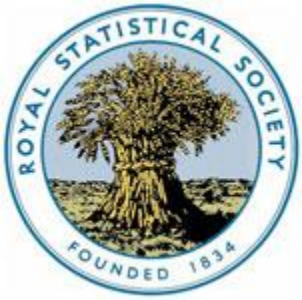

Asymptotically Efficient Rank Invariant Test Procedures
Author(s): Richard Peto and Julian Peto
Source: *Journal of the Royal Statistical Society. Series A (General)*, Vol. 135, No. 2 (1972), pp. 185-207
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2344317
Accessed: 21-04-2017 17:23 UTC

J. R. Statist. Soc. A,
(1972), **135**, *Part* 2, p. 185

Asymptotically Efficient Rank Invariant Test Procedures

By RICHARD PETO AND JULIAN PETO

Radcliffe Infirmary, Oxford University — *Institute of Psychiatry University of London*

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, January 19th, 1972, the President Professor G. A. BARNARD in the Chair]

SUMMARY

Asymptotically efficient rank invariant test procedures for detecting differences between two groups of independent observations are derived. These are generalized to test between two groups of independent censored observations, to test between many groups of observations, and to test between groups after allowance for the effects of concomitant variables.

One of these test procedures—the logrank—is particularly appropriate for comparing life tables, and can therefore be used in the analysis of clinical trials, industrial life-testing experiments and laboratory studies of animal carcinogenesis. It has greater local power than any other rank-invariant test procedure for detecting Lehmann-type differences between groups of independent observations subject to some right-censoring. The logrank test, although a rank test, can be presented in a format which exhibits the physical significance as well as the statistical significance of any important differences between groups of events.

Keywords: LIFE TABLE; EXPERIMENTAL SURVIVAL CURVE; EXPERIMENTAL CDF; PRODUCT-LIMIT ESTIMATE; PERMUTATIONAL TEST; RANK TEST; TWO-GROUP TEST; RIGHT CENSORING; DEATH TIMES; LOGRANK; ASYMPTOTIC EFFICIENCY; RANK INVARIANCE; RELATIVE DEATH RATES; LEHMANN ALTERNATIVES; CLINICAL TRIALS; CENSORING; WILCOXON RANK SUM TEST; FAILURE TIMES.

1. INTRODUCTION

THE principal advantage of any rank test, the absolute reliability of the significance level it generates whatever the distribution functions of the observations, is for many such tests offset by some loss of power. Various rank tests which are asymptotically efficient for particular distributions have been suggested during the last few years, and Hajek and Sidak (1967) have described a general method for the construction of such tests. Since any rank test is invariant under monotonic transformation of the data, distributions fall into disjoint "efficiency classes" within each of which the efficiency of any rank test is constant; the normal and lognormal distributions are evidently members of the same class, for example.

Suppose z_i $(1 \leqslant i \leqslant N)$ are independent observations from the c.d.f. $F(x, \theta_i)$, where $\theta_i = \theta_A$ for $1 \leqslant i \leqslant n$, $\theta_i = \theta_B$ for $n+1 \leqslant i \leqslant N$, and the null hypothesis is H_0: $\theta_A = \theta_B$. If a score is assigned to each observation, the scores being such that the group A sum of scores is an asymptotically efficient test statistic, the null hypothesis distribution of this sum may be derived permutationally. If each score is not calculated exactly from the observation values but is *estimated* from their ranks, the resulting rank test will be asymptotically efficient in the family $F(x, \theta)$; it will of course also be asymptotically efficient in any other family in the same efficiency class as $F(x, \theta)$.

2. SOME DEFINITIONS

Let z be a real-valued random variable with c.d.f. $F(x)$.

(i) Let $G(x) = 1-F(x)$. G is the *survival curve* of z.

(ii) We shall be concerned only with either *exact* or (*interval*) *censored* observations. For the latter, the only information recorded about the random variable is that it lies somewhere in the non-null interval (x_1, x_2). One observation of a random variable may thus generate either one or two *data points* according to whether the observation is exact or censored. If $x_2 = \infty$ then the observation is *right-censored* with *censoring point* or *value* x_1. In data consisting entirely of either exact or right-censored observations, the only data points other than ∞ are the exact observation values and the censored observation values.

(iii) Consider data consisting of observations of N random variables z_i $(1 \leqslant i \leqslant N)$. For these data, the *experimental survival curve*, $H(x)$ is the survival curve under which the product of the likelihoods of the N observations is maximal. H has a discontinuity at each exact observation value, since otherwise the likelihood of such observations would be infinitesimal. At each data point the values of H (or of the top and bottom of the step in H) are well defined† and these values are invariant under monotonic (rank-preserving) transformations of the data points. If there are no tied values and there is no censoring then the steps in H are all of equal size and at the rth exact observation value H decreases from $(N+1-r)/N$ to $(N-r)/N$.

For partially right-censored data $H(x)$ is the familiar life-table estimate of the probability that a random variable will exceed x: let $r(x)$ be the number of (exact or censored) observation values not less than x and let $s(x)$ be the number of exact observations with the value x. $s(x)$ is zero except on the set of exact observation values. Kaplan and Meier (1958) show that in this situation the experimental survival curve is constant except at the exact observation values, where

$$H(x+) = H(x-)\{1-s(x)/r(x)\}.$$

They also show that if the z_i have a common survival curve $G(x)$ then $H(x)$ estimates $G(x)$ unbiasedly; the effect of right censoring is to coarsen this estimate but not to bias it. Fig. 1 compares the experimental survival curve for some partially right-censored data with the experimental survival curve that would have been obtained had all the observations been exact. Ten random variables which actually took the values 4, 8, 11, 15, 20, 26, 34, 45, 57 and 79 were observed subject to some right-censoring, the values 15, 26 and 45 being only known to lie somewhere in $(10, \infty)$ and the value 79 being only observed to lie in $(40, \infty)$.

(iv) Suppose $G(x, \theta)$ is a family of survival curves parametrized by θ, and that the positions of the discontinuities and constant regions of $G(x, \theta)$ with respect to x are independent of θ. If $G(x, \theta_0)$ (θ_0 fixed) is a particular member of the family $G(x, \theta)$ then there exists a function $\gamma_{\theta_0}(., \theta)$ such that for all x, θ, $G(x, \theta) = \gamma_{\theta_0}\{G(x, \theta_0), \theta\}$. For example, the family of exponential survival curves $G(x, \theta) = \exp(-\theta x)$ is generated from any particular member $\exp(-\theta_0 x)$ by the operation of the function $\gamma_{\theta_0}(y, \theta) = y^{\theta/\theta_0}$. If there exists an explicit function $\Psi(\theta_0, \theta)$ and a differentiable function $c(y, \Psi)$ such that for all $y \in [0, 1]$ and for all parameter values θ, θ_0, $c\{y, \Psi(\theta_0, \theta)\} = \gamma_{\theta_0}(y, \theta)$ then we call c the *conversion function* corresponding to the family $G(x, \theta)$ of distributions. (Note that $c\{y, \Psi(\theta, \theta)\} \equiv y$.) In the above exponential

† In "Experimental survival curves for interval-censored data"—a paper by one of the authors, to appear in *Applied Statistics* (1973), **22**, No 1.

example, $\Psi(\theta_0, \theta) = \theta/\theta_0$ and $c(y, \Psi) = y^{\Psi}$. The conversion functions corresponding to two different families of survival curves (the normal and lognormal, for instance) may be identical; a conversion function is more general than the family of survival curves that gave rise to it. In fact, families of distributions with the same conversion function lie in the same "efficiency class".

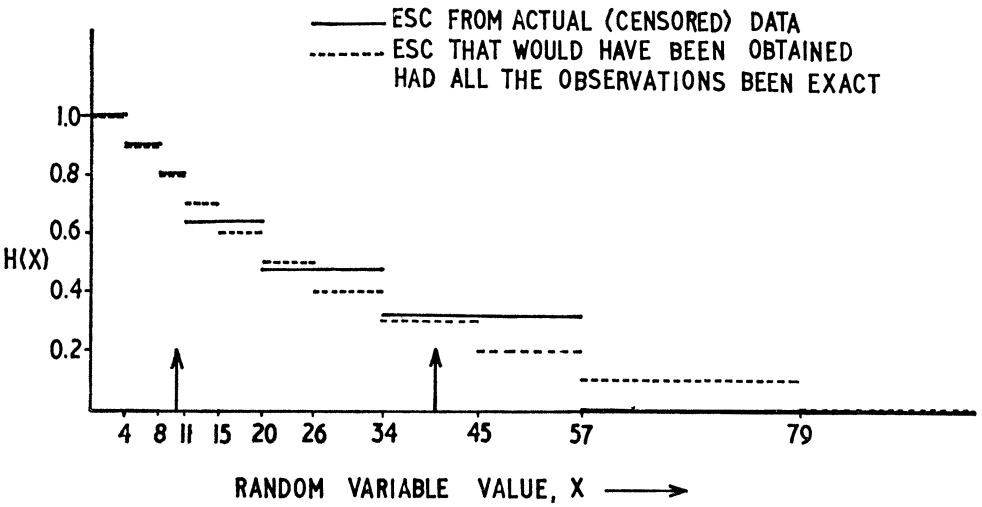

FIG. 1. Experimental survival curve (ESC) for hypothetical data on ten random variables.

3. A PERMUTATIONAL TEST OF H_0—NOT RANK INVARIANT

In the notation of the Introduction, suppose that N independent observations z_i $(1 \leqslant i \leqslant N)$ are subject to censoring. Their total log likelihood is $\sum L_i(\theta_i)$ where $L_i(\theta_i)$ is the log likelihood function for the ith (possibly censored) observation.

If $\theta_i = \theta$ for all i then the total log likelihood is a function $\sum L_i(\theta)$ of θ; let $\hat{\theta}$ be the (ML) value of θ that maximizes this. Defining $U_i = L_i'(\hat{\theta})$ (i.e. $\partial L_i/\partial\theta$ at $\theta = \hat{\theta}$), $\sum U_i = 0$ and it is shown in Appendix A that an asymptotically efficient test of H_0: $\theta_A = \theta_B$ versus H_A: $\theta_A \neq \theta_B$ can be constructed from the statistic $Y = \sum_{\text{group } A} U_i$. If under H_0 the censoring applied to the observations is independent of group membership then under H_0 Y has the distribution of a sum of n chosen at random from $U_1, \ldots, U_N$.

This permutational test of H_0 is to be preferred to the generalized likelihood ratio test since even if, by mistake, the wrong family of distributions has been assumed the permutational test remains valid (although not necessarily asymptotically efficient), whereas the generalized likelihood ratio test does not. (Appendix A also shows how analogous scores U_i can be defined for hypothesis testing in the more general case where nuisance parameters exist with a common but unknown value for both groups.)

Example 1. If each z_i is an uncensored random variable from the exponential distribution with parameter θ_i then $L_i(\theta_i) = \log \theta_i - z_i\,\theta_i$, $\hat{\theta} = 1/\bar{z}$ and $U_i = \bar{z} - z_i$.

Example 2. A family $G(x, \theta)$ of survival curves of the form $g^{\theta}(x)$ is called a Lehmann-type family. If a mixture of exact and right-censored observations are made from a Lehmann-type family, then for an exact observation of t,

$$L(\theta) = \log\theta + \theta \log g(t) + \text{terms independent of } \theta,$$

so $U = L'(\hat{\theta}) = 1/\hat{\theta} + \log g(t)$, while for a right-censored observation with censoring point T $L(\theta) = \theta \log g(T)$ so $U = L'(\hat{\theta}) = \log g(t)$. The test of $\theta_A = \theta_B$ is unaffected by scaling the scores, so we may equally well study the scores $\hat{\theta}U = 1 + \log G(t, \hat{\theta})$ or $\log G(T, \hat{\theta})$.

4. A RANK INVARIANT TEST OF H_0

For a family $G(x, \theta)$ of survival curves such that under H_0 each of the random variables z_i is distributed according to the same member of the family, we can construct an asymptotically efficient rank invariant test of H_0: $\theta_A = \theta_B$. If censoring is present, it is assumed to have been applied at random in a similar way to both groups: this assumption is relaxed in Sections 10–12. The asymptotic efficiency of a test of H_0 is the limit of the efficiencies achieved in a series of experiments of increasing size in which θ_A and θ_B converge to an intermediate value θ_0. Let the conversion function for the family $G(x, \theta)$ be $c(y, \Psi)$ as in paragraph 2, and let $G_0(x)$ denote $G(x, \theta_0)$. The family $c\{G_0(x), \Psi\}$ of distributions is then the original family $G(x, \theta)$ but it is parametrized in a different way, Ψ, a function of θ_0 and θ, now being the parameter of interest instead of θ. As before, the quantities $U_i = \partial L_i/\partial \Psi$ at $\Psi = \hat{\Psi}$ can be defined and an asymptotically efficient test procedure based on them.

However, we may instead consider the family $c\{H(x), \Psi\}$ of survival curves generated from the experimental survival curve by the conversion function. As the sample size increases and θ_A and θ_B tend to θ_0, $H(x)$ tends to $G_0(x)$ at all points where the censoring is not total, and since $c\{H(x), \Psi\}$ therefore tends to $c\{G_0(x), \Psi\}$ test procedures which are asymptotically efficient in the family $c\{H(x), \Psi\}$ will also be asymptotically efficient in the parent family $c\{G_0(x), \Psi\}$ or $G(x, \theta)$. For the family $c\{H(x), \Psi\}$ the U's are derived as follows. For a censored observation of z_i from which z_i is known to lie in an interval over which $H(x)$ drops from a to b or for an exact observation of z_i where H drops from a to b at z_i the likelihood function l is $c(a, \Psi) - c(b, \Psi)$. Thus

$$U_i = \partial \log(l)/\partial \Psi\big|_{\Psi=\hat{\Psi}}$$
$$= \left.\frac{c'(a, \Psi) - c'(b, \Psi)}{c(a, \Psi) - c(b, \Psi)}\right|_{\Psi=\hat{\Psi}}$$

By definition the experimental survival curve $H(x)$ maximizes the total likelihood so that $\hat{\Psi}$ is the value for which $c(y, \hat{\Psi})$ is the identity function $c(y, \hat{\Psi}) \equiv y$. Thus

$$U_i = \{f(a) - f(b)\}/(a - b),$$

where

$$f(y) = \partial c(y, \Psi)/\partial \Psi\big|_{\Psi=\hat{\Psi}}.$$

U_i is therefore a function of a and b only. These are values of $H(x)$ at certain of the data points and are therefore invariant under all rank-preserving transformations of the data.

This is not a unique generalization of the U-test procedure to rank invariance. One may, for example, replace $\{f(a) - f(b)\}/(a - b)$ by $f'\{(a + b)/2\}$ when the observation concerned is exact. Asymptotically $(a - b) \to 0$ for exact observations so this modified rank test would also be asymptotically efficient. If there are no censored observations the experimental survival curve decreases in equal steps from 1 to 0; Hajek and Sidak (1967) suggest in this case the score $f'\{1 - r/(N + 1)\}$ for the observation with rank r, which is also asymptotically efficient. These minor modifications

are unlikely to affect noticeably the power of the test in finite cases. The scores $\{f(a)-f(b)\}/(a-b)$ sum identically to zero and have a definite conceptual source, but in particular cases (e.g. the logrank test) other scores may be simpler to compute or of marginally greater power in small samples.

5. Two Particular Tests: Logrank and Probit

5.1. *The Logrank Test*

Experiments in which the observations are the time to the occurrence of an event are a common source of right-censored data. In this situation, we are generally interested in the incidence rate $-\partial \log G/\partial t$ at which the event is happening over time. The incidence rate suggested by $G(t)$ will in general vary with time, and one obvious family to consider is the family of survival curves $G^\theta(t)$ (Lehmann-type alternatives) since the incidence rate suggested by $G^\theta(t)$ at t is θ times that suggested by $G(t)$ at t. We therefore consider the family $H^\Psi(t)$, for which $\hat{\Psi} = 1$. The conversion function $c(y, \Psi) = y^\Psi$ that corresponds to this family of distributions corresponds to the exponential or Weibull families.

Since $c(y, \Psi) = y^\Psi$,

$$f(y) = \partial c(y, \Psi)/\partial \Psi|_{\Psi=1} = y \log(y)$$

and

$$U = \{f(a)-f(b)\}/(a-b) = \{a \log(a) - b \log(b)\}/(a-b).$$

U is approximately $1+\log\{H(t)\}$ for an exact observation t, and is exactly $\log\{H(T)\}$ for an observation right censored at T. These rank invariant scores are asymptotically efficient in any Lehmann family. They closely resemble the scores $1+\log G(t, \hat{\theta})$ and $\log G(T, \hat{\theta})$ derived in Section 3, Example 2, which are efficient in a particular Lehmann family.

For right-censored data, Altshuler (1970) has suggested for the logarithm of the survival curve the estimator $-e(t) = -\sum_{x \leqslant t} s(x)/r(x)$, summation being taken over all exact observation values up to t and $s(x)$ and $r(x)$ being as in Section 2(iii). As the sample size increases both this and $\log H(t)$ will, if the distribution is continuous, converge to $\log G_0(t)$. The rank invariant scores $W = 1-e(t)$ or $-e(T)$ will therefore also be asymptotically efficient in G^θ, and it can be proved that the test based on group sums of the W-scores is of maximal local power in G^θ among all rank invariant test procedures, even for small sample sizes, if G is continuous.†

These W-scores sum identically to zero for both groups together, and we call the permutational test based on sums of them the *logrank* test. The small-sample efficiency of this test and the physical meaning of the logrank scores are studied in detail in Sections 7 and 10.

5.2. *The Probit Rank Test*

Consider the t-test situation. If F is the c.d.f. of the standardized normal distribution then the family of conversion functions is $c(y, \Psi) = F\{F^{-1}(y)+\Psi\}$. ($F^{-1}(y) \sim N(0, 1)$ if y is uniform over $(0, 1)$ so this corresponds to the family of shifts of a normal mean.) $\hat{\Psi} = 0$ and $f(y) = \partial c(y, \Psi)/\partial \Psi|_{\Psi=0} = (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}F^{-1}(y)^2\}$. Scores based on the related U's give a rank test which is asymptotically efficient

† This is proved in "Rank tests of maximum power against Lehmann-type alternatives"—a paper by one of the authors, to appear in *Biometrika* (1972), **59**, No 2.

against changes in the parameter μ of the normal distribution $N(\mu, \sigma^2)$ or against changes in μ in any monotonic transformation (e.g. lognormal) of the normal distribution.

6. WILCOXON'S RANK SUM TEST

Wilcoxon's two-sample rank sum test (Wilcoxon, 1945) for exact observations is the non-parametric test most commonly used in practice and so comparison of the rank sum test with any other suggested permutational test is clearly of interest.

For uncensored data $H(x) = \frac{1}{2}\{H(x+)+H(x-)\}$ is linearly related to the ranks of the observations, and the permutational test procedure assigning the score $2\bar{H}(z)-1$ to an exact observation of z is equivalent to Wilcoxon's rank sum test.

This scoring system may be generalized to censored data as follows. For censored data $H(x)$ is still an ML estimator of the underlying survival curve and so an exact observation should still score $2\bar{H}(z)-1$. If z is censored to lie in some interval over which $H(.)$ drops from a to b then the appropriate score must lie somewhere between $2a-1$ and $2b-1$; in expectation, it is approximately halfway between them at $a+b-1$, and we suggest this score of $a+b-1$ as appropriate for a censored observation. If we continue to give the score $a+b-1$ to an exact observation of z where H drops from a to b at z then these generalized Wilcoxon scores sum identically to zero.

This scoring system may alternatively be derived from the conversion function

$$c(y, \psi) = \{1+(y^{-1}-1)\psi\}^{-1}$$

for the logistic family

$$G(x, \theta) = \{1+\exp(\theta+x)\}^{-1}$$

of survival curves in which Wilcoxon's rank sum test is asymptotically efficient.

$$\psi(\theta_0, \theta) = \exp(\theta-\theta_0), \hat{\psi} = 1, f(y) = y^2-y$$

and

$$U = \{f(a)-f(b)\}/(a-b) = a+b-1.$$

This generalization of Wilcoxon's test to censored data is to be preferred to Gehan's (1965) generalization, since the relative values of the expectations of the scores Gehan assigns to exact observations at particular times vary according to the pattern of censoring imposed on the observations. If there is no censoring both tests assign a score which is in expectation proportional to $2G(z)-1$ for an exact observation of z where $G(y)$ is the true survival curve under H_0. If the data are right-censored the expectation of our generalized scores is still $2G(z)-1$ for an exact observation of z, but this is no longer true for Gehan's generalized scores. For example, if the censoring rate equals the incidence rate Gehan's scores for exact observations have expectation $3G^2(z)-1$, totally different from the Wilcoxon scores which they are supposed to generalize.

7. POWER OF THE THREE RANK TESTS

The decision whether or not to use a test depends not only on its power against the assumed alternatives but also on its power against other alternatives which might in

fact obtain. The asymptotic efficiencies of the above three rank statistics against normal and Lehmann alternatives are, in the absence of censoring:

	Normal alternatives (%)	*Lehmann alternatives* (%)
Logrank test	82	100
Probit rank test	100	82
Wilcoxon's rank sum test	95	75

When the logrank test is being used between two groups of uncensored Lehmann observations we may compare its power to that of the (uniformly most powerful) likelihood ratio test, which is not rank invariant. The power of the likelihood ratio test may be obtained by noting that the ratio of exponential means has the distribution $F_{2n_1, 2n_2} \times r_1/r_2$, and the power of the logrank test may be obtained by summing the

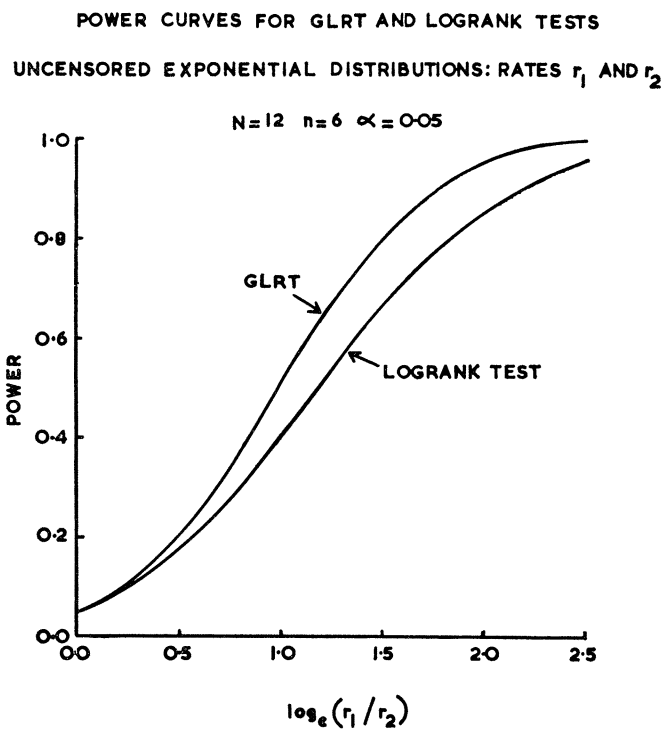

FIG. 2. Loss of power due to using a rank invariant test.

separate probabilities of each of the "significant" rankings for various values of r_1/r_2. Fig. 2 shows this comparison for the special case of two groups each of six uncensored exponentially distributed observations. The comparison indicates that the use of the logrank test can be recommended even with such small samples.

The loss of power relative to the uniformly most powerful test is almost wholly due to the corruption of the data by ranking them; we have at $r_1/r_2 = 1{\cdot}01$, 2, 4, 6, 8, 16 and 32 calculated the power of the most powerful 5% rank test and none of these particular most powerful rank tests has a power greater than 0·006 more than the power of the logrank test. At $r_1/r_2 = 1{\cdot}01$ (i.e. locally) the logrank test is, of course, the best possible 5% rank test. This, incidentally, underlines the fact that no rank test is uniformly most powerful among all rank tests against Lehmann-type alternatives.

8. Calculation of Permutational Significance Levels

Let $U_1, \ldots, U_N$ be a set of scores, and let the number of subjects in group A be n and in group B be $N-n$. Write $X = \sum_A U_j$. There are ${}^NC_n$ ways of choosing n out of the N scores and k of these will have a sum less than or equal to X. The one-sided significance level of X is thus $P = k/{}^NC_n$. We can find P in various ways:

(1) Exactly by counting: this is feasible by computer for any n with N less than about 30 and for certain other special cases. A suitable algorithm is available (Hill and Peto, 1971).

(2) Approximately, using the Pearson family of curves to estimate the c.d.f. of X as described in Appendix B. This method is very accurate: e.g. with the logrank scores on exact observations with $N = 12$, $n = 3$ the true and approximate c.d.f.'s never differ by as much as 0·01. (The normal approximation has been found to be inadequate.)

(3) Approximately, by simulation. This technique is not often necessary since the regions of applicability of (1) and (2) overlap, although it may be preferred to (1) for speed and to (2) for conceptual simplicity.

9. Generalization to Multi-group Experiments

Suppose we have a set of N zero-sum scores $W_1, \ldots, W_N$. Divide them into r groups of sizes $n_1, \ldots, n_r$ ($\sum n_j = N$) and let S_j be the sum of the scores in the jth group. The null hypothesis we wish to test is that the allocation was at random.

Under the null hypothesis $E(S_j) = 0$, and to test it we may calculate

$$X^2 = (\sum S_j^2/n_j)/s^2$$

where $s^2 = \sum W_i^2/(N-1)$. (X^2 can be shown to be $\mathbf{S}^{\mathrm{T}}.\mathrm{var}^{-1}(\mathbf{S}).\mathbf{S}$ where $\mathbf{S}$ is the vector consisting of any $r-1$ of the r zero-sum quantities $S_1, \ldots, S_r$.) Under the null hypothesis $X^2 \sim \chi^2_{r-1}$ approximately by the multivariate normal approximation to the joint distribution of the S_j's.

We may use this result to test for differences between several groups of independent observations rather than just two groups as we have done up to this point, by calculating the scores U_i as before and using the above X^2 (with U_i replacing W_i) as a non-parametric test for variation between the groups.

10. Presentation of the Logrank Scores for Right-censored Data

Consider a situation in which there are r groups of subjects. Let z_i be the observation, either exact or right-censored, on subject i; suppose that all the z_i are independent random variables and that if z_i is from Group j then it is distributed with survival curve $G^{\theta_j}(x)$, where G is known and continuous. No assumption is made about the manner in which censoring is imposed on the data, so it may depend on group membership. The null hypothesis is H_0: $\theta_j = \theta_0$ for all j (θ_0 unknown).

Define d_i to be zero if z_i is right-censored, otherwise 1; and let v_i denote the value (see Section 2) of z_i. Under H_0 the ML estimator of θ_0 is $\hat{\theta} = \sum d_i / \sum -\log G(v_i)$. Writing

$$e_i = -\theta_0 \log G(v_i),$$

we find that under H_0 $E(e_i) = E\{-\log G^{\theta_0}(v_i)\} = E(d_i) = \text{var}(d_i - e_i)$ irrespective of the (possibly unknown) fixed censoring point of z_i. If, therefore, for the observations in Group j we write $O_j = \sum_{i \in j} d_i$ and $E_j = \sum_{i \in j} e_i$, then under H_0 we find $E(O_j - E_j) = 0$ and $\text{var}(O_j - E_j) = E(E_j)$, so that $\sum_j (O_j - E_j)^2 / E_j \sim \chi^2_r$ approximately. Replacing E_j in the above statistic by its ML estimator $\hat{E}_j = -\hat{\theta}_0 \sum_{i \in j} \log G(v_i)$, we have $\sum_j O_j \equiv \sum_j \hat{E}_j$ so $\sum_j (O_j - \hat{E}_j)^2 / \hat{E}_j \sim \chi^2_{r-1}$ approximately.

In a clinical trial where the event recorded is death the observed number of deaths is Group j is O_j and $\hat{E}_j$ is an estimate of the expected number of deaths in Group j given the values of the times at risk v_i but not the values of the indicator variables d_i. The similarities between the properties of the quantities $\hat{E}_j$ and the expected values for the numbers of deaths in the various groups calculated on an exposure-to-risk basis suggest that we define the "expected" number of deaths for Group j to be $\hat{E}_j$. We should note, however, that in a group j of subjects at risk for exceptionally long periods $\hat{E}_j$ may exceed the number of subjects in the group.

If $\hat{e}_i$ denotes the ML estimate $-\hat{\theta}_0 \log G(v_i)$ of e_i then the U-score for the ith subject is given by $U_i = \partial L_i / \partial \theta |_{\theta = \hat{\theta}_0} = d_i - \hat{e}_i$, so the analyses of a set of data using the statistics $\sum_{i \in j} U_i$ and $(O_j - \hat{E}_j)$ to describe the jth group are equivalent. Although actual significance levels will, where possible, be calculated by permutation of the U_i the format involving the O_j and $\hat{E}_j$ is required to help understand the physical significance of the data. The ratios $O_j / \hat{E}_j$ are proportional to the $\hat{\theta}_j$ and thus estimate the *relative incidence rates* in the different groups.

When comparing various combinations of groups one with another the O's and $\hat{E}$'s sum directly when groups are combined, and, if certain groups are eliminated causing $\sum O \neq \sum \hat{E}$ in the remainder, direct scaling of the remaining $\hat{E}$'s making $\sum O = \sum \hat{E}$ is all that is necessary effectively to recalculate all the quantities within this reduced class of subjects. The behaviour of the $\hat{E}_j$ thus closely follows the behaviour of classical expectations.

An analogous presentation of the logrank scores is possible. The logrank score for a subject is defined as $d_i - e(v_i)$ (see Section 5), where $-e(v_i)$ is Altshuler's estimate of the logarithm of the common survival curve. If, analogous to our previous definition of $\hat{e}_i$ as $-\hat{\theta}_0 \log G(v_i)$, we define $\tilde{e}_i = e(v_i)$ and $\tilde{E}_j = \sum_{i \in j} \tilde{e}_i$ then the $\tilde{e}_i$ and $\tilde{E}_j$ have approximately the same properties as the $\hat{e}_i$ and $\hat{E}_j$ above except that the derivation of the $\tilde{e}_i$ does not require knowledge of G and is rank-invariant. The use of the descriptive statistics O_j and $\tilde{E}_j$ as rank-invariant test statistics has already been described by Mantel (1966) from a conditional approach (see Section 12) to the differences between the r groups.

Example. In an experiment (Roe *et al.*, 1970) on cancer rates with many different carcinogenic régimes, each applied to a group of mice, the régimes varied widely in their lethality. Following the method of Pike and Roe (1963) the experimental survival curve for lymphoma incidence was calculated and the logrank scores W_i were derived. The sums $S_j = \sum_{i \in j} W_i$ were calculated for the various treatment groups, but it was not permissible to follow Section 9 and to take $\sum_j S_j^2 / n_j s^2$ as χ^2_{r-1} because the patterns of censoring imposed by the prior deaths of the mice on the observations of the lymphoma occurrence times were different due to the differing lethalities of the

different régimes. It was therefore necessary to calculate O_j and $\tilde{E}_j = (O_j - S_j)$ for each group and to take $\sum_j (O_j - \tilde{E}_j)^2/\tilde{E}_j$ as χ^2_{r-1}, retaining rank invariance but avoiding the assumption of permutability.

11. Nuisance Parameters and Concomitant Variables

If a family $G(x, \mathbf{p}, \theta)$ of survival curves is parametrized not only by the parameter of interest but also by a vector $\mathbf{p}$ of nuisance parameters which by assumption have a common value in groups A and B, then the log likelihood is a function

$$L(\mathbf{p}, \theta) = \sum L_i(\mathbf{p}, \theta)$$

of both $\mathbf{p}$ and θ. Defining $U_i = \partial L_i/\partial\theta$ at $\hat{\mathbf{p}}$, $\hat{\theta}$, the values that jointly maximize L, it is shown in Appendix A that an asymptotically efficient (but not rank invariant) test of H_0: $\theta_A = \theta_B$ is obtained by studying group sums of the U_i. If concomitant observations $\mathbf{w}_i$ on each z_i exist such that the survival curve for z_i depends also on the values of the $\mathbf{w}_i$ then L_i depends on i not only through the observed value of z_i but also through $\mathbf{w}_i$; however, group sums of the scores $U_i = \partial L_i(\hat{\mathbf{p}}, \hat{\theta})/\partial\theta$ still constitute an asymptotically efficient test statistic, where $\mathbf{p}$ now includes any regression coefficients relating $\mathbf{w}_i$ to L_i. The null hypothesis distribution of the group sums can be found permutationally if under H_0 any n of the N observations could equally well have constituted group A. This model would be appropriate, for example, if a normal mean value depended not only on group membership but also linearly on a vector $\mathbf{w}$ of concomitant variables, the coefficients in the linear dependence being the nuisance parameters $\mathbf{p}$ with the same values in each group. In this situation, if z_i is observed exactly $U_i \propto (z_i - \hat{\theta} - \mathbf{w}_i^{\mathrm{T}}.\hat{\mathbf{p}})$. Generalization to rank invariance is possible if functions $\mathbf{g}(\mathbf{p}, \mathbf{p}_0)$, $\Psi(\theta, \theta_0)$ and $c\{y, \mathbf{g}, \Psi\}$ exist such that for any parameter values

$$G(x, \mathbf{p}, \theta) = c\{G(x, \mathbf{p}_0, \theta_0), \mathbf{g}(\mathbf{p}, \mathbf{p}_0), \Psi(\theta, \theta_0)\}.$$

In this case the family $c\{H(x), \mathbf{g}, \Psi\}$ can be studied and permutational rank invariant tests can be derived; however, unless $H(x)$ asymptotically converges to a member $G(x, \mathbf{p}_0, \theta_0)$ of the parent family these may not be asymptotically efficient. This convergence will in general only occur when there is no dependence on concomitant variables, but unless the dependence on concomitant variables is substantial the loss in asymptotic efficiency is second order.

Special Case: Right-censored Observations from a Lehmann-type Distribution

If, in a family $G^{\alpha}(x)$ of survival curves, α depends not only on group membership but also on certain concomitant variable values, a good way to introduce the dependence on these is to assume that $\log\alpha$ is linearly dependent on various functions of the concomitant variables as well as on θ.

Model. If $\mathbf{b}_i$ is a vector of functions of the concomitant variables $\mathbf{w}_i$ then assume that the survival curve for z_i is $G^{\alpha_i}(x)$, where $\log\alpha_i = \theta_A$ or $\theta_B + \mathbf{b}_i^{\mathrm{T}}.\mathbf{p}$. Recall the definition of d_i and v_i from the previous section and let ϵ_i denote $-\log G(v_i)$. Now the log likelihood function L_i is $d_i \log\alpha_i - \epsilon_i\alpha_i$, and the total log likelihood is, if $\theta_A = \theta_B = \theta_0$, a function $L(\theta_0, \mathbf{p})$ which can be maximized explicitly with respect to θ_0, giving a function $L\{\hat{\theta}_0(\mathbf{p}), \mathbf{p}\}$ to be maximized with respect to $\mathbf{p}$. It can be shown that if the values taken by those $\mathbf{b}_i$ for which $d_i = 1$ are linearly independent then this function of $\mathbf{p}$ is everywhere convex with a unique maximum at $|\mathbf{p}| < \infty$ and that there exists a negative upper bound to the second derivatives in any position and direction in $\mathbf{p}$-space, so the location of $\hat{\mathbf{p}}$, $\hat{\theta}_0$ is computationally straightforward. Finally, $U_i \propto \{d_i - \exp(\hat{\theta}_0 + \mathbf{w}_i^{\mathrm{T}}.\hat{\mathbf{p}})\epsilon_i\}$.

Application to the logrank test. The generalization of this to rank invariance with respect to monotonic transformations of the observations of the z_i is possible if the experimental survival curve $H(x)$ for the total data is allowed to generate the family $H^{\alpha}(x)$ of distributions, where as before $\log \alpha_i = \theta + \mathbf{w}_i^{\mathrm{T}}.\mathbf{p}$. For computational ease an approximation identical to the logrank approximation in Section 5 is made, namely that

$$L_i = d_i \log \alpha_i - \alpha_i . \tilde{e}_i,$$

where $\tilde{e}_i$ is, as before, Altshuler's estimator $e(v_i)$ of the log survival curve. The numerical value of $\hat{\mathbf{p}}$, if we let $\log \alpha_i = \theta + \mathbf{b}_i^{\mathrm{T}}.\mathbf{p}$, is of considerable physical interest, as are the changes in the total log likelihood as selected members of $\mathbf{p}$ are constrained to be zero, and the test procedure based on permutational sums of the scores $U_i \propto \{d_i - \exp(\hat{\theta} + \mathbf{b}_i^{\mathrm{T}}.\mathbf{p})\, \tilde{e}_i\}$ is rank invariant and is unaffected by dependence of the distributions on the concomitant parameters. For full asymptotic efficiency the more complex method of fit of this model due to Cox† is recommended, although the dependence of the distributions on the concomitant variables has to be substantial before the loss of asymptotic efficiency of this test procedure becomes important. This, or Cox's, fit of this model and the consequent changes in the total log likelihood enables the arguments of multiple regression to be applied to the rank invariant analysis of right-censored data, and this has proved of importance in clinical trials of cancer therapy. (See, for example, the report on the first M.R.C. (1972) myelomatosis trial.)

12. RIGHT-CENSORING NOT INDEPENDENT OF GROUP MEMBERSHIP

Suppose that two groups (A and B) of independent right-censored observations are being compared using rank invariant statistics based on the experimental survival curve in the manner of Section 4, but that the assumption of the permutability of the statistics does not obtain due to group-dependent differences in the censoring times. This would arise, for example, in clinical trials if withdrawals due to side-effects occurred mainly in one treatment group.

The argument of Section 10 taking $\sum (O-E)^2/E$ as approximately χ_1^2 may still be used. However, the following conditional argument is to be preferred. Let exact observations occur at times $T_1 < T_2 < \ldots < T_w$ and let the number of exact observations taking the value T_k be m_k out of a population of M_k with values greater than or equal to T_k, i.e. at risk at (T_k-). Let a proportion p_k of the M_k be from group A. Then, under the null hypothesis and conditional on the observed p_k, m_k and M_k, the number r_k of the m_k exact observations that are from group A follows the hypergeometric distribution $r(p_k, m_k, M_k)$ (see Appendix C). Thus under these conditions $\sum_k r_k$, the total number of exact observations from group A, is the sum of w independent hypergeometric distributions.

Now let U_i be scores of the type $\{f(a)-f(b)\}/(a-b)$ as suggested in Section 4, where a and b are experimental survival curve values and $f(y)$ is some suitable function. Write $g(y) = f(y)/y$ for $0 < y < 1$ with $g(0) = g(1) = 0$, $H_k = H(T_k+)$ and $H_0 = 1$, and define $\lambda_k = \{g(H_{k-1}) - g(H_k)\} M_k/m_k$. Then it can be shown that $\sum_A U_i = \sum \lambda_k (r_k - p_k m_k)$. Thus our statistic $\sum_A U_i$ is, under the above conditional argument, a weighted sum of independent hypergeometric random variables, and its significance level may be calculated approximately using Pearson curves (see Appendices B and C).

† Described in a paper to be published in *J. R. Statist. Soc.* B, (1972), **34**, No 2.

In the special case of the logrank test, where the scores are not exactly of this form, $\lambda_i = 1$ and so $\sum \lambda_i(r_i - m_i\, p_i) = \sum$(observed − conditionally expected).

Since the way the censoring occurs is independent of the parameter of interest, it should not contribute haphazardly to the assessment of the final significance levels; these should be evaluated conditionally on the actual pattern of censoring we happen to find, and so even if the censoring is known to have been applied symmetrically and at random the conditional approach is apparently to be preferred. However, in practice there will be little difference between the significance levels obtained by the permutational and the conditional arguments when censoring is applied at random to two large groups, and the understanding afforded by the permutational approach to the statistical conclusions is sufficient justification for retaining permutation whenever the censoring is random with respect to group membership.

The conditional variance and the permutational variance attributed to the same quantity $\sum_A U_i$ can differ widely even when the two significance levels are very nearly the same. This underlies the importance of using the Pearson distributions rather than a normal distribution to calculate approximate significance levels.

APPENDIX A

Suppose we have a vector $\mathbf{x}$ of independent observations whose log likelihood function is parametrized by a vector $\boldsymbol{\theta} = \theta_1, \ldots, \theta_q$ of non-degenerate parameters, and that we wish to test H_0: $\theta_1 = 0$ against H_A: $\theta_1 \neq 0$, the null hypothesis being composite because the values of $\theta_2, \ldots, \theta_q$ are unknown. Rao (1965) has considered this problem. He defines the positions of the restricted and unrestricted likelihood maxima to be $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ (where $\theta_1^* = 0$) and defines $\mathbf{V}(\boldsymbol{\theta})$ to be the vector of first derivatives of the total log likelihood function $L(\boldsymbol{\theta}, \mathbf{x})$, noting that if $\mathbf{x}$ is distributed according to the parameter value $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ then $\mathbf{V}(\boldsymbol{\theta}_t)$ has an asymptotically normal distribution with mean $\mathbf{0}$. Writing the variance/covariance matrix of $\mathbf{V}(\boldsymbol{\theta}_t)$ as $\Sigma(\boldsymbol{\theta}_t)$, he suggests the asymptotically efficient statistic $V_1^2(\boldsymbol{\theta}^*)\,\Sigma^{-1}(\boldsymbol{\theta}^*)_{1,1}$, conjecturing that this will be of greater local power than either the Neyman–Pearson likelihood ratio test or Wald's asymptotically efficient test. The actual value of $\boldsymbol{\theta}$ in $\Sigma(\boldsymbol{\theta})$ is not critical, since estimates of Σ^{-1} are asymptotically stable with respect to small variations in $\boldsymbol{\theta}$; Wald uses $\Sigma(\hat{\boldsymbol{\theta}})$, and when discussing this Rao justifies his own choice of $\Sigma(\boldsymbol{\theta}^*)$ by remarking that it saves calculating $\hat{\boldsymbol{\theta}}$ explicitly. The ideal value would presumably be $\boldsymbol{\theta}_t$, but this is not known. However, this ideal statistic $V_1^2(\boldsymbol{\theta}^*).\Sigma^{-1}(\boldsymbol{\theta}_t)_{1,1}$ is a monotonic function of the statistic $\frac{1}{2}V_1(\boldsymbol{\theta}^*)$, and this latter quantity is therefore an asymptotically efficient test statistic for testing H_0: $\theta_1 = 0$ against H_A: $\theta_1 \neq 0$.

Suppose that $\mathbf{x}$ consists of N independent observations $x_1, \ldots, x_N$, each parametrized by a parameter θ of interest and a vector $\mathbf{p}$ of nuisance parameters. (For the purposes of Section 3, there are no parameters other than θ and $\mathbf{p}$ is a null vector.) Let $\hat{\theta}, \hat{\mathbf{p}}$ denote the values of θ and $\mathbf{p}$ that jointly maximize the total log likelihood function $\sum L_i(x_i, \theta, \mathbf{p})$ and define $U_i = \partial L_i/\partial\theta$ at $\hat{\theta}, \hat{\mathbf{p}}$. Suppose that $\mathbf{p}$ is the same for all the observations, that $\theta = \theta_A$ for some (group A) and $\theta = \theta_B$ for the remainder (group B) and that we wish to test H_0: $\theta_A = \theta_B$ against H_A: $\theta_A \neq \theta_B$. Note that

$$\sum_{\text{group } A} U_i + \sum_{\text{group } B} U_i = 0.$$

This problem can be reduced to the previous form by writing $\theta_A = \theta_2 + \theta_1$, $\theta_B = \theta_2 - \theta_1$ and $p_i = \theta_{2+i}$ and testing H_0: $\theta_1 = 0$ against H_A: $\theta_1 \neq 0$. In this case the asymptotically

efficient statistic $\frac{1}{2}V_1(\theta^*)$ equals

$$\frac{1}{2}\left(\sum_{\text{group } A} U_i - \sum_{\text{group } B} U_i\right) = \sum_{\text{group } A} U_i.$$

APPENDIX B

Approximations using Pearson Curves

Let X be the sum of a random sample of size $n>0$ chosen without replacement from the $N>n$ zero-sum scores $U_1, \ldots, U_N$. There are up to ${}^NC_n$ possible values for X, each combination having equal probability.

Writing $m_i = \sum_{j=1}^{N} U_j^i/N$, μ_i for the ith moment of X and r for $n(N-n)/(N-1)$, it can be shown that

$$E(X) = \mu_1 = n.m_1 = 0,$$

$$V(X) = \mu_2 = r.m_2,$$

$$\mu_3 = r(N-2n)\,m_3/(N-2)$$

and

$$\mu_4 = r[m_4+3(n-1)(N-1-n)(N.m_2^2-2m_4)/\{(N-2)(N-3)\}].$$

Define $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$.

The distribution of $X/\sqrt{\mu_2}$ may now be approximated by the Pearson distribution with the same first four moments. This distribution has been tabulated by Johnson *et al.* (1963). Alternatively the significance level of X may be computed as follows.

To avoid trivial complications, assume $N>7$ and that there are more than two possible values for X. Write $C = 6(\beta_2-\beta_1-1)>0$ and $D = 3\beta_1+6-2\beta_2$. We will only describe the computational procedures for the case $D>0$ since $D>0$ for all but possibly a few very large or small values of n. (For a particular set of scores write $w = 3m_3^2+6m_3^2-2m_4m_2$, then $D>0$ if n lies in the interval $r<n<N-r$, where $r=0$ if $w>0$ and $1\leqslant r<N/4$ otherwise.)

With $D>0$, write $R = C/D$, $\phi = \beta_1(R+2)^2/\{(16(R+1)\}$, $\theta = \operatorname{sgn}(\mu_3)\sqrt{\phi}/(1+\phi)$, $p = R(1-\theta)/2$, $q = R(1+\theta)/2$, $a = \{4\mu_2(1+\phi)(1+R)\}^{-\frac{1}{2}}$ and $b = p/R$. Then $p>0, q>0$ and $aX+b$ has the same first four moments as the Pearson type I or II (beta) distribution with p.d.f.

$$f(w) = w^{p-1}(1-w)^{q-1}/B(p,q).$$

The one-sided significance level of a particular value X is then estimated by computing (Ludwig, 1963)

$$\int_0^{aX+b} f(w)\,dw \pm 0{\cdot}5/{}^NC_n,$$

the last term being a continuity correction, applicable in the absence of widespread tied values when most steps in the c.d.f. of X are of size $1/{}^NC_n$. If, as in Wilcoxon's rank sum test, the U_i are mostly small coprime integers then a more appropriate continuity correction is to replace X in this integral by $X\pm 0{\cdot}5$.

APPENDIX C

Weighted Sum of Hypergeometric Random Variables

Let $r(p,m,M)$ be a hypergeometric random variable with real parameter p $(0\leqslant p\leqslant 1)$ and integer parameters m and M $(0\leqslant m\leqslant M$, Mp integral); r takes the

value j (for $j \leqslant Mp$ and $m-j \leqslant M-Mp$) with probability

$$\binom{Mp}{j}\binom{Mq}{m-j}\Big/\binom{M}{m}$$

where $q = 1-p$. $E(r) = mp$ and writing c_k for the kth central moment of r we have

$$c_2 = \frac{mpq(M-m)}{M-1} \qquad \text{(or zero if } M<2\text{)},$$

$$c_3 = \frac{c_2(q-p)(M-2m)}{M-2} \qquad \text{(or zero if } M<3\text{) and}$$

$$c_4 = \frac{c_2[M(M+1)-6m(M-m)+3pq\{M^2(m-2)-Mm^2+6m(M-m)\}]}{(M-2)(M-3)}$$

(or $c_2(1-3pq)$ if $M<4$).

Now let $r_i\ (1 \leqslant i \leqslant w)$ be independent hypergeometric random variables with parameters p_i, m_i and M_i and central moments c_{ij} $(j = 2, 3, 4)$, and let $\lambda_i\ (1 \leqslant i \leqslant w)$ be a set of w constants. Then $Y = \sum \lambda_i(r_i - m_i p_i)$ has mean zero and central moments

$$\mu_2 = \sum \lambda_i^2 c_{i2}, \qquad \mu_3 = \sum \lambda_i^3 c_{i3},$$

and

$$\mu_4 = 3\mu_2^2 + \sum \lambda_i^4 (c_{i4} - 3c_{i2}^2).$$

ACKNOWLEDGEMENTS

We wish to thank the referees for helpful remarks on previous versions of this paper. We are even more grateful to Malcolm Pike, without whose assistance this version would never have existed, and to Virginia Castle and Gale Mead, without whom it would never have been typed.

REFERENCES

ALTSHULER, B. (1970). Theory for the measurement of competing risks in animal experiments. *Math. Biosciences*, **6**, 1–11.

GEHAN, E. A. (1965). A generalised Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203–223.

HAJEK, J. and SIDAK, Z. (1967). *Theory of Rank Tests*. New York: Academic Press.

HILL, I. D. and PETO, R. (1971). Algorithm AS 35. Probabilities derived from finite populations. *Appl. Statist.*, **20**, 99–105.

JOHNSON, N. L., NIXON, E., AMOS, D. E. and PEARSON, E. S. (1963). Table of percentage points of Pearson curves, for given $\sqrt{\beta_1}$ and β_2, expressed in standard measure. *Biometrika*, **50**, 459–498.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.*, **53**, 457–481.

LUDWIG, O. G. (1963). Incomplete beta ratio. *Comm. Ass. Comp. Mach.*, **6**, 314.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50** (3), 163–170.

M.R.C. WORKING PARTY ON THERAPEUTIC TRIALS IN LEUKAEMIA (1972). Report on the first myelomatosis trial (Part 1): to appear in *Brit. J. Haematol.*

PIKE, M. C. and ROE, F. J. C. (1963). An actuarial method of analysis of an experiment in two-stage carcinogenesis. *Brit. J. Cancer*, **17**, 605–610.

RAO, C. R. (1965). *Linear Statistical Inference*, pp. 347–352. New York: Wiley.

ROE, F. J. C., PETO, R., KEARNS, F. and BISHOP, D. (1970). The mechanism of carcinogenesis by the neutral fraction of cigarette smoke condensate. *Brit. J. Cancer*, **24**, 788–806.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**, 80–83.

Discussion on the Paper by R. and J. Peto

Professor R. N. Curnow (University of Reading): Mr President, ladies and gentlemen. I congratulate the two authors on the stimulating and useful paper that they have presented to us this evening. There are occasions when any distributional assumptions about data can be dangerous. This applies particularly to data on survival, whether it be of living organisms or of physical material. The authors have presented a method of analysis that will be very useful in these situations.

I found the authors' written presentation rather too formal for my taste and I wonder if this may result in the paper not being understood and the methods used as much as they should. I would have preferred a more intuitive presentation followed by a tidying-up operation relating the method to other methods and discussing questions of asymptotic efficiency and local power. Certainly the paper puts the methods in the perspective of general theory and this is very useful.

On points of detail, why do the authors suggest in Section 5.1 the approximate U score $1+\log[H(t)]$ for an exact observation when an exact U score could be used? Where does the advantage lie in finite samples? In the same Section, are there any reasons to prefer Altshuler's estimate of the survival curve to Kaplan and Meier's? The expectations and variances of e_i, d_i and $d_i - e_i$ given in Section 10, paragraph 2, follow from general theorems on maximum-likelihood estimation but are given somewhat abruptly. I must comment adversely on the so-called example (Section 10, last paragraph). The example is not an example at all. Where are the data? Where are the inferences to be drawn from applying the method to the data? The examples presented orally this evening should have been included in the written paper.

The authors suggest that, with survival data, the observed and expected numbers of deaths for the different groups provide a useful presentation of the scores. This is true, but the whole shape of the experimental survival curve for the different groups may be of interest, not just the number of deaths to a particular date. We have used the methods of this paper in analysing some data on different treatment régimes for geriatric patients. The survival time distributions for geriatric patients are very far from standard and the methods of this paper have proved very useful in analysing the data. There were about 600 patients in our geriatric trial and this meant that we could present to the consultants meaningful experimental survival curves for the different groups of patients. The curves are, of course, unbiased estimates of the underlying real survival curve and the test statistic provides a test of the differences between a particular feature of these survival curves. I leave it to others to comment on the fact that the experimental survival curve above the point of longest survival to date is always zero. With our large numbers of patients we have so far found it easier to apply the methods of this paper to interesting sub-groups of patients rather than use the procedure suggested for concomitant variables.

With the logrank test, the scores are linearly related to the survival time of the uncensored individual and the expected survival time of the censored individual if we assume the survival curve for the different groups is the same exponential and $H(t)$ is exact, i.e. $H(t) = e^{-\lambda t}$. Does this generalize and do the authors attach any importance to this result? Have the authors considered the possibility of using the experimental survival curve to calculate the expected survival time for censored individuals and using this instead of the logrank scores they suggest?

I hope that the authors or others interested in these methods will carry out more detailed studies of the finite sample properties of these tests. I hope also that some attempt will be made to look at the use of the test statistic in sequential situations. In clinical trials it is clearly imperative that the trial is stopped as soon as possible. Have the authors any ideas on how the method might be used in a sequential manner?

This evening we may be hearing the parametric versus non-parametric arguments. I think that the main question we should consider is at the design or planning stage rather than at analysis. How should we advise experimenters? Assume that we are,

a priori, unable to make sufficient distributional assumptions to specify parameters for estimation. I suggest that we should more strongly advocate the collection of sufficient data to allow some estimate of distributional form followed by estimation of its parameters. This may mean fewer experiments, but they would be larger and therefore more informative ones. One decisive experiment is worth many indecisive ones. In a sequence of experiments, the strength of the argument may depend on the weakest experiment in the sequence. They may all need to be strong. I am really returning to my earlier point about the need to estimate whole distributions. I think that we have oversold the amount that we learn from significance tests that compare some particular feature of unspecified distributions. This evening's paper has removed some of the arbitrariness about the feature chosen for comparison but does this go far enough?

I have much pleasure in proposing a vote of thanks to Richard and Julian Peto for their paper.

Mr W. BRASS (London School of Hygiene & Tropical Medicine): Since this is an Ordinary Meeting of the Society with particular sponsorship from the Medical Section I make no apology for dealing with the practical issues of applied statistics. It also simplifies my task by letting me avoid questions outside my competence. My main difficulty was in being sure what the paper was really about—not the details although these are heavy enough but what advances in knowledge and techniques were being claimed. The authors have been rather too modest in not providing a section which spelled out how useful their work was—and exactly for what. In fact, the necessary drum-beating has been done by them in their oral introduction. I only wish I had heard it before I tried to read the paper.

As I see it they have developed one logical and convenient approach for examining certain important properties of tests based on rank scores in applications to data on survivorship. Their procedure is a good starting point for the construction of useful tests for particular applications. One of the most valuable practical consequences is the neatness with which censored observations are incorporated into the method, whether the censoring is random with respect to the groups being compared or not. In applications of the kind considered censoring is a common nuisance and allowing for it can be a serious complication.

The most interesting feature of the development to me is the elimination of the relation between the survivorship curves and the original time variable as a relevant feature and the concentration instead on the relative behaviour of the curves themselves. There are many analyses of vital statistics where the variation in rates with time is complex but the patterns are similar for different sub-populations, leading to a simple internal relationship, at least to a reasonable approximation. The methods proposed are particularly useful in this situation.

For some years I have been using a model life-table system, extending to mortality at all ages, for which the family of conversion functions, in the terms of this paper, is linear on a logit scale. In general, this is a two-parameter system but if we retain the one measuring a shift in the mean of the hyperbolic $\sec^2$ frequency distribution but fix the corresponding variance we have a set very close to the probit family of Section 5.2. In the notation there $c(y, \psi) = y/\{y+(1-y)\exp\psi\}$ where $y = 1/(1+\exp x)$, $\psi = \theta - \theta_0$. A logit rank test can then be obtained by the procedure given and it turns out to be the same as the Wilcoxon test as generalized by the authors in Section 6. Because of the close similarities between the logit and probit functions it is not surprising that the power of the Wilcoxon test is very close to that of the probit rank analogue in the asymptotic examples of Section 7. We might guess that in practice, with finite samples—and particularly with censoring—the tests would behave virtually identically.

The logrank test is efficient when the forces of mortality in the two-sub-populations compared are in a constant ratio over time; the Wilcoxon test does best when the ratio tends to one in a particular way as time increases. My own prejudice is that the latter pattern for the relationship is generally the more likely.

The main justification of tests based on ranks must be their robustness to deviations from the assumptions. I think it is fair, therefore, to suggest that the authors' might have given this more consideration—anyway it is the standard tactic for a discussant who has run out of ideas to ask for more. Specifically they might have spent some effort on examining the power of the tests for censored data because of its practical importance. The effect of deviations from the assumptions about the relations between forces of mortality in the groups is also worth looking at. Both the logit and probit families can be shifted towards the Lehmann by altering the parameter corresponding to the standard deviation of the frequency distribution. It would be useful to know how the tests behave in such circumstances.

I have pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Dr M. C. PIKE (DHSS Cancer Epidemiology and Clinical Trials Unit, Regius Department of Medicine, Oxford University): One very attractive feature of the method of analysis based on the W scores of Section 5.1 is its simplicity. The W scores can be calculated very easily for each patient and then, dividing the range of any explanatory variable into three or four groups, the analysis can be completed, very nearly optimally, on a desk calculator using the well-known epidemiological technique of indirect standardization. Linear trend terms can be incorporated into this scheme at the cost of some little extra effort (Armitage, 1966; Mantel, 1963).

This method of approach has the advantage, to my way of thinking anyhow, of almost forcing one to be very aware of interactions. These may be the most important things to look for. Consider the attached table showing the results of the Medical Research Council's Concord Trial in acute lymphoblastic leukaemia of childhood (Medical Research Council, 1971). The treatments were chemotherapy with the drug methotrexate (the "best" treatment known at the start of the trial), repeated vaccinations with B.C.G., and no further specific therapy. The last line of the table shows that the best treatment group was that on methotrexate, and the right-hand marginal totals show that a high initial leucocyte count is associated with shorter remissions. The relapse rates of the methotrexate group in relation to the combined B.C.G. plus no treatment groups are, however, as follows:

Initial leucocyte count	*Others* : *Methotrexate*
0–	15·5 : 1
6,000–	3·3 : 1
21,000+	1·1 : 1

Clearly methotrexate does not work for poor prognosis patients. This means that in future trials this group of patients should be treated very much as a separate group. Treatment protocols for them will bear little relation to protocols for patients with low initial leucocyte counts.

Thirty years ago physicians were on the whole against randomized clinical trials, arguing that each patient was different and that this effectively precluded making advances by this method. Bradford Hill and others campaigned against this, and the rapid progress in the treatment of tuberculosis, brought about to a significant extent by clinical trials, is a measure of their success (Hill, 1962). Reports of these trials devote considerable attention to showing that the random allocation to treatment groups had "worked" in the sense that these groups had a similar distribution of "good" and "poor" prognosis patients. I think with today's paper we can now say we have definitely passed this stage and say to the physician that we have met him half-way. We now have the technique which allows us,

Medical Research Council's Immunotherapy (Concord) trial in acute lymphoblastic leukaemia
Remission after randomization

Initial leucocyte count	*Methotrexate*				*B.C.G.*				*No treatment*				*All treatments*			
	N	*O*	*E*	*O/E*	*N*	*O*	*E*	*O/E*	*N*	*O*	*E*	*O/E*	*N*	*O*	*E*	*O/E*
0–	24	1	13·75	0·07	21	12	12·83	0·94	6	6	3·04	1·98	51	19	29·62	0·64
6,000–	13	3	7·69	0·39	16	10	7·65	1·31	6	5	4·01	1·25	35	18	19·35	0·93
21,000+	15	11	6·46	1·70	15	11	6·33	1·74	6	5	2·23	2·24	36	27	15·02	1·80
All	52	15	27·90	0·54	52	33	26·81	1·23	18	16	9·28	1·72	122	64	64·00	1·00

N = No. of patients. *O* = No. of patients relapsing.
E = Expected No. of patients relapsing. *O/E* = Relative relapse rate.

maybe not to regard each patient as different, but to break the patients into sub-groups *whose treatments in trials may well be different.*

Finally, I think it should be emphasized that the logrank test statistic is also generated by considering the data as a sequence of 2 by 2 tables (Mantel, 1966) and in this version has been used for a long time in clinical trials by a number of groups in the United States.

Professor D. R. Cox (Imperial College): It is a pleasure to congratulate the authors warmly on a most original and interesting paper. I have a number of comments and questions on detail.

Censoring operating unequally on the different groups is discussed briefly in Section 12; the exact permutation tests then do not hold. How important in practice is unequal censoring? How far out are the permutation significance levels likely to be?

In the two-sample problem with survivor functions $F(x)$, $\{F(x)\}^\theta$ how are confidence limits for θ best calculated when θ is appreciably different from one?

Are fairly simple procedures available when a matched pair design is used?

In my own unpublished work on regression problems connected with life tables the hazard function (age-specific death rate) for an individual of age x and with regressor variables $z_1, \ldots, z_p$ is assumed to be

$$\exp(\beta_1 z_1 + \ldots + \beta_p z_p)\, \lambda_0(x),$$

where $\lambda_0(x)$ is the unknown hazard function in the "standard" condition $z_1 = \ldots = z_p = 0$ and the β's are regression coefficients. Conditional inference can be used if interest is concentrated on the β's. If the z's are independent of age this is the generalized Lehmann model used by the authors, but in fact the conditional arguments hold also when the z's are functions of age and, at least in theory, this gives some extra flexibility. Can the authors' methods be similarly extended?

Professor J. Durbin (London School of Economics): I wish to add my congratulations to those of the other speakers on what seems to me to be a very solid achievement by the authors. However, I regret that they did not discuss the close relation between the experimental survival curve and the sample distribution function (d.f.). Indeed, if there are no tied values and no censoring the two are equivalent. There is an enormous literature on tests based on the sample d.f. and it would be interesting to see how much light this throws on the problems under consideration.

My own feeling is that one would normally expect to gain greater insight into the effects of treatment differences from a visual comparison of the experimental survival curves than by carrying out a formal test of significance, even an efficient one. From this point of view the function of the significance test is a subsidiary one, i.e. the provision of a yardstick against which the observed differences between two curves can be assessed; one therefore looks for a statistic which one can relate in a fairly direct way to the difference between two curves, e.g. the Kolmogorov–Smirnov statistic.

The comparison with the sample d.f. would also have benefit in the other direction, e.g. in the study of tests on the sample d.f. in the presence of censoring and of tied or grouped data. The authors' work on tests in which ranks are replaced by efficient scores also suggest the possibility of using a modified form of sample d.f. which has jumps with sizes determined by efficient scores instead of taking all jumps equal to 1/n.

Mr I. D. Hill (Medical Research Council): In Dr Pike's contribution to the discussion, he stressed that one of the beauties of the Peto method is that it is so easy to use; but nobody would guess that that was so from looking at the present paper, which of course was not intended as a users' guide.

A "cook-book" users' guide is needed, however, if the method is to achieve any widespread use. May I enquire whether anyone has any plans to produce one?

The following contributions were received in writing, after the meeting:

Professor A. STUART (London School of Economics): The efficiency of the permutational test in Section 3 is essentially implied by the following results, where reference is made, for the sake of convenience, to Volume 2, 2nd edition, of Kendall and Stuart:

1. The fact mentioned just below (22.36) on p. 173, that the one-sided test mentioned there is efficient.
2. The fact that maximum-likelihood estimators converge strongly to the true parameter values, so that the result mentioned above would also hold asymptotically if they were substituted.

Apart from the efficiency point, it seems to me that the null distribution is that of a sum chosen from a finite population on essentially the same grounds as are familiar for the Wilcoxon two-sample test (Exercise 31.15, p. 571) or for Pitman's more general test (Section 31.46, p. 489).

Professor EDMUND A. GEHAN (University of Texas at Houston): This paper is a welcome addition to the literature of non-parametric tests appropriate for comparing survival distributions. It certainly is a common situation, at least in clinical trials for comparing survival distributions, that it would be desirable to test whether the hazard rate in one group is a constant multiple of that in the other. The logrank test seems eminently suitable for this type of problem. Before applying this test in all such circumstances, however, it would be useful to have some further information concerning its characteristics. What is the power of the test when some censored data are present in both groups? It is possible, though perhaps not likely, that censored data would have a more severe effect on the asymptotic efficiency of the logrank test than, say, a version of Wilcoxon's rank sum test so that the latter may become preferable if certain types of censoring are present. Also, it would be of interest to know whether Altshuler's estimate, $\log\{H(t)\} = -\sum_{x \leqslant t} s(x)/r(x)$ is really any better than $\log H(t)$, where $H(t)$ is Kaplan and Meier's estimate of the survivorship function.

It should be pointed out (if it has not already been so) that the logrank statistic is the same as the statistic $U(0)$ which Cox (1972) proposes for the two-sample problem with censored data.

The authors consider the normal approximation to the calculation of permutational significance levels to be "inadequate". Since such an approximation is part of established practice, has surely been applied significantly more often than exact counting or other methods and could be applied here, it would be of great interest to have a more precise statement of how poor the approximation is.

When discussing generalizations of Wilcoxon's rank sum test to censored data, the Petos suggest a scoring system and state "...[it] is to be preferred to Gehan's (1965) generalization, since the relative values of the expectations of the scores Gehan assigns to exact observations at particular times vary according to the pattern of censoring imposed on the observations". Of course, the only really convincing basis for preferring one test to another in a given circumstance is from consideration of the power against alternative hypotheses (and, in this case, when censored data are present). The authors do not supply any information on this point. It might also be mentioned that it can be quite sensible to have a scoring system that is dependent to some degree on the observed censoring pattern. Suppose, for example, in a clinical trial comparing two treatments, A and B, that there is on the average equal exposure to the risk of failure in the two groups, but many more censored observations are observed in the A group. This is an indication of the effectiveness of treatment A in delaying or preventing failure and having a scoring system which is to some extent dependent on the pattern of censoring may well be preferred.

Finally, the authors have proposed a number of tests of high asymptotic efficiency appropriate for testing Lehmann alternatives, shifts of a normal mean and have described a method for generating tests in other circumstances. Suppose that before an experiment

is conducted, a research worker has no good ideas of what types of alternatives to test for (though such an idea may be developed after the data have been collected), what test should be recommended? Might it not be a Wilcoxon type test because of its simplicity and relatively high asymptotic efficiency, even though it is not asymptotically most efficient for any family of distributions?

The authors replied in writing, as follows: Every analysis of survival data should either be efficient against Lehmann alternatives or have a very clear reason for not being so: the Lehmann family is the "normal" distribution of survival theory. In a Lehmann family, the display of "observed" and "expected" event-counts in various subgroups (Mantel, 1966), regression on explanatory variables (Cox, 1972) and logrank-type permutational significance tests (our paper) form as unified a whole as do normal-distribution regression coefficients, degrees of freedom and sums of squares. Some of the contributors to this discussion have suggested various ways in which our logrank test might be modified, but in fact any modification of the logrank test destroys this unified structure and wastes statistical power; the permutational logrank test defined in Section 5.1 is not merely asymptotically efficient against Lehmann alternatives, it is actually of *greater local power* than any other rank test. This is exactly true for any particular finite sample size and for any particular way the censoring happens to occur, and is the best property possible for a rank test that compares groups of similarly censored (or uncensored) event times. We are emphasizing the maximal local power of the logrank test at some length because several discussants have not noticed it, and have suggested that various modifications might improve the efficiency when censoring occurs or the sample size is small. We repeat, no matter what the censoring pattern or the sample size may be, no modification can increase the local power of the permutational logrank test against Lehmann alternatives without sacrificing rank invariance.

The loss of efficiency following local deviations from the assumed relationship between the survival curves of the different groups is difficult to discuss, since no general parametric description of such variations is possible. However, since the loss of asymptotic efficiency when the logrank test is used in the t-test situation is only 18 per cent, we may guess that the effect of less gross deviations is marginal. Professor Cox's question on the effect of erroneously assuming that the censoring patterns in different groups are the same is, for the same reason, difficult to answer generally; the method of Section 12 should be used whenever the symmetry of censoring is in any doubt.

Professor Cox rightly implies that we cannot deal with time-dependent concomitant variables, and that his regression method leads to better parameter estimates and confidence limits. However, our two-group test procedures are better, and the advantages of his regression model and our permutational test can be combined by basing the corrected U-scores of Section 11 on his model, as indicated in the discussion following his paper.

The other question that many people have raised is this: the Wilcoxon generalization and the logrank are equally easy tests to perform. Under what circumstances is the Wilcoxon generalization better? Our answer is that if your data are *times*, then the logrank is probably better, whereas otherwise the Wilcoxon is probably preferable, especially if the distributions have little skewness.

We would like to thank all the contributors to this discussion, especially those who warn that mere significance testing is not good statistics: look for interactions, plot several experimental survival curves, fit various regression models and perhaps derive some confidence intervals. We agree entirely with these suggestions: the more data are looked at, the better they will be understood. In this context, we feel that sequential methods may be misleading if a study is stopped before the pattern of mortality is fully evident. For instance, sequential analysis of a trial in alcoholics of total withdrawal compared with continued drinking might, due to deaths during withdrawal, prove that continued drinking was far safer than abstinence.

We would especially like to answer David Hill, who wants to know where a "cook-book" users' guide to the logrank test can be found. The chapter entitled "Leukaemia

Trials" in the forthcoming edition of *Medical Surveys and Clinical Trials* (Oxford University Press) contains a description for non-statisticians of the logrank methods, and "Three algorithms for doing permutational tests", to appear in *Applied Statistics*, implements the calculation of Kaplan and Meier's "life-table" experimental survival curve, Altshuler's log survival curve estimator, the logrank scores and the Pearson approximation to the distribution of a sum chosen at random from a finite population. Two papers which make extensive practical use of logrank methods are referred to in our text: the Medical Research Council's report on the first myelomatosis trial (Part 1) is on a clinical trial with many interesting explanatory variables, and the 1970 paper by F. J. C. Roe *et al.* is on a multi-group mouse painting experiment, where time to carcinogenesis was being studied.

Although, as Professor Stuart remarks, the proof of the result of Appendix A and Section 3 is not difficult, the consequences of it are interesting for non-rank-invariant testing. The result is that if we ever wish to do a test between two groups of observations which are i.i.d. under H_0, then a *permutational* test can be used which can never generate an unjustifiably high significance level and which is asymptotically efficient against whatever alternative behaviour is hypothesized. As an example of this, suppose we are asked to say whether the means of two groups of supposedly i.i.d. normal observations differ significantly. A t-test assumes normal distributions for its significance level to be valid. Is it not better to argue instead that under H_0 one mean is the average of a random combination of appropriate size selected from the pool of both groups of observations mixed together? This can never be wrong, and if the observations actually are normal it is of full asymptotic efficiency. The general idea of giving a permutational distribution to the Fisher efficient score in any two-group test situation is an old one, but it is rarely applied.

References in the Discussion

ARMITAGE, P. (1966). The χ^2 test for heterogeneity of proportions after adjustment for stratification. *J. R. Statist. Soc.* B, **28**, 150–163.

COX, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B, **34**, Part 2, (in the press).

HILL, A. B. (1962). *Statistical Methods in Clinical and Preventive Medicine.* London: E. & S. Livingstone.

MANTEL, N. (1963). χ^2 tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Ass.*, **58**, 690–700.

MEDICAL RESEARCH COUNCIL (1971). Treatment of acute lymphoblastic leukaemia. *Brit. Med. J.*, **4**, 189–194.

As a result of the ballot held during the meeting the following were elected Fellows of the Society:

ANNABLE, Lawrence, B.Sc.
BAILLIE, Richard Thomas, M.Sc.
BALDWIN, Roger, M.Sc.
BATES, Alan Newton, B.Sc.
BHANSALI, Rajendra Jagmohan, Ph.D.
BONE, Alan John, B.Sc.
CHILVERS, David John, M.Sc. (Econ.)
CLARK, Christopher Richard, M.Sc.
CLARKE, David Allan
CONRAD, Simon Andrew M.Sc.
DAREKAR, Bal Swaruprao, M.Sc.
DAY, William Harold Leonard, B.Sc. (Econ.)
FAN, Shuh-Ching, Ph.D., F.I.S.
FEENY, Michael John, B.A.
GODDARD, Paul Robert, M.Sc.
GRAHAM, Philip Sandford, B.Sc.
HAM, Arthur Frederick, B.Sc.
HANSON, Patrick Robert, H.N.C. Maths. Stats. & Computing, A.I.S.
HARDY, Dennis Leslie, A.I.M.T.A.
HUSSAIN, Mohammed Yassir, B.A.
ISHAM, Valerie Susan, B.Sc.
KOUVATSOS, Drakoulis-Demetrius, M.Sc.
LEWIS, Trevor, B.Sc.
LIM SHIN CHONG, Jean Lim Chee Yiin, B.Sc.
LOCK, Raymond Authur John

MATTHEWS, David John
MITHANI, Almas, M.Sc.
OBASI, Godwin Olu Patrick, D.Sc.
PALMA CARLOS, Ana Isabel, M.Sc.
RIX, John Philip, M.Sc. A.I.S.
ROBINSON, Jeffrey Nicholas, M.Sc.
SEN, Mono Ranjan, M.A.
SHAROT, Trevor, B.Sc.
SWARIS, Roger Neville M.Sc.
SYM, Roger, Ph.D.
TILLING, Richard John, B.Sc. (Econ.)
TULPULE, Ashok Hamumant, M.A.
VANSTONE-WALKER, Christine, B.Sc.
WADDINGHAM, Robert Adrian Joseph, B.Sc.
WALES, Francis Richard, F.I.A.
WARWICK, Kenneth Marshal, Ph.D.
WILLIAMS, Abdul-Rafiu Oladapo
WILLS, Victor John Stormont
YOUNG, Abimbola Sylvester, M.Sc.