5-1-2002

Exact Level And Power Of Permutation, Bootstrap, And Asymptotic Tests Of Trend

Christopher D. Corcoran
Utah State University

Cyrus R. Mehta
Cytel Software Corporation

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Exact Level And Power Of Permutation, Bootstrap, And Asymptotic Tests Of Trend

Christopher D. Corcoran
Department of Mathematics and Statistics
Utah State University

Cyrus R. Mehta
Cytel Software Corporation

We develop computational tools that can evaluate the exact size and power of three tests of trend (e.g., permutation, bootstrap and asymptotic) without resorting to large-sample theory or simulations. We then use these tools to compare the operating characteristics of the three tests. It is seen that the bootstrap test is ultra-conservative relative to the other two tests and as a result suffers from a severe deterioration in power. The power of the asymptotic test is uniformly larger than that of the other two tests, but it fails to preserve the Type I error for most of the range of the baseline response probability. The permutation test, being exact, is guaranteed to preserve the Type I error throughout the range of the baseline response probability. The price paid for this guarantee is a loss of power relative to the asymptotic test. The power loss is, however, small in most situations.

Keywords: Type I error, Power, Trend, Permutation tests, Bootstrapping

Introduction

Forty mice were divided into four equal groups. Each group was treated with a different dose of an animal carcinogen as a result of which some mice developed a tumor. The data are displayed in Table 1. The goal is to test for a dose-response relationship. Specifically, let π_j be the Bernoulli probability that an animal treated at dose d_j develops a tumor. We wish to test the null hypothesis

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 \equiv \pi \qquad (1.1)$$

against the one-sided alternative hypothesis

$$H_1 : \pi_1 \le \pi_2 \le \pi_3 \le \pi_4 \qquad (1.2)$$

Table 1: Dose-Response Data for Animal Carcinogenicity Study

Response Status	Dose d_j assigned to all mice in group j				Total
	$d_1 = 0$	$d_2 = 1$	$d_3 = 5$	$d_4 = 50$	
Tumor	1	0	1	3	5
No Tumor	9	10	9	7	35
Total	10	10	10	10	40

with at least one inequality n equation (1.2) being strict. The value of π, the common response probability under H_0, is typically unknown.

An efficient test of the null hypothesis is the Cochran-Armitage test of trend (Cochran, 1954; Armitage, 1955), in which the test statistic is

$$T(x) = \sum_{j=1}^{4} d_j x_j \qquad (1.3)$$

where x_j is the entry in row 1 and column j of a generic 2×4 contingency table, x, with column sums of 10 in each of the four columns. Substituting the Table 1 data into equation (1.3), the observed value of the test statistic is 155. It is usual to test the null hypothesis by computing a p-value, defined as the probability under H_0 of observing a table whose test statistic equals or exceeds 155. A major difficulty with performing this computation is that even under H_0 the probability of observing any table depends on the unknown nuisance parameter π. We will evaluate three different methods for computing the p-value in the presence of this nuisance parameter. The three methods are bootstrap resampling, permutation resampling and normal approximation. We have two objectives in writing this paper. Our first objective is expository. We wish to clarify the distinction between the bootstrap and permutation resampling methods because these two terms are frequently confused. Our second objective is to compare the performance of all three methods with respect to Type I error and power.

The Bootstrap P-Value

The bootstrap p-value is obtained by resampling from the reference set, Γ, of all 2×4 tables with column sums equal to 10. Under H_0 the probability of observing any $x \in \Gamma$ is

$$f_\pi(x) = \prod_{j=1}^{4} \binom{10}{x_j} \pi^{x_j} (1 - \pi)^{10 - x_j} \qquad (1.4)$$

a product of four binomial probabilities. It is not possible to resample tables from Γ with probabilities given by (1.4)

Christopher D. Corcoran is Assistant Professor, Department of Mathematics and Statistics, 202 LUND, Utah State University, Logan, UT, 84322-3900. Email: corcoran@math.usu.edu. Cyrus R. Mehta is President, Cytel Software Corporation, and Adjunct Professor of Biostatistics, Harvard School of Public Health.

because π, the Bernoulli probability under H_0, is unknown. We can, however, replace π with

$$\hat{\pi} = \frac{5}{40}$$

the maximum likelihood estimate (MLE) under the null hypothesis. The bootstrap p-value is then evaluated by resampling tables from Γ with probabilities given by

$$f_{\hat{\pi}}(x_j) = \prod_{j=1}^{4} \binom{10}{x_j} \hat{\pi}^{x_j}(1 - \hat{\pi})^{10-x_j} \qquad (1.5)$$

Suppose we resample M tables in this manner, denoted by $x_1, x_2, \ldots, x_M$. The bootstrap p-value is evaluated as

$$\tilde{p}_b(M) = \frac{\sum_{i=1}^{M} I\{T(x_i) \geq 155\}}{M} \qquad (1.6)$$

where $I\{.\}$ is the indicator function. In other words we resample tables from $\tilde{A}$ by treating the empirically observed value of π as though it were the true value and estimate the bootstrap p-value as the fraction of resampled tables that are at least as extreme as the observed table with respect to the Cochran-Armitage test statistic. For the data in Table 1 the bootstrap p-value based on $M = 100,000$ samples was found to be $\tilde{p}_b(M) = 0.0941$. In repeated samples this value would vary due to the sampling error associated with $\tilde{p}_b(M)$. The sampling error decreases in proportion to the square root of M. In the limit as $M \to \infty$ the bootstrap p-value converges to the constant

$$p_b = \sum_{\substack{x \in \Gamma \\ T(x) \geq 155}} f_{\hat{\pi}}(x_j) \qquad (1.7)$$

For the data in Table 1, $p_b = 0.0954$, which is almost the same as $\tilde{p}_b(M)$ at $M = 100,000$. Although increasing M eliminates the sampling error associated with $\tilde{p}(M)$, it cannot eliminate the error associated with using $\hat{\pi}$ as an estimate for the unknown nuisance parameter π in equation (1.7). Thus, the accuracy of bootstrap p-value depends on how well $\hat{\pi}$ approximates π rather than on M, the number of times resampled from Γ.

The Permutation P-Value

The permutation p-value is obtained by conditioning on the sum of observed responses. Define the conditional reference set $\Gamma(5)$ to be all contingency tables for which the sum of entries in the first row is 5. Then the conditional probability under H_0 of observing any table $x \in \Gamma(5)$ is given by

$$h_0(x|5) = \frac{f_\pi(x)}{\sum_{y \in \Gamma(5)} f_\pi(y)} \qquad (1.8)$$

which simplifies to

$$h_0(x|5) = \frac{\prod_{j=1}^{4} \binom{10}{x_j}}{\binom{40}{5}} \qquad (1.9)$$

Observe that equation (1.9) does not depend on π. The unknown nuisance parameter has been eliminated by conditioning on its sufficient statistic –the sum of entries in row 1 of Table 1.

The permutation p-value is obtained by resampling tables $x \in \Gamma(5)$ each with probability $h_0(x/5)$. Suppose we resample M tables in this manner, denoted by $x_1, x_2, \ldots, x_M$. Then the permutation p-value is evaluated as

$$\tilde{p}_e(M) = \frac{\sum_{i=1}^{M} I\{T(x_j) \geq 155\}}{M} \qquad (1.10)$$

In other words, we resample tables from $\Gamma(5)$ with probability (1.9) and estimate the permutation p-value as the fraction of resampled tables that are at least as extreme as the observed table with respect to the Cochran-Armitage test statistic. For the data in Table 1 the permutation p-value based on $M = 100,000$ samples was found to be 0.0553. In repeated samples this value would vary due to the sampling error associated with $\tilde{p}_e(M)$. As before, the sampling error decreases in proportion to the square root of M.

For finite values of M the permutation p-value specified in equation (1.10) is also referred to as the Monte Carlo p-value. We can eliminate the sampling error of the permutation or Monte Carlo p-value by letting $M \to \infty$ in which case (StatXact-4, p. 599) $\tilde{p}_e(M)$ converges to the constant quantity

$$p_e = \sum_{\substack{x \in \Gamma(5) \\ T(x) \geq 155}} h_0(x|5) \qquad (1.11)$$

Equation (1.11) does not contain any unknown nuisance parameters, nor is it subject to sampling error. Thus, this probability calculation is exact and p_e is referred to as the exact p-value. For the data in Table 1, $p_e = 0.0546$ which is almost the same as $\tilde{p}_e(M)$ at $M = 100,000$.

The Asymptotic P-Value

Because evaluation of equation (1.11) can be very computationally intensive, one frequently approximates this p-value by appealing to the asymptotic normality of the distribution of $T(x)$. The asymptotic p-value is easily obtained as

$$p_a = 1 - \Phi \left\{ \frac{155 - E(T|5)}{\sqrt{\text{Var}(T|5)}} \right\}$$ (1.12)

where $E(T|5)$ and $\text{Var}(T|5)$ are the conditional mean and conditional variance, respectively, of $T(x)$ given $x \in \Gamma(5)$. Closed form expressions for these two conditional moments in terms of the margins of the observed contingency table are given by equations (2.30) and (2.31). Upon substituting into these equations we obtain E $(T|5)=70$ and var(T $|5)=1954 .52$, where upon $p_a =0 .0273$.

Methodology

We have seen in above that the three p-values, p_b, p_e and p_a, are very different, ranging between 0.02 and 0.09, and thereby leading to different conclusions about the null hypothesis. It is thus important to decide a priori which of the three methods, bootstrap, permutation or asymptotic, we intend to use for testing the null hypothesis. An objective way to compare the three methods is to determine, for a given nominal significance level, the actual significance level and power of each method. In this section we define these quantities and show how they may be computed. Below, we present the results of our comparisons.

We begin by generalizing the dose-response problem discussed in Section 1 to the comparison of K binomial populations with response probabilities $\pi_1, \pi_2, ..., \pi_K$, respectively. We wish to test the null hypothesis

$$H_0 : \pi_1 = \pi_2 = ... = \pi_K \equiv \pi$$ (2.13)

against the one-sided alternative hypothesis

$$H_1 : \pi_1 \leq \pi_2 \leq ... \leq \pi_k$$ (2.14)

with at least one inequality n equation (2.14) being strict. The value of π, the common response probability under H_0, is unknown. Suppose we observe x_j responses and n_j-x_j non-responses from population j. Table 2 displays the observed data in the form of a generic 2×K contingency table, x.

Let Γ denote the set of all 2 ×K contingency tables

Table 2: Data from K Ordered Binomial Populations

Response	Binomial Populations				Total
Status	Pop_1	Pop_2	...	Pop_K	
Response	x_1	x_2	...	x_K	m
Non-Response	$n_1 - x_1$	$n_2 - x_2$	...	$n_K - x_K$	$N - m$
Total	n_1	n_2	...	n_K	N

with column sums of n_j , j =1 ,2 ,...K . For any $x \in \Gamma$ the Cochran-Armitage test statistic is defined as

$$T(x) = \sum_{j=1}^{K} d_j x_j$$ (2.15)

where the d_j's are pre-specified constants that correspond to doses in a dose-response setting. Our objective is to determine, for the bootstrap, permutation and asymptotic procedures, the true significance level and power of a one-sided Cochran-Armitage test conducted at a nominal significance level of Γ. For the bootstrap and permutation procedures we will eliminate sampling error from the comparisons by assuming that the we sample an infinite number of times from the appropriate reference set. That is, we will let $M \to \infty$ and evaluate the performance of p_b rather than $\tilde{p}_b(M)$, and p_e rather than $\tilde{p}_e(M)$. In order to make the size and power comparisons accurately, all the computations are based on exact distribution theory rather than relying on asymptotic approximations. Thus, the formulas presented above, for size and power are very difficult to compute. We use adaptations of the network algorithms described in Mehta, Patel and Senchaudhuri (1998) and Corcoran, Mehta, and Senchaudhuri (2000) to perform these computations.

Size and Power of the Bootstrap Procedure

Suppose we have observed the data displayed in Table 2, where the sum of entries in row 1 is m and the total sample size is N . After eliminating sampling error by letting the number of bootstrap samples M be infinite, the bootstrap distribution of the Cochran-Armitage statistic is

$$\Pr\{T(x) = t|m\} = \sum_{\substack{x \in \Gamma \\ T(x)=t}} \prod_{j=1}^{K} \left(\frac{m}{N}\right)^{x_j} \left(1 - \frac{m}{N}\right)^{n_j - x_j}$$ (2.16)

Suppose we wish to test the null hypothesis (2.13) at a nominal significance level α . Let $t_b(m)$ be the level-α cut-off of the bootstrap distribution (2.16). That is,

$$\sum_{\substack{x \in \Gamma \\ T(x) \geq t_b(m)}} \prod_{j=1}^{K} \left(\frac{m}{N}\right)^{x_j} \left(1 - \frac{m}{N}\right)^{n_j - x_j} \leq \alpha$$ (2.17)

and for any t $<t_b$ (m)

$$\sum_{\substack{x \in \Gamma \\ T(x) \geq t}} \prod_{j=1}^{K} \left(\frac{m}{N}\right)^{x_j} \left(1 - \frac{m}{N}\right)^{n_j - x_j} > \alpha$$ (2.18)

Due to the discreteness of the distribution (2.16) the left hand side of (2.17) will usually be less than α. For notational convenience we have suppressed the dependence of $t_b(m)$ on α.

Conditional on m the true size or Type I error of the bootstrap procedure is

$$S_b(m, \pi) = \sum_{\substack{x \in \Gamma \\ T(x) \geq t_b(m)}} \prod_{j=1}^{K} \binom{n_j}{x_j} \pi^{x_j}(1-\pi)^{n_j - x_j}. \quad (2.19)$$

A priori, the unconditional Type I error of the bootstrap procedure is

$$S_b(\pi) = \sum_{m=0}^{N} S_b(m, \pi) \Pr(m|\pi). \quad (2.20)$$

Let $\pi = (\pi_1, \pi_2, ..., \pi_K)$, where $\{ \pi_1 < \pi_2 < ... < \pi_K \}$, denotes a specific alternative hypothesis. Conditional on m, the power of the bootstrap procedure to reject this alternative is

$$P_b(m, \pi) = \sum_{\substack{x \in \Gamma \\ T(x) \geq t_b(m)}} \prod_{j=1}^{K} \binom{n_j}{x_j} \pi_j^{x_j}(1-\pi_j)^{n_j - x_j}. \quad (2.21)$$

A priori, the unconditional power of the bootstrap procedure is

$$P_b(\pi) = \sum_{m=0}^{N} P_b(m, \pi) \Pr(m|\pi). \quad (2.22)$$

Size and Power of the Permutation Procedure

The permutation procedure differs from bootstrap procedure in the following way. In the bootstrap approach, the nuisance parameter π under the null hypothesis was eliminated by substituting its MLE, $\hat{\pi} = m/N$. In contrast, the permutation approach eliminates π by conditioning on m, its sufficient statistic. Let $\Gamma(m)$ denote all tables $x \in \Gamma$ for which the sum of entries in row 1 equals m. Then, after eliminating sampling error by letting the number of Monte Carlo samples M be infinite, the permutation distribution of the Cochran-Armitage statistic is

$$\Pr\{T(x) = t|m\} = \sum_{\substack{x \in \Gamma(m) \\ T(x) = t}} \frac{\prod_{j=1}^{K} \binom{n_j}{x_j}}{\binom{N}{m}} \quad (2.23)$$

Let $t_e(m)$ be the level-α cut-off of the permutation distribution (2.23). That is,

$$\sum_{\substack{x \in \Gamma(5) \\ T(x) \geq t_e(m)}} \frac{\prod_{j=1}^{K} \binom{n_j}{x_j}}{\binom{N}{m}} \leq \alpha \quad (2.24)$$

and for any $t < t_e(m)$

$$\sum_{\substack{x \in \Gamma(5) \\ T(x) \geq t}} \frac{\prod_{j=1}^{K} \binom{n_j}{x_j}}{\binom{N}{m}} > \alpha \quad (2.25)$$

Conditional on m the true size or Type I error of the permutation procedure is

$$S_e(m, \pi) = \sum_{\substack{x \in \Gamma(m) \\ T(x) \geq t_e(m)}} \frac{\prod_{j=1}^{K} \binom{n_j}{x_j}}{\binom{N}{m}} \quad (2.26)$$

A priori, the unconditional Type I error of the permutation procedure is

$$S_e(\pi) = \sum_{m=0}^{N} S_e(m, \pi) \Pr(m|\pi) \quad (2.27)$$

The conditional power of the permutation procedure to reject the alternative hypothesis $\{ \pi_1 < \pi_2 < ... < \pi_K \}$ given m is

$$P_e(m, \pi) = \sum_{\substack{x \in \Gamma(m) \\ T(x) \geq t_e(m)}} \left[\frac{\prod_{j=1}^{K} \binom{n_j}{x_j}\pi_j^{x_j}(1-\pi_j)^{n_j - x_j}}{\sum_{y \in \Gamma(m)} \prod_{j=1}^{K} \binom{n_j}{y_j}\pi_j^{y_j}(1-\pi_j)^{n_j - y_j}} \right] \quad (2.28)$$

A priori, the unconditional power of the permutation procedure is

$$P_e(\pi) = \sum_{m=0}^{N} P_e(m, \pi) \Pr(m|\pi) \quad (2.29)$$

Size and Power of the Asymptotic Procedure

The asymptotic procedure is very similar to the permutation procedure except that the numerically intensive computation of the level-α cut-off value is no longer required because the exact null permutation distribution (2.23) is replaced by its normal approximation. We have shown in Corcoran, et al. (2000), that the first two moments of this conditional distribution are

$$E(T|m) = \left(\frac{m}{N}\right) \sum_{j=1}^{K} d_j n_j \quad (2.30)$$

and

$$\text{Var}(T|m) = \left[\frac{m(N-m)}{N(N-1)}\right] \sum_{j=1}^{K} \left[d_j - \frac{E(T|m)}{m}\right]^2 n_j \quad (2.31)$$

Assuming that the conditional distribution (2.23) is asymptotically normal with the above two moments, the level-α cut-off is evaluated as

$$t_a(m) = E(T|m) + z_\alpha \text{Var}(T|m) \qquad (2.32)$$

where z_α is the upper α percentile of the standard normal distribution.

Conditional on m the true size or Type I error of the asymptotic procedure is

$$S_a(m, \pi) = \sum_{\substack{x \in \Gamma(m) \\ T(x) \geq t_a(m)}} \frac{\prod_{j=1}^{K} \binom{n_j}{x_j}}{\binom{N}{m}} \qquad (2.33)$$

A priori, the unconditional Type I error of the permutation procedure is

$$S_a(\pi) = \sum_{m=0}^{N} S_a(m, \pi) \Pr(m|\pi) \qquad (2.34)$$

The conditional power of the asymptotic procedure to reject the alternative hypothesis $\{ \pi_1 < \pi_2 < ... < \pi_K \}$ given m is

$$P_a(m, \pi) = \sum_{\substack{x \in \Gamma(m) \\ T(x) \geq t_a(m)}} \left[\frac{\prod_{j=1}^{K} \binom{n_j}{x_j} \pi_j^{x_j} (1 - \pi_j)^{n_j - x_j}}{\sum_{y \in \Gamma(m)} \prod_{j=1}^{K} \binom{n_j}{y_j} \pi_j^{y_j} (1 - \pi_j)^{n_j - y_j}} \right]$$

$$(2.35)$$

A priori, the unconditional power of the asymptotic procedure is

$$P_a(\pi) = \sum_{m=0}^{N} P_a(m, \pi) \Pr(m|\pi) \qquad (2.36)$$

Results

Having formulated the unconditional power for the bootstrap, exact permutation, and asymptotic tests, we revisit the data of Table 1. Recall that the exact bootstrap p-value for these data is 0.0954, the exact permutation p-value is 0.0546, and the asymptotic p-value is 0.0273. At a 0.05 significance level, the three tests yield different substantive results. Observing the operating characteristics of each of the three tests may allow us to understand the apparent contradiction.

Computing the unconditional size and power of these three tests under any specific setting, however, is computationally challenging. Mehta, et al. (1998) first solved the problem of obtaining the quantities (2.27) and (2.29) by using a network algorithm to evaluate the

distribution shown in (2.23), and hence to obtain $t_e(m)$. Corcoran, et al. (2000) extended this algorithm to enable computation of (2.34) and (2.36). These methods are currently available in the software package *StatXact-5* (2001). Without an efficient tool such as a network algorithm, obtaining the critical value $t_b(m)$, as defined by equations (2.17) and (2.18), can likewise pose a difficult problem. In the Appendix we describe how one can use the network approach to find the exact conditional power of the bootstrap test. This algorithm provides a tool that –combined with the algorithms previously developed for the permutation and asymptotic tests – allows a researcher to assess the relative characteristics of these tests, under any set of conditions, without resorting to simulation or approximation.

We apply this approach to the design shown in Table 1. Figure 1 shows the actual Type I error as a function of the quantity π under the null hypothesis (2.13). As the design is perfectly balanced, we need only plot size for $0 < \pi = 0 .50$. Plot (a) of Figure 1 shows the actual Type I error of the three tests when the doses of (1.3) are (d_1, d_2, d_3, d_4) = (0, 1, 2, 3), plot (b) uses doses of (d_1, d_2, d_3, d_4) = (0 ,1 ,2 ,4), and plot (c) uses dose scores of (d_1, d_2, d_3, d_4) = (0, 1, 5, 50). In all three plots, the asymptotic test violates the nominal significance level for most values of π . Under the dose scores used for plots (a) and (b), the asymptotic test violates the nominal significance level for π greater than approximately 0.08. For dose scores of (0, 1, 5, 50), the asymptotic test violates the nominal significance level for π greater than approximately 0.05. As expected, the exact test preserves the nominal significance level – never attaining the 0.05-level exactly due to the discreteness of the tail distribution. The bootstrap method, however,is comparatively very conservative.

Having examined the true significance levels of each test, we now compare the procedures with respect to power. Figures 2, 3, and 4 contain power plots for dose scores of (d_1, d_2, d_3, d_4) = (0, 1, 2, 3), (d_1, d_2, d_3, d_4) = (0, 1, 2 ,4), and (d_1, d_2, d_3, d_4) = (0, 1, 5, 50), respectively. Each of these three figures consists of four plots, corresponding to four values of π_1: 0.01, 0.05, 0.10, and 0.25. For the sake of simplicity, each curve is plotted as a function of the parameter β from the logistic dose-response model given by logit $(\pi_i) = \gamma + \beta d_i$, for $i = 1, 2, 3, 4$, where $\gamma = $ logit (π_1) and logit(x)=log [$x/(1 -x)$].

From Figure 1 we observe – under any of the three sets of dose scores examined here – that the asymptotic test violates the nominal Type I error rate for $\pi_1 = 0.10$ and $\pi_1 = 0 .25$, making this procedure viable only for $\pi_1 = 0 .01$ and $\pi_1 = 0 .05$. For these smaller choices of π, the asymptotic test indeed demonstrates a power advantage over the other two tests. However, with respect to the
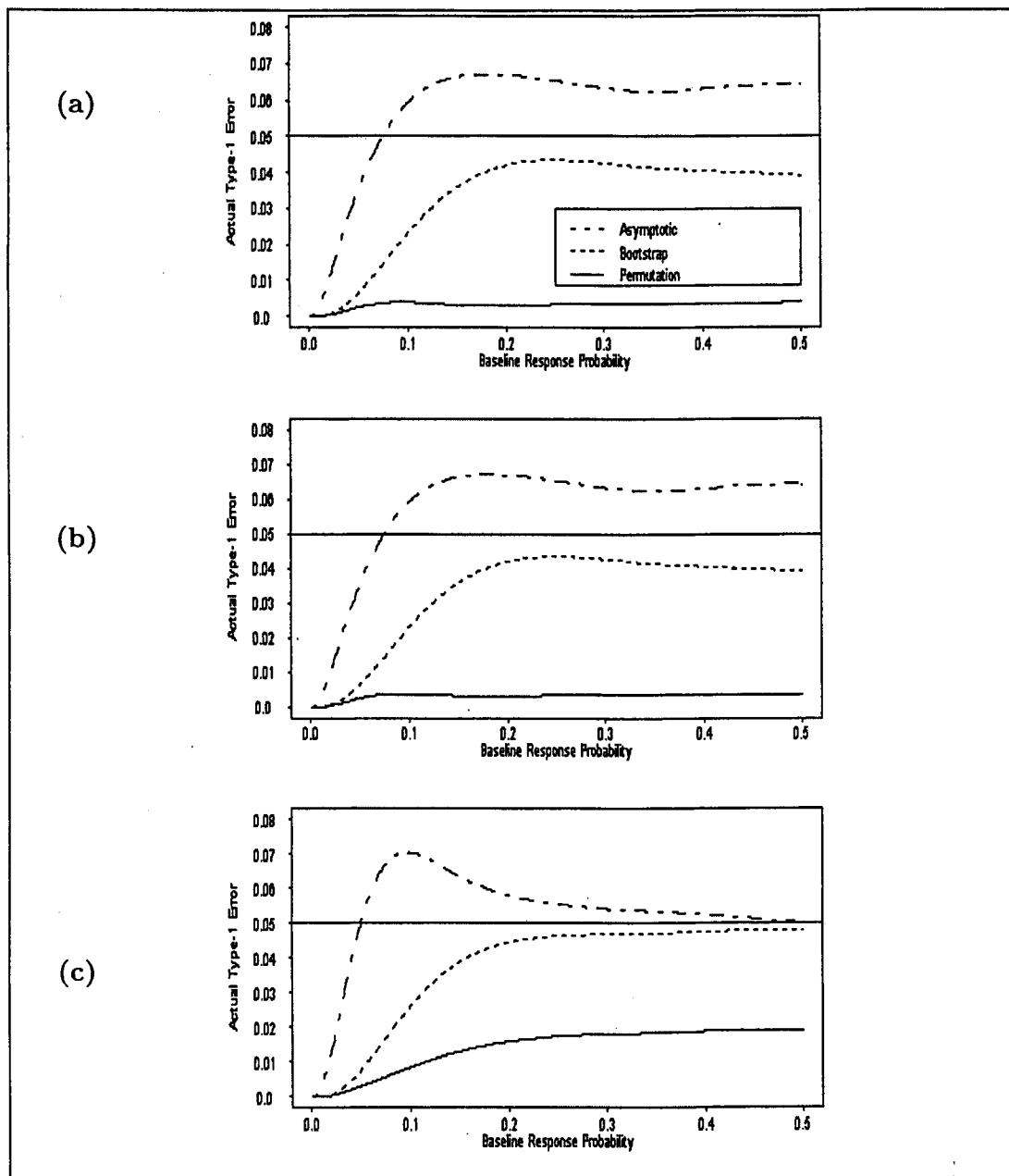
Figure 1: Actual type-1 error rate for asymptotic, permutation, and bootstrap trend tests when $K = 4$, $n_i = 10$ for $i = 1, 2, 3, 4$, and dose scores (d_1, d_2, d_3, d_4) are (a) (0,1,2,3), (b) (0,1,2,4), and (c) (0,1,5,50).

experiment shown in Table 1, under the null hypothesis our best guess at the common response probability π is 5 /40 =0 .125. As the asymptotic test exceeds the nominal significance level for probabilities in this range, one might have less faith in the accuracy of its associated p-value.

The exact power of the bootstrap procedure is clearly dominated by the permutation test under all conditions considered here, particularly when dose scores are equally-spaced (Figure 2) or almost equally-spaced (Figure 3). The conservatism and comparatively low power of the bootstrap procedure explain the relatively high p-value obtained for the data of Table 1.

Conclusion

The primary purpose of a sample size calculation is to ensure that a study has sufficient power to detect a specific effect size. For example, when investigating a dose-response relationship of the form $\text{logit}(\pi_i) = \gamma + \beta d_i$, one would typically have in mind a biologically or clinically meaningful slope, β_σ say, above which one could claim the existence of a trend in the data. The power of any test is the probability that the test will reject the null hypothesis that $\beta = 0$ when in fact $\beta = \beta_\sigma$. We have developed computational tools that, without resorting to approximations or simulations, can provide the exact power of three different tests of trend; permutation, bootstrap and asymptotic. It is seen that the test with the highest power is the asymptotic test, followed closely by the permutation test. The bootstrap test has considerably lower power than the others.

We have also developed computational tools for evaluating the exact Type I error of the three tests of trend. This is necessary because the power comparisons amongst the three tests are only meaningful if their Type I errors, or probabilities of falsely rejecting the null hypothesis, are bounded by the same value. We typically fix the Type I error bound at $\alpha = 0.05$. The permutation test, being exact, is guaranteed to not exceed this error bound. To see this observe from equation (2.27) that the exact Type I error, $S_e(\pi)$, for the permutation test is a weighted sum of terms of the form $S_e(m, \pi)$, where, by equation (2.24) each such term cannot exceed α, and the weights, $\Pr(m \mid \pi)$ are probabilities that sum to 1. Therefore S_e cannot exceed α either, and the Type I error is guaranteed to be preserved. We cannot make the same argument for the bootstrap type-I error, $S_b(\pi)$, or for the asymptotic Type I error, $S_a(\pi)$.

Figure 1 demonstrates that, for the entire range of the baseline response probability π, the Type I error of the permutation test is preserved. Figure 1 also reveals that the Type I error of bootstrap test remains below the stipulated 0.05 level throughout the range of the baseline response probability. This is an interesting finding because
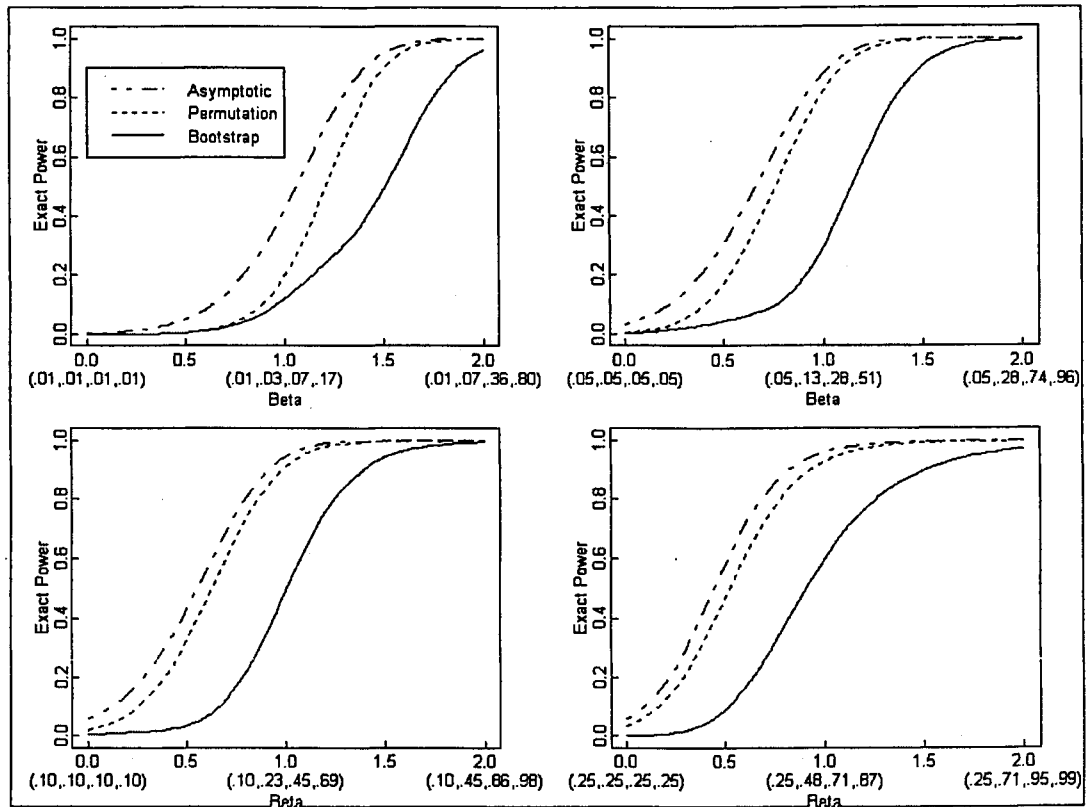

Figure 2: Exact power for asymptotic, permutation, and bootstrap trend tests when $K = 4$, $n_i = 10$ for $i = 1, 2, 3, 4$, and dose scores $(d_1, d_2, d_3, d_4) = (0, 1, 2, 3)$, for (clockwise from upper left) $\pi_1 = 0.01$, $\pi_1 = 0.05$, $\pi_1 = 0.10$, and $\pi_1 = 0.25$. Power is computed as a function of β, based upon the logistic dose-response model $\text{logit}(\pi_i) = \gamma + \beta d_i$, with $\gamma = \text{logit}(\pi_1)$.
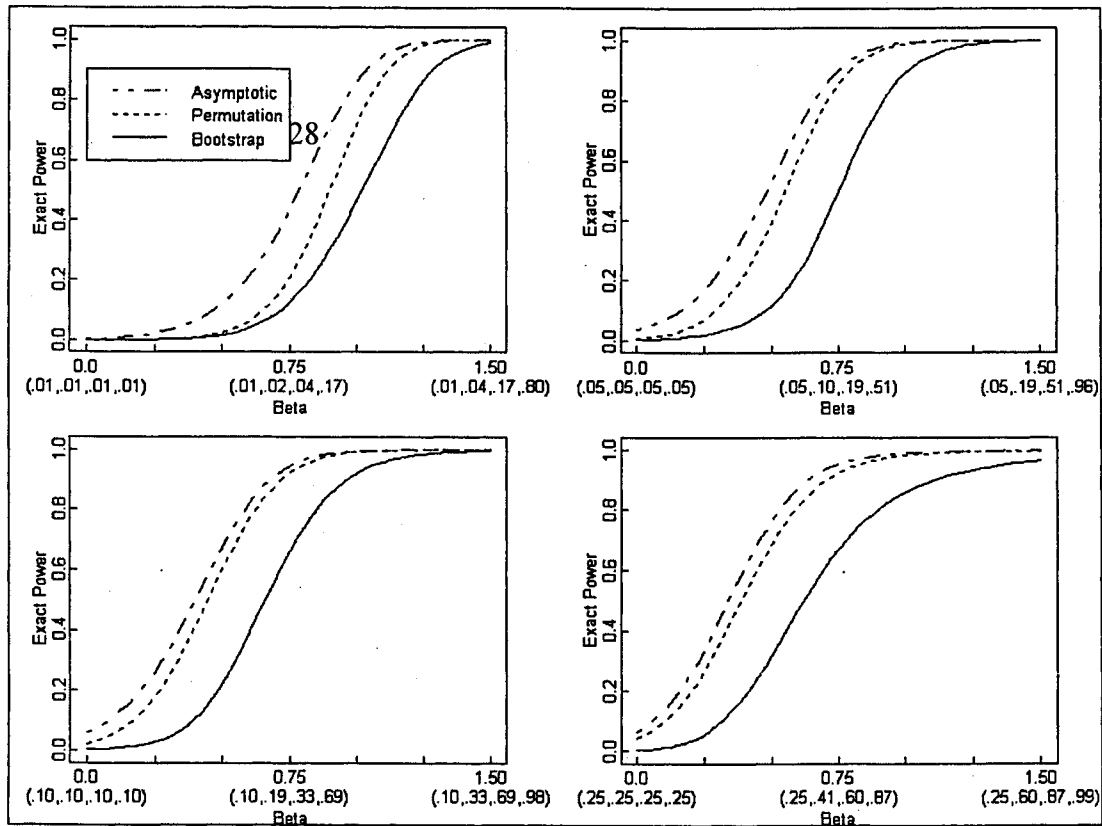
Figure 3: Exact power for asymptotic, permutation, and bootstrap trend tests when $K = 4$, $n_i = 10$ for $i = 1, 2, 3, 4$, and dose scores $(d_1, d_2, d_3, d_4) = (0, 1, 2, 4)$, for (clockwise from upper left) $\pi_1 = 0.01$, $\pi_1 = 0.05$, $\pi_1 = 0.10$, and $\pi_1 = 0.25$. Power is computed as a function of β, based upon the logistic dose-response model $\mathrm{logit}(\pi_i) = \gamma + \beta d_i$, with $\gamma = \mathrm{logit}(\pi_1)$.

the bootstrap test is not exact and therefore not guaranteed to preserve the Type I error. Unfortunately, as shown in Figure 1, the exact Type I error of the bootstrap test is very much below 0.05 for the cases considered here. This extreme conservatism results in a substantial deterioration of power for the bootstrap test relative to the permutation test, as is evident from Figures 2 to 4. Thus, one would never choose the bootstrap test in preference to the permutation test in a dose-response setting. This was not known previously and it was generally held that the two procedures have more or less the same operating characteristics.

The power comparisons between the permutation and asymptotic test are not as unambiguous. Although the asymptotic test is uniformly more powerful than the permutation test, Figure 1 shows that it does not preserve Type I errors. However, for very small values π, the Type I error is preserved and, if one could determine a priori that this nuisance parameter is suitably small, one might be justified in adopting the asymptotic test. This is an important finding that could only be discovered because of the availability of a computational tool for evaluating the exact

power of an asymptotic test. The computational tool that we have developed can therefore be very helpful, not merely for evaluating the power of various exact and asymptotic tests, but also for determining conditions under which one might actually prefer to use an asymptotic test – because of its superior operating characteristics – rather than its exact permutational counterpart.

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375-386.

Cochran, W, G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*, 417-451.

Corcoran, C. D., Mehta, C. R., & Senchaudhuri, P. (2000). Power comparisons for tests of trend in dose-response studies. *Statistics in Medicine*, *19*, 3037-3050.

Mehta, C. R., Patel, N. R., Senchaudhuri, P. (1998). Exact power and sample-size computations for the Cochran-Armitage trend test. *Biometrics*, *54*, 1615-1621.
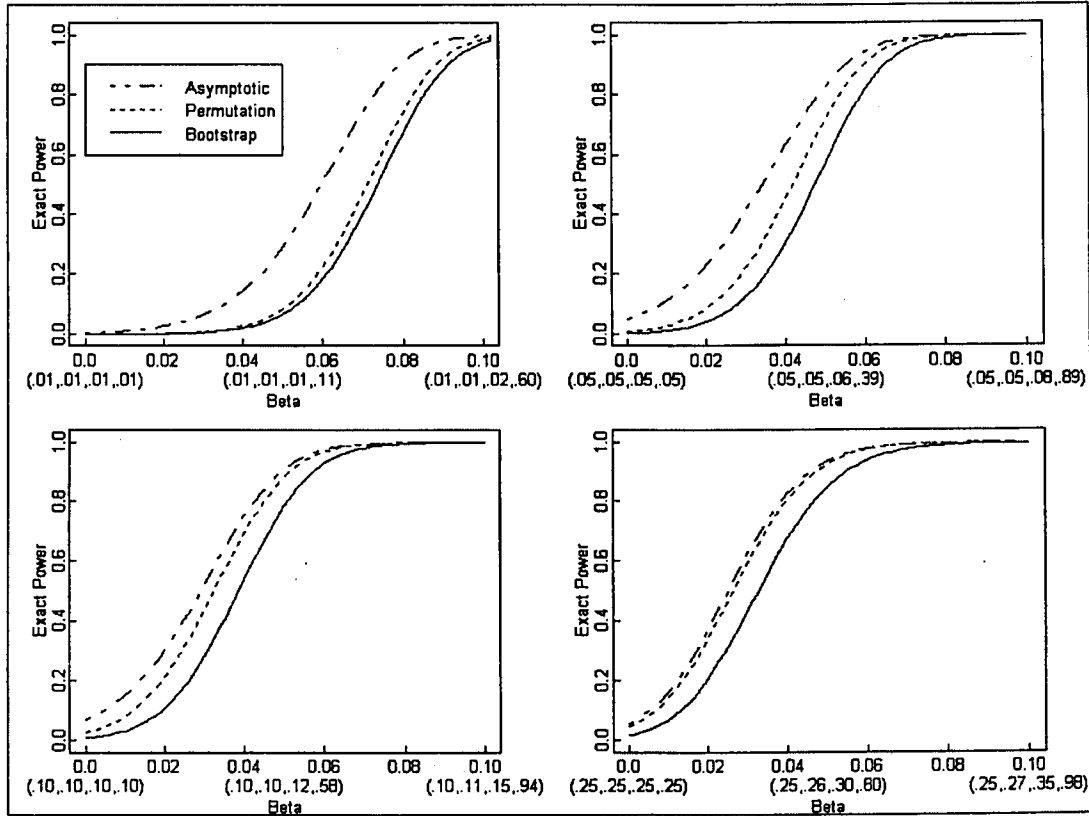
Figure 4: Exact power for asymptotic, permutation, and bootstrap trend tests when $K = 4$, $n_i = 10$ for $i = 1, 2, 3, 4$, and dose scores $(d_1, d_2, d_3, d_4) = (0, 1, 5, 50)$, for (clockwise from upper left) $\pi_1 = 0.01$, $\pi_1 = 0.05$, $\pi_1 = 0.10$, and $\pi_1 = 0.25$. Power is computed as a function of β, based upon the logistic dose-response model $\text{logit}(\pi_i) = \gamma + \beta d_i$, with $\gamma = \text{logit}(\pi_1)$.

StatXact (2001). *A software package for exact non-parametric inference*. Versions 4 & 5. Cytel Software Corporation: Cambridge, MA.

Appendix
Obtaining the Exact Power of the Bootstrap Test using the Network Algorithm

For a given m, it is sufficient to obtain the upper critical value t_b (m) of the exact bootstrap tail distribution. We will follow the notation in the Appendix of Corcoran, et. al (2000), in toto:

1. For any 2 × K table with fixed m, build a network as described with the following exceptions and additions:
   (a) Augment the table with one column – a (K +1)[st] column – such that the marginal total of the new column is N, the total sample size of the new table is 2 N, and the marginal total of the first row

of the new table is m +N , where $N = \sum_{i=1}^{K} n_i$.

(b) For the augmented table, $d_{K+1} = 0$.

(c) For arcs connecting a node of the network to a successor node, define a new probability length where

$$\hat{p}_{0,j+1} = \frac{n_{j+1}}{x_{j+1}}(\hat{\pi})^{x_{j+1}}(1 - \hat{\pi})^{n_{j+1}-x_{j+1}},$$

where $\hat{\pi} = m/N$.

2. Carry out the *Backward induction pass* as prescribed, with the following changes:
   (a) In step 2 of the backward induction pass, define

$$\hat{TP}_0(j, m_j)$$ as the sum of the probability lengths

(computed using the arc lengths $\hat{p}_{0,j+1}$) of all

paths in $\Gamma_m(j, m_j)$.

(b) In step 3 of the backward induction pass, let

$$\hat{TP}_0(K + 1, m + N) = 1.$$

(c)In lieu of steps 4(b) and 4(c) of the backward induction pass, let

$$\hat{TP}_0(j, m_j) = \sum \Gamma_{m(j,m_j)} \hat{p}_{0,j+1} \hat{TP}_0(j+1, m_{j+1})$$

3. Carry out the *Forward pass* as prescribed, with the following alterations:

(a)In step 3 of the forward pass, define

$\hat{p}_0(\tau) = \prod_{l=1}^{j} \hat{p}_{0,l}$ (there is no need to compute $p_0(\tau)$ or $p_1(\tau)$).

(b)In step 4 of the forward pass, define

$$\hat{c}_i(u) = \sum_{\substack{\tau \in T_m(j,m_j), \\ r(\tau)=u}} \hat{p}_i(\tau).$$

(There is no need to keep track of $c_0(u)$ or $c_1(u)$.)

Use $\hat{c}_i(u)$ for steps 5, 6, and 7.)

(c)In step 8 of the forward pass, choose $t_b(m)$ to be the smallest u^*, given the nominal significance level α, such that

$$\sum_{\substack{u \in \Omega(K+1, m+N), \\ u \geq u^*}} \frac{\hat{c}_0(u)}{\hat{TP}_0(0, 0)} \leq \alpha$$

4. Using the original $2 \times K$ table, follow the prescribed backward and forward induction passes exactly, this time replacing $t(m)$ in step 1 of the forward pass with the critical value $t_b(m)$.