\text{within-subjects} \quad : \quad E(Y_{ij} \mid \beta_i) &= \beta_{i,0} + \beta_{i,1}X_{ij} \\
Y_{ij} &= \beta_{i,0} + \beta_{i,1}X_{ij} + \epsilon_{ij} \\
\epsilon_{ij} &\sim N(0, \sigma^2) \\
\text{between-subjects} \quad : \quad \begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix} &\sim N\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix} \right]
\end{aligned}$$

The model can be rewritten using $b_{i,0} = (\beta_{i,0} - \beta_0)$ and $b_{i,1} = (\beta_{i,1} - \beta_1)$:

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 X_{ij}}_{\text{systematic}} + \underbrace{b_{i,0} + b_{i,1}X_{ij} + \epsilon_{ij}}_{\text{random}} \tag{1}$$

In this representation the terms $b_{i,0}$ and $b_{i,1}$ represent deviations from the population average intercept and slope, respectively. These "random effects" now have mean 0 by definition, but their variance and covariance is still given by the elements of the matrix D. For example, $\text{var}(b_{i,0}) = D_{00}$ and $\text{var}(b_{i,0}) = D_{11}$. In equation (1) the "systematic" variation in outcomes is given by the regression parameters β_0 and β_1. These parameters determine how the average for subpopulations differs across distinct values of the covariates, X_{ij}.

In equation (1) the random components are partitioned into the observation-level and subject-level fluctuations:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \underbrace{b_{i,0} + b_{i,1} X_{ij}}_{\text{between-subject}} + \underbrace{\epsilon_{ij}}_{\text{within-subject}}$$

A more general form is

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}_{\text{fixed effects}} + \underbrace{b_{i,0} + b_{i,1} X_{i1} + \cdots + b_{i,q} X_{iq}}_{\text{random effects}} + \epsilon_{ij}$$

$$Y_{ij} = X'_{ij}\beta + Z'_{ij} b_i + \epsilon_{ij}$$

where $X'_{ij} = [X_{ij,1}, X_{ij,2}, \ldots, X_{ij,p}]$ and $Z'_{ij} = [X_{ij,1}, X_{ij,2}, \ldots, X_{ij,q}]$. In general, we assume that the covariates in Z_{ij} are a subset of the variables in X_{ij} and thus $q < p$. In this model the coefficient of covariate k for subject i is given as $(\beta_k + b_{i,k})$ if $k \leq q$ and is simply β_k if $q < k \leq p$. Therefore, in a linear mixed model there may be some regression parameters that vary among subjects, while some regression parameters are common to all subjects. For example, in Figure 18.7 it is apparent that each subject has his or her own intercept, but the subjects may have a common slope. A *random intercept model* assumes parallel trajectories for any two subjects and is given as a special case of the general mixed model:

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + b_{i,0} + \epsilon_{ij}$$

In this model the intercept for subject i is given by $\beta_0 + b_{i,0}$, while the slope for subject i is simply β_1, since there is no additional random slope, $b_{i,1}$, in the random intercept model.

Laird and Ware [1982] discuss the linear mixed model and specific methods to obtain maximum likelihood estimates. Although linear mixed models can be computationally difficult to fit, modern software packages contain excellent numerical routines for estimating parameters and computing standard errors. For example, the SAS package contains the MIXED procedure and S-PLUS has the lme() function.

Example 18.2. (*continued*) In Section 18.3.1 we explored the change over time in CD4 counts for groups of subjects according to their baseline viral load value. Using linear mixed models we can estimate the average rate of decline for each baseline viral load category, and test for differences in the rate of decline.

To test for differences in the rate of decline, we use linear regression with

$$\begin{aligned} E(Y_{ij} \mid X_{ij}) = \beta_0 &+ \\ &\beta_1 \cdot \text{month} + \\ &\beta_2 \cdot I(\text{medium viral load}) + \\ &\beta_3 \cdot I(\text{high viral load}) + \\ &\beta_4 \cdot \text{month} \cdot I(\text{medium viral load}) + \\ &\beta_5 \cdot \text{month} \cdot I(\text{high viral load}) \quad . \end{aligned}$$

Here $X_{ij,3} = I(\text{medium viral load}) = 1$ if subject i has a medium value for baseline viral load and otherwise $= 0$, and $X_{ij,4} = I(\text{high viral load}) = 1$ if subject i has a high baseline viral load and otherwise $= 0$. Using this regression model, the average slope for the low baseline viral category is given by β_1, while the average slope for the other viral load categories are given by $(\beta_1 + \beta_4)$ and $(\beta_1 + \beta_5)$ for the medium- and high-viral-load categories, respectively. If the

estimate of β_4 is not significantly different from 0, we cannot reject equality of the average rates of decline for the low- and medium-viral-load subjects. Similarly, inference regarding β_5 determines whether there is evidence that the rate of decline for high-viral-load subjects is different than for low-viral-load subjects.

The linear mixed model is specified by the regression model for $E(Y_{ij} \mid X_{ij}) = \mu_{ij}$ and assumptions about random effects. We first assume random intercepts, $Y_{ij} = \mu_{ij} + b_{i,0} + \epsilon_{ij}$, and then allow random intercepts and slopes, $Y_{ij} = \mu_{ij} + b_{i,0} + b_{i,1} \cdot \text{month} + \epsilon_{ij}$. Maximum likelihood estimates are presented in Tables 18.5 and 18.6. In Table 18.5 the mixed model assumes that each subject has a random intercept, $b_{i,0}$, but assumes a common slope. In this model there are two estimated variance components: $162.5 = \widehat{\sigma} = \sqrt{\widehat{\text{var}}(\epsilon_{ij})}$ and $219.1 = \sqrt{\widehat{D_{00}}} = \sqrt{\widehat{\text{var}}(b_{i,0})}$. The total variation in CD4 is estimated as $162.5^2 + 219.1^2 = 272.8^2$, and the proportion of total variation that is attributed to within-person variability is $162.5^2/272.8^2 = 35\%$ with $219.1^2/272.8^2 = 65\%$ of total variation attributable to individual variation in their general level of CD4 (e.g., attributable to random intercepts).

Estimates from Table 18.5 are interpreted as follows:

- *(Intercept)* $\widehat{\beta}_0 = 803.4$. The intercept is an estimate of the mean CD4 count at seroconversion (i.e., month = 0) among the low-viral-load subjects.
- *month* $\widehat{\beta}_1 = -5.398$: Among subjects in the low-viral-load group, the mean CD4 declines -5.398 units per month.
- *I[Medium Viral Load]* $\widehat{\beta}_2 = -123.72$. At seroconversion the average CD4 among subjects with a medium value for baseline viral load is 123.72 units lower than the average CD4 among the low-viral-load subjects.
- *I[High Viral Load]* $\widehat{\beta}_3 = -146.40$. At seroconversion the average CD4 among subjects with a high value for baseline viral load is 146.40 units lower than the average CD4 among the low-viral-load subjects.
- *month * I[Medium Viral Load]* $\widehat{\beta}_4 = 0.169$. The rate of decline for subjects in the medium-viral-load category is estimated to be 0.169 count/month higher than the rate of decline among subjects with a low-baseline viral load. The rate of change in mean CD4 is estimated as $-5.398 + 0.169 = -5.229$ counts/month among subjects with medium-baseline viral load.

Table 18.5 Linear Mixed Model Results for the CD4 Data Assuming Random Intercepts[a]

Linear mixed-effects model fit by maximum likelihood

Data: MACS

AIC	BIC	logLik
19838.98	19881.38	−9911.491

Random effects:

Formula: ~ 1 | id

	(Intercept)	Residual
StdDev:	219.1106	162.5071

Fixed effects: cd4 ~ month * vcat

	Value	Std.Error	DF	t-value	p-value
(Intercept)	803.356	29.712	1250	27.04	<.0001
month	−5.398	0.578	1250	−9.34	<.0001
I[Medium Viral Load]	−123.724	42.169	223	−2.93	0.0037
I[High Viral Load]	−146.401	42.325	223	−3.46	0.0006
month * I[Medium Viral Load]	0.169	0.812	1250	0.21	0.8351
month * I[High Viral Load]	−1.968	0.817	1250	−2.41	0.0162

[a]Output from S-PLUS.

Table 18.6 Linear Mixed Model Results for the CD4 Data Assuming Random Intercepts and Slopes[a]

Linear mixed-effects model fit by maximum likelihood

Data: MACS

AIC	BIC	logLik
19719.85	19772.84	−9849.927

Random effects:

Formula: ~ 1 + month | id

Structure: General positive-definite

	StdDev	Corr
(Intercept)	244.05874	(Inter
month	5.68101	−0.441
Residual	142.22835	

Fixed effects: cd4 ~ month * vcat

	Value	Std.Error	DF	t-value	p-value
(Intercept)	803.509	31.373	1250	25.61	<.0001
month	−5.322	0.857	1250	−6.21	<.0001
I[Medium Viral Load]	−125.548	44.536	223	−2.82	0.0053
I[High Viral Load]	−142.177	44.714	223	−3.18	0.0017
month * I[Medium Viral Load]	0.159	1.205	1250	0.13	0.8954
month * I[High Viral Load]	−2.240	1.212	1250	−1.85	0.0648

[a]Output from S-PLUS.

- *month * I[High Viral Load]* $\widehat{\beta}_5 = -1.967$. The rate of decline for subjects in the high-viral-load category is estimated to be -1.967 counts/month lower than the rate of decline among subjects with a low-baseline viral load. The rate of change in mean CD4 is estimated as $-5.398 - 1.967 = -7.365$ counts/month among subjects with a high-baseline viral load.

Although the regression output also includes standard errors for each of the regression estimates, we defer making inference since a model with random intercepts and random slopes appears more appropriate and affects the resulting confidence intervals or tests for the regression estimates (see Table 18.6).

In Table 18.6 we present maximum likelihood estimates assuming random intercepts and random slopes. To assess whether the additional flexibility is warranted, we can evaluate the improvement in the fit to the data as measured by the maximized log likelihood. The maximized log likelihood for random intercepts is -9911.49 (see Table 18.5), while the maximized log likelihood is increased by 61.56 to -9849.93 when also allowing random intercepts. A formal likelihood ratio test is possible since the random intercepts and random intercepts plus slopes form nested models, but since the null hypothesis restriction involves $D_{11} = 0$, which is on the boundary of the allowable values for variance components (i.e., $D_{11} \geq 0$), the null reference distribution is of nonstandard form [Stram and Lee, 1994; Verbeke and Molenberghs, 2000]. However, the increase in maximized log likelihood of 61.56 is quite substantial and statistically significant with $p < 0.001$. Although the variance assumptions can be further relaxed to allow serial correlation in the measurement errors, ϵ_{ij}, the improvement in the maximized log likelihood is small and does not substantially affect the conclusions. We refer the reader to Diggle et al. [2002] and Verbeke and Molenberghs [2000] for further detail regarding linear mixed models that also include serial correlation in the errors.

Table 18.6 gives estimates of the variance components. For example, the standard deviation in intercepts is estimated as $\sqrt{\widehat{D}_{00}} = 244.1$ and the standard deviation of slopes is given

as $\sqrt{\widehat{D}_{11}} = 5.681$. Under the assumption of normally distributed random effects, these estimates imply that 95% of subjects with a low-baseline viral load would have a *mean* CD4 at seroconversion between $803.5 - 1.96 \times 244.1 = 325.1$ and $803.5 + 1.96 \times 244.1 = 1281.9$. We emphasize that this interval is for individual values of the mean CD4 at baseline rather than for individual measurements at baseline. The interval (325.1, 1281.9) does not include the measurement variation attributable to ϵ_{ij} so only describes the variation in the means, $\beta_0 + b_{i,0}$, and not the actual CD4 measurements, $Y_{ij} = \beta_0 + b_{i,0} + \epsilon_{ij}$. Similarly, 95% of low-viral-load subjects are expected to have a slope of $-5.322 \pm 1.96 \times 5.681 = (-16.456, 5.813)$ counts/month.

The estimated regression parameters can be used to make inference regarding the average rate of decline for each of the baseline viral load categories. For example, $\widehat{\beta}_4 = 0.159$ estimates the difference between the rate of decline among medium-viral-load subjects and low-viral-load subjects and is not significantly different from 0 using the standardized regression coefficient as test statistic: $0.159/1.205 = 0.13$ with $p = 0.8954$. Although the estimated rate of decline is lower for the high-viral-load group, $\widehat{\beta}_5 = -2.240$, this is also not significantly different from 0 with p-value 0.0648. It is important to point out that inference using linear mixed models can be quite sensitive to the specific random effects assumptions. If a random intercepts model were used, the comparison of high- versus low-viral-load group slopes over time becomes statistically significant, as seen in Table 18.5, where the p-value for testing $H_0 : \beta_5 = 0$ is $p = 0.0162$, which would naively lead to rejection of the null hypothesis. This inference is invalid, as it assumes that slopes do not vary among individuals, and the data clearly suggest between-subject variation in slopes.

Residual plots can be useful for checking the assumptions made by the linear mixed model. However, there are two types of residuals that can be used. First, the *population residuals* are defined as

$$\begin{aligned} R_{ij}^P &= Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{ij,1} + \cdots + \widehat{\beta}_p X_{ij,p}) \\ &= Y_{ij} - X_{ij}' \widehat{\beta} \end{aligned}$$

The population residuals measure the deviation from the individual measurement to the fitted population mean value. These residuals contain all components of variation, including between- and within-subject deviations since

$$Y_{ij} - X_{ij}' \beta = Z_{ij}' b_i + \epsilon_{ij}$$

The population residuals can be used to evaluate evidence for systematic departures from linear assumptions. Similar to standard multiple regression, plots of residuals versus predictors can be inspected for curvature.

Individual random effects b_i can also be estimated and used to form a second type of residual. Under the linear mixed model, these random effects are typically not estimated simply by using subject i data only to estimate b_i, but rather by using both the individual data $Y_{i1}, Y_{i2}, \ldots, Y_{i,n_i}$ and the assumption that random effects are realizations from a normal distribution among subjects. Empirical Bayes' estimates of b_i balance the assumption that b_i is intrinsic to generating the data Y_{ij} in addition to the assumption that the distribution of b_i is multivariate normal with mean 0. Thus, empirical Bayes' estimates are typically closer to 0 than estimates that would be obtained solely by using individual i data. See Carlin and Louis [1996] for more detail on empirical Bayes' estimation. Using the estimated random effects provides a second residual:

$$\begin{aligned} R_{ij}^W &= Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{ij,1} + \cdots + \widehat{\beta}_p X_{ij,p}) \\ &\quad - (\hat{b}_{i,0} + \hat{b}_{i,1} X_{ij,1} + \cdots + \hat{b}_{i,q} X_{ij,q}) \\ &= Y_{ij} - X_{ij}' \widehat{\beta} - Z_{ij}' \widehat{b}_i \end{aligned}$$

If the regression parameter β and the random effects b were known rather than estimated, the residual R_{ij}^W would equal the within-subject error ϵ_{ij}. The within-subject residuals R_{ij}^W can be used to assess the assumptions regarding the within-subject errors.

Example 18.2. (*continued*) We use the random intercepts and random slopes model for the CD4 data to illustrate residual analysis for linear mixed models. The population residuals are plotted in Figure 18.8, and the within-subject residuals are plotted in Figure 18.9. First, no violation of the linearity assumption for month is apparent in either of these plots. Second, the population residuals are weakly suggestive of an increasing variance over time. However, it is important to note that under the assumption of random intercepts and random slopes, the total variance, $\text{var}(b_{i,0} + b_{i,1} \cdot \text{month} + \epsilon_{ij})$, may be an increasing or decreasing function of time. The population residuals suggest right skewness in the cross-sectional distribution of CD4. Since the within-subject residuals do not appear skewed, the population residuals suggest that the random effects may not be normally distributed. Figure 18.10 presents histograms of the estimated intercepts and slopes obtained using ordinary linear regression for subject i data rather than the empirical Bayes estimates. The histograms for the individual intercepts appear to be right skewed, while the individual slopes appear symmetrically distributed. Therefore, residual analysis coupled with exploratory analysis of individual regression estimates suggests that linearity assumptions appear satisfied, but normality of random effects may be violated. The linear mixed model is known to be moderately robust to distributional assumptions, so large-sample inference regarding the average rate of decline for baseline viral load groups can be achieved.

Mixed models can be adopted for use with categorical and count response data. For example, random effects can be included in logistic regression models for binary outcomes and can be included in log-linear models for count data. Maximum likelihood estimation for these models requires specialized software. Extensions of mixed models to alternate regression contexts is discussed in Chapters 7 and 9 of Diggle et al. [2002].

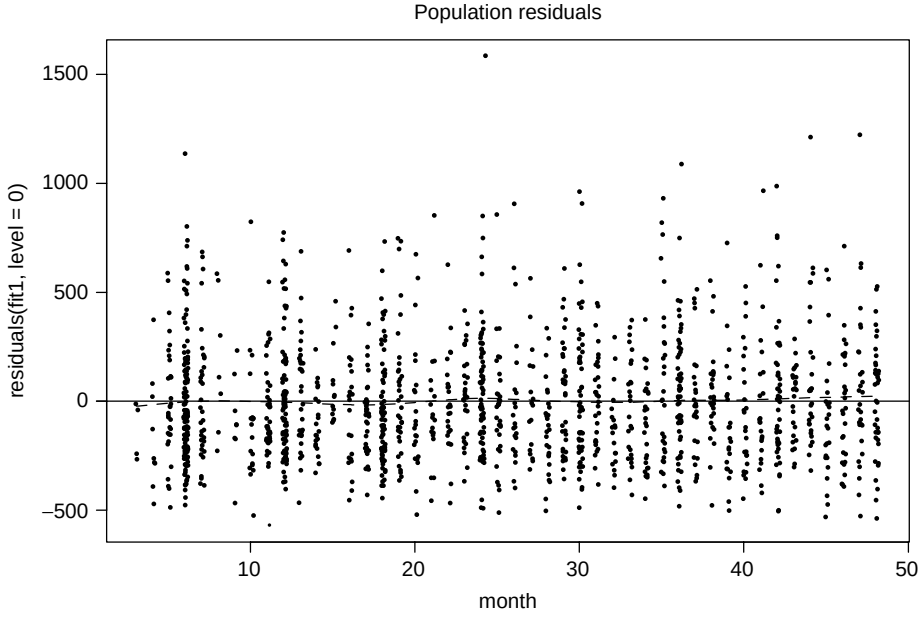


Figure 18.8 Population residuals, R_{ij}^P, vs. visit month for the MACS CD4 data. The dashed line is a smooth curve through the residuals.

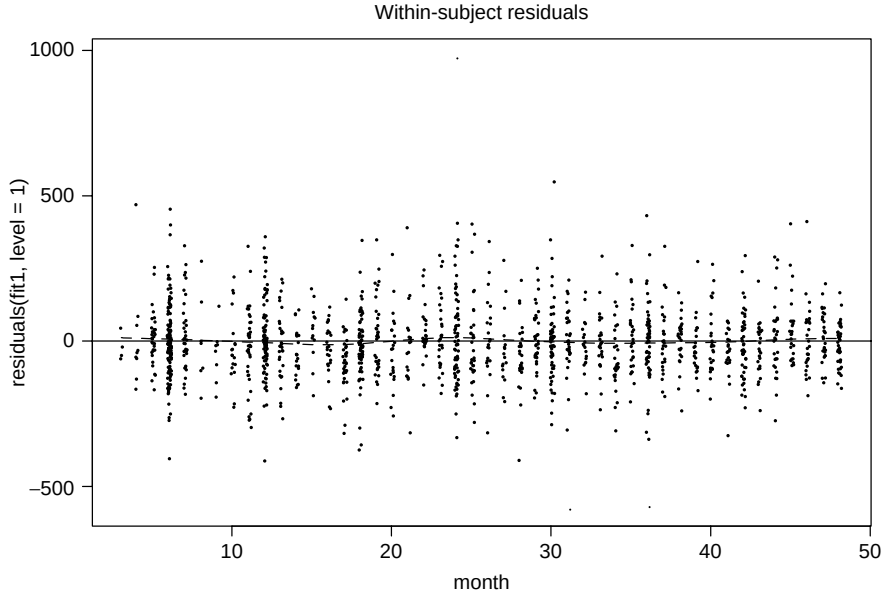


Figure 18.9 Within-subject residuals, R_{ij}^{W}, vs. visit month for the MACS CD4 data. The dashed line is a smooth curve through the residuals.

18.5.1.1 Summary

- Linear mixed models permit regression analysis with correlated data.
- Mixed models specify variance components that represent within-subject variance in outcomes and between-subject variation in trajectories.
- Linear mixed model parameters can be estimated using maximum likelihood.

18.5.2 Generalized Estimating Equations

A second regression approach for inference with longitudinal data is known as *generalized estimating equations* (GEE) [Liang and Zeger, 1986]. In this approach two models are specified. First, a regression model for the mean response is selected. The form of the regression model is completely flexible and can be a linear model, a logistic regression model, a log-linear model, or any generalized linear model [McCullagh and Nelder, 1989]. Second, a model for the within-subject correlation is specified. The correlation model serves two purposes: It is used to obtain weights (covariance inverse) that are applied to the vectors of observations from each subject to obtain regression coefficient estimates; and the correlation model is used to provide model-based standard errors for the estimated coefficients.

A regression model specifies a structure for the mean response, $\mu_{ij} = E(Y_{ij} \mid X_{ij})$, as a function of covariates. For longitudinal data the mean μ_{ij} has been called the *marginal mean* since it does not involve any additional variables, such as random effects, b_i, or past outcomes, Y_{ij-1}. Mixed models consider means conditional on random effects, and transition models include past outcomes as covariates. Adding additional variables leads to subtle changes in the interpretation of covariate coefficients, which becomes particularly important for nonlinear models such as logistic regression. See Diggle et al. [2002, Chaps. 7 and 11] for further discussion.

GEE has two important robustness properties. First, the estimated regression coefficients, $\widehat{\beta}$, obtained using GEE are broadly valid estimates that approach the correct value with increasing

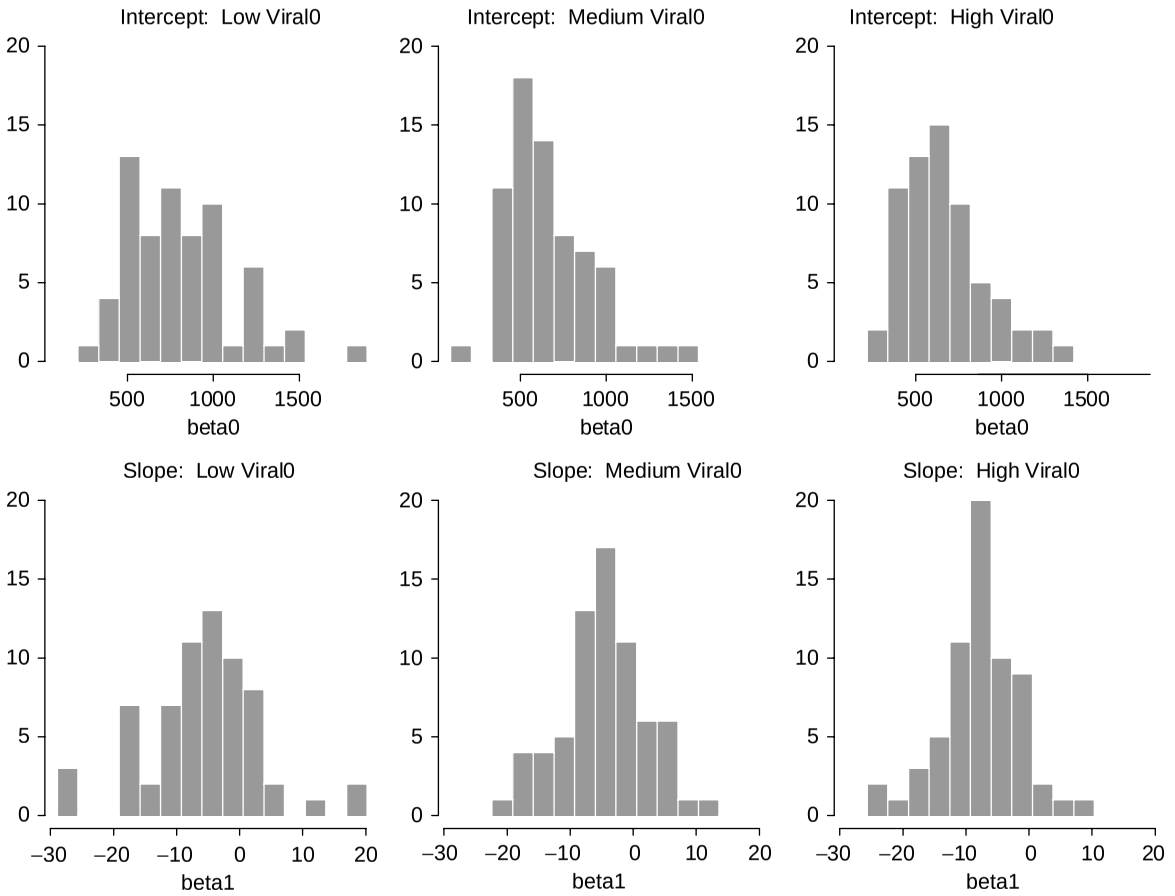


Figure 18.10 Estimates of individual intercepts and slopes by baseline viral load category for the MACS CD4 data.

sample size regardless of the choice of correlation model. In this respect the correlation model is used simply to weight observations, and a good correlation model choice can lead to more precise estimation of regression coefficients than can a poor choice. Based on optimal estimation theory (e.g., Gauss–Markov theory), the best correlation model choice for efficiency of estimation is the true correlation structure. Second, the correlation choice is used to obtain model-based standard errors, and these do require that the correlation model choice is correct in order to use the standard errors for inference. A standard feature of GEE is the additional reporting of *empirical standard errors*, which provide valid estimates of the uncertainty in $\widehat{\beta}$, even if the correlation model is not correct. Therefore, the correlation model can be any model, including one that assumes observations are independent, and proper large-sample standard errors obtained using the empirical estimator. Liang and Zeger [1993] provide an overview of regression methods for correlated data, and Hanley et al. [2003] give an introduction to GEE for an epidemiological audience.

Example 18.2. (*continued*) We return to the CD4 data and use GEE to investigate whether the rate of decline in CD4 over the first 48 months postseroconversion seems to depend on the baseline viral load category. Table 18.7 presents the estimates obtained using GEE and an independence correlation model. Standard errors using the independence correlation model are identical to those obtained from linear regression and are labeled as "model-based." In this application the key feature provided by GEE are the "empirical" standard errors, which are generally valid estimates of the uncertainty associated with the regression estimates. Notice that most of the empirical standard errors are larger than the naive model-based standard errors, which assume that the data are independent. However, corrected standard errors can be either larger or smaller than standard errors obtained under an independence assumption, and the nature of the covariate and the correlation structure interact to determine the proper standard error. It is an oversimplification to state that correction for correlation will lead to larger standard errors. Using GEE we obtain conclusions similar to that obtained using linear mixed models: The high-viral-load group has a steeper estimated rate of decline, but the difference between low and high groups is not statistically significant.

Example 18.1. (*continued*) GEE is particularly useful for binary data and count data. We now turn to analysis of the nurse item from the HIVNET informed consent study. We need to choose a regression model and a correlation model. For our first analysis we assume a common proportion answering correctly after randomization. For this analysis we create the covariate "post," which takes the value 1 if the visit occurs at month 6, 12, or 18, and takes the value 0 for the baseline visit. We use the variable "ICgroup" to denote the intervention and control group, where $\text{ICgroup}_{ij} = 1$ for all visits $j = 1, 2, 3, 4$ if the subject was randomized to the mock informed consent, and $\text{ICgroup}_{ij} = 0$ for all visits, $j = 1, 2, 3, 4$, if the subject was randomized to the control group. Since the response is binary, $Y_{ij} = 1$ if the item was correctly answered by subject i at visit j and 0 otherwise, we use logistic regression to characterize the

Table 18.7 GEE Estimates for the CD4 Data Using an Independence Working Correlation Model

		Standard Error		Z-statistic	
	Estimate	Model	Empirical	Model	Empirical
(Intercept)	792.897	26.847	36.651	29.534	21.633
Month	−4.753	0.950	1.101	−5.001	−4.318
I(Medium viral load)	−121.190	37.872	46.886	−3.200	−2.585
I(high viral load)	−150.705	37.996	45.389	−3.966	−3.320
Month · I(medium viral load)	−0.301	1.341	1.386	−0.224	−0.217
Month · I(high viral load)	−1.898	1.346	1.297	−1.410	−1.464

probability of a correct response as a function of time and treatment group:

$$\begin{aligned}\text{logit}P(Y_{ij} = 1 \mid X_i) = \beta_0 &+ \\ &\beta_1 \cdot \text{post}_{ij} + \\ &\beta_2 \cdot \text{ICgroup}_{ij} + \\ &\beta_3 \cdot \text{ICgroup}_{ij} \cdot \text{post}_{ij}\end{aligned}$$

Since the visits are equally spaced and each subject is scheduled to have a total of four measurements, we choose to use an unstructured correlation matrix. This allows the correlations ρ_{jk} to be different for each pair of visit times (j, k).

In Table 18.8 we provide GEE estimates obtained using the SAS procedure GENMOD. The estimated working correlation is printed and indicates correlation that decreases as the time separation between visits increases. For example, the estimated correlation for Y_{i1} and Y_{i2} is $\widehat{\rho}_{12} = 0.204$, while for Y_{i1} and Y_{i3}, $\widehat{\rho}_{13} = 0.194$, and for Y_{i1} and Y_{i4}, $\widehat{\rho}_{14} = 0.163$. The correlation between sequential observations also appears to increase over time with $\widehat{\rho}_{23} = 0.302$ and $\rho_{34} = 0.351$.

Regression parameter estimates are reported along with the empirical standard error estimates. These parameters are interpreted as follows:

- *(Intercept)* $\widehat{\beta}_0 = 0.1676$. The intercept is an estimate of log odds of a correct response to the nurse item at baseline for the control group. This implies an estimate for the probability of

Table 18.8 GEE Analysis of the Nurse Item from the HIVNET Informed Consent Study[a]

GEE Model Information

Correlation Structure	Unstructured
Subject Effect	id (1123 levels)
Number of Clusters	1123
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	1

Working Correlation Matrix

	Col1	Col2	Col3	Col4
Row1	1.0000	0.2044	0.1936	0.1625
Row2	0.2044	1.0000	0.3022	0.2755
Row3	0.1936	0.3022	1.0000	0.3511
Row4	0.1625	0.2755	0.3511	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.1676	0.0652	0.0398	0.2954	2.57	0.0102
Post	−0.3238	0.0704	−0.4618	−0.1857	−4.60	<.0001
ICgroup	−0.1599	0.1643	−0.4819	0.1622	−0.97	0.3306
ICgroup*Post	1.0073	0.2012	0.6128	1.4017	5.01	<.0001

[a]Output from SAS procedure GENMOD.

a correct response at baseline among controls of $\exp(0.1676)/[1+\exp(0.1676)] = 0.5418$, which agrees closely with the observed proportion presented in Table 18.3.

- *Post* $\widehat{\beta}_1 = -0.3238$. The coefficient of Post is an estimate of the log of the odds ratio comparing the odds of a correct response among control subjects after randomization (either month 6, 12, or 18) relative to the odds of a correct response among the control group at baseline. Since the odds ratio estimate is $\exp(-0.3238) = 0.7234 < 1$, the odds of a correct response is lower after baseline. A test for equality of odds comparing postbaseline to baseline yields a p-value $p < 0.001$.
- *ICgroup* $\widehat{\beta}_2 = -0.1599$. The coefficient of ICgroup is an estimate of the log of the odds ratio comparing the odds of a correct response among intervention subjects at baseline relative to the odds of a correct response among the control subjects at baseline. Since the assignment to treatment and control was based on randomization, we expect this odds ratio to be 1.0, and the log odds ratio estimate is not significantly different from 0.0.
- *ICgroup* * *Post* $\widehat{\beta}_3 = 1.0073$. This interaction coefficient measures the difference between the comparison of treatment and control after randomization and the comparison of treatment and control at baseline. Specifically, $(\beta_3+\beta_2)$ represents the log odds ratio comparing the odds of a correct response among intervention subjects postbaseline to the odds of a correct response among control subjects postbaseline. Since β_2 represents the group comparison at baseline, $\beta_3 = (\beta_3 + \beta_2) - \beta_2$, or β_3 measures the difference between the comparison after baseline and the group comparison at baseline. Therefore, the parameter β_3 becomes the primary parameter of interest in this study, as it assesses the change in the treatment/control comparison that is attributable to the intervention. A test of $\beta_3 = 0$ is statistically significant with $p < 0.001$.

GEE is a convenient analysis tool for the informed consent data, as it allows inference regarding the differences between treatment and control groups over time. A standard logistic regression model is adopted and valid standard errors are calculated that account for the within-subject correlation of outcomes.

In Table 18.8 we used a single time variable that was an indicator for the postbaseline visits at six, 12, and 18 months. However, inspection of crude proportions responding correctly suggest that the treatment/control comparison may be decreasing over time. For example, in Table 18.3 we see (treatment, control) proportions of (72.1%, 44.7%) at month 6, (60.1%, 46.3%) and (66.0%, 48.2%) at months 12 and 18. To assess whether the treatment effect appears to be decreasing over time, we fit a second logistic regression model that uses indicator variables for months 6, 12, and 18. Table 18.9 presents GEE estimates using an exchangeable working correlation model. In this model the coefficient of month6*ICgroup contrasts the treatment/control log odds ratio at the six-month visit and at baseline. Similar to our earlier analysis, this difference in time-specific log odds ratios is the primary treatment effect observed at six months. Similarly, the coefficients of month12*ICgroup and month18*ICgroup represent treatment effects at 12 and 18 months. Each of the estimated differences in log odds ratios are significant as indicated by the individual p-values in Table 18.9. In addition, we contrast the observed treatment effect at six months with the treatment effect observed at 12 and 18 months. The difference between the estimated coefficient of month6*ICgroup and month12*ICgroup assesses the change in the treatment effect and is estimated as $1.3232 - 0.7362 = -0.5871$. A test of this contrast yields a p-value of 0.0035, indicating a different treatment effect at 12 months as compared to the treatment effect at 6 months. A similar analysis for the 18-month effect as compared to 6 months is barely statistically significant with $p = 0.041$. Therefore, there is evidence that the effect of the intervention may be changing over time. Once again GEE provides a general tool for evaluating the evolution of mean outcomes over time for different subgroups of subjects.

There are a number of extensions of the GEE approach introduced by Liang and Zeger [1986]. More flexible and tailored dependence models have been proposed for binary data [Lipsitz et al.,

Table 18.9 GEE Analysis of the Nurse Item from the HIVNET Informed Consent Study[a]

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.1644	0.0653	0.0364	0.2923	2.52	0.0118
month6	−0.3803	0.0839	−0.5448	−0.2158	−4.53	<.0001
month12	−0.3261	0.0854	−0.4934	−0.1587	−3.82	0.0001
month18	−0.2460	0.0886	−0.4197	−0.0723	−2.78	0.0055
ICgroup	−0.1536	0.1639	−0.4748	0.1676	−0.94	0.3487
month6*ICgroup	1.3232	0.2319	0.8687	1.7777	5.71	<.0001
month12*ICgroup	0.7362	0.2358	0.2739	1.1984	3.12	0.0018
month18*ICgroup	0.9101	0.2273	0.4647	1.3556	4.00	<.0001

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square
Effect at 12 versus 6	−0.5871	0.2014	0.05	−0.9817	−0.1924	8.50
Effect at 18 versus 6	−0.4131	0.2023	0.05	−0.8097	−0.0166	4.17

Contrast Estimate Results

Label	Pr > ChiSq
Effect at 12 versus 6	0.0035
Effect at 18 versus 6	0.0412

[a]Output from SAS procedure GENMOD.

1991; Carey et al., 1993], and extension for multiple survival times has been developed [Wei et al., 1989; Lee et al., 1992].

Summary

- GEE permits regression analysis with correlated continuous, binary, or count data.
- GEE requires specification of a regression model and a working correlation model.
- Two standard error estimates are provided with GEE: a model-based standard error that is valid if the correlation model is specified correctly; and empirical standard errors that are valid even if the correlation model is not correct provided that the data contain a large number of independent clusters.
- Estimation with GEE does not involve a likelihood function; rather, it is based on the solution to regression equations that use models only for the mean and covariance.

18.6 MISSING DATA

One of the major issues associated with the analysis of longitudinal data is missing data, or more specifically, *monotone missing data*, which arise when subjects drop out of the study. It is assumed that once a participant drops out, he or she provides no further outcome information. Missing data can lead to biased estimates of means and/or regression parameters when the probability of missingness is associated with outcomes. In this section we first review a standard taxonomy of missing data mechanisms and then briefly discuss methods that can be used to alleviate bias due to attrition. We also discuss some simple exploratory methods that can help determine whether subjects who complete the longitudinal study appear to differ from those who drop out.

18.6.1 Classification of Missing Data Mechanisms

To discuss factors that are associated with missing data, it is useful to adopt the notation $R_{ij} = 1$ if observation Y_{ij} is observed, and $R_{ij} = 0$ if Y_{ij} is missing. Let $R_i = (R_{i1}, R_{i2}, \ldots, R_{in})$. Monotone missing data imply that if $R_{ij} = 0$, then $R_{ij+k} = 0$ for all $k > 0$. Let Y_i^O denote the subset of the outcomes $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ that are observed, and let Y_i^M denote the missing outcomes. For longitudinal data a missing data classification is based on whether observed or unobserved outcomes are predictive of missing data [Laird, 1988]:

$$\text{Missing completely at random (MCAR)}: \quad P(R_i \mid Y_i^O, Y_i^M, X_i) = P(R_i \mid X_i)$$

$$\text{Missing at random (MAR)}: \quad P(R_i \mid Y_i^O, Y_i^M, X_i) = P(R_i \mid Y_i^O, X_i)$$

$$\text{Nonignorable (NI)}: \quad P(R_i \mid Y_i^O, Y_i^M, X_i) \text{ depends on } Y_i^M$$

In Figure 18.7 an example of monotone missing data is presented. For subject 1, all observations after the 7-month visit are missing. If the reason that these observations are missing is purely unrelated to outcomes (observed or not), the missing data are called *MCAR*. However, if the observed data are predictive of missingness, the missing data are called *MAR*, and the mechanism introduces a form of selection bias. MAR data could occur if an attending physician decides to disenroll any participant who appears to be failing treatment, particularly when the decision is based on the value of past measurements or factors associated with the past outcomes, Y_{ij}. Finally, the unobserved outcomes may be associated with missingness if, for example, subjects who are the most ill refuse to travel to attend their scheduled study visit.

The missing data taxonomy translates directly into implications for potential selection bias. If data are MCAR, both the missing and the observed outcomes are representative of the source population. Therefore, when data are MCAR, standard statistical summaries based on the observed data remain valid. However, if data are MAR or NI, summaries based on the available cases may be biased. Returning to Figure 18.7, if the dropout for patient 1 is indicative of a general process by which those subjects who have a high response value do not return for study, the observed mean for the measured outcomes will not be representative of what would be observed had the entire population been followed. In this example, the mean among available subjects would underestimate the population mean for later months.

Formally, we write $E(Y_{ij} \mid X_i, R_{ij} = 1)$ to denote the expected response conditional on responding, and we write $E(Y_{ij} \mid X_i)$ for the target of inference. If the data are MCAR, then $E(Y_{ij} \mid X_i, R_{ij} = 1) = E(Y_{ij} \mid X_i)$. However, if data are either MAR or NI, then $E(Y_{ij} \mid X_i, R_{ij} = 1) \neq E(Y_{ij} \mid X_i)$, implying that the available data, $R_{ij} = 1$, may not provide valid estimates of population parameters.

In any given application, serious thought needs to be given to the types of processes that lead to missing data. External information can help determine whether missingness mechanisms may be classified as MCAR, MAR, or NI. Unfortunately, since NI missingness implies that unobserved data, Y_i^M, predicts dropout, we cannot empirically test whether data are NI vs. MAR or MCAR. Essentially, one would need the unobserved data to check to see if they are associated with missingness, but these data are missing! The observed data can be used to assess whether the missingness appears to be MAR or MCAR. First, the dropout time can be considered a discrete-time "survival" outcome, and methods introduced in Chapter 16 can be used to assess whether past outcomes $Y_{ij-1}, Y_{ij-2}, \ldots$ are predictive of dropout, $R_{ij} = 0$. Second, each subject will have a dropout time, or equivalently, a "last measurement" time, with those completing the study having the final assessment time as their time of last measurement. The longitudinal data can be stratified according to the dropout time. For example, the mean at baseline can be calculated separately for those subjects that dropout at the first visit, second visit, through those that complete the study. Similarly, the mean response at the first follow-up visit can be computed for all subjects who have data for that visit. Such analyses can be used to determine whether the outcomes for the dropout subjects appear to be different from those

of the "completers." Naturally, subjects who are lost can only be compared to others at the visit times prior to their dropout. These exploratory analyses are complementary: The first approach assesses whether outcomes predict dropout, and the second approach evaluates whether the dropout time predicts the outcomes. An example of such modeling can be found in Zhou and Castelluccio [2004].

18.6.2 Approaches to Analysis with Missing Data

There are several statistical approaches that attempt to alleviate bias due to missing data. General methods include:

1. *Data imputation*. See Little and Rubin [1987], Schafer [1997], or Koepsell and Weiss [2003] for more information on imputation methods. Imputation refers to "filling in" missing data. Proper methods of imputation use multiple imputation to account for uncertainty in the missing data. Imputation methods require that a model be adopted that links the missing data to the observed data.

2. *Data modeling*. In this method the missing data process and the longitudinal data are both modeled using maximum likelihood for estimation. Use of a linear mixed model estimated with maximum likelihood is one example of this approach. However, to correct validly for MAR missingness, the mean and the covariance must be specified correctly. See Verbeke and Molenberghs [2000] for more details.

3. *Data weighting*. Nonresponse methods with available data are used to weight the observed data to account for the missing data. Use of inverse probability weighting or nonresponse weighting can be applied to general statistical summaries and has been proposed to allow for use of GEE in MAR situations. See Robins et al. [1995] for the statistical theory and Preisser et al. [2002] for a simulation study of the performance of weighted GEE methods.

However, it is important to note that these methods are designed to address data that are assumed to be MAR rather than the more serious nonignorable (NI) missing data. Nonignorable missing data can lead to bias, which cannot be corrected simply through modeling and estimation of the dropout model and/or the response model since unidentifiable parameters that link the probability of missingness to the unobserved data are needed. Therefore, reliance on statistical methods to correct for bias due to attrition either requires an untestable assumption that the data are MAR or requires some form of sensitivity analysis to characterize plausible estimates based on various missingness assumptions. See Diggle et al. [2002, Chap. 13] for discussion and illustration.

Example 18.1. (*continued*) In the HIVNET Informed Consent Study, there was substantial missing data due to attrition. In Tables 18.2 and 18.3 we see a decreasing number of subjects over time. In the control group there are 946 subjects with baseline data and only 782 with 18-month data. Is the knowledge score for subjects who complete the study different from the score for those who dropout? Figure 18.11 shows the mean response over time stratified by dropout time. For example, among subjects that dropout at the 12-month visit, their mean knowledge score at baseline and 6 months is plotted. This plot suggests that subjects who complete only the baseline interview have a lower mean baseline knowledge score than that of all other subjects. In addition, for subjects who complete the study, the average knowledge score at six and 12 months appears greater than the mean knowledge score among subjects who do not complete the 18-month visit. Thus, Figure 18.11 suggests that the completers and the dropout subjects differ with respect to their knowledge scores. Any analysis that does not account for differential dropout is susceptible to selection bias.

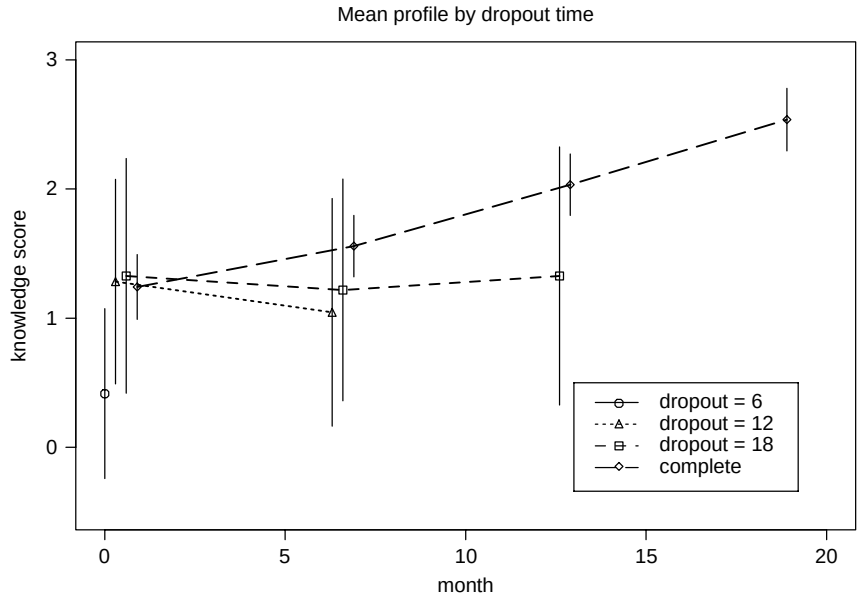


Figure 18.11 Patterns of mean knowledge score by dropout time for the control group. HIVNET informed consent substudy.

18.7 SUMMARY

Longitudinal data provide unique opportunities for inference regarding the effect of an intervention or an exposure. Changes in exposure conditions can be correlated with changes in outcome conditions. However, analysis of longitudinal data requires methods that account for the within-subject correlation of repeated measures. Texts by Diggle et al. [2002], Verbeke and Molenberghs [2000], Brown and Prescott [1999], and Crowder and Hand [1990] provide comprehensive discussions of statistical methods for the analysis of longitudinal data. There are a number of additional issues that warrant attention but are beyond the scope of this book.

NOTES

18.1 Nonlinear Mixed Models

We have introduced linear mixed models and GEE. However, mixed models have also been extended to logistic regression and other nonlinear model settings. See Diggle et al. [2002, Chap. 8 and 11] for illustrations.

18.2 Models for Survival and Repeated Measurements

In many longitudinal studies information on both repeated measurements and on ultimate time until death or key clinical endpoint is collected. Methods have been developed to analyze such data jointly. See Hogan and Laird [1997a, b] for an overview of approaches for the joint analysis of survival and repeated measures.

18.3 Models for Time-Dependent Covariates

In designed experiments the exposures X_{ij} may be controlled by the investigator. However, in many observational studies, exposures or treatments that are selected over time may be related to

past health outcomes. For example, subjects with low values of CD4 may be more likely to be exposed to a therapeutic agent. Analysis of such serial data to assess the effect of the intervention is complicated by the feedback between outcome and exposure. Robins [1986] and Robins et al. [1999] have identified proper causal targets of inference and methods for estimation in settings where time-varying covariates are both causes and effects. See Diggle et al. [2002, Chap. 12].

PROBLEMS

18.1 This exercise considers the interplay between the covariate distribution and the correlation. For each of the following scenarios, assume that there are a total of N pairs of observations, (Y_{i1}, Y_{i2}), with covariates (X_{i1}, X_{i2}). Assume that the covariate is binary: $X_{ij} = 0$ or $X_{ij} = 1$, denoting control and treatment exposures. Let $\overline{Y}_1$ denote the mean of all observations where $X_{ij} = 1$, and let $\overline{Y}_0$ denote the mean of all observations where $X_{ij} = 0$. Assume a constant variance $\sigma^2 = \text{var}(Y_{ij} \mid X_{ij})$ and a correlation $\rho = \text{corr}(Y_{i1}, Y_{i2})$.

- **(a)** Assume that half of the subjects are assigned to control for both visits, $(X_{i1}, X_{i2}) = (0, 0)$, and half of the subjects are assigned to intervention for both visits, $(X_{i1}, X_{i2}) = (1, 1)$. What is the variance of the estimated mean difference, $\widehat{\Delta} = (\overline{Y}_1 - \overline{Y}_0)$?
- **(b)** Assume that subjects change their treatment over time with half of the subjects are assigned to control and then treatment, $(X_{i1}, X_{i2}) = (0, 1)$, and half of the subjects assigned to treatment and then control, $(X_{i1}, X_{i2}) = (1, 0)$. This design is referred to as a *crossover study*. What is the variance of the estimated mean difference $\widehat{\Delta} = (\overline{Y}_1 - \overline{Y}_0)$?
- **(c)** Comment on the advantages and disadvantages of these two study designs.

18.2 Consider a study with a single prerandomization measurement, Y_{i0}, and a single postrandomization measurement, Y_{i1}. For any constant a we can define the average contrast, $\overline{D}(a) = \text{mean}[d_i(a)]$, where $d_i(a) = Y_{i1} - aY_{i0}$. Let $\overline{D}_0(a)$ denote the mean for the control group, and let $\overline{D}_1(a)$ denote the mean for the intervention group. Assume that $\sigma^2 = \text{var}(Y_{ij})$ for $j = 0, 1$, and let $\rho = \text{corr}(Y_{i0}, Y_{i1})$. We assume that the subjects are randomized to treatment and control after randomization at baseline. Therefore, the following table illustrates the mean response as a function of treatment and time:

	Control	**Intervention**
Baseline	μ_0	μ_0
Follow-up	μ_1	$\mu_1 + \Delta$

- **(a)** Show that the expected value of $\widehat{\Delta}(a) = \overline{D}_1(a) - \overline{D}_0(a)$ equals Δ for any choice of a.
- **(b)** When $a = 0$, we effectively do not use the baseline value, and $\widehat{\Delta}(0)$ is the difference of means at follow-up. What is the variance of $\widehat{\Delta}(0)$?
- **(c)** When $a = 1$, we effectively analyze the change in outcomes since $d_i(1) = Y_{i1} - Y_{i0}$. What is the variance of $\widehat{\Delta}(1)$?
- **(d)** What value of a leads to the smallest variance for $\widehat{\Delta}(a)$?

18.3 Use the data from the Web page to perform GEE analysis of the HIVNET Informed Consent Substudy "safety" item.

18.4 For the random intercepts and slopes model given in Table 18.6, the proportion of total variation that is attributable to within-subject variation is not constant over time. Compute estimates of the proportion of total variation at 0, 12, 24, and 36 months that is attributable to within-subject variation, ϵ_{ij}, as opposed to between subject variation, $b_{i,0} + b_{i,1} \cdot \text{month}$.

18.5 For the HIVNET Informed Consent Substudy data, create pairs of plots:

(a) Plot month 12 vs. month 6 knowledge score. Add a pair of lines that show the ordinary least squares estimate for the intervention and the control group.

(b) Plot month 18 vs. month 12 knowledge score. Add a pair of lines that shows the ordinary least squares estimate for the intervention and the control group.

(c) Do these plots suggest that there are additional differences between the intervention and control groups that is not captured by the difference that manifests at the six-month visit?

18.6 For the NURSE and SAFETY items from the HIVNET Informed Consent Substudy, evaluate the transition from incorrect to correct, and from correct to correct again, for the times (six-month $\rightarrow$ 12-month visit) and (12-month $\rightarrow$ 18-month visit). Is there evidence that the intervention and control groups differ in terms of the "correction" and "maintenance" of knowledge at the later times?

REFERENCES

Brown, H., and Prescott, R. [1999]. *Applied Mixed Models in Medicine*. Wiley, New York.

Carlin, B. P., and Louis, T. A. [1996]. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Coletti, A. S., Heagerty, P. J., Sheon, A. R., Gross, M., Koblin, B. A., Metzger, D. S., and Seage G. R. [2003]. Randomized, controlled evaluation of a prototype informed consent process for HIV vaccine efficacy trials. *Journal of Acquired Immune Deficiency Syndrome*, **32**: 161–169.

Carey, V., Zeger, S. L., and Diggle, P. [1993]. Modelling multivariate binary date with alternating logistic regressions. *Biometrika*, **80**: 517–526.

Crowder, M. J., and Hand, D. J. [1990]. *Analysis of Repeated Measures*. Chapman & Hall, New York.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. [2002]. *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Donner, A., and Klar, N. [1994]. Cluster randomization trials in epidemiology: theory and application. *Journal of Statistical Planning and Inference*, **42**: 37–56.

Donner, A., and Klar, N. [1997]. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, **49**: 435–439.

Frison, L. J., and Pocock, S. J. [1992]. Repeated measures in clinical trials: analysis using summary statistics and its implication for design. *Statistics in Medicine*, **11**: 1685–1704.

Frison, L. J., and Pocock, S. J. [1997]. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Statistics in Medicine*, **16**: 2855–2872.

Hanley, J. A., Negassa, A., deB. Edwardes, M. D., and Forrester, J. E. [2003]. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, **157**: 364–375.

Hogan, J. W., and Laird, N. M. [1997a]. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**: 239–257.

Hogan, J. W., and Laird, N. M. [1997b]. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**: 259–272.

Kaslow, R. A., Ostrow, D. G., Detels, R., et al. [1987]. The Multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*, **126**: 310–318.

Koepsell, T. D., and Weiss, N. S. [2003]. *Epidemiological Methods: Studying the Occurrence of Illness*. Oxford University Press, New York.

Koepsell, T. D., Martin, D. C., Diehr, P. H., Psaty, B. M., Wagner, E. H., Perrin, E. B., and Cheadle, A. [1991]. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed model analysis of variance approach. *American Journal of Epidemiology*, **44**: 701–713.

Laird, N. M. [1988]. Missing data in longitudinal studies. *Statistics in Medicine*, **7**: 305–315.

Laird, N. M., and Ware, J. H. [1982]. Random-effects models for longitudinal data. *Biometrics*, **38**: 963–974.

Lebowitz, M. D. [1996]. Age, period, and cohort effects. *American Journal of Respiratory Critical Care Medicine*, **154**: S273–S277.

Lee, E. W., Wei, L. J., and Amato, D. A. [1992]. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J. P. Klein and P. K. Joel (eds.). Kluwer Academic Publishers, Dordrecht.

Liang, K.-Y., and Zeger, S. L. [1986]. Longitudinal data analysis using generalised linear models. *Biometrika*, **73**: 13–22.

Liang, K.-Y., and Zeger, S. L. [1993]. Regression analysis for correlated data. *Annual Review of Public Health*, **14**: 43–68.

Lipsitz, S., Laird, N., and Harrington, D. [1991]. Generalized estimating equations for correlated binary data: using odds ratios as a measure of association. *Biometrika*, **78**: 153–160.

Little, R. J. A., and Rubin, D. B. [2002]. *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

McCullagh, P., and Nelder, J. A. [1989]. *Generalized Linear Models, 2nd ed.* Chapman & Hall, New York.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. [2002]. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, **21**: 3035–3054.

Robins, J. M. [1986]. A new approach to causal inference in mortality studies with sustained exposure periods: application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**: 1393–1512.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. [1995]. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**: 106–121.

Robins, J. M., Greenland, S., and Hu, F.-C. [1999]. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, **94**: 687–712.

Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., and Zeger, S. L. [2000]. Fine particulate air pollution and mortality in 20 US cities. *New England Journal of Medicine*, **343**(24): 1798–1799.

Schafer, J. L. [1997]. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York.

Stram, D. O., and Lee, J. W. [1994]. Variance component testing in the longitudinal mixed model. *Biometrics*, **50**: 1171–1177.

The Childhood Asthma Management Program Research Group [2002]. Long-term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine*, **343**(15): 1054–1063.

Verbeke, G., and Molenberghs, G. [2000]. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

Wei, L. J., Lin, D., and Weissfeld, L. [1989]. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**: 1065–1073.

Weiss, S. T., and Ware, J. H. [1996]. Overview of issues in the longitudinal analysis of respiratory data. *American Journal of Respiratory Critical Care Medicine*, **154**: S208–S211.

Yu, O., Sheppard, L., Lumley, T., Koenig, J., and Shapiro, G. [2000]. Effects of ambient air pollution on symptoms of asthma in Seattle-area children enrolled in the CAMP study. *Environmental Health Perspectives*, **108**: 1209–1214.

Zhou, X.-H., and Castelluccio, P. [2004]. Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine*. **23**: 221–230.

CHAPTER 19

Randomized Clinical Trials

19.1 INTRODUCTION

If Alexander Pope is correct that "the proper study of mankind is man" [Pope, 1733], then the development of new therapeutic and prophylactic measures for humans is one of the more proper uses of biostatistics. In addition, it is one of the most active and highly used areas of biostatistics. In this chapter we consider primarily randomized clinical trials in humans, although we mention other uses of the techniques. The use of *clinical* refers to the evaluation of clinical measures, for example, drug treatments or surgical treatments. If an experiment is randomized—that is, treatment assignments given by some random process—it necessarily implies more than one treatment is being considered or tested. Thus, the trials are comparative. And, of course, the term *trial* means that we test, or try, the treatments considered. The acronym RCT has been used for both a randomized *controlled* trial and a randomized *clinical* trial. Randomized clinical trials are examples of randomized controlled trials, but not necessarily vice versa, as we shall see below. Here we use the abbreviation RCT for both. For the most part we shall be discussing clinical trials, although it will be clear from the context which is referred to.

In addition to the statistical methods we have discussed before there are a number of practical issues in clinical trials that are now accepted as appropriate for the best scientific inference. The issues of trial design to some extent "fall between the cracks" in clinical research. They are not an obvious part of a medical education—not being biological per se—and also not an obvious portion of biostatistics, as they do not explicitly involve the mathematics of probability and statistics. However, the issues are important to successful implementation of good scientific clinical studies (and other studies as well) and are a necessary and appropriate part of biostatistical training. Some of these issues are discussed in less detail in Chapter 2 and in Chapter 8, in which we discuss permutation and randomization tests in Section *8.9. Here we give background on why the design features are needed as well as some discussion of how to implement the design features.

The use of RCTs and new drug development is big business. At the end of 2001, the cost for evaluating an approved new chemical entity was estimated at approximately $800 million [*Wall Street Journal*, 2001], and the time for development is often 10 years or more.

19.2 ETHICS OF EXPERIMENTATION IN HUMANS

The idea of experimenting on humans and other animals is distasteful at first blush. This is especially so in light of the Nazi experiments during the World War II period (see, e.g., Lifton [1986]).

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

Yet it is clear that if new and improved therapies and treatments are to be developed, they must be tried initially at some point in time on humans and/or animals. Whether designated so or not, such use does constitute experimentation. This being the case, it seems best to acknowledge this fact and to try to make such experiments as appropriate, justified, and useful as possible. Considerable work has been devoted to this end. The ethics of experimentation on humans has been the subject of intense study in recent decades. Ethics was touched on in Section 2.5, and because of its importance, we return to the subject here. A good introduction is Beauchamp and Childress [2001]. They review four principles for biomedical ethics: respect for autonomy, nonmaleficence, beneficence, and justice. Briefly summarized:

- The *principle of autonomy* recognizes a person's right to "hold views, make choices, and take actions based on personal values and beliefs."
- The *principle of nonmaleficence* is not to inflict harm to others.
- The *principle of beneficence* "asserts an obligation to help others further their important and legitimate interests."
- The *principle of justice* is more difficult to characterize briefly and may mean different things to different people. As Beauchamp and Childress note: "The only principle common to all theories of justice is a minimal principle traditionally attributed to Aristotle: Equals must be treated equally, and un-equals must be treated unequally."

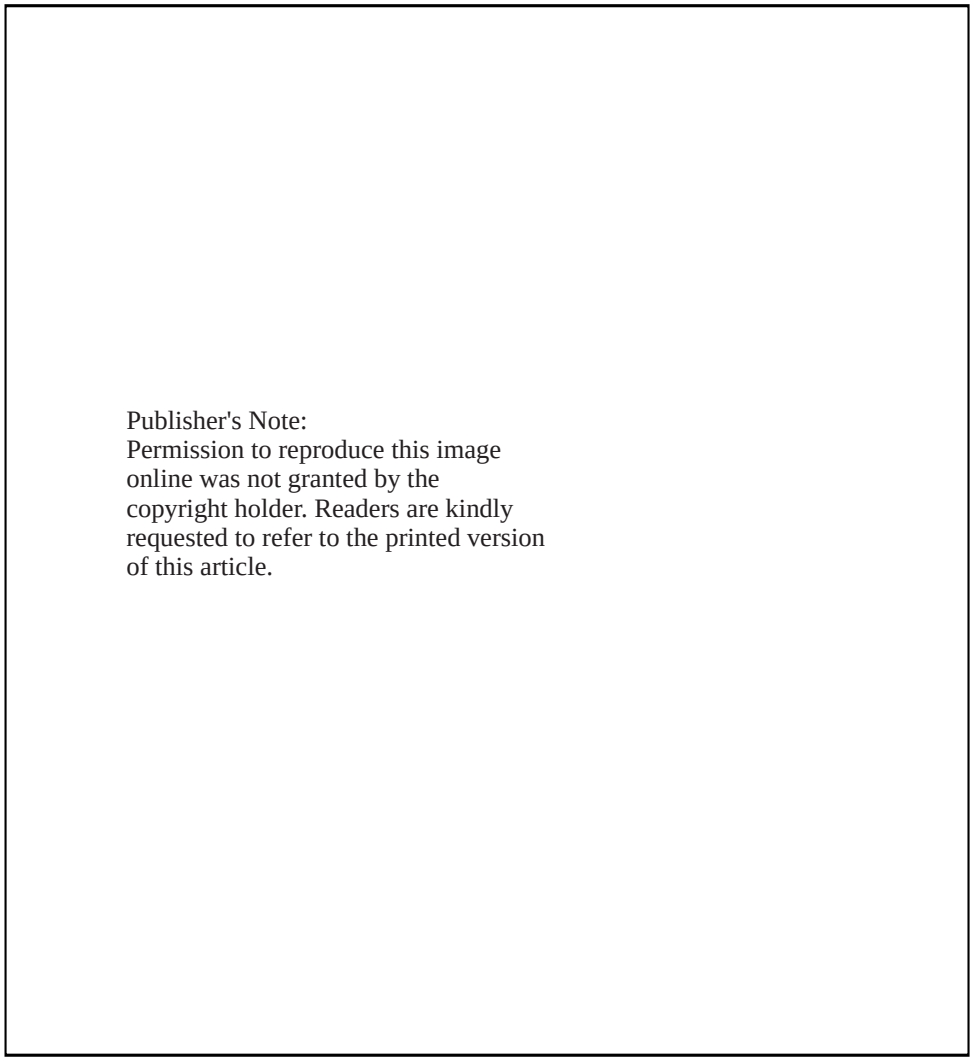
One of the cornerstones of modern clinical research is *informed consent* (consistent with the respect for autonomy). This seemingly simple concept is difficult and complex in application. Can someone near death truly give informed consent? Can prisoners truly give informed consent? Biologically, children are not small adults; drugs may have very different results with children. How can one get informed consent when studying children? Do parents or legal guardians really suffice? How can one do research in emergency settings with unconscious persons who need immediate treatment (e.g., in cardiac arrest)? Do people really understand what they are being told?

The issues have given rise to declarations by professional bodies (e.g., the Declaration of Helsinki, [World Medical Association, 1975], the Nuremberg Code [Reiser et al., 1947], and worldwide regulatory authorities (e.g., Federal Regulations [1988] on Institutional Review Boards). The Health Insurance Portability and Accountability Act (HIPAA) was passed by the U.S. Congress in 1996. The rules resulting from this act have been published and refined since that time. The revised final privacy rules were published in 2002. Much information is protected health information (PHI) and researchers in the United States need to be aware of these regulations and conform to the rules. In the United States, anyone involved in research on humans or animals needs to be familiar with the legal as well as the more general ethical requirements. Without a doubt there is great tension for medical personnel involved in research. Their mandate is to deliver the best possible care to their patients as well as to do good research. See Fisher [1998a] for a brief discussion and some references. In addition, some statistical professional societies have given ethical guidelines for statisticians [Royal Statistical Society, 1993; American Statistical Association, 1999].

All agree that ethical considerations must precede and take precedence over the science. What this means in practice can lead to legitimate differences of opinion. Further continuing scientific advances (such as genetics, cloning, or fetal research) bring up new and important issues that require a societal resolution of what constitutes ethical behavior.

19.3 OBSERVATIONAL AND EXPERIMENTAL STUDIES IN HUMANS

In this section we consider some reasons why randomized studies are usually required by law in the development of new drugs and biologics. Rather than a systematic development, we begin with a few examples and possible lessons to be learned from them.



Example 19.2. If taking a drug helps you survive, it must be effective! During a National Institutes of Health (NIH) Randomized Clinical Trial [Coronary Drug Project Research Group, 1980] a drug was found to have about half the mortality among those who took the drug (defined as taking 80% or more of the assigned medication) vs. those who did not take the drug consistently. The five-year mortality in the men with coronary heart disease was 15.1% of the "good adherers" to drug and 28.2% in the "poor adherers" to drug. Although it certainly seems that the drug is effective (after all counting bodies is not subject to bias), it is possible that those who were good adherers were different when the study started. Fortunately, this was an NIH study with excellent detailed data collected for the known risk factors in this population. There were some differences at baseline between the good and poor adherers. Thus, a multiple linear regression analysis of five-year mortality was run, adjusting for 40 baseline variables in the 2695 patients taking the drug.

The analysis adjusting for these 40 variables led to adjusted five-year mortality of 16.4% for good adherers vs. 25.8% for the poor adherers. This would seem to clearly indicate a survival

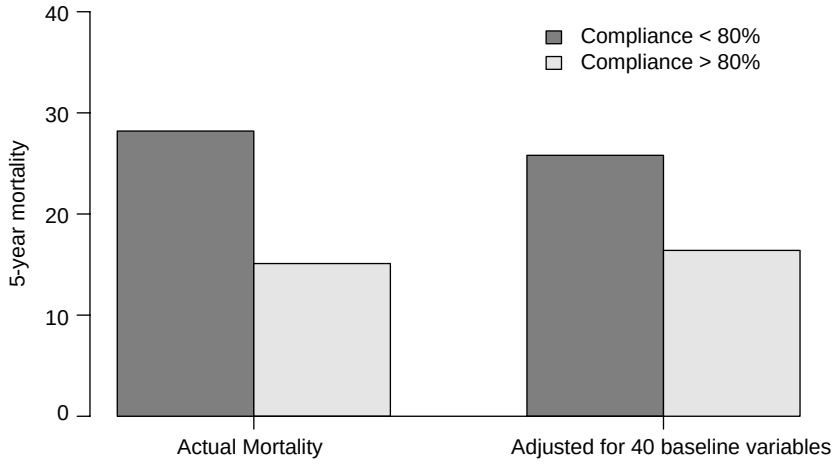


Figure 19.1 Five-year mortality among good and poor adherers to treatment.

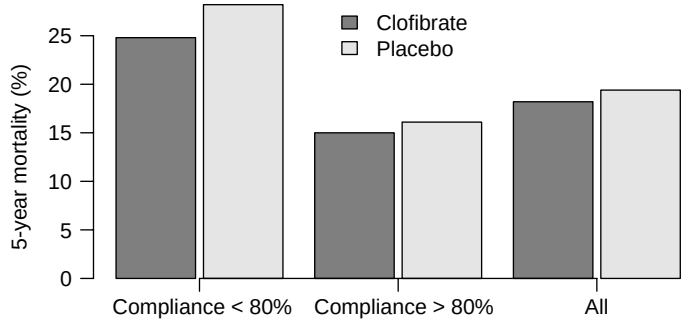


Figure 19.2 Five-year mortality by compliance and treatment in the Coronary Drug Project.

benefit of the drug—thus negating the need for a controlled study, although the data were collected for one arm of a controlled study. The only problem with this result is that the drug above was the placebo! In fact, the good and poor adherers of the active drug, clofibrate, had a very similar pattern. Figures 19.1 and 19.2 give the five-year mortality for the placebo arm of the trial and then for both arms of the trial. The two treatment arms did not differ statistically. The reason for the difference between the placebo mortality for the good and poor adherers was never fully understood.

Results such as this show how difficult it can be to assess a drug effect correctly from observational data. This is one reason why randomized clinical trials are the regulatory gold standard for most drug approvals. This is fine as far as it goes. We are then left with a very difficult consideration. Why, then, does this book give the majority of space to observational data analyses? If we cannot trust such analyses, why bother? The answer is that we do the best we can in any situation. If observational data analyses are the only practical method (due to cost or other feasibility factors), or the only ethical method (as the epidemiology of smoking risk became clear, it would not been considered ethical to randomize to smoking and nonsmoking treatment arms—not to mention the difficulty of execution), observational data must be used.

Example 19.3. If we stop the thing that appears to cause the deaths, we must be prolonging life (or are we?). One of the wonders of the body is our heart; it beats steadily minute

after minute, year after year. If the average number of beats is 60 per minute, there are 86,400 beats/day or 31,536,000 beats/year. In a 65-year-old, the heart may have delivered over 2 billion heartbeats. The contraction of the heart muscle to force blood out into the body is triggered by electrical impulses that depolarize and thus contract the heart in a fixed pattern. As the heart muscle becomes damaged, there can be problems with the electrical trigger that leads to the contraction of the heart. The electrical changes in the heart are monitored when a physician takes an electrocardiogram (ECG) of the heart. If the depolarization starts inappropriately someplace other than the usual trigger point (the sinus atrial node), the heart can contract early; such a resulting irregular heartbeat, or arrhythmic beat, is called a *ventricular premature depolarization* (VPD). Although most people have occasional VPDs, after a heart attack or myocardial infarction (MI), patients may have many more VPDs and complex patterns of irregular heart beats, called *arrhythmias*. The VPDs place patients at an increased risk of sudden cardiac death. To monitor the electrical activity of the heart over longer time periods, ambulatory electrocardiographic monitors (AECGMs) may be used. These units, also called *Holter monitors*, measure and record the electrical activity of the heart over approximately 24-hour periods. In this way, patients' arrhythmic patterns may be monitored over time. Patients have suffered sudden cardiac death, or sudden death, while wearing these monitors, and the electrical sequence of events is usually the following: Patients experience numerous VPDs and then a run of VPDs that occur rapidly in succession (say, at a rate greater than or equal to 120 beats/min); the runs are called *ventricular tachycardia* (VT). Now many coronary patients have runs of VT; however, before death, the VT leads to rapid, irregular, continuous electrical activity of the heart called *ventricular fibrillation* (VF). Observed in a cardiac operation, VF is a fluttering, or quivering, of the heart. This irregular activity interrupts the blood flow and the patient blacks out and if not resuscitated, invariably dies. In hospital monitoring settings and cities with emergency rescue systems, the institution of *cardiopulmonary resuscitation* (CPR) has led to the misnomer of *sudden death survivors*. In a hospital setting and when emergency vehicles arrive, electrical defibrillation with paddles that transmit an electrical shock is used. Individuals with high VPD counts on AECGMs are known to be at increased risk of sudden death, with the risk increasing with the amount and type of arrhythmia.

This being the case, it was natural to try to find drugs that reduced, or even abolished, the arrhythmia in many or most patients. A number of such compounds have been developed. In patients with severe life-threatening arrhythmia, if an antiarrhythmic drug can be found that controls the arrhythmia, the survival is greatly superior to the survival if the arrhythmia cannot be controlled [Graboys et al., 1982]. Graboys and colleagues examined the survival of patients with severe arrhythmia defined as VF (outside the period of an MI) or VT that compromised the blood flow of the heart to the degree that the patients were symptomatic. Figure 19.3 gives the survival from cardiac deaths in 98 patients with the arrhythmia controlled and 25 patients in whom the arrhythmia was not controlled.

Thus, there was a very compelling biological scenario. Arrhythmia leads to runs of VT, which leads to VF and sudden death. Drugs were developed, and could be evaluated using AECGMs, that reduced the amount of arrhythmia and even abolished arrhythmia on AECGMs in many patients. Thus, these people with the reduced or abolished arrhythmia should live longer. One would then rely on the *surrogate endpoint* of the arrhythmia evaluation from an AECGM. A surrogate endpoint is a measurement or event that is thought to be closely associated with the real endpoint of interest such that inducing changes in the surrogate endpoint would imply similar changes in the "real" endpoint of interest. Usually, the surrogate endpoint is a measurement or event that is not of direct benefit to a patient or subject, but that is presumably related to direct benefit and can be used to establish benefit. Prentice [1989] defines the issue statistically: "I define a surrogate endpoint to be *a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint*." Antiarrhythmic drugs were approved by the U.S. Food and Drug Administration (FDA) based on this surrogate endpoint. It is important to point out that antiarrhythmic drugs may have other benefits than preventing sudden death.

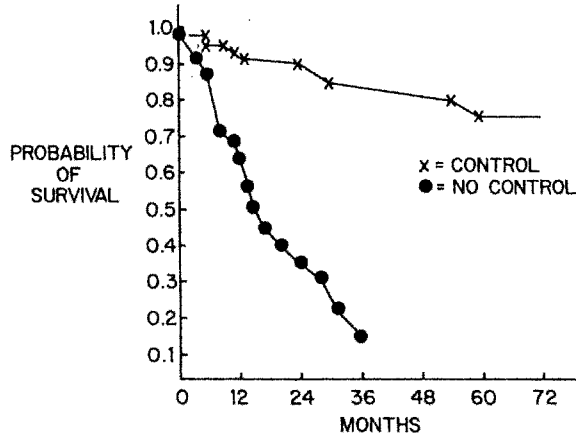


Figure 19.3 Survival free 17 cardiac mortality in patients with severe arrhythmia. The curves are for those whose arrhythmia was controlled by antiarrhythmic drugs and for those in whom the arrhythmia was not controlled by antiarrhythmic drugs.

For example, some patients have such severe runs of VT that they faint. Prevention of fainting spells is of direct benefit to the patient. However, asymptomatic or mildly symptomatic patients with arrhythmia were being prescribed antiarrhythmics with the faith(?), hope(?) that the drugs would prolong their life.

Why, then, would anyone want to perform a randomized survival trial in patients with arrhythmia? How could one perform such a trial ethically? There were a number of reasons: (1) the patients for whom arrhythmia could be controlled by drugs have selected themselves out as biologically different; thus, the survival *even without antiarrhythmic therapy* might naturally be much better than patients for whom no drug worked. That is, modification of the surrogate endpoint of arrhythmia had never been shown to improve the results of the real endpoint of interest (sudden death). (2) Some trials had disturbing results, with adverse trends in mortality on antiarrhythmic drugs [IMPACT Research Group, 1984; Furberg, 1983]. (3) All antiarrhythmic drugs actually produce more arrhythmia in some patients, a *proarrhythmic effect*.

The National Heart, Lung and Blood Institute decided to study the survival benefit of antiarrhythmic drugs in survivors of a myocardial infarction (MI). The study began with a pilot phase to see if antiarrhythmic drugs could be found that reduced arrhythmia by a satisfactory amount. If this could be done, the randomized survival trial would begin. The first study, by the Cardiac Arrhythmia Pilot Study (CAPS) Investigators [1988], showed that three of the drugs studied—encainide, flecainide, and moricizine—suppressed arrhythmias adequately to allow proceeding with the primary survival trial, the Cardiac Arrhythmia Suppression Trial (CAST). Patients within six weeks to two years of an MI needed six VPDs per hour to be eligible for the study. There was an open label, dose titration period where drugs were required to reduce VPDs by at least 80% and runs of VT by at least 90%. (For more detail, see the Cardiac Arrhythmia Suppression Trial (CAST) Investigators [1989] and Echt et al. [1991].) Patients for whom an effective drug was found were then randomized to placebo or to the effective drug (Figure 19.4). Such was the confidence of the investigators that the drugs at least were doing no harm that the test statistic was one-sided to stop for a drug benefit at the 0.025 significance level. The trial was not envisioned as stopping early for excess mortality in the antiarrhythmic drug groups.

The first results to appear were a tremendous shock to the cardiology community. The encainide and flecainide arms were dropped from the study because of excess mortality! Strictly speaking, the investigators could not conclude this with their one-sided design. However, the

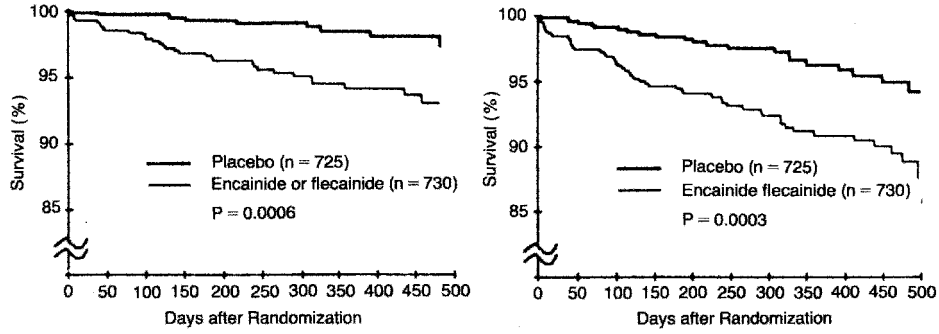


Figure 19.4 The panel on the left shows the survival, free of an arrhythmic death, among 1455 patients randomized to either placebo or one of encainide or flecainide. The second panel is based on all-cause mortality. (From the Cardiac Arrhythmia Suppression Trial (CAST) Investigators [1989].)

evidence was so strong that the investigators, and almost everyone else, were convinced of the harmful effects of these two antiarrhythmic drugs as used in this patient population.

The results of the study have been addressed by Pratt et al. [1990] and Pratt [1990]; the timing of the announcement of the results is described in Bigger [1990]; this paper gives a feeling for the ethical pressure of quickly promulgating the results. Ruskin [1989] conveys some of the impact of the trial results: "The preliminary results ... have astounded most observers and challenge much of the conventional wisdom about antiarrhythmic drugs and some of the arrhythmias they are used to treat. ... Although its basis is not entirely clear, this unexpected outcome is best explained as the result of the induction of lethal ventricular arrhythmias (i.e., a proarrhythmic effect) by encainide and flecainide."

This trial has saved, and will continue to save lives by virtue of changed physician behavior. In addition, it clearly illustrates that consistent, plausible theories and changes in surrogate endpoints cannot be used to replace trials involving the endpoints of importance to the patient, at least not initially. Finally, it is important to note that one should not overextrapolate the results of a trial; the study does not apply directly to patients with characteristics other than those in the trial; it does not imply that other antiarrhythmic drugs have the same effect in this population. However, it does make one more suspicious about the role of antiarrhythmic therapy, with a resulting need for even more well-controlled randomized data for other patient populations and/or drugs.

The trial illustrates the difficulty of relying on very plausible biological theories to generate new drug therapy. New therapies should be tested systematically in a controlled fashion on humans following ethical guidelines and laws. Note also that the arrhythmia itself is not the true focus of the therapy. It was thought to be a good "surrogate" for survival. The use of surrogate endpoints as a guide to approving new therapies is very risky, as the example shows [Temple, 1995; Fleming and DeMets, 1996].

Example 19.4. Epidemiological studies have shown that higher than normal blood pressure in humans is associated with shorter life span [Kesteloot and Joosens, 1980]. The decrease is due especially to increased cardiovascular events, such as a heart attack, stroke, or sudden death due to arrhythmia. Early clinical trials showed that lowering blood pressure by drug therapy resulted in fewer heart attacks, strokes, and cardiovascular deaths. Subsequently, it was considered unethical to treat persons with high blood pressure, called *hypertensive individuals*, with a placebo or sham treatment for a long period of time. Thus blood pressure–lowering drugs, *antihypertensive drugs*, were studied for relatively short periods, six to 12 weeks, in subjects with mild to moderate hypertension. The surrogate endpoint of blood pressure reduction is used for approval of antihypertensive drugs. As blood pressure tends to rise with physical or emotional

stress it is subject to change in response to subtle clues in the environment. For this reason trials use placebo (*inactive*) pills or capsules that are in appearance, smell, and so on, the same as tested *active treatment* pills or capsules, as discussed in Chapter 1. In addition, to prevent the transmission of clues that might affect blood pressure, the subject is not informed if she or he is taking the active drug or the placebo drug. If only the subject does not know the treatment, the trial is a *single-blind trial*.

However, since subtle clues by those treating and/or evaluating the subjects could affect blood pressure, those treating and/or evaluating the subjects also are not told if the subject is getting the active or placebo treatment. A study with both the subject and medical personnel blinded is called a *double-blind study*.

At the beginning of the study, subjects are usually all started on placebo during an initial single-blind period. This period serves multiple purposes: (1) it allows the effect of prior therapy to *wash out* or disappear; (2) it allows identification of subjects who will take their medication to be used in the comparative part of the trial; (3) it lessens the effect of raised blood pressure due to the unsettling medical setting (the *white coat hypertension* effect); (4) it helps to remove a regression to the mean effect of patient selection; and (5) multiple readings can assure relative stability of and measurement of the baseline blood pressure.

Figure 19.5 shows the data of the placebo arm in such a trial. Since subjects were on placebo the entire time, the explanation for the stable mean pressure during the single-blind *run-in period* and the drop during the double-blind portion of the trial was thought to be subtle clues being given to the patients by the medical personnel when they knew that some patients would be getting active therapy. It should be emphasized that subjects were never told in the single-blind portion of the trial that they were not potentially receiving active therapy. (The subjects did sign an informed consent and knew that they might receive placebo or active therapy during portions of the trial.) This figure illustrates the need for blinding in some clinical trials.

Figure 19.6 shows data from a second trial of an antihypertensive drug. The trial was a dose escalation study. That is, the dose of a drug was increased in individual patients until they had a satisfactory blood pressure response. Again the data are from the placebo arm of the trial. The increasing "benefit" observed as the "dose" of placebo escalates illustrates the need for a control group.

Example 19.5. In the United States, the National Institutes of Health (NIH) administers most federal funds for health sciences research as well as having its own (intramural) programs of research. Most of its employees thus value and are aware of the importance of well-conducted medical research. Thus, the NIH population would seem the ideal place to study an intervention

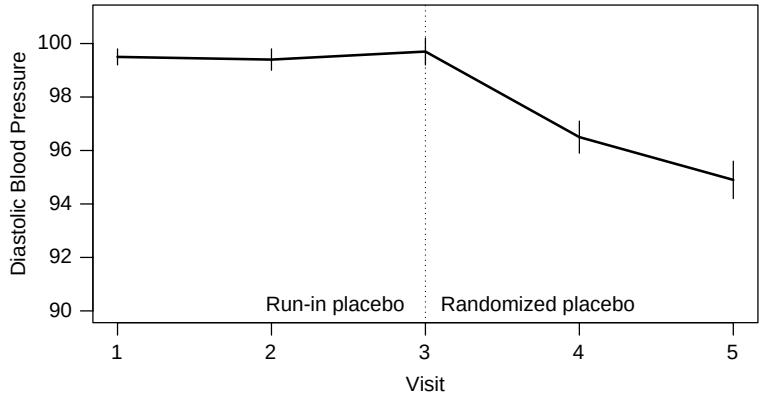


Figure 19.5 Average diastolic blood pressure (±1 standard error) during single-blind run-in and double-blind treatment with placebo.

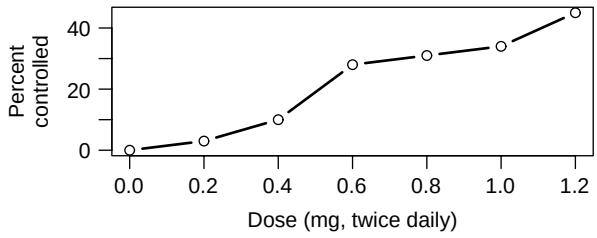


Figure 19.6 Response to escalating doses of placebo antihypertensive.

if there were sufficient numbers of NIH employees experiencing the malady in question. The results of a study on the use of ascorbic acid (vitamin C) for the common cold were published by Karlowski et al. [1975]. Most aspects of the study will not be presented here, in order to concentrate on the difficulty of performing a good experiment. There were four groups in the study. As a preventive (prophylactic) measure there was random assignment to either ascorbic acid or placebo (with capsules containing the study medication), and when a cold was thought to occur (with a clear definition), the study participants were assigned at random (the same for all colds if multiple colds occurred) to either ascorbic acid or placebo. Thus, there were four groups. Three hundred and eleven persons were randomized to therapy (discounting 12 subjects who dropped out early "before taking an appreciable number of capsules"). During the study the investigators learned that some subjects had opened the capsules and tasted the contents to see if they were taking ascorbic acid or placebo. More prophylactic placebo subjects (69) dropped out than ascorbic acid prophylactic subjects (52). At the end of the study the investigators queried the subjects about whether they thought they knew their study drug; of 102 subjects who thought they knew, 79 (77%) guessed correctly. The study results showed no statistical difference in the number of colds, but there was a trend for less severity of a cold if one took ascorbic acid. Unfortunately, this trend disappeared if one took into account those who knew their therapy. The NIH investigators comment under the heading the *power of suggestion*: "Depending upon one's point of view, it is either an unfortunate or fortunate aspect of the study. It would have been gratifying to have performed a flawless clinical trial; on the other hand, it has turned out to be a unique opportunity to gain some insight into the importance of perfect blinding in trials with subjective endpoints. An association between severity and duration of symptoms and knowledge of the medication taken seems to have been clearly established."

These examples above illustrate:

1. The need for a control group to be compared with an active therapy
2. The need for a "fair" or unbiased control, or comparison, group or appropriate mathematical adjustment to make a fair comparison. Appropriate mathematical adjustment is very difficult to do in this setting (as Example 19.2 illustrates)
3. The need for blinding to avoid introducing bias into clinical trials
4. The need for an endpoint of a trial that has clinical relevance (e.g., Temple [1995])

19.4 OBTAINING A FAIR OR UNBIASED COMPARISON: RANDOMIZED CLINICAL TRIAL

We now turn to two aspects of the clinical trial. The first is summarized by the question: How can we assign subjects to unbiased, or comparable, groups at the start of a clinical trial? The idea of random selection to get a "fair" choice or comparison goes back a long time in human history. Lots were used in Old Testament times, the idea of "drawing the short straw," taking a card from a well-shuffled pack, and so on, all show the intuitive appeal of this type of

procedure. However, the formal introduction of *randomization* was made in the 1930s by the British statistician and geneticist Sir Ronald Aylmer Fisher [Box, 1978]. In one of the great intellectual advances of the twentieth century, he combined the methodology of probability theory with the intuitive appeal of randomization to begin the *randomized experiment*. The idea is in some ways counterintuitive. As seen previously in this book, a theme of good observational data analysis or experimentation is to eliminate variability in order to make comparisons as precise as possible. Randomness, or "unexplained noise," does just the opposite. Think of the simplest type of random assignment between two treatments: Each eligible patient has her or his therapy determined (after informed consent) by the flip of an unbiased coin (i.e., the probability of each treatment is $\frac{1}{2}$). The different flips are statistically independent, and if there are n assignments, the number on treatment A, or B for that matter, is a binomial variable. Further, any particular pattern of assignment is equally likely ($1/2^n$).

What are the benefits of this random assignment? First, the assignment to treatment is fair. Human biases, whether conscious or unconscious, are eliminated. Second, on average, the two assignments have the same number of easy or difficult-to-treat assignments; that is, patient characteristics are balanced (statistically). Third, if we assume that treatment is unrelated to our outcome, we can assume that the outcomes were preordained to be good or bad. We can find the probability under this random assignment that each treatment arm had outcomes as extreme or more extreme than that actually observed with the actual assignments because we know that each assignment of cases is equally likely. (see Chapter 8). That is, we can compute a p-value that is not dependent on assumptions about the population we are observing. This is called using the *randomization distribution* (see Edgington [1995]). We do, however, need to be sure that the randomization is done appropriately.

The benefits of the randomized trial are so widely recognized that by law and regulation, in most countries new drugs or biologics need to be evaluated by a randomized clinical trial in order to gain regulatory approval to market the new advance legally. See Note 19.4 for a few references on the need for and benefits of the RCT.

19.4.1 Intent to Treat

There are complications to RCTs in practice. Suppose, in fact, that many patients assigned to one, or both, of the treatments do *not* get the assigned therapy? Does it make sense to compare the treatments as randomized? How can patients who do not receive a therapy benefit from it? Thus, does it not seem odd to keep such patients in a comparison of two therapies? This sticking point has led to some difficult considerations: If we consider only patients who received their assigned or randomized therapy, we can introduce bias since those who do not receive their therapy are usually different (and unfortunately, possibly in unknown ways) from those who do receive their assigned therapy. The issue then becomes one of avoiding bias (include all patients who are randomized into their assigned group) vs. biological plausibility (only count those who actually receive a treatment). At its worst this might pit biostatisticians vs. clinicians. At this point in time, including all subjects in the analysis into the group to which they are randomized is considered standard; such analyses are called *intent-to-treat* (ITT) *analyses*. The name arises from the fact that under the randomized assignment there is an implied initial intent to treat the subject in the manner to which he or she was randomized. The best way to avoid the conflict between bias and biology is to perform an excellent experiment where those randomized to a treatment do receive the treatment. For this reason the assignment to randomization should be accomplished at the last possible moment.

If those subjects who do not begin treatment do so for reasons that cannot have been due to the randomized assignment (e.g., nonbreakable double blinding), the subjects who at least begin therapy can be included into the analysis with all the benefits of the randomization process listed above. Such analyses are called *modified intent to treat* (mITT) and are acceptable provided that one can be assured that the lack of therapeutic delivery *cannot* have been related to the treatment assignment. In practice, modified intent-to-treat analyses are often also called intent to treat.

19.4.2 Blinding

We have seen above that using a randomized assignment does not accomplish the full task of assuring a fair comparison. If the outcome is affected by biased behavior due to the treatment assignment, we can have misleading results despite the fact that the treatments were assigned at random. Bias can still ruin an RCT. We have seen this in both the blood pressure and vitamin C examples above. Wherever possible, double blinding should be used. The more subjective the endpoint, the more important blinding is to a trial. However, even with very "hard" endpoints that would not seem to need blinding (e.g., mortality), blinding can be important. The reason is that if the blinding is not effective, there may be treatment biases that change the way subjects in the assigned groups are treated (e.g., hospitalized, given other medications) and this may affect even hard endpoints such as mortality. It is difficult to blind in many trials [e.g., a drug may induce physiologic changes (in heart rate or blood pressure)] and those seeing and treating a patient may have reasonable guesses as to the therapy. Added steps can be taken. For example, those involved in evaluating a patient for outcome might be required to be different from those treating a patient. Often, outcomes for a trial are evaluated by an external classification committee to reduce bias in the determination of events.

19.4.3 Missing Data

Missing data are one of the most common and difficult issues in the analysis of RCTs. Even a modest discussion of the ways to approach and handle missing data in RTCs goes beyond the scope of this book. However, a few partial solutions, based on the concepts introduced in Section 10.5.2 and Chapter 18 are presented here.

The first and most important thing to understand is that there is no totally satisfactory method of dealing with the issue. The best course is not to have any missing data, but often, that wonderful counsel cannot possibly be implemented. For example, in studies performed in a population of street people with illicit drug use, complete data are virtually unknown if the study requires patient cooperation over a moderate length of time. Subjects simply disappear and are extremely difficult to find. Some turn up in jail or hospitals, but follow-up is difficult. It they are to return for follow-up visits, adherence can be quite low. What are those running such a trial, as well as the general society, with its interest in the outcome, to do? We do the best we can but realize that there will be many missing data. Another example: One studies treadmill walking time in a population of congestive heart failure patients. The primary study endpoint is the change in treadmill time from the baseline measurement to the final visit (at some fixed interval from the time the subject was randomized). Some subjects will die: How should their data be treated in the final analysis? Clearly, the missing information (the impossible final treadmill test) is not independent of patient status. This is known as *informative censoring*. Others may have their heart failure progress to a stage where it is too difficult to come in for the test or to perform the test. Other subjects may become discouraged and exercise their right to withdraw from the study. Others may go on vacation and not be around at the correct time for their evaluation. The possibilities go on and on.

First, one might assume that the missing data do not bias the conclusions and analyze only those who have all appropriate data. This is usually not an acceptable approach unless there are only minimal missing data. However, it is often used as an additional analysis. Data may also be "missing" for legitimate medical reasons. In a trial of blood-pressure-lowering medication, patients may present with greatly elevated blood pressures that require immediate, or perhaps after a week's delay, treatment with known effective drug or drugs. In many trials there are more such subjects in the placebo group. If their data are not taken into account, there is a bias against the active therapy. Further, their data at the end of the scheduled therapy period are not unbiased, as strong active therapy is used to lower blood pressure. In this case the endpoint used is the last observation on the assigned randomized therapy. In effect, the last observation is carried forward to the time for final evaluation. Not surprisingly, such analyses are called *last observation carried*

forward (LOCF). This is often used as a method of analysis when the primary parameter of the study is collected at regularly scheduled visits. Sometimes the missing data are replaced by the mean of the known values for the study. In other cases, more sophisticated methods are used to estimate, *impute*, the missing values. Such strategies can be quite complex. For a discussion of the implications of different reasons that data are missing, the implications for missing data, and analysis methods, see Little and Rubin [2002] and Section 18.6.

If the data are extremely strong, a *worst-case analysis* can be used and an effect still established. For example, in a survival analysis study that is placebo controlled, the comparison to a new therapy, the worst case (for establishing the new therapy), would assume that placebo patients not observed for the full observation period lived to the end of the follow-up period and that those assigned to the new active therapy died immediately after the last time they were known to be alive before being lost to follow-up.

The robustness of the study data to the missing data is sometimes assessed with some type of *sensitivity analysis*. Such analyses make a variety of assumptions about the actual pattern of the missing data and see how extreme it must be to change the study results in some important manner.

19.5 PLANNING AN RCT

19.5.1 Selection of the Study Population

In clinical studies the selection of the study population is critical. The understanding of the drug, biologic, or device mechanism will suggest a population of subjects where efficacy is to be shown. Selection of the highest-risk population is often the most logical choice to demonstrate the effect; however, if only such subjects are studied, the approval for use will usually be limited to such subjects. This may limit use of the new treatment. As a result, this may narrow the range of subjects getting a benefit, as well as lowering the sales potential for the sponsor developing the new therapy.

19.5.2 Special Populations

Historically, many important special populations either were not investigated at all or had very limited data—despite the fact that any realistic appraisal of usage patterns would anticipate such use. For example, women of childbearing potential were avoided; in large part, this was to avoid law suits if there were any birth defects in the children conceived, developing, or born during or close to the trial. The expense of a lifetime of care was avoided by not studying such women. The most infamous example of a drug causing birth defects was thalidomide in Europe and the United States. Nevertheless, many medications are used by pregnant women. Over-the-counter (OTC) products are the most obvious; analgesics (i.e., pain relievers) are one clear class. Now the FDA strongly recommends, and sometimes requires, such studies. Another underevaluated population was children. One might think that from a pharmacological point of view, children are merely small adults and that smaller doses would clearly work if the drug worked in adults. Unfortunately, this idea is simply not true. Children differ in many important ways in addition to size, and care is needed in extrapolating adult results to children. Historically, minorities, especially African-Americans, had limited experimental results in drug development (except in obvious special cases such as sickle cell anemia). In part, this was related to limited access to health care. There are genetic differences in the way that drugs affect humans, and minorities are now studied more systematically. Often, some clinical sites in studies are selected to test a therapy on a more diverse population. The elderly were also underrepresented in RCTs. In part, this is because the elderly have more trouble showing up for clinic visits and complying with their therapy (as they may forget to take their medication). However, the elderly are a particularly important population to study because (1) they take many medications, and drug–drug interactions that cause trouble are more likely to occur in this population; (2) drugs

are often metabolized in the liver, so poor liver function can cause problems (the elderly have more liver impairment); (3) elimination is often through the kidneys, and the elderly are more likely to have kidney problems; and (4) the changing world demography shows that a larger proportion of the world population will be elderly in the next few decades.

19.5.3 Multicenter Clinical Trials

Many clinical trials use multiple clinical centers to enroll patients or subjects. There are several reasons for this. The most obvious is the need to enroll many patients in a timely fashion. There are also other reasons, perhaps not as obvious. Most new drugs are developed to be registered (approved for marketing) in many markets around the world: the United States, the European Union, Japan, and Canada, among others. Thus, the studies often have clinical sites from around the world to aid in approval under the various regulatory authorities. Using "influential" physicians at different centers as investigating clinicians in the research program can also be an aid to marketing when approval is granted. Other benefits of using multiple clinics include (1) showing that there is a benefit in different settings, and (2) assessing therapy under a variety of concomitant medical therapeutic settings.

In addition to the benefits, there are numerous additional challenges to multicenter clinical trials. Standardization of treatment and data recording often require extensive education and monitoring. The randomization process needs to be available over a wide range of times if subjects are enrolled around the world. Forms and data collection may be complicated by the number of languages and cultures involved. Data are analyzed for clinical site heterogeneity in response; often, this is done for different delivery settings (e.g., North America, Europe, and the rest of the world). Security of data, monitoring of the raw data (often in clinical files), and investigator and staff training are all quite complicated.

19.5.4 Practical Aspects of Randomization

The process of randomizing subjects in an RCT involves choices. To simplify the discussion we consider only *two-arm trials*, but similar considerations can be used with more than two treatment arms. The simplest random allocation is a fair coin flip, allocating each subject to one arm or the other. (In practice, the "flips" are done using a *pseudorandom number generator* on a computer.) There are drawbacks to the coin-flip approach. If there are clinical sites, each enrolling a small number of subjects, a number of such sites may involve only one treatment. This makes it impossible to see the variability in treatment effect within such sites. Therefore, the randomization is done using randomized blocks. If the ratio of subjects randomized to each arm is to be the same, even-numbered blocks are used. If the size is $2n$, then among each $2n$ randomizations, n will be to one arm and n to the other. Potentially, this can lead to bias, since if the study is unblinded or one can unblind with a reasonable probability, the probabilities for subsequent patients is no longer $\frac{1}{2}$ to $\frac{1}{2}$. To see this, consider an unblinded study: If we know the first $2n - 1$ treatment assignments, we know what the next subject will receive as a treatment. To get around this problem partially, blocks of different size are sometimes used, being chosen with some probability. For example, one might choose a block of size 4 half the time and a block of size 6 half the time.

Often, the blocks are not used to get balance within a site. If there is an important factor that determines the risk of the trial outcome, blocks with some strata for the risk factor may be used. This "forces" some balance with respect to the important prognostic factor. If more than one factor exists, combinations of two or more factors might be used. There is a limitation, however; if one had five factors, each of which had three levels, and we took all combinations, there would be $5^3 = 125$ possible strata. As the number goes up, we tend to get cells with zero or one subject actually randomized within a cell. When we are using only the first element of each block, randomization is the same as if we did not block at all! For this reason, more complex schemes have been developed for forcing balance on a number of factors; this technique

is known as *adaptive randomization*. For blocking and adaptive randomization, one needs to know selected information about a subject before an assignment can be given. This is often done through either an interactive voice randomization system that uses touchtone phones or through the Internet. In either case, the needed information will be entered, eligibility may be checked, and the database is quickly informed of the randomization, and may check for subsequently expected data. See Efron [1971], Friedman et al. [1999, Chap. 5], or Meinert [1986, Sec. 10.2].

19.5.5 Data Management and Processing

Data management of randomized clinical trials is challenging, particularly so for international multicenter trials. In most instances, data are entered on *case report forms* (CRFs). Often, clinical sites are visited to compare the forms with the official medical records for consistency and documentation. Inspections are made by those sponsoring a study as well as by regulatory authorities if the trial aims to register a drug. Forms are usually submitted to a central data processing unit. They may be carried by hand using monitors, faxed after data entry at the clinical site (*remote data entry*), transferred electronically, entered via the Internet, or (more and more rarely) mailed in batches. To minimize data-entry errors, the data are often entered twice by two different people, and the entries compared for consistency with resolution in the case of disagreement. Entered files usually undergo extensive *consistency checks* [e.g., are the dates possible? Is a datum plausible (in that it is in a reasonable range)? If a discrete variable, is the code a legal one?] One of the worst errors is to have an incorrect patient identifier for a form or forms; for this reason, patient-identifying information (which only identifies the patient uniquely, not allowing the actual person to be identified) often has redundant checking information. When an entry fails a check, a process is instituted to resolve the problem. Tracking the resolution and any changes is documented for possible subsequent review. Problem resolution can be quite extensive and time consuming.

The database often allows identification of the timing of needed follow-up visits, examinations, or contact. For complex studies the database is sometimes used for notifying clinical sites of the expected upcoming data collection. Missing forms (i.e., those expected from subsequent visits) are asked for after some time interval. Some possible inconsistencies may arise externally (e.g., from a blinded committee used to classify endpoints that need resolution), and these are also tracked and recorded. Before a study is analyzed, or unblinded, all outstanding data issues are resolved to the extent possible, and the data file is then *frozen* for the analysis and interpretation of data. In studies that need ongoing monitoring for ethical reasons, there may be an independent *data and safety monitoring board* to review interim data (see Ellenberg et al. [2002]). To avoid introducing bias into the study, a group, independent of the sponsor, often provides tables, lists, and materials. The complexity and effort needed for such processes is hard to appreciate unless one has been through it. (See also Sections 2.6 to 2.9.)

19.6 ANALYSIS OF AN RCT

19.6.1 Preservation of the Validity of Type I Error

Because drug development costs so much and because the financial reward for a successful new drug in the right setting is so great, there is an apparent conflict between the sponsors and regulators. Stated statistically, the sponsors want to maximize the power of a study (i.e., minimize Type II error), and the regulators want to minimize and preserve the appropriateness and interpretability of the Type I error or p-value. Some areas of particular related concern are discussed below.

19.6.2 Interim Analysis of an Ongoing Clinical Trial

New investigational therapies hold potential for both benefit and harm. Experience has shown that no matter how thorough the prior work in other animal species, the results in humans may

differ in unexpected ways. This is especially true with respect to adverse events. This requires looking at outcomes during the study—carrying out *interim analyses*. Similarly, when serious irreversible endpoints, such as death or permanent disability, are being considered, if a therapy is beneficial, there is an ethical requirement to stop the trial. But repeated interim analyses inflate the Type I error. This problem has been dealt with extensively in the biostatistical literature under the rubric of *sequential analysis*. Boundaries for values of a test statistic that would stop the trial at different times have been studied extensively (e.g., O'Brien and Fleming [1979]; Whitehead [1983]; Jennison and Turnbull [2000]; Lan and DeMets [1983]). In recent years, methods have been developed that allow examination of the results by treatment arm, with resulting modifications of the trial that still preserve the Type I error (e.g., Fisher [1998b]; Cui et al. [1999]). A basic strategy is to parcel out the Type I error over the trial. For example, suppose that two interim analyses are planned during the course of a study. Then test the results for the two interim analyses at the 0.001 level and the final analysis at the 0.048 level. This still ensures an overall level of 0.05.

19.6.3 Multiple Endpoints, Multivariate Endpoints, and Composite Endpoints

In some situations, multiple endpoints may be used to demonstrate the benefit of a new therapy. Of course, one cannot simply look at all of them and claim success if any one of them meets the significance level used in the RCT because the multiple comparisons inflate the Type I error. Several strategies have been used:

1. Select one of the possible beneficial endpoints to be the primary analysis for trial.
2. Adjust the p-value to account for the multiple comparisons. A conservative adjustment is to use the Bonferroni inequality and its refinements (Chapter 12) [Wright, 1992]. If the possible endpoints are positively correlated, as is usually the case, less severe adjustments are possible using the randomization distribution for the RCT.
3. The various components of possible endpoints can be considered to be a vector (i.e., arranged in sequence), and methods are available to test all the endpoints at once.
4. Sometimes an index, a weighted sum of the endpoints, is used as the one primary endpoint (see Schouten [2000]).
5. When a number of endpoints occur as distinct events in time, the first occurrence of any of them can be used as one event. Comparisons may be made using the methods of survival, or time to event, analysis (Chapter 16).

These issues are discussed in more detail in Chapter 12.

19.7 DRUG DEVELOPMENT PARADIGM

The following points introduce some of the ideas and terminology used in the development of drugs and biologics (see Mathieu [2002] for more). The first step is to identify a potential drug (a molecule). This used to be accomplished largely by chance (e.g., the discovery of penicillin) or through large screening programs, but because of recent substantial advances in genetics, molecular biology, and computer modeling, more and more compounds are being designed for specific purposes. Compounds may be screened for *in vitro* (i.e., "in glass") reaction with known molecules to identify candidates.

The first testing is carried out in several animal species. This *preclinical phase* of drug development accomplishes several purposes. Among the purposes are the following: The first is to identify if a drug is toxic at most possible doses (both short-term and longer-term studies in at least two species are done). Second, a range of doses can be evaluated. Are there doses that are not toxic (that have efficacy at the lower doses)? Third, use of an animal species will sometimes allow examination of an efficacy assessment vs. toxicity as a function of the dose. Other tasks

performed are to look for the formation of fetal and birth abnormalities (*teratogenicity studies*), to see if drugs cause cancer (*carcinogenicity*, as a function of animal species and dose), and to see if gene abnormality results (*mutagenicity testing*). Of course, usually, the more drug one takes, the greater the amount that enters the body. One studies the time course of the drug [whether administered as a pill or capsule, by injection (intravenously or intramuscularly), by inhalation, etc.] within the body. Almost all drugs change into other molecules (metabolites) when in the body. Study of the time course of adsorption, distribution, metabolism, and elimination of the drug molecule and its metabolites comprises the field of drug *pharmacokinetics*. The relationship of the drug time–concentration value to the magnitude of effect is the field of *pharmacodynamics*.

After the preclinical data have been reviewed (in the United States) and approved by appropriate authorities, testing may begin in humans. *Phase I* of drug development is initial use of the drug in humans. Unfortunately, the preclinical animal testing gives only a rough idea of possible appropriate doses in humans. The animal data are often predictive only to an order of magnitude, so testing in humans usually begins at a very low dose and is slowly escalated. If the drug is anticipated to be well tolerated by normal subjects, the initial testing is usually done in healthy, normal volunteers. Drugs that are harmful by their nature [e.g., cancer (oncology) drugs that kill cells] are tested initially in patients. Some idea of activity may be gained in this initial phase. Slow escalation of the dose given helps to establish a preliminary dose range for the compound.

Phase II studies are reasonably large studies that give preliminary evidence of the efficacy of a drug in humans, to determine reasonable doses, and to get evidence on safety and tolerability in a patient population. These studies are often not blinded.

Phase III studies are large, randomized clinical trials to establish efficacy and safety. For most drugs it is expected that there will be at least two independent RCTs, double-blinded where possible, that establish efficacy at the 0.05 significance level. An increasing number of active control trials are being conducted in which noninferiority is established by showing that the new compound does not differ from the active control by more than a small equivalence margin (see, e.g., Temple and Ellenberg [2000]; Ellenberg and Temple [2000]). Often, the Phase III trials for efficacy do not provide adequate experience to evaluate patient safety. There often are *open label* (i.e., patient and physician know what treatment the subject is getting) extensions, where all patients get the new therapy if they consent to continue in the study. These trials may enroll more subjects, to get additional safety data.

After drugs are approved, *postmarketing*, or *Phase IV*, studies are sometimes performed for a variety of purposes: to collect more safety data, to do additional evaluation of efficacy (sometimes using a different endpoint), or to study efficacy in a broader, representative population.

19.8 SUMMARY

RCTs are difficult, expensive, ethically challenging, and require great attention to planning and monitoring operationally. Still the benefits are generally agreed to be worth the effort. This type of human experimentation gives the most cogent and convincing proof of the benefit of a new therapy. Further, the control group (whether a placebo or a proven active therapy) provides a better comparison of the safety of a new therapy. The benefit and risk must be traded off in the approval of new therapies.

NOTES

19.1 Interventions Other Than Drugs

In the discussion above we have discussed RCTs primarily as if they were for new drugs or biologics. Many interventions, such as medical devices, have been and/or could be investigated using RCTs or analogs. A variety of surgical interventions have been investigated by RCTs.

Prevention programs, such as smoking-cessation programs, can be investigated by randomizing larger experimental units. For example, in an NIH study of smoking prevention, the school district was the unit of randomization [Peterson et al., 2000]. One could randomize to different health care strategies, different modes of psychotherapy, and so on. In these studies the unit of randomization may be much larger; such group randomization is discussed by Feng et al. [2001].

19.2 Drug Approval and Physician Use of Drugs

In the United States, drugs are approved by the Food and Drug Administration. The approval includes labeling that specifies the population the drug is to benefit (i.e., the *indication*) as well as dosing information and warnings about safety, interactions with other drugs, and so on. Physicians may then legally use the drug for other indications (other diseases or patient populations) without violating the law (*off-label use*). If this use is in accord with the practice norms of the community, adequate malpractice defense can often be established. Drug companies selling the drugs are prohibited by law from advertising such off-label use of their product. One suspects that implied off-label uses are sometimes promoted.

19.3 Generic Drugs

In the United States, from the time that human experimentation begins, a sponsor has exclusive rights to sell the drug (assuming approval) for a limited period of time. The rights are for 17 years from the time the application is approved for experimentation on humans. Thus, there is a limited time to recoup research costs and make a profit. After this time, others may manufacture and sell the drug provided that they establish that it is the same drug (*bioequivalence*). These are called *generic drugs*. Equivalence is shown by establishing that the pharmacokinetics is the same for the new version and the original approved version. We do not address the topic of bioequivalence further here (see Chow and Liu [2000]).

19.4 Further Reading: Specific Topics

For more information on informed consent see, for example, Faden and Beauchamp [1986]. For a mathematical discussion of what constitutes an appropriate surrogate endpoint, see the paper of Prentice [1989]. For nice discussions of the history of blinding, see the papers by Kaptchuk [1998] and Chalmers [2001]. Some references on the benefits of the randomized clinical trial are Ederer [1975], Green [1982], Greenberg [1951], and Kempthorne [1977].

Since the 1970s, the number of articles and books about RCTs and statistical analysis has grown exponentially (e.g., books on clinical trials: Bulpitt, 1996; Cato and Sutton, 2002; Chow and Liu, 2003; Cleophas et al., 2002; Duley and Farrell, 2002; Friedman et al., 1999; Matthews, 2000; Meinert, 1986; Mulay, 2001; Norleans, 2001; Piantadosi, 1997; Pocock, 1996; Spilker, 1991).

There are numerous books about particular disease areas (e.g., AIDS [Finkelstein and Schoenfeld, 1999]; cardiology and cardiovascular disease [Hennekens and Zorab, 2000; Pitt et al., 1997]; epilepsy [French et al., 1997]; hypertension [Black, 2001]; multiple sclerosis [Goodkin and Rudick, 1998]; neurology [Guilogg, 2001; Porter and Schoenberg, 1990]; oncology [Green et al., 2002]; opthamology [Kertes and Conway, 1998]); and for material for patients [Giffels, 1996; Slevin and Wood, 1996]; aspects of trials, such as quality of life and pharmacoeconomics [Fairclough, 2002; Spilker, 1995]; data management [McFadden, 1997]; combining data from trials (metaanalysis: [Whitehead, 2002]; evaluating the literature [Ascione, 2001]; and dictionary or encyclopedic entries [Day, 1999; Redmond et al., 2001]).

REFERENCES

American Statistical Association [1999]. *Ethical Guidelines for Statistical Practice.* ASA, Alexandria, VI.

Ascione, F. J. [2001]. *Principles of Scientific Literature Evaluation: Critiquing Clinical Drug Trials*. American Pharmaceutical Association, Washington, DC.

Beauchamp, T. L., and Childress, J. F. [2001]. *Principles of Biomedical Ethics*, 5th ed. Oxford University Press, New York.

Bigger, J. T., Jr., [1990]. Editorial: the events surrounding the removal of encainide and flecainide from the Cardiac Arrhythmia Suppression Trial (CAST) and why CAST is continuing with moricizine. *Journal of the American College of Cardiology*, **15**: 243–245.

Black, H. R. [2001]. *Clinical Trials in the Pharmacologic Management of Hypertension*. Marcel Dekker, New York.

Box, J. F. [1978]. *R. A. Fisher: The Life of a Scientist*. Wiley, New York, p. 146.

Bulpitt, C. J. [1996]. *Randomized Controlled Clinical Trials*. Kluwer Academic, New York.

Cardiac Arrhythmia Pilot Study (CAPS) Investigators [1988]. Effect of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: the CAPS. *American Journal of Cardiology*, **61**: 501–509.

Cardiac Arrhythmia Suppression Trial (CAST) Investigators [1989]. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine,* **321**: 406–412.

Cato, A. E., and Sutton, L. [2002]. *Clinical Drug Trials and Tribulations*, 2nd ed. Marcel Dekker, New York.

Chalmers, I. [2001]. Comparing like with like: some historical milestones in the evolution of methods to create unbiased groups in therapeutic experiments. *International Journal of Epidemiology*, **30**: 1156–1164.

Chow, S.-C., and Liu, J.-P. [2003]. *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd ed. Wiley, New York.

Chow, S.-C., and Liu, J.-P. [2000]. *Design and Analysis of Bioavailability and Bioequivalence Studies*, rev. ed. Marcel Dekker, New York.

Cleophas, T. J., Zwinderman, A. H., and Cleophas, T. F. [2002]. *Statistics Applied to Clinical Trials*. Kluwer Academic, New York.

Coronary Drug Project Research Group [1980]. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine*, **303**: 1038–1041.

Cui, L., Hung, H. M. J., and Wang, S.-J. [1999]. Modification of sample size in group sequential clinical trials. *Biometrics*, **55**: 853–857.

Day, S. [1999]. *Dictionary for Clinical Trials*. Wiley, New York.

Duley, L., and Farrell, B. (eds.) [2002]. *Clinical Trials*. British Medical Association, London.

Echt, D. S., Liebson, P. R., Mitchell, B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A. Friedman, L., Greene, H. L., Huther, M. L., Richardson, D. W., and the CAST Investigators [1991]. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. *New England Journal of Medicine*, **324**: 781–788.

Ederer, F. [1975]. Why do we need controls? Why do we need to randomize? *American Journal of Ophthalmology*, **79**: 758–762.

Edgington, E. S. [1995]. *Randomization Tests*, 3rd rev. exp. ed. Marcel Dekker, New York.

Efron, B. [1971]. Forcing a sequential experiment to be balanced. *Biometrika*, **58**: 403–417.

Ellenberg, S. S., and Temple, R. [2000]. Placebo-controlled trials and active-control trials in the evaluation of new treatments: 2. Practical issues and specific cases. *Annals of Internal Medicine*, **133**: 464–470.

Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. [2002]. *Data Monitoring Committees in Clinical Trials*. Wiley, New York.

Faden, R. R., and Beauchamp, T. L. [1986]. *A History and Theory of Informed Consent*. Oxford University Press, New York.

Fairclough, D. L. [2002]. *Design and Analysis of Quality of Life Studies in Clinical Trials*. CRC Press, Boca Raton, FL.

Federal Regulations [1988]. 21 CFR Ch. I, Part 56: Institutional Review Boards (4-1-88 ed.). U.S. Government Printing Office, Washington, DC.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. [2001]. Selected statistical issues in group randomized trials. *Annual Review of Public Health*, **22**: 167–187.

Finkelstein, D. M., and Schoenfeld, D. A. [1999]. *AIDS Clinical Trials*. Wiley, New York.

Fisher, L. D. [1998a]. Ethics of randomized clinical trials. In *Encyclopedia of Biostatistics*, Vol. 2, P. Armitage and T. Colton (eds.). Wiley, New York, pp. 1394–1398.

Fisher, L. D. [1998b]. Self-designing clinical trials. *Statistics in Medicine,* **17**: 1551–1562.

Fisher, L. D, Dixon, D. O., Herson, J., Frankowski, R. F., Hearron, M. S., and Peace, K. E. [1990]. Intention to treat in clinical trials. In *Statistical Issues in Drug Research and Development*, K. E. Peace (ed.). Marcel Dekker, New York, pp. 331–350.

Fleming, T. R., and DeMets, D. L. [1996]. Surrogate endpoints in clinical trials: are we being mislead? *Annals of Internal Medicine*, **125**: 605–613.

French, J. A., Leppik, I. E., and Dichter, M. A. [1997]. *Antiepileptic Drug Trials*, Vol. 76. Lippincott Williams & Wilkins, Philadelphia.

Friedman, L., Furberg, C., and DeMets, D. L. [1999]. *Fundamentals of Clinical Trials*, 3rd ed. Springer-Verlag, New York.

Furberg, C. D. [1983]. Effect of antiarrhythmic drugs on mortality after myocardial infarction. *American Journal of Cardiology*, **52**: 32C–36C.

Giffels, J. J. [1996]. *Clinical Trials: What You Should Know before Volunteering to Be a Research Subject*. Demos Medical Publishing, New York.

Goodkin, D. E., and Rudick, R. A. [1998]. *Multiple Sclerosis: Advances in Clinical Trial Design, Treatment and Perspectives*. Springer, New York.

Graboys, T. B., Lown, B., Podrid, P. J., and DeSilva, R. [1982]. Long-term survival of patients with malignant ventricular arrhythmia treated with antiarrhythmic drugs. *American Journal of Cardiology*, **50**: 437–443.

Green, S. B. [1982]. Patient heterogeneity and the need for randomized clinical trials. *Controlled Clinical Trials*, **3**: 189–198.

Green, S., Benedetti, J., and Crowley, J. [2002]. *Clinical Trials in Oncology*, 2nd ed. CRC Press, Boca Raton, FL.

Greenberg, B. G. [1951]. Why randomize? *Biometrics*, **7**: 309–322.

Guilogg, R. J. (ed.) [2001]. *Clinical Trials in Neurology*. Springer, New York.

Hennekens, C. H., and Zorab, R. [2000]. *Clinical Trials in Cardiovascular Disease: A Companion to Braunwald's Heart Disease*. W. B. Saunders, Philadelphia.

IMPACT Research Group [1984]. International Mexiletine and placebo antiarrhythmic coronary trial: I. Report on arrhythmias and other findings. *Journal of the American College of Cardiology*, **4**: 1148–1163.

Jennison, C., and Turnbull, B. W. [2000]. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, New York.

Kaptchuk, T. J. [1998]. Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine*, **72**: 389–433.

Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Zapikian, A. Z., Lewis, T. L., and Lynch, J. M. [1975]. Ascorbic acid for the common cold: a prophylactic and therapeutic trial. *Journal of the American Medical Association*, **231**: 1038–1042.

Kempthorne, O. [1977]. Why randomize? *Journal of Statistical Planning and Inference*, **1**: 1–25.

Kertes, P. J., and Conway, M. D. [1998]. *Clinical Trials in Opthamology: A Summary and Practice Guide*. Lippincott Williams & Wilkins, Philadelphia.

Kesteloot, H., and Joosens, J. V. [1980]. *Epidemiology of Arterial Blood Pressure: Developments in Cardiovascular Medicine*, Vol. 8. Martinus Nijhoff, Dordrecht, The Netherlands.

Lan, K. K. G., and DeMets, D. L. [1983]. Discrete sequential boundaries for clinical trials. *Biometrika*, **70**: 659–663.

Lifton, R. J. [1986]. *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. Basic Books, New York.

Little, R. J. A., and Rubin, D. B. [2002]. *Statistical Analysis of Missing Data*, 2nd ed. Wiley, New York.

Mathieu, M. [2002]. *New Drug Development: Regulation Overview*. Parexel International Corporation, Cambridge, MA.

Matthews, J. N. [2000]. *Introduction to Randomized Controlled Clinical Trials*. Edward Arnold, London.

McFadden, E. [1997]. *Management of Data in Clinical Trials*. Wiley, New York.

Meinert, C. L. [1986]. *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

Mulay, M. [2001]. *A Step-by-Step Guide to Clinical Trials*. Jones & Bartlett, Boston.

Norleans, M. X. [2001]. *Statistical Methods for Clinical Trials*. Marcel Dekker, New York.

O'Brien, P. C., and Fleming, T. R. [1979]. A multiple testing procedure for clinical trials. *Biometrics*, **35**: 549–556.

Peterson, A. V., Mann, S. L., Kealey, K. A., and Marek, P. M. [2000]. Experimental design and methods for school-based randomized trials: experience from the Hutchinson smoking prevention project (HSPP). *Controlled Clinical Trials*, **21**: 144–165.

Piantadosi, S. [1997]. *Clinical Trials: A Methodologic Perspective*. Wiley, New York.

Pitt, B., Julian, D., and Pocock, S. J. [1997]. *Clinical Trials in Cardiology*. W. B. Saunders, Philadelphia.

Pocock, S. J. [1982]. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, **38**: 153–162.

Pocock, S. J. [1996]. *Clinical Trials: A Practical Approach*. Wiley, New York.

Pope, A. [1733]. *An Essay on Man*. Cited in [1968] *Bartlett's Familiar Quotations*, 14th ed. Little, Brown, Boston.

Porter, R. J., and Schoenberg, B. S. [1990]. *Controlled Clinical Trials in Neurological Disease.* Kluwer Academic, New York.

Pratt, C. M. (ed.) [1990]. A symposium: the Cardiac Arrhythmia Suppression Trial—does it alter our concepts of and approaches to ventricular arrhythmias? *American Journal of Cardiology*, **65**: 1B–42B.

Pratt, C. M., Brater, D. C., Harrell, F. E., Jr., Kowey, P. R., Leier, C. V., Lowenthal, D. T., Messerlie, F., Packer, M., Pritchett, E. L. C., and Ruskin, J. N. [1990]. Clinical and regulatory implications of the Cardiac Arrhythmia Suppression Trial. *American Journal of Cardiology*, **65**: 103–105.

Prentice, R. L. [1989]. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, **8**: 431–440.

Redmond, C. K., Colton, T., and Stephenson, J. [2001]. *Biostatistics in Clinical Trials*. Wiley, New York.

Reiser, S. J., Dyck, A. J., and Curran, W. J. (eds.). [1947]. The Nuremberg Code. in *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*. MIT Press, Cambridge, MA, pp. 272–274.

Royal Statistical Society [1993]. *Code of Conduct*. RSS, London.

Ruskin, J. N. [1989]. The cardiac arrhythmia suppression trial (CAST) (editorial). *New England Journal of Medicine*, **321**: 386–388.

Schouten, H. J. A. [2000]. Combined evidence from multiple outcomes in a clinical trial. *Journal of Clinical Epidemiology*, **53**: 1137–1144.

Slevin, M., and Wood, S. [1996]. *Understanding Clinical Trials*. Cancer BACUP, London. *http://www.cancerbacup.org.uk/info/trials.htm.* Accessed June 10, 2003.

Spilker, B. [1991]. *Guide to Clinical Trials*. Lippincott Williams & Wilkins, Philadelphia.

Spilker, B. [1995]. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Lippincott Williams & Wilkins, Philadelphia.

Student [1931]. The Lanarkshire milk experiment. *Biometrika*, **23**: 398–406.

Temple, R. J. [1995]. A regulatory authority's opinion about surrogate endpoints. In *Clinical Measurement in Drug Evaluation,* W. S. Nimmo and G. T. Tucker, (eds.). Wiley, New York, pp. 3–22.

Temple, R., and Ellenberg, S. S. [2000]. Placebo-controlled trials and active-control trials in the evaluation of new treatments: 1. Ethical and scientific issues. *Annals of Internal Medicine*, **133**: 455–463.

Thomas, L. [1983]. *The Youngest Science*. pp. 30–31. New York, NY, The Viking Press.

Wall Street Journal [2001]. Cost of drug development found to rise, p. B14, Dec. 3, 2001, taken from the Tufts Center for Drug Development. Also given in Tufts Center for the Study of Drug Development, *Outlook 2002*, Boston.

Whitehead J. [1983]. *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood, Chichester, West Sussex, England.

Whitehead, A. [2002]. *Meta-analysis of Controlled Clinical Trials*. Wiley, New York.

World Medical Association [1975]. Declaration of Helsinki, revision of original 1964 version. In *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*, S. J., Reiser, A. J. Dyck, and W. J., Curran, (eds.). MIT Press, Cambridge, MA, pp. 328–330.

Wright, S. P. [1992]. Adjusted p-values for simultaneous inference. *Biometrics*, **48**: 1005–1013.

CHAPTER 20

Personal Postscript

20.1 INTRODUCTION

One reviewer of this book felt that it would be desirable to have a final chapter that ended the book with more interesting material than yet another statistical method. This stimulated us to think about all the exciting, satisfying, and interesting things that had occurred in our own careers as biostatisticians. We decided to try to convey some of these feelings through our own experiences. This chapter is unabashedly written from a first-person point of view. The examples do not represent a random sample of our experiences but rather, the most important and/or interesting experiences of our careers. There is some deliberate duplication of background material that appears in other chapters so that this chapter may be self-contained (except for the statistical methods used). We have not made an effort to choose experiences that illustrate the use of many different statistical methods (although this would have been possible). Rather, we want to entertain, and in doing so, show the important collaborative role of biostatistics in biomedical research.

20.2 IS THERE TOO MUCH CORONARY ARTERY SURGERY?

The National Institutes of Health in the United States funds much of the health research in the country. During the late 1960s and early 1970s, an exciting new technique for dealing with anginal chest pain caused by coronary artery disease was developed. Recall that coronary artery disease is caused by fibrous fatty deposits building up within the arteries that supply blood to the heart muscle (i.e., the coronary arteries). As the arteries narrow, the blood supply to the heart is inadequate when there are increased demands because of exercise and/or stress; the resulting pain is called *angina*. Further, the narrowed arteries tend to close with blood clots, which results in the death (infarction) of heart muscle (myocardium), whose oxygen and nutrients are supplied by the blood coming through the artery; these heart attacks are also called *myocardial infarctions* (Mls). *Coronary artery bypass graft* (CABG; pronounced "cabbage") surgery replumbs the system. Either saphenous veins from the leg or the internal mammary arteries already in the chest are used to supply blood beyond the narrowing, that is, bypassing the narrowing. Figure 20.1 shows the results of bypass surgery. A key measure of damaged arteries is the *ejection fraction* (EF), the proportion of blood pushed out of the pumping chamber of the heart, the left ventricle. A normal value is 0.5 or greater. EF values between 0.35 and 0.49 are considered evidence of mild to moderate impairment. When the heart muscle is damaged, say by an MI, or has a limited blood supply, the EF decreases.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

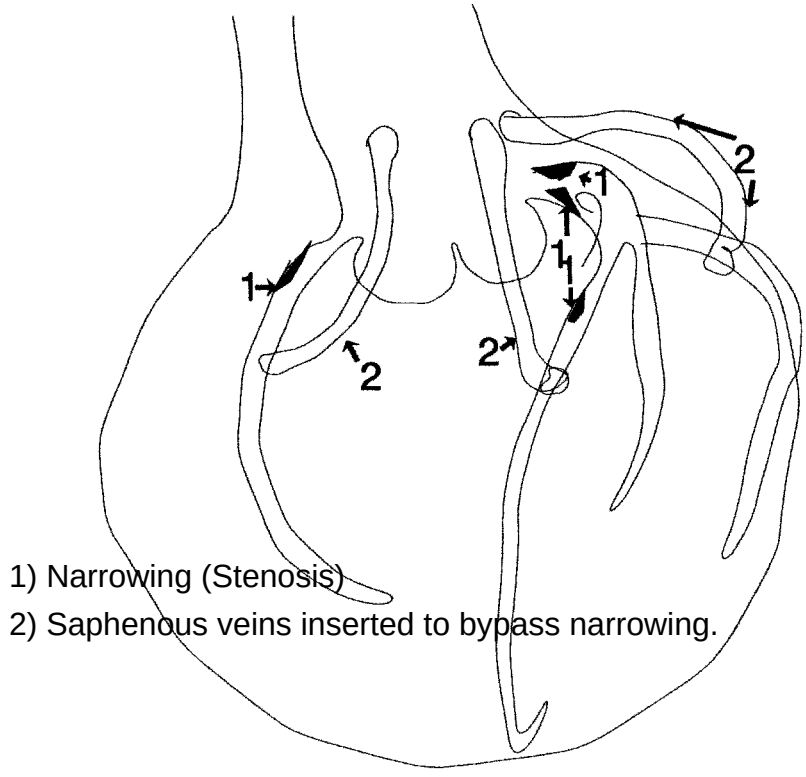


Figure 20.1 Schematic display of coronary artery bypass graft surgery. Here saphenous veins from the leg are sewn into the aorta where the blood is pumped out of the heart and then sewn into coronary arteries beyond narrowings in order to deliver a normal blood supply.

Because the restored blood flow should allow normal function, it was conjectured that surgery would both remove the anginal pain and also prolong life by reducing both the stress on the heart and the number of myocardial infarctions. It became clear early on that surgery did help to relieve angina pain (although even this has been debated; see Preston [1977]). However, the issue of prolonging life was more debatable. The amount of surgery had important implications for the health care budget, since in the early 1970s the cost per operation ranged between $12,000 and $50,000, depending on the location of the clinic, complexity of the surgery, and a variety of other factors. The number of surgeries by year up to 1972 is shown in Figure 20.2.

Because of the potential savings in lives and the large health resources requirements, the National Heart, Lung and Blood Institute (NHLBI; at that time the National Heart Institute) decided that it was appropriate to obtain firm information about which patients have improved survival with CABG surgery. Such therapeutic comparisons are best addressed through a randomized clinical trial, and that was the approach taken here with randomization to early surgery or early medical treatment. However, because not all patients could ethically be randomized, it was also decided to have a registry of patients studied with coronary angiography so that observational data analyses could be performed on other subsets of patients to compare medical and surgical therapy. When the NHLBI has internally sponsored initiatives, they are developed through a request for proposals (RFP), which recruits investigators to perform the collaborative research. This trial and registry, called the Coronary Artery Surgery Study (CASS), had two RFPs; one was for clinical sites and the other for a coordinating center. The RFP for the

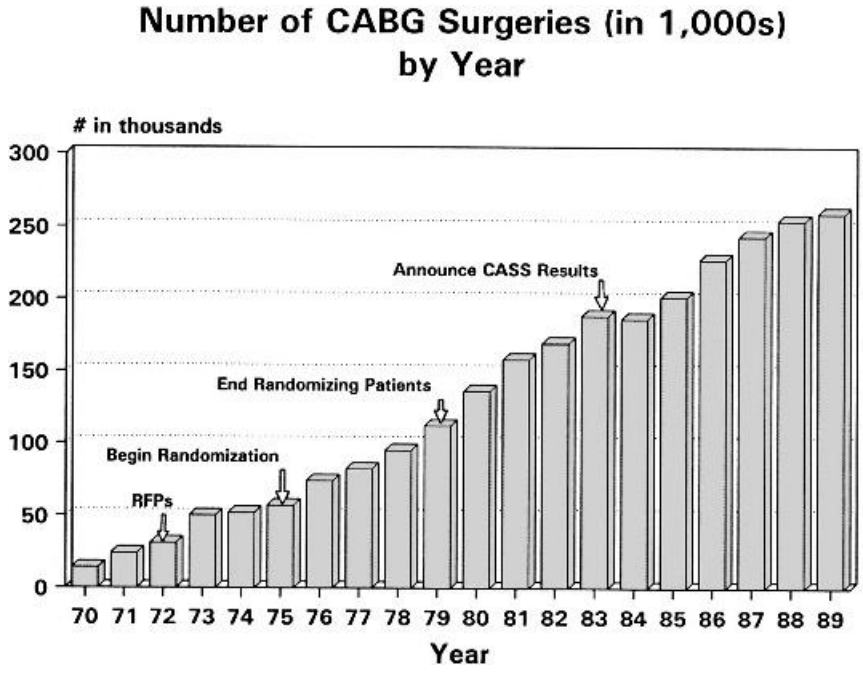


Figure 20.2 Number of coronary artery bypass graft surgeries in thousands of operations by year, 1970–1989. Marked are some of the key time points in the Coronary Artery Surgery Study. (Data courtesy of the cardiac diseases branch of the National Heart, Lung and Blood Institute from the National Hospital Discharge Survey, National Center for Health Services.)

clinical sites was issued in November 1972 and described the proposed study, both randomized and registry components, and asked for clinics to help complete the design and to enroll patients in the randomized and registry components of the study. The coordinating center RFP requested applications for a center to help with the statistical design and analysis of the study, to receive and process the study forms with a resultant database, to produce reports for monitoring the progress of the study and to otherwise participate in the quality assurance of the study, and finally, to collaborate in the analysis and publication of the randomized study and registry results. The organization of such a large multicenter study had a number of components: The NHLBI had a program office with medical, biostatistical, and financial expertise to oversee operation of the study; there were 15 cooperating clinical sites in the United States and Canada; the Coordinating Center was at the University of Washington under the joint direction of Lloyd Fisher and Richard Kronmal; a laboratory to read electrocardiograms (ECG lab) was established at the University of Alabama.

The randomized study enrolled 780 cases with mild angina or no angina with a prior MI, and significant disease (defined as a 70% or greater narrowing of the internal diameter of a coronary artery that was suitable for bypass surgery). There were a variety of other criteria for eligibility for randomization. The registry, including the patients randomized, enrolled 24,959 patients. Extensive data were collected on all patients. The first patients were enrolled in July 1974, with randomization beginning in August 1975 [CASS Principal Investigators and Their Associates, 1981]. Follow-up of patients within the randomized study ended in 1992. Needless to say, such a large effort cost a considerable amount of money, over $30,000,000. It will be shown that the investment was very cost-effective.

Results of the survival analysis and indicators of the quality of life were made public in 1983 [CASS Investigators, 1983a,b, 1984b]. The survival estimates for the subjects randomized to

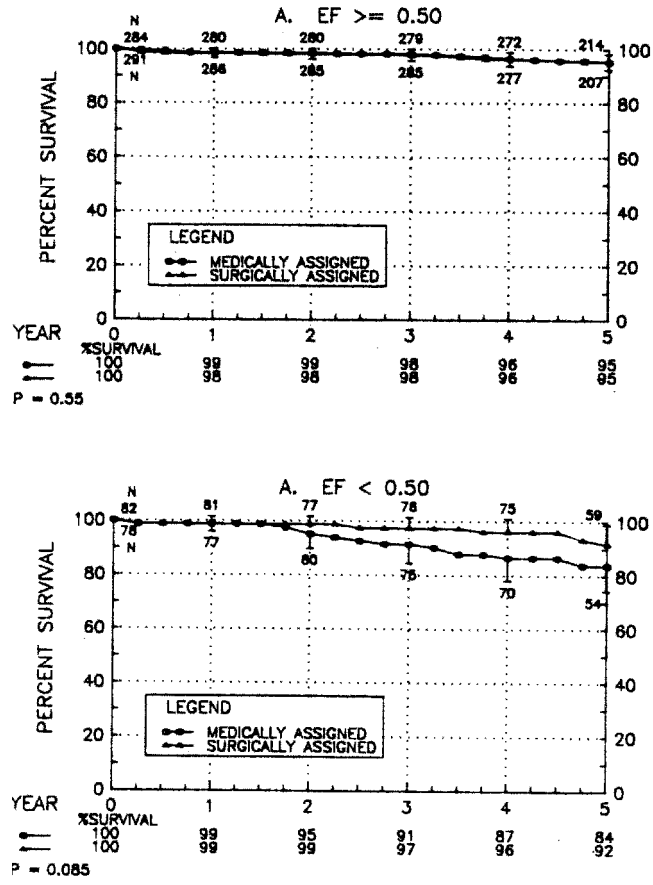


Figure 20.3 Data from the CASS randomized clinical trial; the bottom panel is for patients with ejection fractions less than 0.50; the top panel is for patients with ejection fractions of 0.50 or above. The p-values are the log-rank statistic for the comparison.

initial medical and surgical treatment are given in Figure 20.3. For patients with an EF of 0.5 or more, the survival curves were virtually identical; for subjects with lower EF values, there was a trend toward favorable mortality in the surgery group ($p = 0.085$ by the log-rank test).

A number of points were important in interpreting these data:

1. The CASS investigators agreed before the study started that the surgery was efficacious in relieving angina. Thus, if a patient started to have severe angina that could not be controlled by medication, the patient was allowed to "cross over" to surgery. By year 5, 24% of the patients assigned to initial medical therapy had crossed over to the CABG surgery group. If surgery is, in fact, having a beneficial effect and there is much crossover, the statistical power of the comparison is reduced. Is this a bad thing? The issue is a complex one (see Peto et al. [1977]; Weinstein and Levin [1989]; Fisher et al. [1989, 1990]). We know that one of the benefits of randomization is that we are assured of comparable groups (on average) even with respect to unrecorded and unknown variables. If we manipulated people, or parts of their experience, between groups by using events that occurred after the time of randomization, bias can enter the analysis. Thus, people should be included only in the group to which they are randomized; this is called an *intent-to-treat analysis* since they are counted with the group whose treatment

was intended. (Does such an approach avoid bias? Does it always make biological sense?) The CASS investigators favored an intent-to-treat analysis not only because it avoided possible bias but also because of the ethical imperative to perform CABG surgery for pain relief when the pain became intractable under medical treatment. Thus, including all the experience of those assigned to initial medical treatment, including CABG surgery and subsequent events, mirrored what would happen to such a group in real life. This is the question that the trial should answer: Is early surgery helpful when patients will receive it anyway when the pain becomes too severe? However, the power of such a comparison will be diminished by the crossovers. The interpretation of such intent-to-treat analyses must acknowledge that without the crossover, the results could have been substantially different.

2. Because bypass surgery is such a big industry (e.g., 200,000 surgeries per year at $30,000 per operation adds up to $6 billion per year), with many careers and much professional prestige committed to the field, one could expect a counter reaction if surgery did not look beneficial. Such reactions did occur, and a number of editorials, reviews, and sessions at professional meetings were given to consideration of the results. One of the authors (LF) appeared on the CBS national news as well as going to New York City to be interviewed by Mike Wallace and appearing on the TV program *60-minutes*. Based largely on the CASS results, the program suggested that there was too much CABG surgery.

3. It is important to keep the findings in context. They did not apply to all, or even most, patients. The CASS was one of three major randomized trials of CABG surgery. One study showed definitively that the surgery prolonged life in patients with left main disease [Takaro et al., 1976]. This study excluded patients with severe angina and thus had nothing to say about differential survival in such patients. In fact, there is observational data to suggest that early elective CABG surgery prolonged life in such patients [Kaiser et al., 1985; Myers et al., 1989].

4. Even though the findings may apply to a *relatively* small number of patients, the results could have a very substantial impact on the national health scene. Subsequent CASS papers showed that the trend toward increased survival with surgery in the low ejection fraction patients was real [Passamani et al., 1985; Alderman et al., 1990]. Thus, suppose that we restrict ourselves to those patients with EFs of at least 0.5. This accounted for 575 of the 780 randomized patients. Suppose that the randomized study had not been in effect; how many of these patients might have received early surgery? In the CASS study, there were 1315 patients who met the eligibility criteria and might have been randomized but in fact were not randomized [CASS Principal Investigators, 1984a; Chaitman et al., 1990]; these patients were called the *randomizable patients*. In this group, 43% (570/1315) received early elective surgery. Of those who did not receive early surgery and had good ejection fractions, by 10 years, 38% had received surgery. That is, 60% or so did not receive surgery. Assuming that the CASS clinics were representative of the surgical practice in the country (they may have been more conservative than many centers because they were willing to participate in research to assess the appropriate role of bypass surgery), about 4.4% of the surgery in the United States might be prevented by applying the results of the study. In a year with 188,000 CABGs costing $30,000 each, this would lead to a savings of over $245 million. Over a 4-year period over $1 billion could be saved in surgical costs. However, because the patients treated medically have more anginal pain, they have higher drug costs; they might have higher hospitalization costs (but they do not; see CASS Principal Investigators [1983b] and Rogers et al. [1990]). Without going into detail, it is my (L.F.) opinion that the study saved several billion dollars in health care costs without added risk to patient lives.

5. The issues are more complex than presented here; we have not discussed the findings and integration of results with the other major randomized studies of CABG surgery. Further, it is important to note that a number of other proven and/or promising techniques for dealing with coronary artery disease (CAD) have been developed. These include drug and/or dietary therapy; blowing up balloons in the artery to "squish" the narrowing into the walls of the artery [**p**ercutaneous **t**ransluminal **c**oronary **a**ngioplasty (PTCA)]; introducing lasers into the coronary

arteries to disintegrate the plaques that narrow the arteries; using a roto-rooter in the arteries to replumb by grinding up the plaques; and stents. Although all of these alternatives have been or are being used, the number of CABG surgeries did not decrease but leveled off up to 1989.

6. The surgery may improve with time as techniques and skills improve. Further, it became apparent that the results of the surgery deteriorated at 10 to 12 years or so. The disease process at work in the coronary arteries also was at work in the grafts that bypassed the narrowed areas; thus the grafts themselves narrow and close, often requiring repeat CABG surgery. Internal mammary grafts have a longer lifetime and are now used more often, suggesting that current long-term results will be better.

In summary, the CASS study showed that in patients with selected characteristics, CABG surgery is not needed immediately to prolong life and can often be avoided. The study was a bargain both in human and economic terms, illustrating the need and benefits of careful evaluation of important health care procedures.

20.3 SCIENCE, REGULATION, AND THE STOCK MARKET

In the United States, foods, drugs, biologics, devices, and cosmetics are regulated by the Food and Drug Administration (FDA). To get a new drug or biologic approved for marketing within the United States, the sponsor (usually, a pharmaceutical company or biotechnology company) must perform adequate and well-controlled clinical trials that show the efficacy and safety of the product. The FDA is staffed with personnel who have expertise in a number of areas, including pharmacology, medicine, and biostatistics. The FDA staff reviews materials submitted and rules on the approval or nonapproval of a product. The FDA also regulates marketing of the compounds. Marketing before approval is not allowed. The FDA uses the services of a number of advisory committees composed of experts in the areas considered. The deliberations of the advisory committees are carried out in public, often with large audiences in attendance. At the meetings, the sponsor makes a presentation, usually with both company and clinical experts, and answers questions from the committee. The FDA has a presence, asks questions, particularly of the advisory committee, but usually does not play a dominant role. At the end of its deliberations the committee votes on whether the drug or biologic should be approved, should be disapproved, or should be disapproved at least temporarily because further information is needed before final approval or disapproval is appropriate.

Two of the authors have been members of FDA advisory committees, G.vB. with the peripheral and central nervous system drugs advisory committee and L.F. with the cardiovascular and renal drugs advisory committee. Here we discuss the consideration of one biologic: tissue plasminogen activator (tPA). A *biologic* is a compound that occurs naturally in the human body, whereas a *drug* is a compound that does not occur naturally but is introduced artificially, solely for therapeutic purposes. For example, insulin is a biologic, whereas aspirin is a drug. Here we will use the term *drug* for tPA because that is the more common usage, although within the FDA, drugs and biologics go to different divisions. We turn next to the background and rationale for the use of tPA.

As discussed above, when coronary artery disease occurs, it narrows the arteries, changing the fluid flow properties of the blood, leading to clotting within the coronary arteries. These clots then block the blood supply to the heart muscle, resulting in heart attacks, or myocardial infarctions (MIs), as discussed above. The clot is composed largely of fibrin. When converted to plasmin, plasminogen converts insoluble fibrin into soluble fragments. One conceptual way to treat a heart attack would be to dissolve the blood clot, thus reestablishing blood flow to the heart muscle and preventing the death of the muscle, saving the heart and often saving the life. Should a drug be approved for dissolving blood clots alone? Although biologically plausible, does this assure that the drug will work? In other words, is this an acceptable surrogate endpoint?

Returning to the thrombolytic (i.e., to *lyse*, or break up, the blood clot, the thrombosis) tPA therapy, it is clear that lysing the coronary arterial blood clot is a surrogate endpoint. Should this surrogate endpoint be appropriate for approving the drug? After all, there is such a clearcut biological rationale: Coronary artery clots cause heart attacks; heart attacks damage the heart, often either impairing the heart function, and thus lowering exercise capacity, or killing the person directly. But experience has shown that very convincing biological scenarios do not always deliver the benefits expected; below we present an important example of a situation where an obvious surrogate endpoint did not work out.

Let us return now to the tPA cardiorenal advisory committee meeting and decision. In addition to tPA, another older thrombolytic drug, streptokinase, was also being presented for approval for the same indication. Prior to the meeting, there was considerable publicity over the upcoming meeting and possible approval of the drug tPA. The advisory committee meeting was to take place on Friday, May 29, 1987. On Thursday, May 28, 1987, the *Wall Street Journal* published an editorial entitled "The TPA Decision." The editorial read as follows:

> Profile of a heart-attack victim: 49 years old, three children, middle-manager, in seemingly good health. Cutting the grass on a Saturday afternoon, he is suddenly driven to the ground with severe chest pain. An ambulance takes him to the nearest emergency room, where he receives drugs to reduce shock and pain.
>
> At this point, he is one of approximately 4000 people who suffer a heart attack each day. If he has indeed had a heart attack, he will experience one of two possible outcomes. Either he will be dead, joining the 500,000 Americans killed each year by heart attack. Or, if he's lucky, he will join the one million others who go on to receive some form of therapy for his heart disease.
>
> Chances of survival will depend in great part on the condition of the victim's heart, that is, how much permanent muscular damage the heart sustained during the time a clot prevented the normal flow of blood into the organ. Heart researchers have long understood that if these clots can be broken up early after a seizure's onset, the victim's chances of staying alive increase significantly. Dissolving the clot early enhances the potential benefits of such post-attack therapies as coronary bypass surgery or balloon angioplasty.
>
> Tomorrow morning, a panel of the Food and Drug Administration will review the data on a blood-clot dissolver called TPA, for tissue-type plasminogen activator. In our mind, TPA—not any of the pharmaceutical treatments for AIDS—is the most noteworthy, unavailable drug therapy in the United States. Put another way, the FDA's new rules permitting the distribution of experimental drugs for life-threatening diseases came under pressure to do something about the AIDS epidemic. But isn't it as important for the government to move with equal speed on the epidemic of heart attacks already upon us?
>
> This isn't to say that TPA is more important than AIDS treatments. Both have a common goal: keeping people alive. The difference is that while the first AIDS drug received final approval in about six months, TPA remains unapproved and unavailable to heart-attack victims despite the fact that the medical community has known for more than two years that it can save lives.
>
> How many lives? Obviously no precise projection is possible, but the death toll is staggering, with about 41,000 individuals killed monthly by heart attacks.
>
> In its April 4, 1985, issue, the *New England Journal of Medicine* carried the first report on the results of the National Institutes of Health's TIMI study comparing TPA's clot dissolving abilities with a drug already approved by the FDA. NIH prematurely ended that trial because TPA's results were so significantly better than the other drug.
>
> In an accompanying editorial, the *Journal*'s editor, Dr. Arnold Relman, said a safe and effective thrombolytic "might be of immense clinical value." In October 1985, a medical-policy committee of California's Blue Shield recommended that TPA be recognized "as acceptable medical practice." The following month at the American Heart Association's meeting, Dr. Eugene Braunwald, chairman of the department of medicine at Harvard Medical School, said, "If R-TPA were available on a wide basis, I would select that drug today." In its original TIMI report, the NIH said TPA would next be

> tested against a placebo; later, citing ethical reasons, the researchers dropped the placebo and now all heart patients in the TIMI trial are receiving TPA.
>
> It is for these reasons that we call TPA the most noteworthy unavailable drug in the U.S. The FDA may believe it is already moving faster than usual with the manufacturer's new-drug application. Nonetheless, bureaucratic progress [*sic*] must be measured against the real-world costs of keeping this substance out of the nation's emergency rooms. The personal, social and economic consequences of heart disease in this country are immense. The American Heart Association estimates the total costs of providing medical services for all cardiovascular disease at $71 billion annually.
>
> By now more than 4,000 patients have been treated with TPA in clinical trials. With well over a thousand Americans going to their deaths each day from heart attack, it is hard to see what additional data can justify the government's further delay in making a decision about this drug. If tomorrow's meeting of the FDA's cardio-renal advisory committee only results in more temporizing, some in Congress or at the White House should get on the phone and demand that the American public be given a reason for this delay.

The publicity before the meeting of the advisory committee was quite unusual since companies are prohibited from preapproval advertising; thus the impetus presumably came from other sources.

The cardiorenal advisory committee members met and considered the two thrombolytic drugs, streptokinase and tPA. They voted to recommend approval of streptokinase but felt that further data were needed before tPA could be approved. The reactions to the decision were extreme, but probably predictable given the positions expressed prior to the meeting.

The *Wall Street Journal* responded with an editorial on Tuesday, June 2, 1987, entitled "Human Sacrifice." It follows in its entirety:

> Last Friday an advisory panel of the Food and Drug Administration decided to sacrifice thousands of American lives on an altar of pedantry.
>
> Under the klieg lights of a packed hearing room at the FDA, an advisory panel picked by the agency's Center for Drugs and Biologics declined to recommend approval of TPA, a drug that dissolves blood clots after heart attacks. In a 1985 multicenter study conducted by the U.S. National Heart, Lung and Blood Institute, TPA was so conclusively effective at this that the trial was stopped. The decision to withhold it from patients should be properly viewed as throwing U.S. medical research into a major crisis.
>
> Heart disease dwarfs all other causes of death in the industrialized world, with some 500,000 Americans killed annually; by comparison, some 20,000 have died of AIDS. More than a thousand lives are being destroyed by heart attacks every day. In turning down treatment with TPA, the committee didn't dispute that TPA breaks up the blood clots impeding blood flow to the heart. But the committee asked that Genentech, which makes the genetically engineered drug, collect some more mortality data. Its submission didn't include enough statistics to prove to the panel that dissolving blood clots actually helps people with heart attacks.
>
> Yet on Friday, the panel also approved a new procedure for streptokinase, the less effective clot dissolver—or thrombolytic agent—currently in use. Streptokinase previously had been approved for use in an expensive, specialized procedure called intracoronary infusion. An Italian study, involving 11,712 randomized heart patients at 176 coronary-care units in 1984–1985, concluded that administering streptokinase intravenously reduced deaths by 18%. So the advisory panel decided to approve intravenous streptokinase, but not approve the superior thrombolytic TPA. This is absurd.
>
> Indeed, the panel's suggestion that it is necessary to establish the efficacy of thrombolysis stunned specialists in heart disease. Asked about the committee's justification for its decision, Dr. Eugene Braunwald, chairman of Harvard Medical School's department of medicine, told us: "The real question is, do you accept the proposition that the proximate cause of a heart attack is a blood clot in the coronary artery? The evidence is overwhelming, *overwhelming*. It is sound, basic medical knowledge. It is in every textbook of medicine. It has been firmly established in the past decade beyond any reasonable question. If you accept the fact that a drug [TPA] is twice as effective as

> streptokinase in opening closed vessels, and has a good safety profile, then I find it baffling how that drug was not recommended for approval."
>
> Patients will die who would otherwise live longer. Medical research has allowed statistics to become the supreme judge of its inventions. The FDA, in particular its bureau of drugs under Robert Temple, has driven that system to its absurd extreme. The system now serves itself first and people later. Data supersede the dying.
>
> The advisory panel's suggestion that TPA's sponsor conduct further mortality studies poses grave ethical questions. On the basis of what medicine already knows about TPA, what U.S. doctor will give a randomized placebo or even streptokinase? We'll put it bluntly: Are American doctors going to let people die to satisfy the bureau of drugs' chi-square studies?
>
> Friday's TPA decision should finally alert policy makers in Washington and the medical-research community that the theories and practices now controlling drug approval in this country are significantly flawed and need to be rethought. Something has gone grievously wrong in the FDA bureaucracy. As an interim measure FDA Commissioner Frank Young, with Genentech's assent, could approve TPA under the agency's new experimental drug rules. Better still, Dr. Young should take the matter in hand, repudiate the panel's finding and force an immediate reconsideration. Moreover, it is about time Dr. Young received the clear, public support of Health and Human Services Secretary Dr. Otis Bowen in his efforts to fix the FDA.
>
> If on the other hand Drs. Young and Bowen insist that the actions of bureaucrats are beyond challenge, then perhaps each of them should volunteer to personally administer the first randomized mortality trials of heart-attack victims receiving the TPA clot buster or nothing. Alternatively, coronary-care units receiving heart-attack victims might use a telephone hotline to ask Dr. Temple to randomize the trial himself by flipping a coin for each patient. The gods of pedantry are demanding more sacrifice.

Soon after joining the Cardiovascular and Renal Drugs Advisory Committee, L.F. noticed that a number of people left the room at what seemed inappropriate times, near the end of some advisory deliberations. I was informed that often, stock analysts with expertise in the pharmaceutical industry attended meetings about key drugs; when the analysts thought they knew how the vote was going to turn out, they went out to the phones to send instructions. That was the case during the tPA deliberations (and made it particularly appropriate that the *Wall Street Journal* take an interest in the result). Again we convey the effect of the deliberations through quotations taken from the press. On June 1, 1978, the *Wall Street Journal* had an article under the heading "FDA Panel Rejection of Anti-Clot Drug Set Genentech Back Months, Perils Stock." The article said in part:

> A Food and Drug Administration advisory panel rejected licensing the medication TPA, spoiling the summer debut of what was touted as biotechnology's first billion-dollar drug. . . . Genentech's stock—which reached a high in March of $64.50 following a 2-for-1 split—closed Friday at $48.25, off $2.75, in national over-the-counter trading, even before the close of the FDA panel hearing attended by more than 400 watchful analysts, scientists and competitors. Some analysts expect the shares to drop today. . . . Wall Street bulls will also be rethinking their forecasts. For example, Kidder Peabody & Co.'s Peter Drake, confident of TPA's approval, last week predicted sales of $51 million in the second half of 1987, rising steeply to $205 million in 1988, $490 million in 1989 and $850 million in 1990.

USA Today, on Tuesday, June 2, 1987, on the first page of the Money section, had an article headed "Biotechs Hit a Roadblock, Investors Sell." The article began:

> Biotechnology stocks, buoyed more by promise than products, took one of their worst beatings Monday. Leading the bad-news pack: Biotech giant Genentech Inc., dealt a blow when its first blockbuster drug failed to get federal approval Friday. Its stock plummeted $11\frac{1}{2}$ points to $\$36\frac{3}{4}$, on 14.2 million shares traded—a one-day record for Genentech. "This is very serious, dramatically serious," said analyst Peter Drake, of Kidder, Peabody & Co., who Monday changed his recommendations for the

group from buy to "unattractive." His reasoning: The stocks are driven by "a blend of psychology and product possibilities. And right now, the psychology is terrible."

Biotechnology stocks as a group dropped with the Genentech panel vote. This seemed strange to me because the panel had not indicated that the drug, tPA, was bad but only that in a number of areas the data needed to be gathered and analyzed more appropriately (as described below). The panel was certainly not down on thrombolysis (as the streptokinase approval showed); it felt that the risk/benefit ratio of tPA needed to be clarified before approval could be made.

The advisory committee members replied to the *Wall Street Journal* editorials both individually and in groups, explaining the reasons for the decision [Borer, 1987; Kowey et al., 1988; Fisher et al., 1987]. This last response to the *Wall Street Journal* was submitted with the title "The Prolongation of Human Life"; however, after the review of the article by the editor, the title was changed by the *Wall Street Journal* to "The FDA Cardio-Renal Committee Replies." The reply:

The evaluation and licensing of new drugs is a topic of legitimate concern to not only the medical profession but our entire populace. Thus it is appropriate when the media, such as the *Wall Street Journal*, take an interest in these matters. The Food and Drug Administration recognizes the public interest by holding open meetings of advisory committees that review material presented by pharmaceutical companies, listen to expert opinions, listen to public comment from the floor and then give advice to the FDA. The Cardiovascular and Renal Drugs Advisory Committee met on May 29 to consider two drugs to dissolve blood clots causing heart attacks. The *Journal* published editorials prior to the meeting ("The TPA Decision," May 28) and after the meeting ("Human Sacrifice," June 2 and "The Flat Earth Committee," July 13). The second editorial began with the sentence: "Last Friday an advisory committee of the Food and Drug Administration decided to sacrifice thousands of American lives on an altar of pedantry." How can such decisions occur in our time? This reply by members of the advisory panel presents another side to the story. In part the reply is technical, although we have tried to simplify it. We first discuss drug evaluation in general and then turn to the specific issues involved in the evaluation of the thrombolytic drugs streptokinase and TPA.

The history of medicine has numerous instances of well-meaning physicians giving drugs and treatments that were harmful rather than beneficial. For example, the drug thalidomide was widely marketed in many countries—and in West Germany without a prescription—in the late 1950s and early 1960s. The drug was considered a safe and effective sleeping pill and tranquilizer. Marketing was delayed in the U.S. despite considerable pressure from the manufacturer upon the FDA. The drug was subsequently shown to cause birth defects and thousands of babies world-wide were born with grotesque malformations, including seal-like appendages and lack of limbs. The FDA physician who did not approve the drug in the U.S. received an award from President Kennedy. One can hardly argue with the benefit of careful evaluation in this case. We present this, not as a parallel to TPA, but to point out that there are two sides to the approval coin—early approval of a good drug, with minimal supporting data, looks wise in retrospect; early approval, with minimal supporting data, of a poor drug appears extremely unwise in retrospect. Without adequate and well-controlled data one cannot distinguish between the two cases. Even with the best available data, drugs are sometimes found to have adverse effects that were not anticipated. Acceptance of unusually modest amounts of data, based on assumptions and expectations rather than actual observation is very risky. As will be explained below, the committee concluded there were major gaps in the data available to evaluate TPA.

The second editorial states that "Medical research has allowed statistics to become the supreme judge of its inventions." If this means that data are required, we agree; people evaluate new therapies with the hope that they are effective—again, before licensing, proof of effectiveness and efficacy is needed. If the editorial meant that the TPA decision turned on some arcane mathematical issue, it is incorrect. Review of the transcript shows that statistical issues played no substantial role.

We now turn to the drug of discussion, TPA. Heart attacks are usually caused by a "blood clot in an artery supplying the heart muscle with blood." The editorial quotes Dr. Eugene Braunwald, "The real question is, do you accept the proposition that the proximate cause of a heart attack is a blood clot in the coronary artery?" We accept the statement, but there is still a significant question: "What can one then do to benefit the victim?" It is not obvious that modifying the cause after the event

occurs is in the patient's best interest, especially when the intervention has toxicity of its own. Blood clots cause pulmonary embolism; it is the unusual patient who requires dissolution of the clot by streptokinase. Several trials show the benefit does not outweigh the risk.

On May 29 the Cardiovascular and Renal Drugs Advisory Committee reviewed two drugs that "dissolve" blood clots. The drug streptokinase had been tested in a randomized clinical trial in Italy involving 11,806 patients. The death rate in those treated with streptokinase was 18% lower than in patients not given streptokinase; patients treated within six hours did even better. Review of 10 smaller studies, and early results of a large international study, also showed improved survival. It is important to know that the 18% reduction in death rate is a reduction of a few percent of the patients studied. The second drug considered—recombinant tissue plasminogen activator (TPA)—which also was clearly shown to dissolve blood clots, was not approved. Why? At least five issues contributed, to a greater or lesser amount, to the vote not to recommend approval for TPA at this time. These issues were: the safety of the drug, the completeness and adequacy of the data presented, the dose to be used, and the mechanism of action by which streptokinase (and hopefully TPA) saves lives.

Safety was the first and most important issue concerning TPA. Two formulations of TPA were studied at various doses; the highest dose was 150 milligrams. At this dose there was an unacceptable incidence of cerebral hemorrhage (that is, bleeding in the brain), in many case leading to both severe stroke and death. The incidence may be as high as 4% or as low as 1.5% to 2% (incomplete data at the meeting made it difficult to be sure of the exact figure), but in either case it is disturbingly high; this death rate due to side effects is of the same magnitude as the lives saved by streptokinase. This finding led the National Heart, Lung and Blood Institute to stop the 150-milligram treatment in a clinical trial. It is important to realize that this finding was unexpected, as TPA was thought to be relatively unlikely to cause such bleeding. Because of bleeding, the dose of TPA recommended by Genentech was reduced to 100 milligrams. The safety profile at doses of 100 milligrams looks better, but there were questions of exactly how many patients had been treated and evaluated fully. Relatively few patients getting this dose had been reported in full. Without complete reports from the studies there could be smaller strokes not reported and uncertainty as to how patients were examined. The committee felt a substantially larger database was needed to show safety.

The TPA used to evaluate the drug was manufactured by two processes. Early studies used the double-stranded (roller bottle) form of the drug; the sponsor then changed to a predominantly single-stranded form (suspension culture method) for marketing and production reasons. The second drug differed from the first in how long the drug remained in the blood, in peak effect, in the effect on fibrinogen and in the dose needed to cause lysis of clots. Much of the data was from the early form; these data were not considered very helpful with respect to the safety of the recommended dose of the suspension method drug. This could perhaps be debated, but the intracranial bleeding makes the issue an important one. The excessive bleeding may well prove to be a simple matter of excessive dose, but this is not yet known unequivocally.

Data were incomplete in that many of the patients' data had not been submitted yet and much of the data came from treatment with TPA made by the early method of manufacture. There was uncertainty about the data used to choose the 100-milligram dose, i.e., perhaps a lower dose is adequate. When there is a serious dose-related side effect it is crucial that the dose needed for effectiveness has been well-defined and has acceptable toxicity.

Let us turn to the mechanism of action, the means by which the beneficial effect occurs. There may be a number of mechanisms. The most compelling is clot lysis (dissolution). However, experts presented data that streptokinase changes the viscosity of the blood that could improve the blood flow; the importance is uncertain. Streptokinase also lowers blood pressure, which may decrease tissue damage during a heart attack. While there is convincing evidence that TPA (at least by the first method of manufacture) dissolves clots faster than streptokinase (at least after a few hours from the onset of the heart attack), we do not have adequate knowledge to know what portion of the benefit of streptokinase comes from dissolving the clot. TPA, thus, may differ in its effect on the heart or on survival. The drugs could differ in other respects, such as how often after opening a vessel they allow reclosure, and, of course, the frequency of important adverse effects.

These issues delay possible approval. Fortunately, more data are being collected. It is our sincere hope that the drug lives up to its promise, but should the drug prove as valuable as hoped, that would

not imply the decision was wrong. The decision must be evaluated as part of the overall process of drug approval.

The second editorial suggests that if the drug is not approved, Dr. Temple (director of the Bureau of Drugs, FDA), Dr. Young (FDA commissioner) and Dr. Bowen (secretary of health and human services) should administer "randomized mortality trials of heart-attack victims receiving the TPA clot buster or nothing." This indignant rhetoric seems inappropriate on several counts. First, the advisory committee has no FDA members; our votes are independent and in the past, on occasion, we have voted against the FDA's position. It is particularly inappropriate to criticize Drs. Temple and Young for the action of an independent group. The decision (by a vote of eight against approval, one for and two abstaining) was made by an independent panel of experts in cardiovascular medicine and research from excellent institutions. These unbiased experts reviewed the data presented and arrived at this decision; the FDA deserves no credit or blame. Second, we recommend approval of streptokinase; we are convinced that the drug saves lives of heart-attack victims (at least in the short term). To us it would be questionable to participate in a trial without some treatment in patients of the type shown to benefit from streptokinase. A better approach is to use streptokinase as an active control drug in a randomized trial. If it is as efficacious or better than streptokinase, we will rejoice. We have spent our adult lives in the care of patients and/or research to develop better methods for treatment. Both for our patients and our friends, our families and ourselves, we want proven beneficial drugs available.

In summary, with all good therapeutic modalities the benefits must surely outweigh the risks of treatment. In interpreting the data presented by Genentech in May 1987 the majority of the Cardiovascular and Renal Drugs Advisory Committee members could not confidently identify significant benefits without concomitant significant risk. The review was clouded by issues of safety, manufacturing process, dose size and the mechanism of action. We are hopeful these issues will be addressed quickly, allowing more accurate assessment of TPA's risk-benefit ratio with conclusive evidence that treatment can be recommended that allows us to uphold the physician's credo, *primum non nocere* (first do no harm).

The July 28 1987, *USA Today's* Life section carried an article on the first page entitled "FDA Speeds Approval of Heart Drug." The article mentioned that the FDA commissioner Frank Young was involved in the data gathering. Within a few months of the advisory committee meeting, tPA was approved for use in treating myocardial infarctions. The drug was 5 to 10 times more expensive than streptokinase; however, it opened arteries faster and that was thought to be a potential advantage. A large randomized comparison of streptokinase and tPA was performed (ISIS 3); the preliminary results were presented at the November 1990 American Heart Association meeting. The conclusion was that the efficacy of the two drugs was essentially equivalent. Thus by approving streptokinase, even in retrospect, no period of the lack of availability of a clearly superior drug occurred because of the time delay needed to clear up the questions about tPA. This experience shows that biostatistical collaboration has consequences above and beyond the scientific and humanitarian aspects; large political and financial issues also are often involved.

20.4 OH, MY ACHING BACK!

One of the most common maladies in the industrialized world is the occurrence of low-back problems. By the age of 50, nearly 85% of humans can recall back symptoms; and as someone has said, the other 15% probably forgot. Among persons in the United States, back and spine impairment are the chronic conditions that most frequently cause activity limitation. The occurrence of industrial back disability is one of the most expensive health problems afflicting industry and its employees. The cost associated with back injury in 1976 was $14 billion; the costs are greatly skewed, with a relatively low percent of the cost accrued by a few chronic back injury cases [Spengler et al., 1986]. The costs and human price associated with industrial back injury prompted the Boeing Company to contact the orthopedics department at the University of Washington to institute a collaborative study of back injury at a Boeing factory in western Washington

State. Collaboration was obtained from the Boeing company management, the workers and their unions, and a research group at the University of Washington (including one of the authors, L.F.). The study was supported financially by the National Institutes of Health, the National Institute for Occupational Safety and Health, the Volvo Foundation, and the Boeing Company. The study was designed in two phases. The first phase was a retrospective analysis of past back injury reports and insurance costs from already existing Boeing records; the second phase was a prospective study looking at a variety of possible predictors (to be described below) of industrial back injury.

The retrospective Boeing data were analyzed and presented in a series of three papers [Spengler et al., 1986; Bigos et al., 1986a,b]. The analysis covered 31,200 employees who reported 900 back injuries among 4645 claims filed by 3958 different employees. The data emphasized the cost to Boeing of this malady, and as in previous studies, showed that a small percentage of the back injury reports lead to most of the cost; for example, 10% of the cases accounted for 79% of the cost. The incurred costs of back injury claims was 41% of the Boeing total, although only 19% of the claims were for the back. The most expensive 10% of the back injury claims accounted for 32% of all the Boeing injury claims. Workers were more likely to have reported an acute back injury if they had a poor employee appraisal rating from their supervisor within 6 months prior to the injury.

The prospective study was unique and had some very interesting findings (the investigators were awarded the highest award of the American Academy of Orthopedic Surgeons, the Kappa Delta award, for excellence in orthopedic research). Based on previously published results and investigator conjectures, data were collected in a number of areas with potential ability to predict reports of industrial back injury. Among the information obtained prospectively from the 3020 aircraft employees who volunteered to participate in the study were the following:

- *Demographics:* race, age, gender, total education, marital status, number in family, method, and time spent in commuting to work.
- *Medical history:* questions about treatment for back pain by physicians and by chiropractors; hospitalization for back pain; surgery for back injury; smoking status.
- *Physical examination:* flexibility; spinal canal size by ultrasonography; and anthropometric measures such as height and weight.
- *Physical capacities:* arm strength; leg strength; and aerobic capacity measured by a submaximal treadmill test.
- *Psychological testing:* the MMPI (Minnesota Multiphasic Inventory and its subscales); a schedule of recent life change events; a family questionnaire about interactions at home; a health locus of control questionnaire.
- *Job satisfaction:* subjects were asked a number of questions about their job: did they enjoy their job almost always, some of the time, hardly ever; do they get along well with their supervisor; do they get along well with their fellow employees, etc.

The details of the design and many of the study results may be found in Battie et al. [1989, 1990a,b] and Bigos et al. [1991, 1992a,b]. The extensive psychological questionnaires were given to the employees to be taken home and filled out; 54% of the 3020 employees returned completed questionnaires, and some data analyses were necessarily restricted to those who completed the questionnaire(s). Figure 20.4 summarizes graphically some of the important predictive results.

The results of several stepwise, step-up multivariate Cox models are presented in Table 20.1. There are some substantial risk gradients among the employees. However, the predictive power is not such that one can conclusively identify employees likely to report an acute industrial back injury report. Of more importance, given the traditional approaches to this field, which have been largely biomechanical, work perception and psychological variables are important predictors, and the problem cannot be addressed effectively with only one factor in mind. This is emphasized in Figure 20.5, which represents the amount of information (in a formal sense)

in each of the categories of variables as given above. The figure is a Venn diagram of the estimated amount of predictive information for variables in each of the data collection areas [Fisher and Zeh, 1991]. The job perception and psychological areas are about as important as the medical history and physical examination areas. To truly understand industrial back injury, a multifactorial approach must be used.

Among the more interesting aspects of the study is speculation on the meaning and implications of the findings. Since, as mentioned above, most people experience back problems at

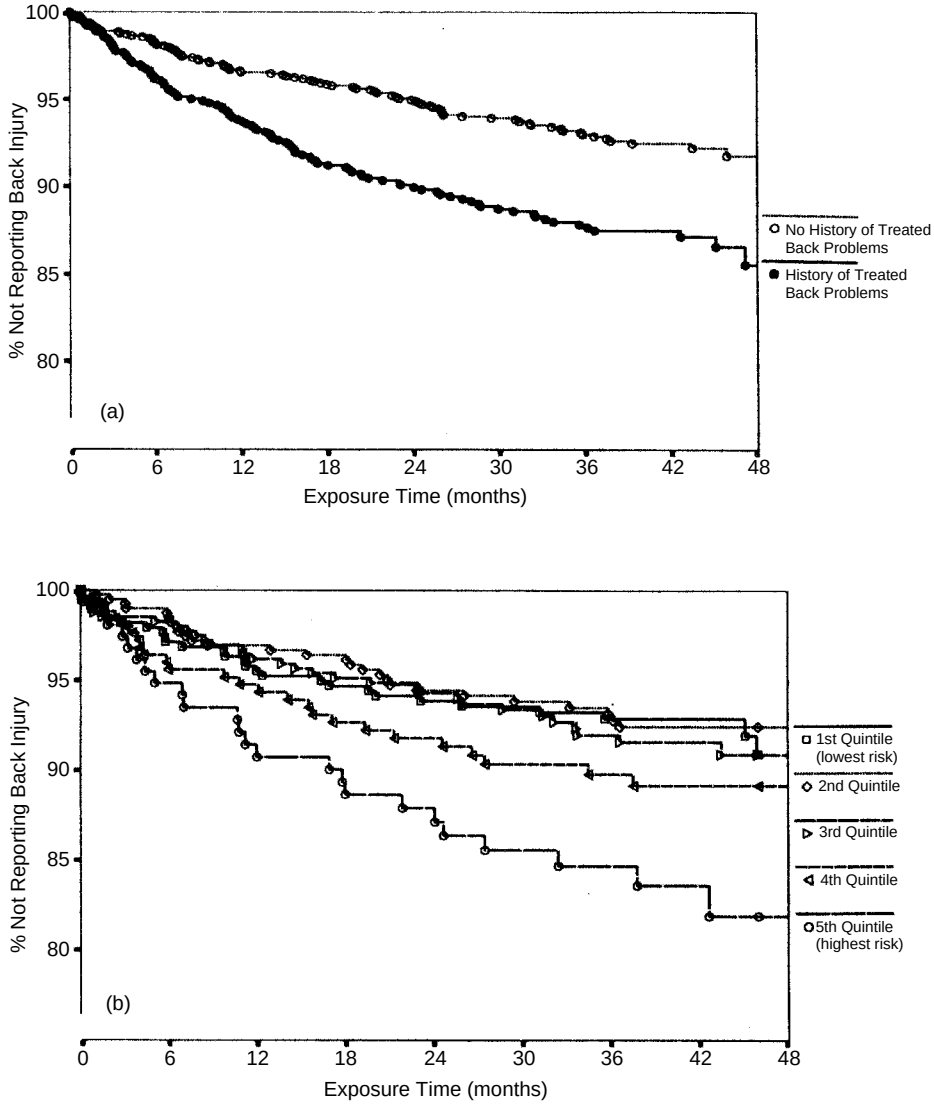


Figure 20.4 Panel (*a*) shows the product limit curves for the time to a subsequent back injury report for those reporting previous back problems and those who did not report such problems. Panel (*b*) divides the MMPI scale 3 (hysteria) values by cut points taken from the quintiles of those actually reporting events. Panel (*c*) divides the subjects by their response to the question: "Do you enjoy your job (1) almost always; (2) some of the time; or (3) hardly ever?" Panel (*d*) gives the results of the multivariate Cox model of Table 20.1; the predictive equation uses the variables from the first three panels. (From Bigos et al. [1991].)

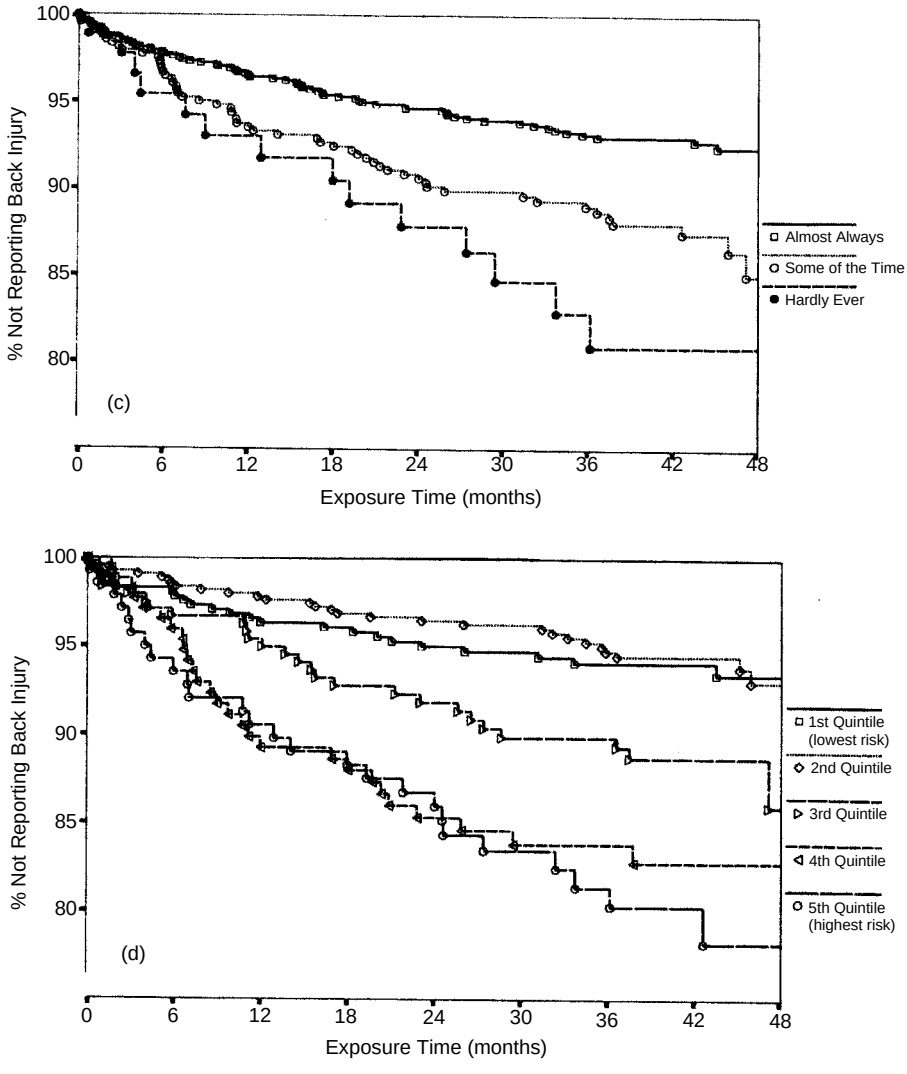


Figure 20.4 *(continued)*

some time in their lives, could legitimate back discomfort be used as an escape if one does not enjoy his or her job? Can the problem be reduced by taking measures to make workers more satisfied with their employment, or do a number of people tend to be unhappy no matter what? Is the problem a mixture of these? The results invite systematic, randomized intervention studies. Because of the magnitude of the problem, such approaches may be effective in both human and financial terms; however, this remains for the future.

20.5 SYNTHESIZING INFORMATION ABOUT MANY COMPETING TREATMENTS

Randomized controlled trials, discussed in Chapter 19, are the gold standard for deciding if a drug is effective and are required before new drugs are marketed. These trials may compare a

Table 20.1 Predicting Acute Back Injury Reports[a]

Variable	Univariate Analysis *p*-Value	Multivariate Analysis *p*-Value	Relative Risk	(95% Confidence Interval)
		Entire Population (n = 1326, injury = 117)		
Enjoy job[b]	0.0001	0.0001	1.70	(1.31, 2.21)
MMPI 3[c]	0.0003	0.0032	1.37	(1.11, 1.68)
Prior back pain[d]	0.0010	0.0050	1.70	(1.17, 2.46)
		Those with a History of Prior Back Injury (n = 518, injury = 63)		
Enjoy job[b]	0.0003	0.0006	1.85	(1.30, 2.62)
MMPI 3[c]	0.0195	0.0286	1.34	(1.17, 1.54)
		Those without a History of Prior Back Pain (n = 808, injury = 54)		
Enjoy jobs[b]	0.0220	0.0353	1.53	(1.09, 2.29)
MMPI 3[c]	0.0334	0.0475	1.41	(1.19, 1.68)

[a]Using the Cox proportional hazards regression model.
[b]Only subjects with complete information on the enjoy job question, MMPI, and history of back pain were included in these analyses.
[c]For an increase of one unit.
[d]For an increase of 10 units.

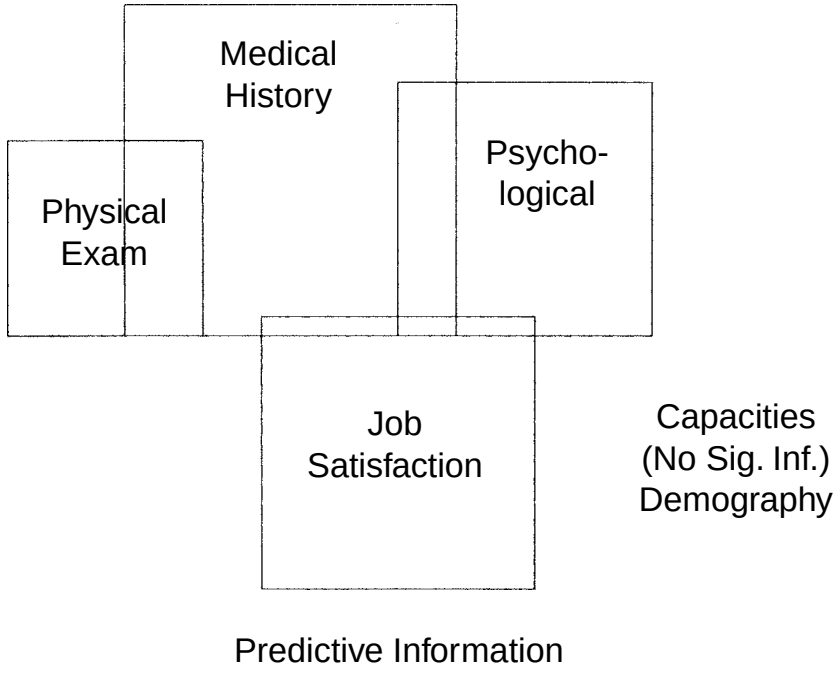


Figure 20.5 Predictive information by type of variable collected. Note that the job satisfaction and psychological areas contribute the same order of magnitude as the more classical medical history and physical examination variables. The relative lack of overlap in predictive information means that at least these areas must be considered if the problem is to be fully characterized. Capacities and demography variables added no information and so have no boxes.

new treatment to a placebo or to an accepted treatment. When many different treatments are available, however, it is not enough to know that they are all better than nothing, and it is often not feasible to compare all possible pairs of treatments in large randomized trials.

Clinicians would find it helpful to be able to use information from "indirect" comparisons. For example, if drug A reduces mortality by 20% compared to placebo, and drug B reduces mortality by 10% compared to drug A, it would be useful to conclude that B was better than placebo. However, indirect comparisons may not be reliable. The International Conference on Harmonisation, a project of European, Japanese, and U.S. regulators and industry experts, says in its document E10 on choice of control groups [2000, Sec. 2.1.7.4]

> "Placebo-controlled trials lacking an active control give little useful information about comparative effectiveness, information that is of interest and importance in many circumstances. Such information cannot reliably be obtained from cross-study comparisons, as the conditions of the studies may have been quite different."

The major concern with cross-study comparisons is that the populations being studied may be importantly different. People who participate in a trial of drug A when no other treatment is available may be very different from those who participate in a trial comparing drug A as an established treatment with a new experimental drug, B. For example, people for whom drug A is less effective may be more likely to participate in the hope of getting a better treatment. The ICH participants are certainly correct that cross-study comparisons *may* be misleading, but it would be very useful to know if they *are* actually misleading in a particular case.

An important example of this comes from the treatment of high blood pressure. There are many classes of drugs to treat high blood pressure, working in different ways on the heart, the blood vessels, and the kidneys. These include α-blockers, β-blockers, calcium channel blockers, angiotensin-converting enzyme (ACE) inhibitors, angiotensin receptor blockers, and diuretics. The availability of multiple treatments is useful because they have different side effects and because a single drug may not reduce blood pressure sufficiently. Some of the drug classes have the advantage of also treating other conditions that may be present in some people (β-blockers or calcium channel blockers for angina, α-blockers for the symptoms of prostatic hyperplasia). However, in many cases it is not obvious which drug class to try first.

Many clinical trials have been done, but these usually compare a single pair of treatments, and many important comparisons have not been done. For example, until late 2002, there had been only one trial in previously healthy people designed to measure clinical outcomes comparing ACE inhibitors with diuretics, although these drug classes are both useful in congestive heart failure and so seem a natural comparison. In a situation such as this, where there is reliable information from within-study comparisons of many, but not all, pairs of drugs, it should be possible to assess the reliability of cross-study comparisons and decide whether they can be used. That is, the possible cross-study comparisons of, say, ACE inhibitors and calcium channel blockers can be compared with each other and with any direct within-study comparisons. The better the agreement, the more confidence we will have in the cross-study comparisons. This technique is called *network metaanalysis* [Lumley, 2002]. The name comes from thinking of each randomized trial as a link connecting two treatments. A cross-study comparison is a path between two treatments composed of two or more links. If there are many possible paths joining two treatments, we can obtain an estimate along each path and see how well they agree.

The statistical model behind network metaanalysis is similar to the random-effects models discussed in Chapter 18. Write Y_{ijk} for a summary of the treatment difference in trial k of drugs i and j, for example, the logarithm of the estimated relative risk. If we could simply assume that trials were comparable, we could model this log relative risk by

$$Y_{ijk} = \beta_i - \beta_j + \epsilon_{ijk}$$

where β_i and β_j measure the effectiveness of drugs i and j, and ϵ_{ijk} represents the random sampling error.

When we say that trials of different sets of treatments are not comparable, we mean precisely that the average log relative risk when comparing drugs i and j is not simply given by $\beta_i - \beta_j$: there is some extra systematic difference. These differences can be modeled as random intercepts belonging to each pair of drugs:

$$Y_{ijk} = \beta_i - \beta_j + \xi_{ij} + \epsilon_{ijk}$$
$$\xi \sim N(0, \omega^2)$$

So, comparing two drugs i and j gives on average $\beta_i - \beta_j - \xi_{ij}$. If ξ_{ij} is large, the metaanalysis is useless, since the true differences between treatments $(\beta_i - \beta_j)$ are masked by the biases ξ_{ij}. The random effects standard deviation, ω, also called the *incoherence*, measures how large these biases are, averaged over all the trials. If the incoherence is large, the metaanalysis should not be done. If the incoherence is small, the metaanalysis may be worthwhile. Confidence intervals for $\beta_i - \beta_j$ will be longer because of the uncertainty in ξ_{ij}, slightly longer if the incoherence is very small, and substantially longer if the incoherence is moderately large.

Clearly, it would be better to have a single large trial that compared all the treatments, but this may not be feasible. There is no particular financial incentive for the pharmaceutical companies to conduct such a trial, and the cost would make even the National Institutes of Health think twice. In the case of antihypertensive treatments, a trial of many of the competing treatments was eventually done. This trial, ALLHAT [ALLHAT, 2002] compared a diuretic, a calcium channel blocker, an ACE inhibitor, and an α-blocker. It found that α-blockers were distinctly inferior (that portion of the trial was stopped early), and that diuretics were perhaps slightly superior to the other treatments.

Before the results of ALLHAT were available, Psaty et al. performed a network metaanalysis of the available randomized trials, giving much the same conclusions but also including comparisons with β-blockers, placebo, and angiotensin receptor blockers. This analysis, updated to include the results of ALLHAT, strengthens the conclusion that diuretics are probably slightly superior to the other options in preventing serious cardiovascular events [Psaty et al., 2003]. The cross-study comparisons showed good agreement except for the outcome of congestive heart failure, where there seemed to be substantial disagreement (perhaps due to different definitions over time). The network metaanalysis methodology incorporates this disagreement into confidence intervals, so the conclusions are weaker than they would otherwise be, but still valid.

The most important limitation of network metaanalysis is that it requires many paths and many links to assess the reliability of the cross-study comparisons. If each new antihypertensive drug had been compared only to placebo, there would be only a single path between any two treatments, and no cross-checking would be possible. Reliability of cross-study comparisons would then be an unsupported (and unsupportable) assumption.

20.6 SOMETHING IN THE AIR?

Fine particles in the air have long been known to be toxic in sufficiently high doses. Recently, there has been concern that even the relatively low exposures permitted by European and U.S. law may be dangerous to sensitive individuals. These fine particles come from smoke (wood smoke, car exhaust, power stations), dust from roads or fields, and haze formed by chemical reactions in the air. They have widely varying physical and chemical characteristics, which are incompletely understood, but the legal limits are based simply on the total mass per cubic meter of air.

Most of the recent concern has come from *time-series studies*, which are relatively easy and inexpensive to carry out. These studies examine the associations between total number of deaths, hospital admissions, or emergency room visits in a city with the average pollution levels.

As the EPA requires regular monitoring of air pollution and other government agencies collect information on deaths and hospital attendance, the data merely need to be extracted from the relevant databases.

This description glosses over some important statistical issues, many of which were pointed out by epidemiologists when the first studies were published:

1. There is a lot of variation in exposure among a group of people.
2. The monitors may be deliberately located in dirty areas to detect problems (or in clean areas so as not to detect problems).
3. The day-to-day outcome measurements are not independent.
4. There is a large seasonal variation in both exposure and outcome, potentially confounding the results.
5. We don't know how much time should be expected between exposure to fine particles and death or illness.

You should be able to think of several other potential problems, but a more useful exercise for the statistician is to classify the problems by whether they are important and whether they are soluble. It turns out that the first two are not important because they are more or less constant from day to day and so cancel out of our comparisons. The third problem is potentially important and led to some interesting statistical research, but it turns out that addressing it does not alter the results.

The fourth problem, seasonal variation, is important, as Figure 20.6 shows. In Seattle, mortality and air pollution peak in the winter. In many other cities the pattern is slightly different, with double peaks in winter and summer, but some form of strong seasonality is the rule. The

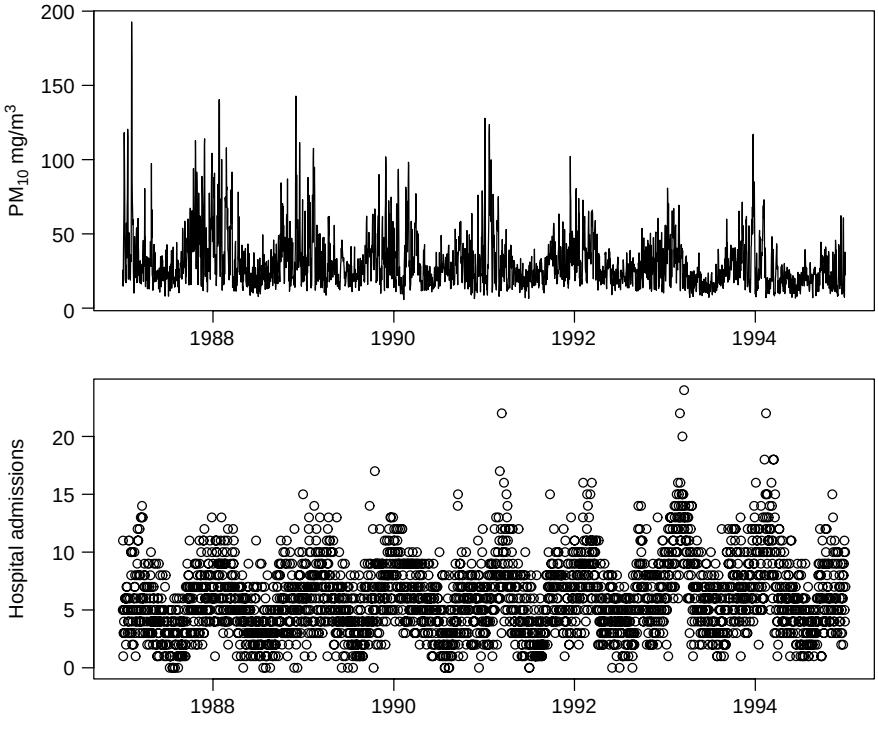


Figure 20.6 Particulate air pollution concentrations and hospital admissions for respiratory disease in Seattle.

solution to this confounding problem is to include these seasonal effects in our regression model. This is complicated: As gardeners and skiers well know, the seasons are not perfectly regular from year to year. Epidemiologists found a statistical solution, the generalized additive model (GAM), which had been developed for completely different problems, and adapted it to these time series. The GAM models allow the seasonal variation to be modeled simply by saying how smooth it should be:

$$\log(\text{mortality rate on day } t) = \alpha(t) + \beta \times \text{fine-particle concentration}$$

The smooth function $\alpha(t)$ absorbs all the seasonal variation and leaves only the short-term day-to-day fluctuations for evaluating the relationship between air pollution and mortality summarized by the log relative risk β. Computationally, $\alpha(t)$ is similar to the scatter plot smoothers discussed in Chapter 3.

With the problem of seasonal variation classified as important but soluble, analyses proceeded using data from many different U.S. cities and cities around the world. Shortly after the EPA had compiled a review of all the relevant research as a prelude to setting new standards, some bad news was revealed. Researchers at Johns Hopkins School of Public Health, who had compiled the largest and most systematic set of time-series studies, reported that they and everyone else had been using the GAM software incorrectly. The software had been written many years before, when computers were much slower, and had been intended for simpler examples than these time-series studies. The computations for a GAM involve iterative improvements to an estimate until it stops changing, and the default criterion for "stops changing" was not tight enough for the air pollution time-series models. At about the same time, researchers in Canada noticed that one of the approximations used in calculating confidence intervals and p-values was also not quite good enough in these time-series models [Ramsay et al., 2003]. When the dust settled, it became clear that the problem of seasonal variation was still soluble—fixes were found for these two problems, many studies were reanalyzed, and the conclusions remained qualitatively the same.

The final problem, the fact that the latency is not known, is just one special case of the problem of model uncertainty—choosing a regression model is much harder than fitting it. It is easy to estimate the association between mortality and today's pollution, or yesterday's pollution, or the previous day's, or the average of the past week, or any other choice. It is very hard to choose between these models. Simply reporting the best results is clearly biased, but is sometimes done. Fitting all the possible models may obscure the true associations among all the random noise. Specifying a particular model a priori allows valid inference but risks missing the true association. This final problem is important, but there is no simple mathematical solution.

20.7 ARE TECHNICIANS AS GOOD AS PHYSICIANS?

The neuropathological diagnosis of Alzheimer's disease (AD) is time consuming and difficult, even for experienced neuropathologists. Work in the late 1960s and early 1970s found that the presence of senile neuritic plaques in the neocortex and hippocampus justified a neuropathological diagnosis of Alzheimer's disease [Tomlinson et al., 1968, 1970]. Plaques are proteins associated with degenerating nerve cells in the brain; they tend to be located near the points of contact between cells. Typically, they are found in the brains of older persons.

These studies also found that large numbers of neurofibrillary tangles were often present in the neocortex and the hippocampus of brains from Alzheimer's disease victims. A tangle is another protein in the shape of a paired helical fragment found in the nerve cell. Neurofibrillary tangles are also found in other diseases. Later studies showed that plaques and tangles could be found in the brains of elderly persons with preserved mental status. Thus, the quantity and distribution of plaques and tangles, rather than their mere presence, are important in distinguishing Alzheimer's brains from the brains of normal aging persons.

A joint conference of 1985 [Khachaturian, 1985] stressed the need for standardized clinical and neuropathological diagnoses for Alzheimer's disease. We wanted to find out whether subjects with minimal training can count plaques and tangles in histological specimens of patients with Alzheimer's disease and controls [van Belle et al., 1997]. Two experienced neuropathologists trained three student helpers to recognize plaques and tangles in slides obtained from autopsy material. After training, the students and pathologists examined coded slides from patients with Alzheimer's disease and controls. Some of the slides were repeated to provide an estimate of reproducibility. Each reader read four fields, which were then averaged.

Ten sequential cases with a primary clinical and neuropathological diagnosis of Alzheimer's disease were chosen from the Alzheimer's Disease Research Center's (ADRC) brain autopsy registry. Age at death ranged from 67 years to 88 years, with a mean of 75.7 years and a standard deviation of 5.9 years.

Ten controls were examined for this study. Nine controls were selected from the ADRC registry of patients with brain autopsy, representing all subjects in the registry with no neuropathological evidence of AD. Four of these did have a clinical diagnosis of Alzheimer's disease, however. One additional control was drawn from files at the University of Washington's Department of Neuropathology. This control, aged 65 years at death, had no clinical history of Alzheimer's disease.

For each case and control, sections from the hippocampus and from the temporal, parietal, and frontal lobes were viewed by two neuropathologists and three technicians. The three technicians were a first-year medical school student, a graduate student in biostatistics with previous histological experience, and a premedical student. The technicians were briefly trained (for several hours) by a neuropathologist. The training consisted of looking at brain tissue (both Alzheimer's cases and normal brains) with a double-headed microscope and at photographs of tissue. The neuropathologist trained the technicians to identify plaques and tangles in the tissue samples viewed. The training ended when the neuropathologist was satisfied that the technicians would be able to identify plaques and tangles in brain tissue samples on their own for the purposes of this study. The slides were masked to hide patient identity and were arbitrarily divided into batches of five subjects, with cases and controls mixed. Each viewer was asked to scan the entire slide to find the areas of the slide with the highest density of plaques and tangles (implied by Khachaturian [1985]). The viewer then chose the four fields on the slide that appeared to contain the highest density of plaques and tangles when viewed at 25×. Neurofibrillary tangles and senile plaques were counted in these four fields at 200×. If the field contained more than 30 plaques or tangles, the viewer scored the number of lesions in that field as 30.

The most important area in the brain for the diagnosis of Alzheimer's is the hippocampus, and the results are presented for that region. Results for other regions were similar. In addition, we deal here only with cases and plaques. Table 20.2 contains results for the estimated number of plaques per field for cases; each reading is the average of readings from four fields. The estimated number of plaques varied considerably, ranging from zero to more than 20. Inspection of Table 20.2 suggests that technician 3 tends to read higher than the other technicians and the neuropathologists, that is, tends to see more plaques. An analysis of variance confirms this impression:

Source of Variation	d.f.	Mean Square	*F*-Ratio
Patients	9	102.256	—
Observers	4	—	—
Technicians vs. neuropathologists	1	21.31	2.70
Within technicians	2	42.53	5.39
Neuropathologist A vs. neuropathologist B	1	2.556	0.32
Patients × observers	36	7.888	—

Table 20.2 Average Number of Plaques per Field in the Hippocampus as Estimated by Three Technicians and Two Neuropathologists[a]

						Correlations:			
	Technician			Neuropathologist		Technician		Neuropathologist	
Case	1	2	3	A	B		2	3	A	B
1	0.75	0.00	0.00	0.00	0.00	1	0.69	0.63	0.65	0.76
2	7.25	6.50	7.50	4.75	3.75	2		0.77	0.79	0.84
3	5.50	7.25	5.50	5.75	8.75	3			0.91	0.67
4	5.25	8.00	14.30	5.75	6.50	A				0.82
5	10.00	8.25	9.00	3.50	7.75					
6	7.25	7.00	21.30	13.00	8.50					
7	5.75	15.30	18.80	10.30	8.00					
8	1.25	4.75	3.25	3.25	4.00					
9	1.75	5.00	7.25	2.50	3.50					
10	10.50	16.00	18.30	13.80	19.00					
Mean	5.25	7.80	10.50	6.26	6.98					
SD	3.44	4.76	7.21	4.60	5.08					

[a] Averages are over four fields.

You will recognize from Chapter 10 the idea of partitioning the variance attributable to observers into three components; there are many ways of partitioning this variance. The table above contains one useful way of doing this. The analysis suggests that the average levels of response do not vary within neuropathologists. There is a highly significant difference among technicians. We would conclude that technician 3 is high, rather than technician 1 being low, because of the values obtained by the two neuropathologists. Note also that the residual variability is estimated to be $\sqrt{7.888} = 2.81$ plaques per patient. This represents considerable variability since the values represent averages of four readings. Using a single reading as a basis produces an estimated standard deviation of $(\sqrt{4})(2.81) = 5.6$ plaques per reading.

But how shall agreement be measured or evaluated? Equality of the mean levels suggests only that the raters tended to count the same number of plaques on average. We need a more precise formulation of the issue. A correlation between the technicians and the neuropathologists will provide some information but is not sufficient because the correlation is invariant under changes in location and scale. In Chapter 4 we distinguished between precision and accuracy. *Precision* is the degree to which the observations cluster around a line; *accuracy* is the degree to which the observations are close to some standard. In this case the standard is the score of the neuropathologist and accuracy can be measured by the extent to which a technician's readings are from a 45° line. A paper by Lin [1989] nicely provides a framework for analyzing these data. In our case, the data are analyzed according to five criteria: location shift, scale shift, precision, accuracy, and concordance. *Location shift* refers to the degree to which the means of the data differ between technician and neuropathologist. A *scale shift* measures the differences in variability. *Precision* is quantified by a measure of correlation (Pearson's in our case). *Accuracy* is estimated by the distance that the observations are from the 45° line. *Concordance* is defined as the product of the precision and the accuracy. In symbols, denote two raters by subscripts 1 and 2. Then we define

$$\text{location shift} = u = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1 \sigma_2}}$$

$$\text{scale shift} = v = \frac{\sigma_1}{\sigma_2}$$

Table 20.3 Characteristics of Ratings of Three Technicians and Two Neuropathologists[a]

Technician	Pathologist	Location Shift	Scale Shift	Precision	Accuracy	Concordance
1	A	−0.18	0.75	0.95	0.94	0.89
	B	−0.35	0.68	0.76	0.88	0.67
2	A	0.33	1.03	0.79	0.95	0.75
	B	0.17	0.94	0.84	0.98	0.83
3	A	0.74	1.57	0.91	0.73	0.66
	B	0.58	1.42	0.67	0.81	0.55
A	B	−0.14	0.98	0.82	0.99	0.81

[a]Estimated numbers of plaques in the hippocampus of 10 cases, based on data from Table 20.2.

$$\text{precision} = r$$

$$\text{accuracy} = A = \left(\frac{v + 1/v + u^2}{2}\right)^{-1}$$

$$\text{concordance} = rA$$

We discuss these briefly. The location shift is a standardized estimate of the difference between the two raters. The quantity $\sqrt{\sigma_1\sigma_2}$ is the geometric mean of the two standard deviations. If there is no location difference between the two raters, this quantity is centered around zero. The scale shift is a ratio; if there is no scale shift, this quantity is centered around 1. The precision is the usual correlation coefficient; if the paired data fall on a straight line, the correlation is 1. The accuracy is made up of a mixture of the means and the standard deviations. Note that if there is no location or scale shift, the accuracy is 1, the upper limit for this statistic. The concordance is the product of the accuracy and the precision; it is also bounded by 1. The data in Table 20.2 are analyzed according to the criteria above and displayed in Table 20.3. This table suggests that all the associations between technicians and neuropathologists are comparable. In addition, the comparisons between neuropathologists provide an internal measure of consistency. The "location shift" column indicates that, indeed, technician 3 tended to see more plaques than the neuropathologists. Technician 3 was also more variable, as indicated in the "scale shift" column. Technician 1 tended to be less variable than the neuropathologists. The precision of the technicians was comparable to that of the two neuropathologists compared with each other. The neuropathologists also displayed very high accuracy, almost matched by technician 1 and 2. The concordance, the product of the precision and the accuracy, averaged over the two neuropathologists is comparable to their concordance. As usual, it is very important to graph the data to confirm these analytical results by a graphical display. Figure 20.7 displays the seven possible graphs.

In summary, we conclude that it is possible to train relatively naive observers to count plaques in a manner comparable to that of experienced neuropathologists, as defined by the measures above. By this methodology, we have also been able to isolate the strengths and weaknesses of each technician.

20.8 RISKY BUSINESS

Every day of our lives we meet many risks: the risk of being struck by lightning, getting into a car accident on the way to work, eating contaminated food, and getting hepatitis. Many risks have associated moral and societal values. For example, what is the risk of being infected by AIDS through an HIV-positive health practitioner? How does this risk compare with getting infectious hepatitis from an infected worker? What is the risk to the health practitioner in being identified as HIV positive? As we evaluate risks, we may ignore them, despite their being real

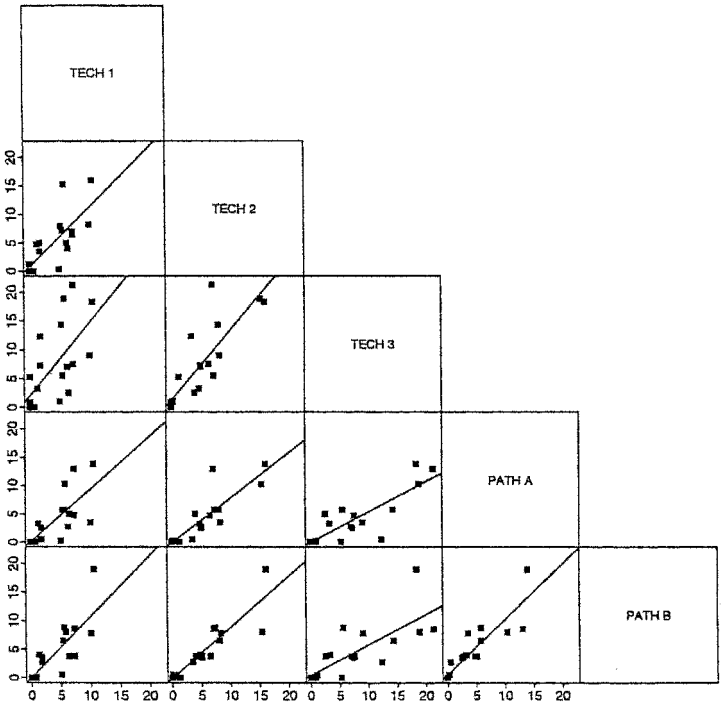


Figure 20.7 Seven possible graphs for the data in Table 20.3, prepared by SYSTAT, a very comprehensive software package. (From Wilkinson [1989].)

and substantial: for example, smoking in the face of the evidence in the Surgeon General's reports. Or we may react to risks even though they are small: for example, worry about being hit by a falling airplane.

What is a risk? A *risk* is usually an event or the probability of the event. Thus, the risk of being hit by lightning is defined to be the probability of this event. The word *risk* has an unfavorable connotation. We usually do not speak of the risk of winning the lottery. For purposes of this chapter, we relate the risk of an event to the probability of the occurrence of the event. In Chapter 3 we stated that all probabilities are conditional probabilities. When we talk about the risk of breast cancer, we usually refer to its occurrence among women. Probabilities are modified as we define different groups at risk. R. A. Fisher talked about *relevant subsets*, that is, what group or set of events is intended when a probability is specified.

In the course of thinking about environmental and occupational risks, one of us (G.vB.) wanted to develop a scale of risks similar to the Richter scale for earthquakes. The advantages of such a scale is to present risks numerically in such a way that the public would have an intuitive understanding of the risks. This, despite not understanding the full basis of the scale (it turns out to be fairly difficult to find a complete description of the Richter scale).

What should be the characteristics of such a scale? It became clear very quickly that the scale would have to be logarithmic. Second, it seemed that increasing risks should be associated with increasing values of the scale. It would also be nice to have the scale have roughly the same numerical range as the Richter scale. Most of its values are in the range 3 to 7. The *risk scale* for events is defined as follows: Let $P(E)$ be the probability of an event; then the risk units, $\mathrm{RU}(E)$, for this event are defined to be

$$\mathrm{RU}(E) = 10 + \log_{10}[P(E)]$$

Table 20.4 Relationship of Risk Units to Probabilities

Probability of Event	Risk Units
1	10
1/10	9
1/100	8
1/1000	7
1/10,000	6
1/100,000	5
1/1,000,000	4
1/10,000,000	3
1/100,000,000	2
1/1,000,000,000	1
1/10,000,000,000	0
1/100,000,000,000	−1

This scale has several nice properties. First, the scale is logarithmic. Second, if the event is certain, $P(E) = 1$ and $\mathrm{RU}(E) = 10$. Given two independent events, E_1 and E_2, the difference in their risks is

$$\mathrm{RU}(E_1) - \mathrm{RU}(E_2) = \log_{10} \frac{P(E_1)}{P(E_2)}$$

that is, the difference in the risk units is related to the relative risk of the events in a logarithmic fashion, that is, a logarithm of the odds (see Table 20.4). Third, the progression is terms of powers of 10 is very simple; and so on. So a shift of 2 risk units represents a 100-fold change in probabilities. Events with risk units of the order of 1 to 4 are associated with relatively rare events. Note that the scale can go below zero.

As with the Richter scale, familiarity with common events will help you get a feeling for the scale. Let us start by considering some random events; next we deal with some common risks and locate them on the scale; finally, we give you some risks and ask you to place them on the scale (the answers are given at the end of the chapter). The simplest case is the coin toss. The probability of, say, a head is 0.5. Hence the risk units associated with observing a head with a single toss of a coin is $\mathrm{RU}(\text{heads}) = 10 - \log_{10}(0.5) = 9.7$ (expressing risk units to one decimal place is usually enough). For a second example, the risk units of drawing at random a specified integer from the digits 0, 1, 2, 3, . . . , 9 is 1/10 and the RU value is 9. Rolling a pair of sevens with two dice has a probability of 1/36 and are RU value of 8.4. Now consider some very small probabilities. Suppose that you dial at random; what is the chance of dialing your own phone number? Assume that we are talking about the seven-digit code and we allow all zeros as a possible number. The RU value is 3. If you throw in the area code as well, you must deduct three more units to get the value $\mathrm{RU} = 0$. There are clearly more efficient ways to make phone calls.

The idea of a logarithmiclike scale for probabilities appears in the literature quite frequently. In a delightful, little-noticed book, *Risk Watch*, Urquhart and Heilmann [1984] defined the safety unit of an event, E, as

$$\text{safety unit of } E = -\log_{10}[P(E)]$$

The drawback of this definition is that it calibrates events in terms of safety rather than risk. People are more inclined to think in terms of risk; they are "risk avoiders" rather than safety

Table 20.5 The Risk Unit Scale and Some Associated Risks

Risk Unit	Event
10	Certain event
9	Pick number 3 at random from 0 to 3
8	Car accident with injury (annual)
7	Killed in hang gliding (annual)
6	EPA action (life time risk)
5	Cancer from 4 tbsp peanut butter/day (annual)
4	Cancer from one transcontinental trip
3	Killed by falling aircraft
2	Dollar bill has specified set of eight numbers
1	Pick spot on earth at random and land within $\frac{1}{4}$ mile of your house
0	Your phone number picked at random (+ area code)
−0.5	Killed by falling meteorite (annual)

Table 20.6 Events to Be Ranked and Placed on Risk Units Scale[a]

a.	Accidental drowning
b.	Amateur pilot death
c.	Appear on the *Johnny Carson Show* (1991)
d.	Death due to smoking
e.	Die in mountain climbing accident
f.	Fatality due to insect bite or sting
g.	Hit by lightning (in lifetime)
h.	Killed in college football
i.	Lifetime risk of cancer due to chlorination
j.	Cancer from one diet cola per day with saccharin
k.	Ace of spades in one draw from 52-card deck
l.	Win the *Reader's Digest* Sweepstakes
m.	Win the Washington State lottery grand prize (with one ticket)

[a] All risks are annual unless otherwise indicated. Events not ordered by risk.

seekers. But it is clear that risk units and safety units very simply related:

$$\mathrm{RU}(E) = 10 - \mathrm{SU}(E)$$

Table 20.5 lists the risk units for a series of events. Most of these probabilities were gleaned from the risk literature. Beside the events mentioned already, the risk unit for a car accident with injury in a 1-year time interval has a value of 8. This corresponds to a probability of 0.01, or 1/100. The Environmental Protection Agency takes action on lifetime risks of risk unit 6. That is, if the lifetime probability of death is 1/10,000, the agency will take some action. This may seem rather anticonservative, but there are many risks, and some selection has to be made. All these probabilities are estimates with varying degrees of precision. Crouch and Wilson [1982] include references to the data set upon which the estimate is based and also indicate whether the risk is changing. Table 20.6 describes some events for which you are asked to estimate the risk units. The answers are given in Table 20.7, preceding the References.

Table 20.7 Activities Estimated to Increase the Annual Probability of Death by One in a Million[a]

Activity	Cause of Death
Smoking 1.4 cigarettes	Cancer, heart disease
Drinking 0.5 liter of wine	Cirrhosis of the liver
Living 2 days in New York or Boston	Air pollution
Traveling 10 miles by bicycle	Accident
Living 2 months with a cigarette smoker	Cancer, heart disease
Drinking Miami drinking water for 1 year	Cancer from chloroform
Living 150 years within 5 miles of a nuclear power plant	Cancer from radiation
Eating 100 charcoal-broiled steaks	Cancer from benzopyrene

Source: Condensed from Wynne [1991].
[a]All events have a risk unit value of 4.

How do we evaluate risks? Why do we take action on some risks but not on others? The study of risks has been become a separate science with its own journals and society. The Borgen [1990] and Slovic [1986] articles in the journal *Risk Analysis* are worth examining. The following dimensions about evaluating risks have been mentioned in the literature:

Voluntary	Involuntary
Immediate effect	Delayed effect
Exposure essential	Exposure a luxury
Common hazard	"Dread" hazard
Affects average person	Affects special group
Reversible	Irreversible

We discuss these briefly. Recreational scuba diving has an annual probability of death of 4/10,000, or a risk unit of 6.6 [Crouch and Wilson, 1982, Table 7.4]. Compare this with some of the risks in Table 20.5. Another dimension is the timing of the effect. If the effect is delayed, we are usually willing to take a bigger risk; the most obvious example is smoking (which also is a voluntary behavior). If the exposure is essential, as part of one's occupation, then again, larger risks are acceptable. A "dread" hazard is often perceived as of greater risk than a common hazard. The most conspicuous example is an airplane crash vs. an automobile accident. But perversely, we are less likely to be concerned about hazards that affect special groups to which we are not immediately linked. For example, migrant workers have high exposures to pesticides and resulting increased immediate risks of neurological damage and long-term risks of cancer. As a society, we are not vigorous in reducing those risks. Finally, if the effects of a risk are reversible, we are willing to take larger risks.

Table 20.7 lists some risks with the same estimated value: Each one increases the annual risk of death by 1 in a million; that is, all events have a risk unit value of 4. These examples illustrate that we do not judge risks to be the same even though the probabilities are equal. Some of the risks are avoidable; others may not be. It may be possible to avoid drinking Miami drinking water by drinking bottled water or by moving to Alaska. Most of the people who live in New York or Boston are not aware of the risk of living in those cities. But even if they did, it is unlikely that they would move. A risk of 1 in a million is too small to act on.

How can risks be ranked? There are many ways. The primary one is by the probability of occurrence as we have discussed so far. Another is by the expected loss (or gain). For example, the probability of a fire destroying your home is fairly small but the loss is so great that it pays to make the unfair bet with the insurance company. An unfair bet is one where the expected gain is negative. Another example is the lottery. A typical state lottery takes more than 50 cents from every dollar that is bet (compared to about 4 cents for roulette play in a casino). But the reward is so large (and the investment apparently small) that many people gladly play this unfair game.

Table 20.8 Answers to Evaluation of Risks in Table 20.5

	Risk Units	Source/Comments
a.	5.6	Crouch and Wilson [1982, Table 7.2]
b.	7.0	Crouch and Wilson [1982, Table 7.4]
c.	4.3	Siskin et al. [1990]
d.	7.5	Slovic [1986, Table 1]
e.	6.8	Crouch and Wilson [1982, Table 7.4]
f.	3.4	Crouch and Wilson [1982, Table 7.2]
g.	4.2	Siskin et al. [1990]
h.	5.5	Crouch and Wilson [1982, Table 7.4]
i.	4.0	Crouch and Wilson [1982, Table 7.5 and pp. 186–187]
j.	5.0	Slovic [1986, Table 1]
k.	8.3	$10 + \log(1/52)$
l.	1.6	From back of announcement; $10 + \log(1/250{,}000{,}000)$
m.	3.0	From back of lottery ticket; $10 + \log(1/10{,}000{,}000)$

How can risks be changed? It is clearly possible to stop smoking, to give up scuba diving, quit the police force, never drive a car. Many risks are associated with specific behaviors and changing those behaviors will change the risks. In the language of probability we have moved to another subset. Some changes will not completely remove the risks because of lingering effects of the behavior. But a great deal of risk reduction can be effected by changes in behavior. It behooves each one of us to assess the risks we take and to decide whether they are worth it.

The *Journal of the Royal Statistical Society*, Series A devoted the June 2003 issue (Volume 166) to statistical issues in risk communication. The journal *Risk Analysis* address risk analysis, risk assessment, and risk communication.

REFERENCES

Alderman, E. L., Bourassa, M. G., Cohen, L. S., Davis, K. B., Kaiser, G. C., Killip, T., Mock, M. B., Pettinger, M., and Robertson, T. L. [1990]. Ten-year follow-up of survival and myocardial infarction in the randomized Coronary Artery Surgery Study. *Circulation*, **82**: 1629–1646.

ALLHAT Officers and Coordinators [2002]. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic. The antihypertensive and lipid-lowering treatment to prevent heart attach trial (ALLHAT). *JAMA*, **288**: 2981–2997.

Battie, M. C., Bigos, S. J., Fisher, L. D., Hansson, T. H., Nachemson, A. L., Spengler, D. M., Wortley, M. D., and Zeh, J. [1989]. A prospective study of the role of cardiovascular risk factors and fitness in industrial back pain complaints. *Spine*, **14**: 141–147.

Battie, M. C., Bigos, S. J., Fisher, L. D., Spengler, D. M., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. [1990a]. Anthropometric and clinical measures as predictors of back pain complaints in industry: a prospective study. *Journal of Spinal Disorders*, **3**: 195–204.

Battie, M. C., Bigos, S. J., Fisher, L. D., Spengler, D. M., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. [1990b]. The role of spinal flexibility in back pain complaints within industry: a prospective study. *Spine*, **15**: 768–773.

Bigos, S. J., Spengler, D. M., Martin, N. A., Zeh, J., Fisher, L., Nachemson, A., and Wang, M. H. [1986a]. Back injuries in industry—a retrospective study: II. Injury factors. *Spine*, **11**: 246–251.

Bigos, S. J., Spengler, D. M., Martin, N. A., Zeh, J., Fisher, L., Nachemson, A., and Wang, M. H. [1986b]. Back injuries in industry—a retrospective study: III. Employee-related factors. *Spine*, **11**: 252–256.

Bigos, S. J., Battie, M. C., Spengler, D. M., Fisher, L. D., Fordyce, W. E., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. [1991]. A prospective study of work perceptions and psychosocial factors affecting the report of back injury. *Spine*, **16**: 1–6.

Bigos, S. J., Battie, M. C., Fisher, L. D., Fordyce, W. E., Hansson, T. H., Nachemson, A. L., and Spengler, D. M. [1992a]. A longitudinal, prospective study of industrial back injury reporting in industry. *Clinical Orthopaedics*, **279**: 21–34.

Bigos, S. J., Battie, M. C., Fisher, L. D., Hansson, T. H., Spengler, D. M., and Nachemson, A. L. [1992b]. A prospective evaluation of commonly used pre-employment screening tools for acute industrial back pain. *Spine*, **17**: 922–926.

Borer, J. S. [1987]. t-PA and the principles of drug approval (editorial). *New England Journal of Medicine*, **317**: 1659–1661.

Borgen, K. T. [1990]. Of apples, alcohol, and unacceptable risks. *Risk Analysis*, **10**: 199–200.

CASS Principal Investigators and Their Associates [1981]. *National Heart, Lung, Blood Institute Coronary Artery Surgery Study*, T. Killip, L. D. Fisher, and M. B. Mock (eds.). American Heart Association Monograph 79. *Circulation*, **63**(p. II): I-1 to I-81.

CASS Principal Investigators and Their Associates: Coronary Artery Surgery Study (CASS) [1983a]. A randomized trial of coronary artery bypass surgery: survival data. *Circulation*, **68**: 939–950.

CASS Principal Investigators and Their Associates: Coronary Artery Surgery Study (CASS) [1983b]. A randomized trial of coronary artery bypass surgery: quality of life in patients randomly assigned to treatment groups. *Circulation*, **68**: 951–960.

CASS Principal Investigators and Their Associates: Coronary Artery Surgery Study (CASS) [1984a]. A randomized trial of coronary artery bypass surgery: comparability of entry characteristics and survival in randomized patients and nonrandomized patient meeting randomization criteria. *Journal of the American College of Cardiology*, **3**: 114–128.

CASS Principal Investigators and Their Associates [1984b]. Myocardial infarction and mortality in the Coronary Artery Surgery Study (CASS) randomized trial. *New England Journal of Medicine*, **310**: 750–758.

Chaitman, B. R., Ryan, T. J., Kronmal, R. A., Foster, E. D., Frommer, P. L., Killip, T., and the CASS Investigators [1990]. Coronary Artery Surgery Study (CASS): comparability of 10 year survival in randomized and randomizable patients. *Journal of the American College of Cardiology*, **16**: 1071–1078.

Crouch, E. A. C., and Wilson, R. [1982]. *Risk Benefit Analysis*. Ballinger, Cambridge, MA.

Fisher, L. D., Giardina, E.-G., Kowy, P. R., Leier, C. V., Lowenthal, D. T., Messerli, F. H., Pratt, C. M., and Ruskin, J. [1987]. The FDA Cardio-Renal Committee replies (letter to the editor). *Wall Street Journal*, Wed., Aug. 12, p. 19.

Fisher, L. D., Kaiser, G. C., Davis, K. B., and Mock, M. [1989]. Crossovers in coronary bypass grafting trials: desirable, undesirable, or both? *Annals of Thoracic Surgery*, **48**: 465–466.

Fisher, L. D., Dixon, D. O., Herson, J., and Frankowski, R. F. [1990]. Analysis of randomized clinical trials: intention to treat. In *Statistical Issues in Drug Research and Development*, K. E. Peace (ed.). Marcel Dekker, New York, pp. 331–344.

Fisher, L. D., and Zeh, J. [1991]. An information theory approach to presenting predictive value in the Cox proportional hazards regression model (unpublished).

International Conference on Harmonisation [2000]. *ICH Harmonised Tripartite Guideline: E10. Choice of Control Group and Related Issues in Clinical Trials. http://www.ich.org*

Kaiser, G. C., Davis, K. B., Fisher, L. D., Myers, W. O., Foster, E. D., Passamani, E. R., and Gillespie, M. J. [1985]. Survival following coronary artery bypass grafting in patients with severe angina pectoris (CASS) (with discussion). *Journal of Thoracic and Cardiovascular Surgery*, **89**: 513–524.

Khachaturian, Z. S. [1985]. Diagnosis of Alzheimer's disease. *Archives of Neurology*, **42**: 1097–1105.

Kowey, P. R., Fisher, L. D., Giardina, E.-G., Leier, C. V., Lowenthal, D. T., Messerli, F. H., and Pratt, C. M. [1988]. The TPA controversy and the drug approval process: the view of the Cardiovascular and Renal Drugs Advisory Committee. *Journal of the American Medical Association*, **260**: 2250–2252.

Lin, L. I. [1989]. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**: 255–268.

Lumley, T. [2002]. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, **21**: 2313–2324

Myers, W. O., Schaff, H. V., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Ryan, T. J., Kaiser, G. C., and CASS Investigators [1989]. Improved survival of surgically treated patients with triple vessel coronary disease and severe angina pectoris: a report from the Coronary Artery Surgery Study (CASS) registry. *Journal of Thoracic and Cardiovascular Surgery*, **97**: 487–495.

Passamani, E., Davis, K. B., Gillespie, M. J., Killip, T., and the CASS Principal Investigators and Their Associates [1985]. A randomized trial of coronary artery bypass surgery: survival of patients with a low ejection fraction. *New England Journal of Medicine*, **312**: 1665–1671.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. L., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. [1977]. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples. *British Journal of Cancer*, **35**: 1–39.

Preston, T. A. [1977]. *Coronary Artery Surgery: A Critical Review*. Raven Press, New York.

Psaty, B., Lumley, T., Furberg, C., Schellenbaum, G., Pahor, M., Alderman, M. H., and Weiss, N. S. [2003]. Health outcomes associated with various anti-hypertensive therapies used as first-line agents: a network meta-analysis. *Journal of the American Medical Association*, **289**: 2532–2542.

Ramsay, T. O., Burnett, R. T., and Krewski, D. [2003]. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**: 18–23.

Rogers, W. J., Coggin, C. J., Gersh, B. J., Fisher, L. D., Myers, W. O., Oberman, A., and Sheffield, L. T. [1990]. Ten-year follow-up of quality of life in patients randomized to receive medical therapy or coronary artery bypass graft surgery. *Circulation*, **82**: 1647–1658.

Siskin, B., Staller, J., and Rornik, D. [1990]. *What Are the Chances? Risk, Odds and Likelihood in Everyday Life*. Crown Publishers, New York.

Slovic, P. [1986]. Informing and educating the public about risk. *Risk Analysis*, **6**: 403–415.

Spengler, D. M., Bigos, S. J., Martin, N. A., Zeh, J., Fisher, L. D., and Nachemson, A. [1986]. Back injuries in industry: a retrospective study: I. Overview and cost analysis. *Spine*, **11**: 241–245.

Takaro, T., Hultgren, H., Lipton, M., Detre, K., and participants in the Veterans Administration Cooperative Study Group [1976]. VA cooperative randomized study for coronary arterial occlusive disease: II. Left main disease. *Circulation*, **54**(suppl. 3): III-107.

Tomlinson, B. E., Blessed, G., and Roth, M. [1968]. Observations on the brains of non-demented old people. *Journal of Neurological Science*, **7**: 331–356.

Tomlinson, B. E., Blessed, G., and Roth, M. [1970]. Observations on the brains of demented old people. *Journal of Neurological Science*, **11**: 205–242.

Urquhart, J., and Heilmann, K. [1984]. *Risk Watch: The Odds of Life*. Facts on File Publications, New York.

van Belle, G., Gibson, K., Nochlin, D., Sumi, M., and Larson, E. B. [1997]. Counting plaques and tangles in Alzheimer's disease: concordance of technicians and pathologists. *Journal of neurological Science*, **145**: 141–146.

Wall Street Journal [1987a]. The TPA decision (editorial). *Wall Street Journal*, Thurs., May 28, p. 26.

Wall Street Journal [1987b]. Human sacrifice (editorial). *Wall Street Journal*, Tues., June 2, p. 30.

Weinstein, G. S., and Levin, B. [1989]. Effect of crossover on the statistical power of randomized studies. *Annals of Thoracic Surgery*, **48**: 490–495.

Wilkinson, L. [1989]. *SYGRAPH: The System for Graphics*. SYSTAT, Inc., Evanston, IL.

Wynne, B. [1991]. Public perception and communication of risk: what do we know? *NIH Journal of Health*, **3**: 65–71.

Appendix

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

Table A.1 Standard Normal Distribution

Let Z be a normal random variable with mean zero and variance 1. For selected values of Z, three values are tabled: (1) the two-sided p-value, or $P[|Z| \geq z]$; (2) the one-sided p-value, or $P[Z \geq z]$; and (3) the cumulative distribution function at Z, or $P[Z \leq z]$.

z	Two-sided	One-sided	Cum-dist.	z	Two-sided	One-sided	Cum-dist.	z	Two-sided	One-sided	Cum-dist.
0.00	1.0000	.5000	.5000	1.30	.1936	.0968	.9032	1.80	.0719	.0359	.9641
0.05	.9601	.4801	.5199	1.31	.1902	.0951	.9049	1.81	.0703	.0351	.9649
0.10	.9203	.4602	.5398	1.32	.1868	.0934	.9066	1.82	.0688	.0344	.9656
0.15	.8808	.4404	.5596	1.33	.1835	.0918	.9082	1.83	.0673	.0336	.9664
0.20	.8415	.4207	.5793	1.34	.1802	.0901	.9099	1.84	.0658	.0329	.9671
0.25	.8026	.4013	.5987	1.35	.1770	.0885	.9115	1.85	.0643	.0322	.9678
0.30	.7642	.3821	.6179	1.36	.1738	.0869	.9131	1.86	.0629	.0314	.9686
0.35	.7263	.3632	.6368	1.37	.1707	.0853	.9147	1.87	.0615	.0307	.9693
0.40	.6892	.3446	.6554	1.38	.1676	.0838	.9162	1.88	.0601	.0301	.9699
0.45	.6527	.3264	.6736	1.39	.1645	.0823	.9177	1.89	.0588	.0294	.9706
0.50	.6171	.3085	.6915	1.40	.1615	.0808	.9192	1.90	.0574	.0287	.9713
0.55	.5823	.2912	.7088	1.41	.1585	.0793	.9207	1.91	.0561	.0281	.9719
0.60	.5485	.2743	.7257	1.42	.1556	.0778	.9222	1.92	.0549	.0274	.9726
0.65	.5157	.2578	.7422	1.43	.1527	.0764	.9236	1.93	.0536	.0268	.9732
0.70	.4839	.2420	.7580	1.44	.1499	.0749	.9251	1.94	.0524	.0262	.9738
0.75	.4533	.2266	.7734	1.45	.1471	.0735	.9265	1.95	.0512	.0256	.9744
0.80	.4237	.2119	.7881	1.46	.1443	.0721	.9279	1.96	.0500	.0250	.9750
0.85	.3953	.1977	.8023	1.47	.1416	.0708	.9292	1.97	.0488	.0244	.9756
0.90	.3681	.1841	.8159	1.48	.1389	.0694	.9306	1.98	.0477	.0239	.9761
0.95	.3421	.1711	.8289	1.49	.1362	.0681	.9319	1.99	.0466	.0233	.9767
1.00	.3173	.1587	.8413	1.50	.1336	.0668	.9332	2.00	.0455	.0228	.9772
1.01	.3125	.1562	.8438	1.51	.1310	.0655	.9345	2.01	.0444	.0222	.9778
1.02	.3077	.1539	.8461	1.52	.1285	.0643	.9357	2.02	.0434	.0217	.9783
1.03	.3030	.1515	.8485	1.53	.1260	.0630	.9370	2.03	.0424	.0212	.9788
1.04	.2983	.1492	.8508	1.54	.1236	.0618	.9382	2.04	.0414	.0207	.9793
1.05	.2937	.1469	.8531	1.55	.1211	.0606	.9394	2.05	.0404	.0202	.9798
1.06	.2891	.1446	.8554	1.56	.1188	.0594	.9406	2.06	.0394	.0197	.9803
1.07	.2846	.1423	.8577	1.57	.1164	.0582	.9418	2.07	.0385	.0192	.9808
1.08	.2801	.1401	.8599	1.58	.1141	.0571	.9429	2.08	.0375	.0188	.9812
1.09	.2757	.1379	.8621	1.59	.1118	.0559	.9441	2.09	.0366	.0183	.9817
1.10	.2713	.1357	.8643	1.60	.1096	.0548	.9452	2.10	.0357	.0179	.9821
1.11	.2670	.1335	.8665	1.61	.1074	.0537	.9463	2.11	.0349	.0174	.9826
1.12	.2627	.1314	.8686	1.62	.1052	.0526	.9474	2.12	.0340	.0170	.9830
1.13	.2585	.1292	.8708	1.63	.1031	.0516	.9484	2.13	.0332	.0166	.9834
1.14	.2543	.1271	.8729	1.64	.1010	.0505	.9495	2.14	.0324	.0162	.9838
1.15	.2501	.1251	.8749	1.65	.0989	.0495	.9505	2.15	.0316	.0158	.9842
1.16	.2460	.1230	.8770	1.66	.0969	.0485	.9515	2.16	.0308	.0154	.9846
1.17	.2420	.1210	.8790	1.67	.0949	.0475	.9525	2.17	.0300	.0150	.9850
1.18	.2380	.1190	.8810	1.68	.0930	.0465	.9535	2.18	.0293	.0146	.9854
1.19	.2340	.1170	.8830	1.69	.0910	.0455	.9545	2.19	.0285	.0143	.9857
1.20	.2301	.1151	.8849	1.70	.0891	.0446	.9554	2.20	.0278	.0139	.9861
1.21	.2263	.1131	.8869	1.71	.0873	.0436	.9564	2.21	.0271	.0136	.9864
1.22	.2225	.1112	.8888	1.72	.0854	.0427	.9573	2.22	.0264	.0132	.9868
1.23	.2187	.1093	.8907	1.73	.0836	.0418	.9582	2.23	.0257	.0129	.9871
1.24	.2150	.1075	.8925	1.74	.0819	.0409	.9591	2.24	.0251	.0125	.9875
1.25	.2113	.1056	.8944	1.75	.0801	.0401	.9599	2.25	.0244	.0122	.9878
1.26	.2077	.1038	.8962	1.76	.0784	.0392	.9608	2.26	.0238	.0119	.9881
1.27	.2041	.1020	.8980	1.77	.0767	.0384	.9616	2.27	.0232	.0116	.9884
1.28	.2005	.1003	.8997	1.78	.0751	.0375	.9625	2.28	.0226	.0113	.9887
1.29	.1971	.0985	.9015	1.79	.0735	.0367	.9633	2.29	.0220	.0110	.9890

Table A.1 (***continued***)

z	Two-sided	One-sided	Cum-dist.	z	Two-sided	One-sided	Cum-dist.	z	Two-sided	One-sided	Cum-dist.
2.30	.0214	.0107	.9893	2.80	.0051	.0026	.9974	3.30	.0010	.0005	.9995
2.31	.0209	.0104	.9896	2.81	.0050	.0025	.9975	3.31	.0009	.0005	.9995
2.32	.0203	.0102	.9898	2.82	.0048	.0024	.9976	3.32	.0009	.0005	.9995
2.33	.0198	.0099	.9901	2.83	.0047	.0023	.9977	3.33	.0009	.0004	.9996
2.34	.0193	.0096	.9904	2.84	.0045	.0023	.9977	3.34	.0008	.0004	.9996
2.35	.0188	.0094	.9906	2.85	.0044	.0022	.9978	3.35	.0008	.0004	.9996
2.36	.0183	.0091	.9909	2.86	.0042	.0021	.9979	3.36	.0008	.0004	.9996
2.37	.0178	.0089	.9911	2.87	.0041	.0021	.9979	3.37	.0008	.0004	.9996
2.38	.0173	.0087	.9913	2.88	.0040	.0020	.9980	3.38	.0007	.0004	.9996
2.39	.0168	.0084	.9916	2.89	.0039	.0019	.9981	3.39	.0007	.0003	.9997
2.40	.0164	.0082	.9918	2.90	.0037	.0019	.9981	3.40	.0007	.0003	.9997
2.41	.0160	.0080	.9920	2.91	.0036	.0018	.9982	3.41	.0006	.0003	.9997
2.42	.0155	.0078	.9922	2.92	.0035	.0018	.9982	3.42	.0006	.0003	.9997
2.43	.0151	.0075	.9925	2.93	.0034	.0017	.9983	3.43	.0006	.0003	.9997
2.44	.0147	.0073	.9927	2.94	.0033	.0016	.9984	3.44	.0006	.0003	.9997
2.45	.0143	.0071	.9929	2.95	.0032	.0016	.9984	3.45	.0006	.0003	.9997
2.46	.0139	.0069	.9931	2.96	.0031	.0015	.9985	3.46	.0005	.0003	.9997
2.47	.0135	.0068	.9932	2.97	.0030	.0015	.9985	3.47	.0005	.0003	.9997
2.48	.0131	.0066	.9934	2.98	.0029	.0014	.9986	3.48	.0005	.0003	.9997
2.49	.0128	.0064	.9936	2.99	.0028	.0014	.9986	3.49	.0005	.0002	.9998
2.50	.0124	.0062	.9938	3.00	.0027	.0013	.9987	3.50	.0005	.0002	.9998
2.51	.0121	.0060	.9940	3.01	.0026	.0013	.9987	3.51	.0004	.0002	.9998
2.52	.0117	.0059	.9941	3.02	.0025	.0013	.9987	3.52	.0004	.0002	.9998
2.53	.0114	.0057	.9943	3.03	.0024	.0012	.9988	3.53	.0004	.0002	.9998
2.54	.0111	.0055	.9945	3.04	.0024	.0012	.9988	3.54	.0004	.0002	.9998
2.55	.0108	.0054	.9946	3.05	.0023	.0011	.9989	3.55	.0004	.0002	.9998
2.56	.0105	.0052	.9948	3.06	.0022	.0011	.9989	3.56	.0004	.0002	.9998
2.57	.0102	.0051	.9949	3.07	.0021	.0011	.9989	3.57	.0004	.0002	.9998
2.58	.0099	.0049	.9951	3.08	.0021	.0010	.9990	3.58	.0003	.0002	.9998
2.59	.0096	.0048	.9952	3.09	.0020	.0010	.9990	3.59	.0003	.0002	.9998
2.60	.0093	.0047	.9953	3.10	.0019	.0010	.9990	3.60	.0003	.0002	.9998
2.61	.0091	.0045	.9955	3.11	.0019	.0009	.9991	3.61	.0003	.0002	.9998
2.62	.0088	.0044	.9956	3.12	.0018	.0009	.9991	3.62	.0003	.0001	.9999
2.63	.0085	.0043	.9957	3.13	.0017	.0009	.9991	3.63	.0003	.0001	.9999
2.64	.0083	.0041	.9959	3.14	.0017	.0008	.9992	3.64	.0003	.0001	.9999
2.65	.0080	.0040	.9960	3.15	.0016	.0008	.9992	3.65	.0003	.0001	.9999
2.66	.0078	.0039	.9961	3.16	.0016	.0008	.9992	3.66	.0003	.0001	.9999
2.67	.0076	.0038	.9962	3.17	.0015	.0008	.9992	3.67	.0002	.0001	.9999
2.68	.0074	.0037	.9963	3.18	.0015	.0007	.9993	3.68	.0002	.0001	.9999
2.69	.0071	.0036	.9964	3.19	.0014	.0007	.9993	3.69	.0002	.0001	.9999
2.70	.0069	.0035	.9965	3.20	.0014	.0007	.9993	3.70	.0002	.0001	.9999
2.71	.0067	.0034	.9966	3.21	.0013	.0007	.9993	3.71	.0002	.0001	.9999
2.72	.0065	.0033	.9967	3.22	.0013	.0006	.9994	3.72	.0002	.0001	.9999
2.73	.0063	.0032	.9968	3.23	.0012	.0006	.9994	3.73	.0002	.0001	.9999
2.74	.0061	.0031	.9969	3.24	.0012	.0006	.9994	3.74	.0002	.0001	.9999
2.75	.0060	.0030	.9970	3.25	.0012	.0006	.9994	3.75	.0002	.0001	.9999
2.76	.0058	.0029	.9971	3.26	.0011	.0006	.9994	3.76	.0002	.0001	.9999
2.77	.0056	.0028	.9972	3.27	.0011	.0005	.9995	3.77	.0002	.0001	.9999
2.78	.0054	.0027	.9973	3.28	.0010	.0005	.9995	3.78	.0002	.0001	.9999
2.79	.0053	.0026	.9974	3.29	.0010	.0005	.9995	3.79	.0002	.0001	.9999

Table A.2 Critical Values (Percentiles) for the Standard Normal Distribution
The fourth column is the $N(0, 1)$ percentile for the percent given in column one. It is also the upper one-sided $N(0, 1)$ critical value and two-sided $N(0, 1)$ critical value for the significance levels given in columns two and three, respectively.

Percent	One-sided	Two-sided	z	Percent	One-sided	Two-sided	z
50	.50	1.00	0.00	99.59	.0041	.0082	2.64
55	.45	.90	0.13	99.60	.0040	.0080	2.65
60	.40	.80	0.25	99.61	.0039	.0078	2.66
65	.35	.70	0.39	99.62	.0038	.0076	2.67
70	.30	.60	0.52	99.63	.0037	.0074	2.68
75	.25	.50	0.67	99.64	.0036	.0072	2.69
80	.20	.40	0.84	99.65	.0035	.0070	2.70
85	.15	.30	1.04	99.66	.0034	.0068	2.71
90	.10	.20	1.28	99.67	.0033	.0066	2.72
91	.09	.18	1.34	99.68	.0032	.0064	2.73
92	.08	.16	1.41	99.69	.0031	.0062	2.74
93	.07	.14	1.48	99.70	.0030	.0060	2.75
94	.06	.12	1.55	99.71	.0029	.0058	2.76
95	.05	.10	1.64	99.72	.0028	.0056	2.77
95.5	.045	.090	1.70	99.73	.0027	.0054	2.78
96.0	.040	.080	1.75	99.74	.0026	.0052	2.79
96.5	.035	.070	1.81	99.75	.0025	.0050	2.81
97.0	.030	.060	1.88	99.76	.0024	.0048	2.82
97.5	.025	.050	1.96	99.77	.0023	.0046	2.83
98.0	.020	.040	2.05	99.78	.0022	.0044	2.85
98.5	.015	.030	2.17	99.79	.0021	.0042	2.86
99.0	.010	.020	2.33	99.80	.0020	.0040	2.88
99.05	.0095	.0190	2.35	99.81	.0019	.0038	2.89
99.10	.0090	.0180	2.37	99.82	.0018	.0036	2.91
99.15	.0085	.0170	2.39	99.83	.0017	.0034	2.93
99.20	.0080	.0160	2.41	99.84	.0016	.0032	2.95
99.25	.0075	.0150	2.43	99.85	.0015	.0030	2.97
99.30	.0070	.0140	2.46	99.86	.0014	.0028	2.99
99.35	.0065	.0130	2.48	99.87	.0013	.0026	3.01
99.40	.0060	.0120	2.51	99.88	.0012	.0024	3.04
99.45	.0055	.0110	2.54	99.89	.0011	.0022	3.06
99.50	.0050	.0100	2.58	99.90	.0010	.0020	3.09
99.51	.0049	.0098	2.58	99.91	.0009	.0018	3.12
99.52	.0048	.0096	2.59	99.92	.0008	.0016	3.16
99.53	.0047	.0094	2.60	99.93	.0007	.0014	3.19
99.54	.0046	.0092	2.60	99.94	.0006	.0012	3.24
99.55	.0045	.0090	2.61	99.95	.0005	.0010	3.29
99.56	.0044	.0088	2.62	99.96	.0004	.0008	3.35
99.57	.0043	.0086	2.63	99.97	.0003	.0006	3.43
99.58	.0042	.0084	2.64	99.98	.0002	.0004	3.54
				99.99	.0001	.0002	3.72

Table A.3 Critical Values (Percentiles) for the Chi-Square Distribution

For each degree of freedom (d.f.) in the first column, the table entries are the critical values for the upper one-sided significance levels in the column headings or, equivalently, the percentiles for the corresponding percentages.

	Percentage								
	2.5	5	50	75	90	95	97.5	99	99.9
	Upper One-Sided α								
d.f.	.975	.95	.50	.25	.10	.05	.025	.01	.001
1	.001	.004	.455	1.32	2.71	3.84	5.02	6.63	10.83
2	.051	.103	1.39	2.77	4.61	5.99	7.38	9.21	13.82
3	.216	.352	2.37	4.11	6.25	7.82	9.35	11.34	16.27
4	.484	.711	3.36	5.39	7.78	9.49	11.14	13.28	18.47
5	.831	1.15	4.35	6.63	9.24	11.07	12.83	15.09	20.52
6	1.24	1.64	5.35	7.84	10.64	12.59	14.45	16.81	22.46
7	1.69	2.17	6.35	9.04	12.02	14.07	16.01	18.47	24.32
8	2.18	2.73	7.34	10.22	13.36	15.51	17.53	20.09	26.12
9	2.70	3.33	8.34	11.39	14.68	16.92	19.02	21.67	27.88
10	3.25	3.94	9.34	12.55	15.99	18.31	20.48	23.21	29.59
11	3.82	4.57	10.34	13.70	17.27	19.68	21.92	24.72	31.26
12	4.40	5.23	11.34	14.85	18.55	21.03	23.34	26.22	32.91
13	5.01	5.89	12.34	15.98	19.81	22.36	24.74	27.69	34.53
14	5.63	6.57	13.34	17.12	21.06	23.68	26.12	29.14	36.12
15	6.26	7.26	14.34	18.25	22.31	25.00	27.49	30.58	37.70
16	6.91	7.96	15.34	19.37	23.54	26.30	28.85	32.00	39.25
17	7.56	8.67	16.34	20.49	24.77	27.59	30.19	33.41	40.79
18	8.23	9.39	17.34	21.60	25.99	28.87	31.53	34.81	42.31
19	8.91	10.12	18.34	22.72	27.20	30.14	32.85	36.19	43.82
20	9.59	10.85	19.34	23.83	28.41	31.41	34.17	37.57	45.31
21	10.28	11.59	20.34	24.93	29.62	32.67	35.48	38.93	46.80
22	10.98	12.34	21.34	26.04	30.81	33.92	36.78	40.29	48.27
23	11.69	13.09	22.34	27.14	32.01	35.17	38.08	41.64	49.73
24	12.40	13.85	23.34	28.24	33.20	36.42	39.36	42.98	51.18
25	13.12	14.61	24.34	29.34	34.38	37.65	40.65	44.31	52.62
26	13.84	15.38	25.34	30.43	35.56	38.89	41.92	45.64	54.05
27	14.57	16.15	26.34	31.53	36.74	40.11	43.19	46.96	55.48
28	15.31	16.93	27.34	32.62	37.92	41.34	44.46	48.28	56.89
29	16.05	17.71	28.34	33.71	39.09	42.56	45.72	49.59	58.30
30	16.79	18.49	29.34	34.80	40.26	43.77	46.98	50.89	59.70
35	20.57	22.47	34.34	40.22	46.06	49.80	53.20	57.34	66.62
40	24.43	26.51	39.34	45.62	51.81	55.76	59.34	63.69	73.40
45	28.37	30.61	44.34	50.98	57.51	61.66	65.41	69.96	80.08
50	32.36	34.76	49.33	56.33	63.17	67.50	71.42	76.15	86.66
55	36.40	38.96	54.33	61.66	68.80	73.31	77.38	82.29	93.17
60	40.48	43.19	59.33	66.98	74.40	79.08	83.30	88.38	99.61
65	44.60	47.45	64.33	72.28	79.97	84.82	89.18	94.42	105.99
70	48.76	51.74	69.33	77.58	85.53	90.53	95.02	100.43	112.32
75	52.94	56.05	74.33	82.86	91.06	96.22	100.84	106.39	118.60
80	57.15	60.39	79.33	88.13	96.58	101.88	106.63	112.33	124.84
85	61.39	64.75	84.33	93.39	102.08	107.52	112.39	118.24	131.04
90	65.65	69.13	89.33	98.65	107.57	113.15	118.14	124.12	137.21
95	69.92	73.52	94.33	103.90	113.04	118.75	123.86	129.97	143.34
100	74.22	77.93	99.33	109.14	118.50	124.34	129.56	135.81	149.45

For more than 100 degrees of freedom chi-square critical values may be found in terms of the degrees of freedom and the corresponding two-sided critical value for a standard normal deviate Z by the equation $X^2 = 0.5 \cdot (Z + \sqrt{2 \cdot D - 1})^2$.

Table A.4 Critical Values (Percentiles) for the t-Distribution

The table entries are the critical values (percentiles) for the t-distribution. The column headed d.f. (degrees of freedom) gives the degrees of freedom for the values in that row. The columns are labeled by "percent," "one-sided," and "two-sided." "Percent" is 100 × cumulative distribution function—the table entry is the corresponding percentile. "One-sided" is the significance level for the one-sided upper critical value—the table entry is the critical value. "Two-sided" gives the two-sided significance level—the table entry is the corresponding two-sided critical value.

	Percent											
	75	90	95	97.5	99	99.5	99.75	99.9	99.95	99.975	99.99	99.995
	One-Sided α											
	.25	.10	.05	.025	.01	.005	.0025	.001	.0005	.00025	.0001	.00005
	Two-Sided α											
	.50	.20	.10	.05	.02	.01	.005	.002	.001	.0005	.0002	.0001
d.f.												
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62	1273.24	3183.10	6366.20
2	.82	1.89	2.92	4.30	6.96	9.22	14.09	22.33	31.60	44.70	70.70	99.99
3	.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92	16.33	22.20	28.00
4	.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61	10.31	13.03	15.54
5	.73	1.48	2.02	2.57	3.37	4.03	4.77	5.89	6.87	7.98	9.68	11.18
6	.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96	6.79	8.02	9.08
7	.71	1.42	1.90	2.37	3.00	3.50	4.03	4.79	5.41	6.08	7.06	7.88
8	.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04	5.62	6.44	7.12
9	.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78	5.29	6.01	6.59
10	.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59	5.05	5.69	6.21
11	.70	1.36	1.80	2.20	2.72	3.11	3.50	4.03	4.44	4.86	5.45	5.92
12	.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32	4.72	5.26	5.69
13	.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22	4.60	5.11	5.51
14	.69	1.35	1.76	2.15	2.63	2.98	3.33	3.79	4.14	4.50	4.99	5.36
15	.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07	4.42	4.88	5.24
16	.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02	4.35	4.79	5.13
17	.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.97	4.29	4.71	5.04
18	.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92	4.23	4.65	4.97
19	.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88	4.19	4.59	4.90
20	.69	1.33	1.73	2.09	2.53	2.85	3.15	3.55	3.85	4.15	4.54	4.84
21	.69	1.32	1.72	2.08	2.52	2.83	3.14	3.53	3.82	4.11	4.49	4.78
22	.69	1.32	1.72	2.07	2.51	2.82	3.12	3.51	3.79	4.08	4.45	4.74
23	.68	1.32	1.71	2.07	2.50	2.81	3.10	3.49	3.77	4.05	4.42	4.69
24	.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.75	4.02	4.38	4.65
25	.68	1.32	1.71	2.06	2.49	2.79	3.08	3.45	3.73	4.00	4.35	4.62
26	.68	1.32	1.71	2.06	2.48	2.78	3.07	3.44	3.71	3.97	4.32	4.59
27	.68	1.31	1.70	2.05	2.47	2.77	3.06	3.42	3.69	3.95	4.30	4.56
28	.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67	3.94	4.28	4.53
29	.68	1.31	1.70	2.05	2.46	2.76	3.04	3.40	3.66	3.92	4.25	4.51
30	.68	1.31	1.70	2.04	2.46	2.75	3.03	3.39	3.65	3.90	4.23	4.48
35	.68	1.31	1.69	2.03	2.44	2.72	3.00	3.34	3.59	3.84	4.15	4.39
40	.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55	3.79	4.09	4.32
45	.68	1.30	1.68	2.01	2.41	2.69	2.95	3.28	3.52	3.75	4.05	4.27
50	.68	1.30	1.68	2.01	2.40	2.68	2.94	3.26	3.50	3.72	4.01	4.23
55	.68	1.30	1.67	2.00	2.40	2.67	2.93	3.25	3.48	3.70	3.99	4.20
60	.68	1.30	1.67	2.00	2.39	2.66	2.91	3.23	3.46	3.68	3.96	4.17
65	.68	1.29	1.67	2.00	2.39	2.65	2.91	3.22	3.45	3.66	3.94	4.15
70	.68	1.29	1.67	1.99	2.38	2.65	2.90	3.21	3.44	3.65	3.93	4.13
75	.68	1.29	1.67	1.99	2.38	2.64	2.89	3.20	3.43	3.64	3.91	4.11
80	.68	1.29	1.66	1.99	2.37	2.64	2.89	3.20	3.42	3.63	3.90	4.10
85	.68	1.29	1.66	1.99	2.37	2.64	2.88	3.19	3.41	3.62	3.89	4.08
90	.68	1.29	1.66	1.99	2.37	2.63	2.88	3.18	3.40	3.61	3.88	4.07
95	.68	1.29	1.66	1.99	2.37	2.63	2.87	3.18	3.40	3.60	3.87	4.06
100	.68	1.29	1.66	1.98	2.36	2.63	2.87	3.17	3.39	3.60	3.86	4.05
200	.68	1.29	1.65	1.97	2.35	2.60	2.84	3.13	3.34	3.54	3.79	3.97
500	.68	1.28	1.65	1.97	2.33	2.59	2.82	3.11	3.31	3.50	3.75	3.92
∞	.67	1.28	1.65	1.96	2.33	2.58	2.81	3.10	3.30	3.49	3.73	3.91

Table A.5 Critical Values (Percentiles) for the F-Distribution

Upper one-sided 0.05 significance levels; two-sided 0.10 significance levels; 95% percentiles. Tabulated are critical values for the F-distribution. The column headings give the numerator degrees of freedom and the row headings the denominator degrees of freedom. Lower one-sided critical values may be found from these tables by reversing the degrees of freedom and using the reciprocal of the tabled value at the same significance level (100 minus the percent for the percentile).

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	6.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96

(*continued overleaf*)

Table A.5 (*continued*)

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table A.6 Critical Values (Percentiles) for the F-Distribution

Upper one-sided 0.01 significance levels; two-sided 0.02 significance levels; 99% percentiles.

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65

(*continued overleaf*)

Table A.6 (***continued***)

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table A.7 Fisher's Exact Test for 2 × 2 Tables

Consider a 2 × 2 table: $\begin{matrix} aA-a|A \\ bB-b|B \end{matrix}$ with rows and/or columns exchanged so that (1) $A \geq B$ and (2) $(a/A) \geq (b/B)$. The table entries are ordered lexicographically by A (ascending). B (descending) and a (descending). For each triple (A, B, a) the table presents critical values for one-sided tests of the hypothesis that the true proportion corresponding to a/A is greater than the true proportion corresponding to b/B. Significance levels of 0.05, 0.025, and 0.01 are considered. For $A \leq 15$ all values where critical values exist are tabulated. For each significance level two columns give (1) the nominal critical value for b (i.e., reject the null hypothesis if the observed b is less than or equal to the table entry) and (2) the p-value corresponding to the critical value (this is less than the nominal significance level in most cases due to the discreteness of the distribution).

A	B	a	b	p	b	p	b	p	A	B	a	b	p	b	p	b	p
3	3	3	0	.050	—	—	—	—	8	7	5	0	.019	0	.019	—	—
4	4	4	0	.014	0	.014	—	—	8	6	8	2	.015	2	.015	1	.003
4	3	4	0	.029	—	—	—	—	8	6	7	1	.016	1	.016	0	.002
5	5	5	1	.024	1	.024	0	.004	8	6	6	0	.009	0	.009	0	.009
5	5	4	0	.024	0	.024	—	—	8	6	5	0	.028	—	—	—	—
5	4	5	1	.048	0	.008	0	.008	8	5	8	2	.035	1	.007	1	.007
5	4	4	0	.040	—	—	—	—	8	5	7	1	.032	0	.005	0	.005
5	3	5	0	.018	0	.018	—	—	8	5	6	0	.016	0	.016	—	—
5	2	5	0	.048	—	—	—	—	8	5	5	0	.044	—	—	—	—
6	6	6	2	.030	1	.008	1	.008	8	4	8	1	.018	1	.018	0	.002
6	6	5	1	.040	0	.008	0	.008	8	4	7	0	.010	0	.010	—	—
6	6	4	0	.030	—	—	—	—	8	4	6	0	.030	—	—	—	—
6	5	6	1	.015	1	.015	0	.002	8	3	8	0	.006	0	.006	0	.006
6	5	5	0	.013	0	.013	—	—	8	3	7	0	.024	0	.024	—	—
6	5	4	0	.045	—	—	—	—	8	2	8	0	.022	0	.022	—	—
6	4	6	1	.033	0	.005	0	.005	9	9	9	5	.041	4	.015	3	.005
6	4	5	0	.024	0	.024	—	—	9	9	8	3	.025	3	.025	2	.008
6	3	6	0	.012	0	.012	—	—	9	9	7	2	.028	1	.008	1	.008
6	3	5	0	.048	—	—	—	—	9	9	6	1	.025	1	.025	0	.005
6	2	6	0	.036	—	—	—	—	9	9	5	0	.015	0	.015	—	—
7	7	7	3	.035	2	.010	1	.002	9	9	4	0	.041	—	—	—	—
7	7	6	1	.015	1	.015	0	.002	9	8	9	4	.029	3	.009	3	.009
7	7	5	0	.010	0	.010	—	—	9	8	8	3	.043	2	.013	1	.003
7	7	4	0	.035	—	—	—	—	9	8	7	2	.044	1	.012	0	.002
7	6	7	2	.021	2	.021	1	.005	9	8	6	1	.036	0	.007	0	.007
7	6	6	1	.025	0	.004	0	.004	9	8	5	0	.020	0	.020	—	—
7	6	5	0	.016	0	.016	—	—	9	7	9	3	.019	3	.019	2	.005
7	6	4	0	.049	—	—	—	—	9	7	8	2	.024	2	.024	1	.006
7	5	7	2	.045	1	.010	0	.001	9	7	7	1	.020	1	.020	0	.003
7	5	6	1	.045	0	.008	0	.008	9	7	6	0	.010	0	.010	—	—
7	5	5	0	.027	—	—	—	—	9	7	5	0	.029	—	—	—	—
7	4	7	1	.024	1	.024	0	.003	9	6	9	3	.044	2	.011	1	.002
7	4	6	0	.015	0	.015	—	—	9	6	8	2	.047	1	.011	0	.001
7	4	5	0	.045	—	—	—	—	9	6	7	1	.035	0	.006	0	.006
7	3	7	0	.008	0	.008	0	.008	9	6	6	0	.017	0	.017	—	—
7	3	6	0	.033	—	—	—	—	9	6	5	0	.042	—	—	—	—
7	2	7	0	.028	—	—	—	—	9	5	9	2	.027	1	.005	1	.005
8	8	8	4	.038	3	.013	2	.003	9	5	8	1	.023	1	.023	0	.003
8	8	7	2	.020	2	.020	1	.005	9	5	7	0	.010	0	.010	—	—
8	8	6	1	.020	1	.020	0	.003	9	5	6	0	.028	—	—	—	—
8	8	5	0	.013	0	.013	—	—	9	4	9	1	.014	1	.014	0	.001
8	8	4	0	.038	—	—	—	—	9	4	8	0	.007	0	.007	0	.007
8	7	8	3	.026	2	.007	2	.007	9	4	7	0	.021	0	.021	—	—
8	7	7	2	.035	1	.009	1	.009	9	4	6	0	.049	—	—	—	—
8	7	6	1	.032	0	.006	0	.006	9	3	9	1	.045	0	.005	0	.005

(*continued overleaf*)

Table A.7 (*continued*)

A	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*	*A*	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*
9	3	8	0	.018	0	.018	—	—	11	11	8	3	.043	2	.015	1	.004
9	3	7	0	.045	—	—	—	—	11	11	7	2	.040	1	.012	0	.002
9	2	9	0	.018	0	.018	—	—	11	11	6	1	.032	0	.006	0	.006
10	10	10	6	.043	5	.016	4	.005	11	11	5	0	.018	0	.018	—	—
10	10	9	4	.029	3	.010	3	.010	11	11	4	0	.045	—	—	—	—
10	10	8	3	.035	2	.012	1	.003	11	10	11	6	.035	5	.012	4	.004
10	10	7	2	.035	1	.010	1	.010	11	10	10	4	.021	4	.021	3	.007
10	10	6	1	.029	0	.005	0	.005	11	10	9	3	.024	3	.024	2	.007
10	10	5	0	.016	0	.016	—	—	11	10	8	2	.023	2	.023	1	.006
10	10	4	0	.043	—	—	—	—	11	10	7	1	.017	1	.017	0	.003
10	9	10	5	.033	4	.011	3	.003	11	10	6	1	.043	0	.009	0	.009
10	9	9	4	.050	3	.017	2	.005	11	10	5	0	.023	0	.023	—	—
10	9	8	2	.019	2	.019	1	.004	11	9	11	5	.026	4	.008	4	.008
10	9	7	1	.015	1	.015	0	.002	11	9	10	4	.038	3	.012	2	.003
10	9	6	1	.040	0	.008	0	.008	11	9	9	3	.040	2	.012	1	.003
10	9	5	0	.022	0	.022	—	—	11	9	8	2	.035	1	.009	1	.009
10	8	10	4	.023	4	.023	3	.007	11	9	7	1	.025	1	.025	0	.004
10	8	9	3	.032	2	.009	2	.009	11	9	6	0	.012	0	.012	—	—
10	8	8	2	.031	1	.008	1	.008	11	9	5	0	.030	—	—	—	—
10	8	7	1	.023	1	.023	0	.004	11	8	11	4	.018	4	.018	3	.005
10	8	6	0	.011	0	.011	—	—	11	8	10	3	.024	3	.024	2	.006
10	8	5	0	.029	—	—	—	—	11	8	9	2	.022	2	.022	1	.005
10	7	10	3	.015	3	.015	2	.003	11	8	8	1	.015	1	.015	0	.002
10	7	9	2	.018	2	.018	1	.004	11	8	7	1	.037	0	.007	0	.007
10	7	8	1	.013	1	.013	0	.002	11	8	6	0	.017	0	.017	—	—
10	7	7	1	.036	0	.006	0	.006	11	8	5	0	.040	—	—	—	—
10	7	6	0	.017	0	.017	—	—	11	7	11	4	.043	3	.011	2	.002
10	7	5	0	.041	—	—	—	—	11	7	10	3	.047	2	.013	1	.002
10	6	10	3	.036	2	.008	2	.008	11	7	9	2	.039	1	.009	1	.009
10	6	9	2	.036	1	.008	1	.008	11	7	8	1	.025	1	.025	0	.004
10	6	8	1	.024	1	.024	0	.003	11	7	7	0	.010	0	.010	—	—
10	6	7	0	.010	0	.010	—	—	11	7	6	0	.025	0	.025	—	—
10	6	6	0	.026	—	—	—	—	11	6	11	3	.029	2	.006	2	.006
10	5	10	2	.022	2	.022	1	.004	11	6	10	2	.028	1	.005	1	.005
10	5	9	1	.017	1	.017	0	.002	11	6	9	1	.018	1	.018	0	.002
10	5	8	1	.047	0	.007	0	.007	11	6	8	1	.043	0	.007	0	.007
10	5	7	0	.019	0	.019	—	—	11	6	7	0	.017	0	.017	—	—
10	5	6	0	.042	—	—	—	—	11	6	6	0	.037	—	—	—	—
10	4	10	1	.011	1	.011	0	.001	11	5	11	2	.018	2	.018	1	.003
10	4	9	1	.041	0	.005	0	.005	11	5	10	1	.013	1	.013	0	.001
10	4	8	0	.015	0	.015	—	—	11	5	9	1	.036	0	.005	0	.005
10	4	7	0	.035	—	—	—	—	11	5	8	0	.013	0	.013	—	—
10	3	10	1	.038	0	.003	0	.003	11	5	7	0	.029	—	—	—	—
10	3	9	0	.014	0	.014	—	—	11	4	11	1	.009	1	.009	1	.009
10	3	8	0	.035	—	—	—	—	11	4	10	1	.033	0	.004	0	.004
10	2	10	0	.015	0	.015	—	—	11	4	9	0	.011	0	.011	—	—
10	2	9	0	.045	—	—	—	—	11	4	8	0	.026	—	—	—	—
11	11	11	7	.045	6	.018	5	.006	11	3	11	1	.033	0	.003	0	.003
11	11	10	5	.032	4	.012	3	.004	11	3	10	0	.011	0	.011	—	—
11	11	9	4	.040	3	.015	2	.004	11	3	9	0	.027	—	—	—	—

Table A.7 (*continued*)

A	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*	*A*	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*
11	2	11	0	.013	0	.013	—	—	12	6	11	2	.022	2	.022	1	.004
11	2	10	0	.038	—	—	—	—	12	6	10	1	.013	1	.013	0	.002
12	12	12	8	.047	7	.019	6	.007	12	6	9	1	.032	0	.005	0	.005
12	12	11	6	.034	5	.014	4	.005	12	6	8	0	.011	0	.011	—	—
12	12	10	5	.045	4	.018	3	.006	12	6	7	0	.025	0	.025	—	—
12	12	9	4	.050	3	.020	2	.006	12	6	6	0	.050	—	—	—	—
12	12	8	3	.050	2	.018	1	.005	12	5	12	2	.015	2	.015	1	.002
12	12	7	2	.045	1	.014	0	.002	12	5	11	1	.010	1	.010	1	.010
12	12	6	1	.034	0	.007	0	.007	12	5	10	1	.028	0	.003	0	.003
12	12	5	0	.019	0	.019	—	—	12	5	9	0	.009	0	.009	0	.009
12	12	4	0	.047	—	—	—	—	12	5	8	0	.020	0	.020	—	—
12	11	12	7	.037	6	.014	5	.005	12	5	7	0	.041	—	—	—	—
12	11	11	5	.024	5	.024	4	.008	12	4	12	2	.050	1	.007	1	.007
12	11	10	4	.029	3	.010	2	.003	12	4	11	1	.027	0	.003	0	.003
12	11	9	3	.030	2	.009	2	.009	12	4	10	0	.008	0	.008	0	.008
12	11	8	2	.026	1	.007	1	.007	12	4	9	0	.019	0	.019	—	—
12	11	7	1	.019	1	.019	0	.003	12	4	8	0	.038	—	—	—	—
12	11	6	1	.045	0	.009	0	.009	12	3	12	1	.029	0	.002	0	.002
12	11	5	0	.024	0	.024	—	—	12	3	11	0	.009	0	.009	0	.009
12	10	12	6	.029	5	.010	5	.010	12	3	10	0	.022	0	.022	—	—
12	10	11	5	.043	4	.015	3	.005	12	3	9	0	.044	—	—	—	—
12	10	10	4	.048	3	.017	2	.005	12	2	12	0	.011	0	.011	—	—
12	10	9	3	.046	2	.015	1	.004	12	2	11	0	.033	—	—	—	—
12	10	8	2	.038	1	.010	0	.002	13	13	13	9	.048	8	.020	7	.007
12	10	7	1	.026	0	.005	0	.005	13	13	12	7	.037	6	.015	5	.006
12	10	6	0	.012	0	.012	—	—	13	13	11	6	.048	5	.021	4	.008
12	10	5	0	.030	—	—	—	—	13	13	10	4	.024	4	.024	3	.008
12	9	12	5	.021	5	.021	4	.006	13	13	9	3	.024	3	.024	2	.008
12	9	11	4	.029	3	.009	3	.009	13	13	8	2	.021	2	.021	1	.006
12	9	10	3	.029	2	.008	2	.008	13	13	7	2	.048	1	.015	0	.003
12	9	9	2	.024	2	.024	1	.006	13	13	6	1	.037	0	.007	0	.007
12	9	8	1	.016	1	.016	0	.002	13	13	5	0	.020	0	.020	—	—
12	9	7	1	.037	0	.007	0	.007	13	13	4	0	.048	—	—	—	—
12	9	6	0	.017	0	.017	—	—	13	12	13	8	.039	7	.015	6	.005
12	9	5	0	.039	—	—	—	—	13	12	12	6	.027	5	.010	5	.010
12	8	12	5	.049	4	.014	3	.004	13	12	11	5	.033	4	.013	3	.004
12	8	11	3	.018	3	.018	2	.004	13	12	10	4	.036	3	.013	2	.004
12	8	10	2	.015	2	.015	1	.003	13	12	9	3	.034	2	.011	1	.003
12	8	9	2	.040	1	.010	1	.010	13	12	8	2	.029	1	.008	1	.008
12	8	8	1	.025	1	.025	0	.004	13	12	7	1	.020	1	.020	0	.004
12	8	7	0	.010	0	.010	—	—	13	12	6	1	.046	0	.010	0	.010
12	8	6	0	.024	0	.024	—	—	13	12	5	0	.024	0	.024	—	—
12	7	12	4	.036	3	.009	3	.009	13	11	13	7	.031	6	.011	5	.003
12	7	11	3	.038	2	.010	2	.010	13	11	12	6	.048	5	.018	4	.006
12	7	10	2	.029	1	.006	1	.006	13	11	11	4	.021	4	.021	3	.007
12	7	9	1	.017	1	.017	0	.002	13	11	10	3	.021	3	.021	2	.006
12	7	8	1	.040	0	.007	0	.007	13	11	9	3	.050	2	.017	1	.004
12	7	7	0	.016	0	.016	—	—	13	11	8	2	.040	1	.011	0	.002
12	7	6	0	.034	—	—	—	—	13	11	7	1	.027	0	.005	0	.005
12	6	12	3	.025	3	.025	2	.005	13	11	6	0	.013	0	.013	—	—

(*continued overleaf*)

Table A.7 (*continued*)

A	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*	*A*	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*
13	11	5	0	.030	—	—	—	—	13	4	11	0	.006	0	.006	0	.006
13	10	13	6	.024	6	.024	5	.007	13	4	10	0	.015	0	.015	—	—
13	10	12	5	.035	4	.012	3	.003	13	4	9	0	.029	—	—	—	—
13	10	11	4	.037	3	.012	2	.003	13	3	13	1	.025	1	.025	0	.002
13	10	10	3	.033	2	.010	1	.002	13	3	12	0	.007	0	.007	0	.007
13	10	9	2	.026	1	.006	1	.006	13	3	11	0	.018	0	.018	—	—
13	10	8	1	.017	1	.017	0	.003	13	3	10	0	.036	—	—	—	—
13	10	7	1	.038	0	.007	0	.007	13	2	13	0	.010	0	.010	0	.010
13	10	6	0	.017	0	.017	—	—	13	2	12	0	.029	—	—	—	—
13	10	5	0	.038	—	—	—	—	14	14	14	10	.049	9	.020	8	.008
13	9	13	5	.017	5	.017	4	.005	14	14	13	8	.038	7	.016	6	.006
13	9	12	4	.023	4	.023	3	.007	14	14	12	6	.023	6	.023	5	.009
13	9	11	3	.022	3	.022	2	.006	14	14	11	5	.027	4	.011	3	.004
13	9	10	2	.017	2	.017	1	.004	14	14	10	4	.028	3	.011	2	.003
13	9	9	2	.040	1	.010	0	.001	14	14	9	3	.027	2	.009	2	.009
13	9	8	1	.025	1	.025	0	.004	14	14	8	2	.023	2	.023	1	.006
13	9	7	0	.010	0	.010	—	—	14	14	7	1	.016	1	.016	0	.003
13	9	6	0	.023	0	.023	—	—	14	14	6	1	.038	0	.008	0	.008
13	9	5	0	.049	—	—	—	—	14	14	5	0	.020	0	.020	—	—
13	8	13	5	.042	4	.012	3	.003	14	14	4	0	.049	—	—	—	—
13	8	12	4	.047	3	.014	2	.003	14	13	14	9	.041	8	.016	7	.006
13	8	11	3	.041	2	.011	1	.002	14	13	13	7	.029	6	.011	5	.004
13	8	10	2	.029	1	.007	1	.007	14	13	12	6	.037	5	.015	4	.005
13	8	9	1	.017	1	.017	0	.002	14	13	11	5	.041	4	.017	3	.006
13	8	8	1	.037	0	.006	0	.006	14	13	10	4	.041	3	.016	2	.005
13	8	7	0	.015	0	.015	—	—	14	13	9	3	.038	2	.013	1	.003
13	8	6	0	.032	—	—	—	—	14	13	8	2	.031	1	.009	1	.009
13	7	13	4	.031	3	.007	3	.007	14	13	7	1	.021	1	.021	0	.004
13	7	12	3	.031	2	.007	2	.007	14	13	6	1	.048	0	.010	—	—
13	7	11	2	.022	2	.022	1	.004	14	13	5	0	.025	0	.025	—	—
13	7	10	1	.012	1	.012	0	.002	14	12	14	8	.033	7	.012	6	.004
13	7	9	1	.029	0	.004	0	.004	14	12	13	6	.021	6	.021	5	.007
13	7	8	0	.010	0	.010	—	—	14	12	12	5	.025	4	.009	4	.009
13	7	7	0	.022	0	.022	—	—	14	12	11	4	.026	3	.009	3	.009
13	7	6	0	.044	—	—	—	—	14	12	10	3	.024	3	.024	2	.007
13	6	13	3	.021	3	.021	2	.004	14	12	9	2	.019	2	.019	1	.005
13	6	12	2	.017	2	.017	1	.003	14	12	8	2	.042	1	.012	0	.002
13	6	11	2	.046	1	.010	1	.010	14	12	7	1	.028	0	.005	0	.005
13	6	10	1	.024	1	.024	0	.003	14	12	6	0	.013	0	.013	—	—
13	6	9	1	.050	0	.008	0	.008	14	12	5	0	.030	—	—	—	—
13	6	8	0	.017	0	.017	—	—	14	11	14	7	.026	6	.009	6	.009
13	6	7	0	.034	—	—	—	—	14	11	13	6	.039	5	.014	4	.004
13	5	13	2	.012	2	.012	1	.002	14	11	12	5	.043	4	.016	3	.005
13	5	12	2	.044	1	.008	1	.008	14	11	11	4	.042	3	.015	2	.004
13	5	11	1	.022	1	.022	0	.002	14	11	10	3	.036	2	.011	1	.003
13	5	10	1	.047	0	.007	0	.007	14	11	9	2	.027	1	.007	1	.007
13	5	9	0	.015	0	.015	—	—	14	11	8	1	.017	1	.017	0	.003
13	5	8	0	.029	—	—	—	—	14	11	7	1	.038	0	.007	0	.007
13	4	13	2	.044	1	.006	1	.006	14	11	6	0	.017	0	.017	—	—
13	4	12	1	.022	1	.022	0	.002	14	11	5	0	.038	—	—	—	—

Table A.7 (***continued***)

A	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*	*A*	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*
14	10	14	6	.020	6	.020	5	.006	14	5	8	0	.040	—	—	—	—
14	10	13	5	.028	4	.009	4	.009	14	4	14	2	.039	1	.005	1	.005
14	10	12	4	.028	3	.009	3	.009	14	4	13	1	.019	1	.019	0	.002
14	10	11	3	.024	3	.024	2	.007	14	4	12	1	.044	0	.005	0	.005
14	10	10	2	.018	2	.018	1	.004	14	4	11	0	.011	0	.011	—	—
14	10	9	2	.040	1	.011	0	.002	14	4	10	0	.023	0	.023	—	—
14	10	8	1	.024	1	.024	0	.004	14	4	9	0	.041	—	—	—	—
14	10	7	0	.010	0	.010	0	.010	14	3	14	1	.022	1	.022	0	.001
14	10	6	0	.022	0	.022	—	—	14	3	13	0	.006	0	.006	0	.006
14	10	5	0	.047	—	—	—	—	14	3	12	0	.015	0	.015	—	—
14	9	14	6	.047	5	.014	4	.004	14	3	11	0	.029	—	—	—	—
14	9	13	4	.018	4	.018	3	.005	14	2	14	0	.008	0	.008	0	.008
14	9	12	3	.017	3	.017	2	.004	14	2	13	0	.025	0	.025	—	—
14	9	11	3	.042	2	.012	1	.002	14	2	12	0	.050	—	—	—	—
14	9	10	2	.029	1	.007	1	.007	15	15	15	11	.050	10	.021	9	.008
14	9	9	1	.017	1	.017	0	.002	15	15	14	9	.040	8	.018	7	.007
14	9	8	1	.036	0	.006	0	.006	15	15	13	7	.025	6	.010	5	.004
14	9	7	0	.014	0	.014	—	—	15	15	12	6	.030	5	.013	4	.005
14	9	6	0	.030	—	—	—	—	15	15	11	5	.033	4	.013	3	.005
14	8	14	5	.036	4	.010	4	.010	15	15	10	4	.033	3	.013	2	.004
14	8	13	4	.039	3	.011	2	.002	15	15	9	3	.030	2	.010	1	.003
14	8	12	3	.032	2	.008	2	.008	15	15	8	2	.025	1	.007	1	.007
14	8	11	2	.022	2	.022	1	.005	15	15	7	1	.018	1	.018	0	.003
14	8	10	2	.048	1	.012	0	.002	15	15	6	1	.040	0	.008	0	.008
14	8	9	1	.026	0	.004	0	.004	15	15	5	0	.021	0	.021	—	—
14	8	8	0	.009	0	.009	0	.009	15	15	4	0	.050	—	—	—	—
14	8	7	0	.020	0	.020	—	—	15	14	15	10	.042	9	.017	8	.006
14	8	6	0	.040	—	—	—	—	15	14	14	8	.031	7	.013	6	.005
14	7	14	4	.026	3	.006	3	.006	15	14	13	7	.041	6	.017	5	.007
14	7	13	3	.025	2	.006	2	.006	15	14	12	6	.046	5	.020	4	.007
14	7	12	2	.017	2	.017	1	.003	15	14	11	5	.048	4	.020	3	.007
14	7	11	2	.041	1	.009	1	.009	15	14	10	4	.046	3	.018	2	.006
14	7	10	1	.021	1	.021	0	.003	15	14	9	3	.041	2	.014	1	.004
14	7	9	1	.043	0	.007	0	.007	15	14	8	2	.033	1	.009	1	.009
14	7	8	0	.015	0	.015	—	—	15	14	7	1	.022	1	.022	0	.004
14	7	7	0	.030	—	—	—	—	15	14	6	1	.049	0	.011	—	—
14	6	14	3	.018	3	.018	2	.003	15	14	5	0	.025	—	—	—	—
14	6	13	2	.014	2	.014	1	.002	15	13	15	9	.035	8	.013	7	.005
14	6	12	2	.037	1	.007	1	.007	15	13	14	7	.023	7	.023	6	.009
14	6	11	1	.018	1	.018	0	.002	15	13	13	6	.029	5	.011	4	.004
14	6	10	1	.038	0	.005	0	.005	15	13	12	5	.031	4	.012	3	.004
14	6	9	0	.012	0	.012	—	—	15	13	11	4	.030	3	.011	2	.003
14	6	8	0	.024	0	.024	—	—	15	13	10	3	.026	2	.008	2	.008
14	6	7	0	.044	—	—	—	—	15	13	9	2	.020	2	.020	1	.005
14	5	14	2	.010	2	.010	1	.001	15	13	8	2	.043	1	.013	0	.002
14	5	13	2	.037	1	.006	1	.006	15	13	7	1	.029	0	.005	0	.005
14	5	12	1	.017	1	.017	0	.002	15	13	6	0	.013	0	.013	—	—
14	5	11	1	.038	0	.005	0	.005	15	13	5	0	.031	—	—	—	—
14	5	10	0	.011	0	.011	—	—	15	12	15	8	.028	7	.010	7	.010
14	5	9	0	.022	0	.022	—	—	15	12	14	7	.043	6	.016	5	.006

(*continued overleaf*)

Table A.7 (*continued*)

A	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*	*A*	*B*	*a*	*b*	*p*	*b*	*p*	*b*	*p*
15	12	13	6	.049	5	.019	4	.007	15	8	10	1	.019	1	.019	0	.003
15	12	12	5	.049	4	.019	3	.006	15	8	9	1	.038	0	.006	0	.006
15	12	11	4	.045	3	.017	2	.005	15	8	8	0	.013	0	.013	—	—
15	12	10	3	.038	2	.012	1	.003	15	8	7	0	.026	—	—	—	—
15	12	9	2	.028	1	.007	1	.007	15	8	6	0	.050	—	—	—	—
15	12	8	1	.018	1	.018	0	.003	15	7	15	4	.023	4	.023	3	.005
15	12	7	1	.038	0	.007	0	.007	15	7	14	3	.021	3	.021	2	.004
15	12	6	0	.017	0	.017	—	—	15	7	13	2	.014	2	.014	1	.002
15	12	5	0	.037	—	—	—	—	15	7	12	2	.032	1	.007	1	.007
15	11	15	7	.022	7	.022	6	.007	15	7	11	1	.015	1	.015	0	.002
15	11	14	6	.032	5	.011	4	.003	15	7	10	1	.032	0	.005	0	.005
15	11	13	5	.034	4	.012	3	.003	15	7	9	0	.010	0	.010	—	—
15	11	12	4	.032	3	.010	2	.003	15	7	8	0	.020	0	.020	—	—
15	11	11	3	.026	2	.008	2	.008	15	7	7	0	.038	—	—	—	—
15	11	10	2	.019	2	.019	1	.004	15	6	15	3	.015	3	.015	2	.003
15	11	9	2	.040	1	.011	0	.002	15	6	14	2	.011	2	.011	1	.002
15	11	8	1	.024	1	.024	0	.004	15	6	13	2	.031	1	.006	1	.006
15	11	7	1	.049	0	.010	0	.010	15	6	12	1	.014	1	.014	0	.002
15	11	6	0	.022	0	.022	—	—	15	6	11	1	.029	0	.004	0	.004
15	11	5	0	.046	—	—	—	—	15	6	10	0	.009	0	.009	0	.009
15	10	15	6	.017	6	.017	5	.005	15	6	9	0	.017	0	.017	—	—
15	10	14	5	.023	5	.023	4	.007	15	6	8	0	.032	—	—	—	—
15	10	13	4	.022	4	.022	3	.007	15	5	15	2	.009	2	.009	2	.009
15	10	12	3	.018	3	.018	2	.005	15	5	14	2	.032	1	.005	1	.005
15	10	11	3	.042	2	.013	1	.003	15	5	13	1	.014	1	.014	0	.001
15	10	10	2	.029	1	.007	1	.007	15	5	12	1	.031	0	.004	0	.004
15	10	9	1	.016	1	.016	0	.002	15	5	11	0	.008	0	.008	0	.008
15	10	8	1	.034	0	.006	0	.006	15	5	10	0	.016	0	.016	—	—
15	10	7	0	.013	0	.013	—	—	15	5	9	0	.030	—	—	—	—
15	10	6	0	.028	—	—	—	—	15	4	15	2	.035	1	.004	1	.004
15	9	15	6	.042	5	.012	4	.003	15	4	14	1	.016	1	.016	0	.001
15	9	14	5	.047	4	.015	3	.004	15	4	13	1	.037	0	.004	0	.004
15	9	13	4	.042	3	.013	2	.003	15	4	12	0	.009	0	.009	0	.009
15	9	12	3	.032	2	.009	2	.009	15	4	11	0	.018	0	.018	—	—
15	9	11	2	.021	2	.021	1	.005	15	4	10	0	.033	—	—	—	—
15	9	10	2	.045	1	.011	0	.002	15	3	15	1	.020	1	.020	0	.001
15	9	9	1	.024	1	.024	0	.004	15	3	14	0	.005	0	.005	0	.005
15	9	8	1	.048	0	.009	0	.009	15	3	13	0	.012	0	.012	—	—
15	9	7	0	.019	0	.019	—	—	15	3	12	0	.025	0	.025	—	—
15	9	6	0	.037	—	—	—	—	15	3	11	0	.043	—	—	—	—
15	8	15	5	.032	4	.008	4	.008	15	2	15	0	.007	0	.007	0	.007
15	8	14	4	.033	3	.009	3	.009	15	2	14	0	.022	0	.022	—	—
15	8	13	3	.026	2	.006	2	.006	15	2	13	0	.044	—	—	—	—
15	8	12	2	.017	2	.017	1	.003									
15	8	11	2	.037	1	.008	1	.008	23	10	21	5	.016	5	.016	4	.004
									32	13	32	10	.020	10	.020	9	.005

Table A.8 Sample Sizes for Comparing Two Proportions with a One-Sided Fisher's Exact Test in 2 × 2 Tables

Let P_A and P_B be the true proportions in two populations. The sample size, N, for two equally sized groups is tabulated for one-sided significance level α and probability β of not rejecting the null hypothesis. Each rectangular portion of the table contains sample sizes for *two* pairs of α and β values, one above the diagonal and one below it. The arcsine approximation was used to estimate N.

P_A	P_B							$\alpha = .01$ and $\beta = .01$						
	.001	.01	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
.001	—	2305	288	129	81	58	45	37	26	20	15	12	10	8
.01	1679	—	689	221	123	82	61	48	32	24	18	14	11	9
.05	210	502	—	1169	366	191	122	87	52	35	25	19	14	11
.10	94	161	852	—	1877	538	266	163	83	51	34	25	18	13
.15	59	90	266	1368	—	2489	683	327	132	73	46	31	22	15
.20	43	60	140	392	1814	—	3012	805	222	105	61	39	27	18
.25	33	44	89	194	498	2194	—	3447	417	158	83	50	32	21
.30	27	35	63	119	239	587	2511	—	981	256	116	64	39	25
.40	19	24	38	60	96	162	304	715	—	1068	267	116	61	34
.50	14	17	26	37	53	77	116	187	778	—	1068	256	105	51
.60	11	13	19	25	34	45	61	84	195	778	—	981	222	83
.70	9	10	14	18	23	29	37	47	84	187	715	—	805	163
.80	7	8	11	13	16	20	24	29	45	77	162	587	—	538
.90	6	6	8	10	11	13	15	18	25	37	60	119	392	—
							$\alpha = .01$ and $\beta = .05$ (or $\alpha = .05$ and $\beta = .01$)							

P_A	P_B						$\alpha = .025$ and $\beta = .05$ (or $\alpha = .05$ and $\beta = .025$)							
	.001	.01	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
.001	—	1384	173	78	49	35	27	22	16	12	9	8	6	5
.01	1119	—	414	133	74	50	37	29	20	14	11	9	7	5
.05	140	335	—	702	220	115	74	52	31	21	15	12	9	7
.10	63	108	568	—	1127	323	160	98	50	31	21	15	11	8
.15	40	60	178	911	—	1494	410	197	79	44	28	19	13	9
.20	29	40	93	261	1208	—	1808	483	133	63	37	24	16	11
.25	22	30	60	129	332	1462	—	2069	251	95	50	30	20	13
.30	18	23	42	79	159	391	1673	—	589	154	70	39	24	15
.40	13	16	25	40	64	108	203	476	—	641	161	70	37	21
.50	10	12	17	25	35	51	77	125	519	—	641	154	63	31
.60	8	9	13	17	23	30	40	56	130	519	—	589	133	50
.70	6	7	9	12	15	19	25	32	56	125	476	—	483	98
.80	5	6	7	9	11	13	16	19	30	51	108	391	—	323
.90	4	4	6	7	8	9	10	12	17	25	40	79	261	—
							$\alpha = .025$ and $\beta = .10$ (or $\alpha = .10$ and $\beta = .025$)							

P_A	P_B							$\alpha = .05$ and $\beta = .05$						
	.001	.01	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
.001	—	1152	144	65	41	29	23	19	13	10	8	6	5	4
.01	912	—	345	111	62	41	31	24	16	12	9	7	6	5
.05	114	273	—	585	183	96	61	44	26	18	13	10	7	6
.10	51	88	463	—	939	269	133	82	42	26	17	13	9	7
.15	32	49	145	743	—	1245	342	164	66	36	23	16	11	8
.20	23	33	76	213	985	—	1506	403	111	53	31	20	14	9
.25	18	24	49	106	271	1192	—	1723	209	79	42	25	16	11
.30	15	19	35	65	130	319	1364	—	491	128	58	32	20	13
.40	11	13	21	33	52	88	165	388	—	534	134	58	31	17
.50	8	10	14	20	29	42	63	102	423	—	534	128	53	26
.60	6	7	10	14	18	24	33	46	106	423	—	491	111	42
.70	5	6	8	10	13	16	20	26	46	102	388	—	403	82
.80	4	5	6	7	9	11	13	16	24	42	88	319	—	269
.90	3	4	5	5	6	7	9	10	14	20	33	65	213	—
							$\alpha = .05$ and $\beta = .10$ (or $\alpha = .10$ and $\beta = .05$)							

(*continued overleaf*)

Table A.8 (***continued***)

P_A	P_B	$\alpha = .10$ and $\beta = .10$												
	.001	.01	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
.001	—	700	88	40	25	18	14	11	8	6	5	4	3	3
.01	480	—	210	67	38	25	19	15	10	7	6	5	4	3
.05	60	144	—	355	111	58	37	27	16	11	8	6	5	4
.10	27	46	244	—	570	164	81	50	25	16	11	8	6	4
.15	17	26	77	391	—	756	208	100	40	22	14	10	7	5
.20	13	18	40	112	519	—	914	245	68	32	19	12	8	6
.25	10	13	26	56	143	628	—	1046	127	48	25	16	10	7
.30	8	10	18	34	69	168	718	—	298	78	35	20	12	8
.40	6	7	11	18	28	47	87	205	—	325	82	35	19	11
.50	4	5	8	11	15	22	33	54	223	—	325	78	32	16
.60	4	4	6	8	10	13	18	25	56	223	—	298	68	25
.70	3	3	4	6	7	9	11	14	25	54	205	—	245	50
.80	2	3	3	4	5	6	7	9	13	22	47	168	—	164
.90	2	2	3	3	4	4	5	6	8	11	18	34	112	—
	$\alpha = .10$ and $\beta = .20$ (or $\alpha = .20$ and $\beta = .10$)													

Table A.9 Critical Values for the Signed Ranks Test

For the given n, critical values for the signed ranks test are tabled corresponding to the upper one- and two-sided significance levels in the column headings.

	One-Sided α												
	.05	.025	.01	.005	.05	.025	.01	.005	.05	.025	.01	.005
	Two-Sided α												
	.10	.05	.02	.01	.10	.05	.02	.01	.10	.05	.02	.01
n					n				n				
5	1	—	—	—	20	60	52	43	37	35	214	195	174	160
6	2	1	—	—	21	68	59	49	43	36	228	208	186	171
7	4	2	0	—	22	75	66	56	49	37	242	222	198	183
8	6	4	2	0	23	83	73	62	55	38	256	235	211	195
9	8	6	3	2	24	92	81	69	61	39	271	250	224	208
10	11	8	5	3	25	101	90	77	68	40	287	264	238	221
11	14	11	7	5	26	110	98	85	76	41	303	279	252	234
12	17	14	10	7	27	120	107	93	84	42	319	295	267	248
13	21	17	13	10	28	130	117	102	92	43	336	311	281	262
14	26	21	16	13	29	141	127	111	100	44	353	327	297	277
15	30	25	20	16	30	152	137	120	109	45	371	344	313	292
16	36	30	24	19	31	163	148	130	118	46	389	361	329	307
17	41	35	28	23	32	175	159	141	128	47	408	379	345	323
18	47	40	33	28	33	188	171	151	138	48	427	397	362	339
19	54	46	38	32	34	201	183	162	149	49	446	415	380	356
										50	466	434	398	373

Table A.10 Critical Values for the Mann–Whitney (Wilcoxon) Statistic

This table presents upper one- and two-sided critical values for the Mann–Whitney U statistic. Lower one-sided critical values are computed from the upper one-sided critical value (at the same significance level) as $(M \cdot N) - U$. The Wilcoxon two-sample statistic, W, is related to U by the equation $W = (M \cdot N) + (M \cdot (M + 1)/2) - U$, where W is the sum of the ranks of the sample of size M in the combined sample.

		One-Sided α								*One-Sided α*					
		.10	.05	.025	.01	.005	.001			.10	.05	.025	.01	.005	.001
		Two-Sided α								*Two-Sided α*					
		.20	.10	.05	.02	.01	.002			.20	.10	.05	.02	.01	.002
n	*m*							*n*	*m*						
3	2	6	—	—	—	—	—	10	1	10	—	—	—	—	—
3	3	8	9	—	—	—	—	10	2	17	19	20	—	—	—
								10	3	24	26	27	29	30	—
4	2	8	—	—	—	—	—	10	4	30	33	35	37	38	40
4	3	11	12	—	—	—	—	10	5	37	39	42	44	46	49
4	4	13	15	16	—	—	—	10	6	43	46	49	52	54	57
								10	7	49	53	56	59	61	65
5	2	9	10	—	—	—	—	10	8	56	60	63	67	69	74
5	3	13	14	15	—	—	—	10	9	62	66	70	74	77	82
5	4	16	18	19	20	—	—	10	10	68	73	77	81	84	90
5	5	20	21	23	24	25	—								
								11	1	11	—	—	—	—	—
6	2	11	12	—	—	—	—	11	2	19	21	22	—	—	—
6	3	15	16	17	—	—	—	11	3	26	28	30	32	33	—
6	4	19	21	22	23	24	—	11	4	33	36	38	40	42	44
6	5	23	25	27	28	29	—	11	5	40	43	46	48	50	53
6	6	27	29	31	33	34	—	11	6	47	50	53	57	59	62
								11	7	54	58	61	65	67	71
7	2	13	14	—	—	—	—	11	8	61	65	69	73	75	80
7	3	17	19	20	21	—	—	11	9	68	72	76	81	83	89
7	4	22	24	25	27	28	—	11	10	74	79	84	88	92	98
7	5	27	29	30	32	34	—	11	11	81	87	91	96	100	106
7	6	31	34	36	38	39	42								
7	7	36	38	41	43	45	48	12	1	12	—	—	—	—	—
								12	2	20	22	23	—	—	—
8	2	14	15	16	—	—	—	12	3	28	31	32	34	35	—
8	3	19	21	22	24	—	—	12	4	36	39	41	43	45	48
8	4	25	27	28	30	31	—	12	5	43	47	49	52	54	58
8	5	30	32	34	36	38	40	12	6	51	55	58	61	63	68
8	6	35	38	40	42	44	47	12	7	58	63	66	70	72	77
8	7	40	43	46	49	50	54	12	8	66	70	74	79	81	87
8	8	45	49	51	55	57	60	12	9	73	78	82	87	90	96
								12	10	81	86	91	96	99	106
9	1	9	—	—	—	—	—	12	11	88	94	99	104	108	115
9	2	16	17	18	—	—	—	12	12	95	102	107	113	117	124
9	3	22	23	25	26	27	—								
9	4	27	30	32	33	35	—	13	1	13	—	—	—	—	—
9	5	33	36	38	40	42	44	13	2	22	24	25	26	—	—
9	6	39	42	44	47	49	52	13	3	30	33	35	37	38	—
9	7	45	48	51	54	56	60	13	4	39	42	44	47	49	51
9	8	50	54	57	61	63	67	13	5	47	50	53	56	58	62
9	9	56	60	64	67	70	74	13	6	55	59	62	66	68	73

(*continued overleaf*)

Table A.10 (***continued***)

		One-Sided α												
		.10	.05	.025	.01	.005	.001		.10	.05	.025	.01	.005	.001
		Two-Sided α												
		.20	.10	.05	.02	.01	.002		.20	.10	.05	.02	.01	.002
n	*m*							*n*	*m*					
13	7	63	67	71	75	78	83	16	12	125	132	139	146	151	161
13	8	71	76	80	84	87	93	16	13	134	143	149	157	163	173
13	9	79	84	89	94	97	103	16	14	144	153	160	168	174	185
13	10	87	93	97	103	106	113	16	15	154	163	170	179	185	197
13	11	95	101	106	112	116	123	16	16	163	173	181	190	196	208
13	12	103	109	115	121	125	133								
13	13	111	118	124	130	135	143	17	1	17	—	—	—	—	—
								17	2	28	31	32	34	—	—
14	1	14	—	—	—	—	—	17	3	39	42	45	47	49	51
14	2	23	25	27	28	—	—	17	4	50	53	57	60	62	66
14	3	32	35	37	40	41	—	17	5	60	65	68	72	75	80
14	4	41	45	47	50	52	55	17	6	71	76	80	84	87	93
14	5	50	54	57	60	63	67	17	7	81	86	91	96	100	106
14	6	59	63	67	71	73	78	17	8	91	97	102	108	112	119
14	7	67	72	76	81	83	89	17	9	101	108	114	120	124	132
14	8	76	81	86	90	94	100	17	10	112	119	125	132	136	145
14	9	85	90	95	100	104	111	17	11	122	130	136	143	148	158
14	10	93	99	104	110	114	121	17	12	132	140	147	155	160	170
14	11	102	108	114	120	124	132	17	13	142	151	158	166	172	183
14	12	110	117	123	130	134	143	17	14	153	161	169	178	184	195
14	13	119	126	132	139	144	153	17	15	163	172	180	189	195	208
14	14	127	135	141	149	154	164	17	16	173	183	191	201	207	220
								17	17	183	193	202	212	219	232
15	1	15	—	—	—	—	—								
15	2	25	27	29	30	—	—	18	1	18	—	—	—	—	—
15	3	35	38	40	42	43	—	18	2	30	32	34	36	—	—
15	4	44	48	50	53	55	59	18	3	41	45	47	50	52	54
15	5	53	57	61	64	67	71	18	4	52	56	60	63	66	69
15	6	63	67	71	75	78	83	18	5	63	68	72	76	79	84
15	7	72	77	81	86	89	95								
15	8	81	87	91	96	100	106	18	6	74	80	84	89	92	98
15	9	90	96	101	107	111	118	18	7	85	91	96	102	105	112
15	10	99	106	111	117	121	129	18	8	96	103	108	114	118	126
15	11	108	115	121	128	132	141	18	9	107	114	120	126	131	139
15	12	117	125	131	138	143	152	18	10	118	125	132	139	143	153
15	13	127	134	141	148	153	163								
15	14	136	144	151	159	164	174	18	11	129	137	143	151	156	166
15	15	145	153	161	169	174	185	18	12	139	148	155	163	169	179
								18	13	150	159	167	175	181	192
16	1	16	—	—	—	—	—	18	14	161	170	178	187	194	206
16	2	27	29	31	32	—	—	18	15	172	182	190	200	206	219
16	3	37	40	42	45	46	—	18	16	182	193	202	212	218	232
16	4	47	50	53	57	59	62	18	17	193	204	213	224	231	245
16	5	57	61	65	68	71	75	18	18	204	215	225	236	243	258
16	6	67	71	75	80	83	88								
16	7	76	82	86	91	94	101	19	1	18	19	—	—	—	—
16	8	86	92	97	102	106	113	19	2	31	34	36	37	38	—
16	9	96	102	107	113	117	125	19	3	43	47	50	53	54	57
16	10	106	112	118	124	129	137	19	4	55	59	63	67	69	73
16	11	115	122	129	135	140	149	19	5	67	72	76	80	83	88

Table A.10 (*continued*)

		One-Sided α												
		.10	.05	.025	.01	.005	.001		.10	.05	.025	.01	.005	.001
		Two-Sided α												
		.20	.10	.05	.02	.01	.002		.20	.10	.05	.02	.01	.002
n	*m*							*n*	*m*					
19	6	78	84	89	94	97	103	20	4	58	62	66	70	72	77
19	7	90	96	101	107	111	118	20	5	70	75	80	84	87	93
19	8	101	108	114	120	124	132	20	6	82	88	93	98	102	108
19	9	113	120	126	133	138	146	20	7	94	101	106	112	116	124
19	10	124	132	138	146	151	161	20	8	106	113	119	126	130	139
19	11	136	144	151	159	164	175	20	9	118	126	132	140	144	154
19	12	147	156	163	172	177	188	20	10	130	138	145	153	158	168
19	13	158	167	175	184	190	202	20	11	142	151	158	167	172	183
19	14	169	179	188	197	203	216	20	12	154	163	171	180	186	198
19	15	181	191	200	210	216	230	20	13	166	176	184	193	200	212
19	16	192	203	212	222	230	244	20	14	178	188	197	207	213	226
19	17	203	214	224	235	242	257	20	15	190	200	210	220	227	241
19	18	214	226	236	248	255	271	20	16	201	213	222	233	241	255
19	19	226	238	248	260	268	284	20	17	213	225	235	247	254	270
20	1	19	20	—	—	—	—	20	18	225	237	248	260	268	284
20	2	33	36	38	39	40	—	20	19	237	250	261	273	281	298
20	3	45	49	52	55	57	60	20	20	249	262	273	286	295	312

Table A.11 Critical Values of the Bivariate Normal Sample Correlation Coefficient ρ

When $\rho = 0$, the distribution is symmetric about zero; thus, one-sided lower critical values are -1 times the tabled one-sided upper critical values. Column headings are also labeled for the corresponding two-sided significance level and the percentage of the distribution less than the tabled value. N is the number of observations; the degrees of freedom is two less than this.

	Percent								Percent						
	90	95	97.5	99	99.5	99.9	99.95		90	95	97.5	99	99.5	99.9	99.95
	One-Sided α								One-Sided α						
	.10	.05	.025	.01	.005	.001	.0005		.10	.05	.025	.01	.005	.001	.0005
	Two-Sided α								Two-Sided α						
	.20	.10	.05	.02	.01	.002	.001		.20	.10	.05	.02	.01	.002	.001
N								*N*							
3	.951	.988	.997	1.000	1.000	1.000	1.000	20	.299	.378	.444	.516	.562	.648	.679
4	.800	.900	.950	.980	.990	.998	.999	25	.265	.337	.396	.462	.505	.588	.618
5	.687	.805	.878	.934	.959	.986	.991	30	.241	.306	.361	.423	.463	.542	.570
6	.608	.729	.811	.882	.917	.963	.974								
7	.551	.669	.755	.833	.875	.935	.951	35	.222	.283	.334	.392	.430	.505	.532
								40	.207	.264	.312	.367	.403	.474	.501
8	.507	.622	.707	.789	.834	.905	.925	45	.195	.248	.294	.346	.380	.449	.474
9	.472	.582	.666	.750	.798	.875	.898	50	.184	.235	.279	.328	.361	.427	.451
10	.443	.549	.632	.716	.765	.847	.872	55	.176	.224	.266	.313	.345	.408	.432
11	.419	.522	.602	.685	.735	.820	.847								
12	.398	.497	.576	.658	.708	.795	.823	60	.168	.214	.254	.300	.330	.391	.414
								65	.161	.206	.244	.288	.317	.376	.399
13	.380	.476	.553	.634	.684	.772	.801	70	.155	.198	.235	.278	.306	.363	.385
14	.365	.458	.533	.612	.661	.750	.780	75	.150	.191	.227	.268	.296	.351	.372
15	.351	.441	.514	.592	.641	.730	.760	80	.145	.185	.220	.260	.286	.341	.361
16	.338	.426	.497	.574	.623	.711	.742								
17	.327	.412	.482	.558	.606	.694	.725	85	.140	.180	.213	.252	.278	.331	.351
								90	.136	.175	.207	.245	.270	.322	.341
18	.317	.400	.468	.543	.590	.678	.708	95	.133	.170	.202	.238	.263	.313	.332
19	.308	.389	.456	.529	.575	.662	.693	100	.129	.165	.197	.232	.257	.305	.324

Table A.12 Critical Values for Spearman's Rank Correlation Coefficient

For a sample of size n, two-sided critical values are given for significance levels .10, .05, and .01. Reject the null hypothesis of independence if the absolute value of the sample Spearman correlation coefficient exceeds the tabled value.

	Two-Sided α		
	.10	.05	.01
n			
5	.900	—	—
6	.829	.886	—
7	.714	.786	.929
8	.643	.738	.881
9	.600	.700	.833
10	.564	.648	.794
11	.536	.618	.818
12	.497	.591	.780
13	.475	.566	.745
14	.457	.545	.716
15	.441	.525	.689
16	.425	.507	.666
17	.412	.490	.645
18	.399	.476	.625
19	.388	.462	.608
20	.377	.450	.591
21	.368	.438	.576
22	.359	.428	.562
23	.351	.418	.549
24	.343	.409	.537
25	.336	.400	.526
26	.329	.392	.515
27	.323	.385	.505
28	.317	.377	.496
29	.311	.370	.487
30	.305	.364	.478

Table A.13 Expected Values of Normal Order Statistics

A sample of $NN(0, 1)$ observations is ranked from largest (rank 1) to smallest (rank N). The expected values of the order statistics (the ranked values) are given. Only the expected values for the upper half of the order statistics are given since the expected values are symmetric about zero. The column headings give the size of the sample and the row headings the rank of the order statistic.

	Sample Size												
Rank	2	3	4	5	6	7	8	9	10	11	12	13	14
1	.56419	.34628	1.02938	1.16296	1.26721	1.35218	1.42360	1.48501	1.53875	1.58644	1.62923	1.66799	1.70338
2		.00000	.29701	.49502	.64176	.75737	.85222	.93230	1.00136	1.06192	1.11573	1.16408	1.20790
3				.00000	.20155	.35271	.47282	.57197	.65606	.72884	.79284	.84983	.90113
4						.00000	.15251	.27453	.37576	.46198	.53684	.60285	.66176
5								.00000	.12267	.22489	.31225	.38833	.45557
6										.00000	.10259	.19052	.26730
7												.00000	.08816

	Sample Size												
Rank	15	16	17	18	19	20	21	22	23	24	25	26	27
1	1.73591	1.76599	1.79394	1.82003	1.84448	1.86748	1.88917	1.90969	1.92916	1.94767	1.96531	1.98216	1.99827
2	1.24794	1.28474	1.31878	1.35041	1.37994	1.40760	1.43362	1.45816	1.48137	1.50338	1.52430	1.54423	1.56326
3	.94769	.99027	1.02946	1.06573	1.09945	1.13095	1.16047	1.18824	1.21445	1.23924	1.26275	1.28511	1.30641
4	.71488	.76317	.80738	.84812	.88586	.92098	.95380	.98459	1.01356	1.04091	1.06679	1.09135	1.11471
5	.51570	.57001	.61946	.66479	.70661	.74538	.78150	.81527	.84697	.87682	.90501	.93171	.95705
6	.33530	.39622	.45133	.50158	.54771	.59030	.62982	.66667	.70115	.73354	.76405	.79289	.82021
7	.16530	.23375	.29519	.35084	.40164	.44833	.49148	.53157	.56896	.60299	.63690	.66794	.69727
8	.00000	.07729	.14599	.20774	.26374	.31493	.36203	.40559	.44609	.48391	.51935	.55267	.58411
9			.00000	.06880	.13072	.18696	.23841	.28579	.32965	.37047	.40860	.44436	.47801
10					.00000	.06200	.11836	.16997	.21755	.26163	.30268	.34105	.37706
11							.00000	.05642	.10813	.15583	.20006	.24128	.27983
12									.00000	.05176	.09953	.14387	.18520
13											.00000	.04781	.09220
14													.00000

(*continued overleaf*)

Table A.13 (*continued*)

	Sample Size											
Rank	28	29	30	31	32	33	34	35	36	37	38	39
1	2.01371	2.02852	2.04276	2.05646	2.06967	2.08241	2.09471	2.10661	2.11812	2.12928	2.14009	2.15059
2	1.58145	1.59888	1.61560	1.63166	1.64712	1.66200	1.67636	1.69023	1.70362	1.71659	1.72914	1.74131
3	1.32674	1.34619	1.36481	1.38268	1.39985	1.41637	1.43228	1.44762	1.46244	1.47676	1.49061	1.50402
4	1.13697	1.15822	1.17855	1.19803	1.21672	1.23468	1.25196	1.26860	1.28466	1.30016	1.31514	1.32964
5	.98115	1.00414	1.02609	1.04709	1.06721	1.08652	1.10509	1.12295	1.14016	1.15677	1.17280	1.18830
6	.84615	.87084	.89439	.91688	.93841	.95905	.97886	.99790	1.01624	1.03390	1.05095	1.06741
7	.72508	.75150	.77666	.80066	.82359	.84555	.86660	.88681	.90625	.92496	.94300	.96041
8	.61385	.64205	.66885	.69438	.71875	.74204	.76435	.78574	.80629	.82605	.84508	.86343
9	.50977	.53982	.56834	.59545	.62129	.64596	.66954	.69214	.71382	.73465	.75468	.77398
10	.41096	.44298	.47329	.50206	.52943	.55552	.58043	.60427	.62710	.64902	.67009	.69035
11	.31603	.35013	.38235	.41287	.44185	.46942	.49572	.52084	.54488	.56793	.59005	.61131
12	.22389	.26023	.29449	.32686	.35755	.38669	.41444	.44091	.46620	.49042	.51363	.53592
13	.13361	.17240	.20885	.24322	.27573	.30654	.33582	.36371	.39032	.41576	.44012	.46348
14	.04442	.08588	.12473	.16126	.19572	.22832	.25924	.28863	.31663	.34336	.36892	.39340
15		.00000	.04148	.08037	.11695	.15147	.18415	.21515	.24463	.27272	.29954	.32520
16				.00000	.03890	.07552	.11009	.14282	.17388	.20342	.23159	.25849
17						.00000	.03663	.07123	.10399	.13509	.16469	.19292
18								.00000	.03461	.06739	.09853	.12817
19										.00000	.03280	.06395
20												.00000

Author Index

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

Subject Index

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

Symbol Index

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*
BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Third Edition*
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
*COCHRAN and COX · Experimental Designs, *Second Edition*
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Statistical Modelling
CONOVER · Practical Nonparametric Statistics, *Third Edition*
COOK · Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods

*Now available in a lower priced paperback edition in the Wiley Classics Library.

*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, *Third Edition*
*DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
*DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, *Second Edition*
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

*Now available in a lower priced paperback edition in the Wiley Classics Library.

GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
*HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HELLER · MACSYMA for Statisticians
HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
*HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design

*Now available in a lower priced paperback edition in the Wiley Classics Library.

KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON · Statistical Methods in Spatial Epidemiology
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP · Directional Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
McFADDEN · Management of Data in Clinical Trials
McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
*MILLER · Survival Analysis, *Second Edition*
MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Third Edition*
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
MURTHY, XIE, and JIANG · Weibull Models
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Second Edition*
MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models, and Applications
PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
*RAO · Linear Statistical Inference and Its Applications, *Second Edition*
RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RENCHER · Linear Models in Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

RENCHER · Methods of Multivariate Analysis, *Second Edition*
RENCHER · Multivariate Statistical Inference with Applications
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, *Second Edition*
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER and LEE · Linear Regression Analysis, *Second Edition*
SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, *Second Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEISBERG · Applied Linear Regression, *Second Edition*
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
YANG · The Construction Theory of Denumerable Markov Processes
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine