

Determining the Appropriate Sample Size for Nonparametric Tests for Location Shift

Martin A. Hamilton

Department of Mathematical Sciences
Montana State University
Bozeman, MT 59717

Bruce Jay Collings

Department of Statistics
Brigham Young University
Provo, UT 84602

For the case in which two independent samples are to be compared using a nonparametric test for location shift, we propose a bootstrap technique for estimating the sample sizes required to achieve a specified power. The estimator (called BOOT) uses information from a small pilot experiment. For the special case of the Wilcoxon test, a simulation study is conducted to compare BOOT to two other sample-size estimators. One method (called ANPV) is based on the assumption that the underlying distribution is normal with a variance estimated from the pilot data. The other method (called NOETHER) adapts the sample size formula of Noether for use with a location-shift alternative. The BOOT and NOETHER sample-size estimators are particularly appropriate for this nonparametric setting because they do not require assumptions about the shape of the underlying continuous probability distribution. The simulation study shows that (a) sample size estimates can have large uncertainty, (b) BOOT is at least as accurate as and can be much more accurate than ANPV, and (c) BOOT and NOETHER achieve similar accuracy, although NOETHER is prone to underestimation.

KEY WORDS: Bootstrap; Hypothesis tests; Power; Wilcoxon two-sample test.

Collings and Hamilton (1988) described a bootstrap method for estimating the power of a test for location shift. The procedure does not require knowledge of the shape of the underlying continuous distribution. We discuss here the feasibility of adapting that bootstrap procedure to the problem of determining the sample size required to achieve desired power at a specified shift alternative. Although the method is quite general, we focus on determining the sample size for an experiment in which significance is to be determined by the one-sided, .05 level, two-sample Wilcoxon (Mann-Whitney) test.

We plan to observe two independent samples, each of size n . (Equal sample sizes are not required for the methodology of this article; as is shown later, the bootstrap algorithm can easily be adjusted for unequal-sample-size experiments.) Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ denote the potential observations in the two samples. The X 's are a random sample from a population having continuous cdf F_X and the Y 's are a random sample from a population having continuous cdf F_Y . We consider situations in which these distributions have the same shape but the locations may differ. Thus, if δ denotes the location difference [i.e., median (Y) - median (X) = δ], then $F_Y(y) = F_X(y - \delta)$, for every y . We will often drop the subscript X from F_X and denote the underlying continuous distribution function simply by F .

The sample information will be used to test the null hypothesis $H_0: \delta = 0$ against the alternative $H_a: \delta > 0$. The Wilcoxon test statistic, denoted by T , is the sum of the ranks of the Y 's among the combined $2n$ observations. Large T is evidence for the alternative hypothesis. Let T have distribution function $F_T(\cdot; F, \delta, n)$, a function completely determined by F , δ , and n . Under the null hypothesis, F_T does not depend on F and, for this reason, the Wilcoxon test is said to be "distribution-free." For level of significance α , let T_α denote the critical value of T ; that is, for any F , $F_T(T_\alpha; F, 0, n) = 1 - \alpha$. For alternative δ , let the functional $\kappa(F; \delta, \alpha, n)$ denote the power of the test. Thus $\kappa(F; \delta, \alpha, n) = 1 - F_T(T_\alpha; F, \delta, n)$ is the probability that T exceeds the critical value T_α when the distributions are δ units apart. The mathematical expression for κ is usually complicated. Because $\kappa(F; \delta, \alpha, n)$ depends on F , power is not distribution-free. If F were known exactly, one might well choose to use the optimal test for distribution F instead of the nonparametric Wilcoxon test.

For specified κ_0 and δ_0 , we wish to determine the smallest sample size n_0 such that $\kappa(F; \delta_0, \alpha, n_0) \geq \kappa_0$. In practice, the traditional approach for determining the sample size entails three steps: (1) Assume that F is a member of a location/scale parametric family (e.g., normal), (2) estimate the scale parameter using pre-experiment information (e.g.,

the pooled variance calculated from pilot data), and (3) either calculate the sample size approximately using a formula based on asymptotic theory or approximate the sample size by computer simulation. Step (2) can be avoided when F is assumed to be normal with standard deviation σ if it is appropriate to express δ_0 in σ units—for example, to find the sample size required to achieve power κ_0 at $\delta_0 = .5\sigma$, σ not specified. Step (3) could be accomplished using a search procedure, as follows. Pick a value of n , calculate $\kappa(F; \delta, \alpha, n)$. If $\kappa < \kappa_0$, try a larger n ; if $\kappa > \kappa_0$, try a smaller n . Continue until n_0 is correct. Because the value of n_0 determined by steps (1)–(3) depends on F and the estimated scale parameter, it is an estimate of the appropriate sample size. A poor estimate of the scale parameter or misspecification of F may produce a poor estimate of n_0 .

Some information is required to make a reasonable guess of F . Potential sources of such information include historical data, stochastic models, or small pilot experiments. We are concerned here with methods for sample-size determination that use the results of a small pilot experiment. The pilot data are the observed values of a random sample of size m from each of F_X and F_Y . (We use equal sample sizes in the pilot experiment only for notational convenience. The sample size estimators are calculable when the pilot data sample sizes are unequal, even when there are pilot data on only one of the treatments.) Denote the $2m$ realized values from the pilot experiment by $x_1, \dots, x_m$, and $y_1, \dots, y_m$.

We believe that practicing statisticians need a generally applicable, reliable method for determining the sample size in this setting. The method should require no stronger distributional assumptions than needed to calculate the critical value or p value for the nonparametric test. The bootstrap method proposed in this article satisfies this need.

Section 1 contains descriptions of three methods for estimating sample sizes, the bootstrap method, a popular version of the traditional method, and a method adapted from a formula of Noether (1987). We illustrate the use of the proposed methodology by application to a test of time-to-failure data for two air-conditioner systems. We then describe a simulation study conducted to compare the methods. Section 4 is discussion.

1. METHODS FOR ESTIMATING THE SAMPLE SIZE

1.1 The Bootstrap Method

The bootstrap strategy is to use pilot data to produce a nonparametric estimate of F and to use a simulation method for finding the power of the test

associated with any specified sample size n if the data follow the estimated distribution function. Denote the distribution function estimate by G . Because F is assumed to be continuous, we require that G must also be continuous. At alternative δ , estimate the true power by $\kappa(G; \delta, \alpha, n)$, the functional κ evaluated at the estimated distribution function. Because κ is so complicated, we suggest a computer simulation technique (described in detail later) for approximating $\kappa(G; \delta, \alpha, n)$.

Calculating an Empirical Distribution Function. Several methods for calculating an empirical distribution function were considered in preliminary work (Collings and Hamilton 1988). The following method was most satisfactory. Suppose that there are k distinct data points denoted by $z_1, \dots, z_k$. (As explained later, the z 's will be either the x 's or y 's available from the pilot experiment.) Let $z_{(1)}, \dots, z_{(k)}$ denote the associated order statistics. Define $z_{(0)} = 2z_{(1)} - z_{(2)}$ and $z_{(k+1)} = 2z_{(k)} - z_{(k-1)}$. Let G be the continuous distribution function obtained by assigning probability $1/(k+1)$ uniformly to each interval $(z_{(i)}, z_{(i+1)})$, $i = 0, \dots, k$. Henceforth, *empirical distribution function* means this specific calculation. Let G_X denote the empirical distribution function based solely on the preliminary x 's, and let G_Y denote the empirical distribution function based solely on the y 's.

Approximating $\kappa(G; \delta, \alpha, n)$ by Resampling From G . Given an estimate G , $\kappa(G; \delta, \alpha, n)$ can be closely approximated by the following computer simulation procedure:

1. Draw a random sample of size $2n$ from G . The first n observations in the sample form a simulated sample of X 's, denoted by $X_1^o, \dots, X_n^o$. Then δ is added to each of the other n observations in the sample to form the simulated sample of Y 's, denoted by $Y_1^o, \dots, Y_n^o$. (The Y 's and X 's have been generated from the same distribution except that the distribution of the Y 's is shifted δ units to the right.)

2. The Wilcoxon statistic is calculated for the X 's and Y 's, yielding T^o . If $T^o \geq T_\alpha$, a success is recorded; otherwise, a failure is recorded.

3. Steps 1 and 2 are repeated J times. The power, $\kappa(G; \delta, \alpha, n)$, is approximated by the proportion of successes among the J repetitions. This estimate will be denoted by $\hat{\kappa}(G; \delta, \alpha, n)$.

Note that the n used in step 1 need not equal the sample size m for the pilot data set. For example, to estimate the power for an experiment using a sample size of $n = 18$, one would use 18 for n at step 1, regardless of the size of m . The BOOT technique can also be used to choose unequal sample sizes. If,

for example, the costs of observation are such that we wish to take twice as many X 's as Y 's in the main experiment, then one could alter step 1 to pick $3n$ observations, letting the first $2n$ be X 's, then shifting the others by adding δ to create the Y 's.

The Bootstrap (BOOT) Estimator. Among the several versions of bootstrap power estimators that we tried, the following method was the best (Collings and Hamilton 1988). The bootstrap power estimator, denoted by $\hat{\kappa}_{\text{BOOT}}(\delta, \alpha, n)$, is the average

$$\hat{\kappa}_{\text{BOOT}}(\delta, \alpha, n) = \{\hat{\kappa}(G_X; \delta, \alpha, n) + \hat{\kappa}(G_Y; \delta, \alpha, n)\}/2.$$

If G_X and G_Y were based on different sample sizes, then $\hat{\kappa}_{\text{BOOT}}$ is a weighted average (Collings and Hamilton 1988); $\hat{\kappa}_{\text{BOOT}}$ can be calculated even if only G_X or G_Y is available. Because $\hat{\kappa}(G_X; \delta, \alpha, n)$ and $\hat{\kappa}(G_Y; \delta, \alpha, n)$ are consistent estimators of $\kappa(F; \delta, \alpha, n)$ (Bickel and Freedman 1981), $\hat{\kappa}_{\text{BOOT}}(\delta, \alpha, n)$ is the average of two independent, consistent estimators. The BOOT method estimated sample size for given α , δ_0 , and κ_0 is found by calculating $\hat{\kappa}_{\text{BOOT}}(\delta_0, \alpha, n)$ iteratively on n to locate $\hat{n}_0 = \min\{n: \hat{\kappa}_{\text{BOOT}}(\delta_0, \alpha, n) \geq \kappa\}$.

We have studied the following alternative strategy for estimating $\kappa(F; \delta, \alpha, n)$ and have found that it is unsatisfactory: (a) Estimate δ with $\hat{\delta}$, (b) find the empirical distribution function of the $2m$ values $x_1, \dots, x_m, y_1 - \hat{\delta}, \dots, y_m - \hat{\delta}$, and (c) find the power for this empirical distribution function (see Collings and Hamilton 1988). In our investigations, we calculated $\hat{\delta}$ in various ways—for example, by the difference in sample means, by the difference in sample medians, and by the Hodges–Lehmann estimator. For each method, the empirical distribution of the combined data was too steep at the center, a characteristic that led to positively biased power estimates. This happened because the procedure forces the first m observations in the combined sample (the x 's) to have the same location as the second m observations (the shifted y 's); that is, the $y - \hat{\delta}$ values are nested too tightly within the x 's.

We also tried some elaborate shift ideas that allow $\hat{\delta}$ to be perturbed randomly at each bootstrap iteration. These schemes essentially eliminated the bias, but none of them estimated power as well as simply averaging $\hat{\kappa}(G_X; \delta, \alpha, n)$ and $\hat{\kappa}(G_Y; \delta, \alpha, n)$.

1.2 The Assume Normal, Pool Variance Method

The assume normal, pool variance (ANPV) method is included in this investigation to provide a standard to which we can compare the reliability of the bootstrap method. If one were willing to assume that the

underlying distribution is normal with unknown variance, one could calculate the preliminary sample variance and substitute it into the approximate power formula (2.31) of Lehmann (1975) and solve for the n required to achieve the target κ_0 . When the true variance is known, the formula is quite accurate, even for small sample sizes. Let s^2 denote the pooled sample variance for the x 's and y 's of the pilot experiment, $\Phi(\cdot)$ denote the standard normal cdf, and $C = [\Phi^{-1}(\kappa_0) + \Phi^{-1}(1 - \alpha)]^2 [s/\delta_0]^2 [\pi/3]$. Then the ANPV sample-size estimate is

$$\hat{n}_0 = C[1 + \{1 + (1/C)\}^{1/2}], \quad (1.1)$$

which, for large C , is essentially $\hat{n}_0 = 2C$ as in formula (2.34) of Lehmann (1975). Because this procedure is simple and because power calculations for other parametric families are quite difficult, we believe ANPV is commonly used by practicing statisticians.

1.3 The NOETHER Method

Noether (1987) gave a sample-size formula for the Wilcoxon two-sample test in which the alternative is defined in terms of $P = \Pr(Y > X)$ rather than in terms of the shift δ . It is appropriate for any F and can be used whenever the sampling distribution F_T can be closely approximated by the normal distribution, an approximation that is usually quite good, except for very small n . Noether's formula is

$$n_0 = [\Phi^{-1}(\kappa_0) + \Phi^{-1}(1 - \alpha)]^2 / [6(P - .5)^2]. \quad (1.2)$$

Although this formula is specific to the Wilcoxon test, Noether's derivation could possibly be imitated to find a sample-size formula for any nonparametric test in which the test statistic is asymptotically normal.

Equation (1.2) requires that one express the alternative of importance in terms of P rather than the location shift δ_0 . Given δ_0 , the associated P is $P = \int F(z)F'(z - \delta_0)dz$, which can be estimated by substituting G_X or G_Y for F . To use only the x 's to estimate P , form x^* 's by adding δ_0 to each of the x 's. Let c_x equal the number of i, j pairs such that $x_i^* > x_j$. Then the estimate of P based on G_X is $p_x = c_x/m^2$. Similarly, define p_y , the estimate of P based on G_Y . Let n_x and n_y be the results from substituting p_x and p_y , respectively, into Equation (1.2). Then $\hat{n}_0 = (n_x + n_y)/2$ is the sample-size estimate called NOETHER in this article.

We also studied an alternative method of calculation in which P is estimated by $p = (p_x + p_y)/2$ and then p is substituted into (1.2) to find the estimated sample size. This alternative approach proved to be inferior, and we will not discuss it further.

2. EXAMPLE

Bain and Engelhardt (1987, p. 448) listed times-to-failure of airplane air conditioners for two different airplanes, x and y . These data are actually a subset of data published by Prochan (1963) showing a total of 213 times-to-failure observed on 13 different Boeing 720 jet airplanes. For illustration purposes, we shall treat the Bain and Engelhardt data as the results of a pilot experiment. Let U and V be random variables denoting times-to-failure of air conditioners on airplanes x and y , respectively. Suppose that the air conditioner on airplane y is at least as reliable as the air conditioner on airplane x . Then, for some $a \geq 0$, we assume that V is identically distributed as $(1 + a)U$, or equivalently, for any time z , $\Pr(\ln(V) < z) = \Pr(\ln(U) < z - \ln(1 + a))$. If X and Y are defined by $X = \ln(U)$ and $Y = \ln(V)$, then the distributions of X and Y follow the location-shift model.

Denote the log-transformed observed times-to-failure for the air conditioners by $x_1, \dots, x_{12}$ and $y_1, \dots, y_{11}$. Plots of G_X and G_Y are shown in Figure 1. The two empirical distribution functions are roughly similar in shape, with G_Y shifted slightly to the right.

Suppose that the investigator plans to conduct a main experiment, which will be analyzed using a Wilcoxon two-sample test to determine whether the distribution of logarithms of times-to-failure for y is shifted to the right of the distribution for x . The investigator needs to know the appropriate sample size to use in the main experiment. Because the analysis is conducted on a log scale, the shift is a difference of logarithms that is the logarithm of a ratio, and one can re-express the shift δ as a percentage increase ($= 100a$) in median time-to-failure. If $\alpha = .05$, $\delta_0 = .35$ (corresponding to a 42% increase in time-to-failure) and the target power is $\kappa_0 = .70$, then the BOOT estimated sample size is 59 per group, the ANPV estimated sample size is 84, and the NOETHER estimated sample size is 50. Table 1 shows estimated sample sizes based on the pilot data for various choices of δ_0 and κ_0 . Because there are 12

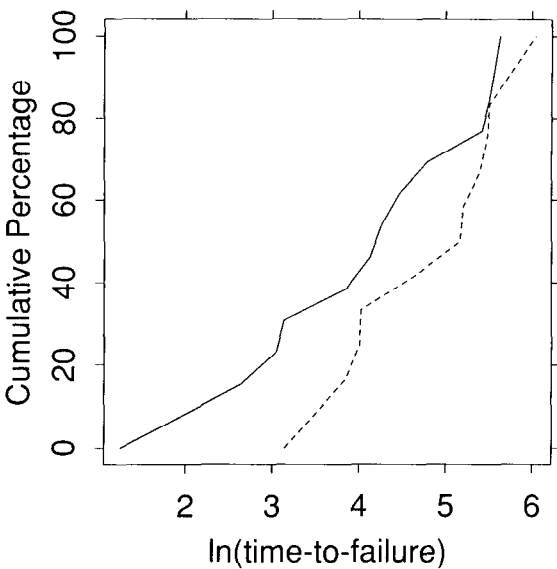


Figure 1. Empirical Distribution Functions G_X and G_Y Based on $m_x = 12$ and $m_y = 11$ Pilot Observations, Where x and y Denote Two Different Airplanes: —, Airplane x ; ---, Airplane y . The response variable is time-to-failure (hours of service) for the airplane air conditioner. The data are from Bain and Engelhardt (1987, p. 448).

x 's and 11 y 's in the pilot data, we used $\hat{\kappa}_{\text{BOOT}}(\delta, \alpha, n) = \{12\hat{\kappa}(G_X; \delta, \alpha, n) + 11\hat{\kappa}(G_Y; \delta, \alpha, n)\}/23$, a weighted average; see section 2.3 of Collings and Hamilton (1988).

The ANPV method produces much larger sample sizes than the BOOT or NOETHER methods. The NOETHER sample sizes are always the smallest for this example. Based on our simulation study of the bias, variability, and robustness of each method, we would recommend the BOOT estimate.

3. THE MONTE CARLO SIMULATION STUDY

3.1 Design of the Simulation Study

Pilot data sets were generated from the normal, exponential, uniform, and Cauchy distributions. Two sample sizes, $m = 10$ and 20 per treatment, were

Table 1. Estimated Sample Size $\hat{n}_0$ According to the BOOT, ANPV, and NOETHER Methods as Applied to the Pilot Data of Figure 1; Two-Sample, .05-Level Wilcoxon Test of Target Power κ_0 at Alternative δ_0 ($J = 500$ resamples were used for the BOOT method)

Method	Associated increase in time-to-failure								
	28%			42%			57%		
	$\kappa_0 = .5$	$\kappa_0 = .7$	$\kappa_0 = .9$	$\kappa_0 = .5$	$\kappa_0 = .7$	$\kappa_0 = .9$	$\kappa_0 = .5$	$\kappa_0 = .7$	$\kappa_0 = .9$
BOOT method $\hat{n}_0$	57	100	195	33	59	120	23	43	84
ANPV method $\hat{n}_0$	94	164	298	48	84	152	29	51	92
NOETHER method $\hat{n}_0$	48	84	152	27	50	91	23	39	72

NOTE: Shift δ_0 (ln scale) is .25 for first three columns, .35 for second three columns, and .45 for last three columns.

studied. For each combination of distribution and sample sizes (eight total combinations), 200 data sets were drawn. The BOOT, ANPV, and NOETHER sample-size-determination methods were applied to each data set to produce sample-size estimates at 15 combinations of δ_0 and κ_0 . The target power was set at $\kappa_0 = .5, .6, .7, .8$, and $.9$. Three equally spaced values of δ_0 were chosen for each distribution. The δ_0 values were chosen so that the exact associated sample sizes were about the same across the four distributions. Throughout this investigation, α was set at $.05$. The exact α may be slightly less than $.05$ because of the discrete distribution of the test statistic T . We used $J = 500$ resamples when performing steps 1–3 of Section 1.1.

The simulation programs were written in FORTRAN 77, and computations were performed on the Brigham Young University Convex C210 computer. The random-number generator was a compound generator with multiplicative component generators and a multiplicative mixing generator (Collings 1987).

3.2 Exact Sample Size Required to Achieve the Target Power

To determine the accuracy of the estimators, it is necessary to know the exact sample sizes required at

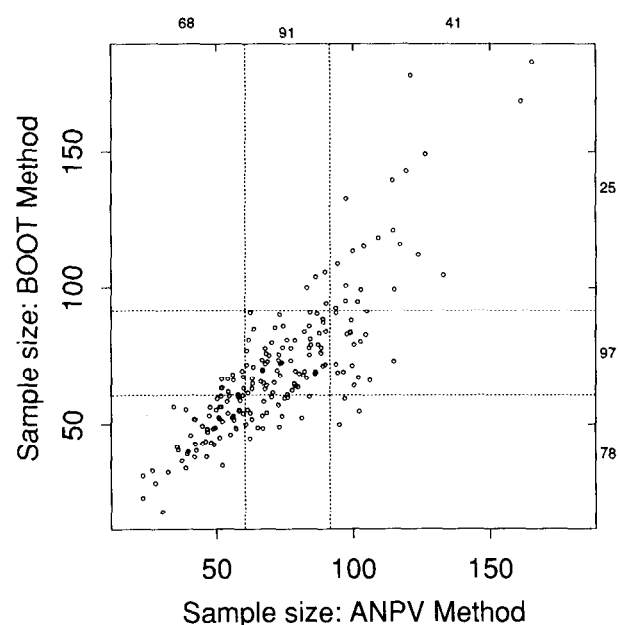


Figure 2. Sample Sizes Estimated by the ANPV and BOOT Methods for each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From a Normal Distribution. Target power = $.90$ at alternative $\delta = .5$. The exact sample size is $n_0 = 73$. The horizontal (vertical) dashed lines form a band enclosing BOOT (ANPV) estimated sample sizes such that the power achieved is between $.85$ and $.95$; numbers in the right (top) margin show the number of points in the band as well as the number of points in the region of samples sizes less than the band and the number greater than the band.

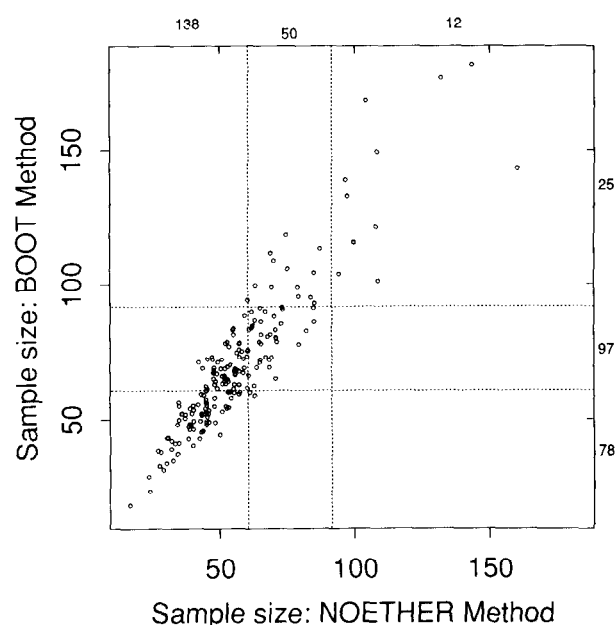


Figure 3. Sample Sizes Estimated by the NOETHER and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From a Normal Distribution. Target power = $.90$ at alternative $\delta = .5$. The exact sample size is $n_0 = 73$. See Figure 2 for an explanation of the dashed lines.

designated values of α , δ_0 , κ_0 , and F . We determined these sample sizes by using the simulation procedure of steps 1–3 in Section 1.1 to approximate closely the power $\kappa(F; \delta_0, \alpha, n)$ over an appropriate range of values of n . In this article, the phrase “exact sample size” is the smallest value of n for which the simulated power exceeds κ_0 , and the phrase “power achieved at n ” denotes the simulated power at n . The simulation involved $J = 1,000,000$ resamples at each chosen combination of F , δ_0 , α , and n . In the least favorable case in which the exact power is $.5$, we can be 95% confident that the simulated power is within $.001$ of the exact power. With very high likelihood, the exact sample size found by our simulations is correct to within one unit. When F is the normal distribution, the exact sample sizes found by this simulation procedure agree with Formula (1.1) when evaluated using the exact variance.

3.3 Analysis of the Monte Carlo Simulation Experiment

For $\alpha = .05$ and each combination of δ_0 , κ_0 , and F , the simulation program produced 200 triplets of sample sizes, one sample size by each of the three methods, BOOT, ANPV, and NOETHER. The estimation methods are compared using both visual displays and quantitative summary statistics. The visual comparison is a scatterplot of the 200 BOOT and ANPV pairs or of the BOOT and NOETHER pairs.

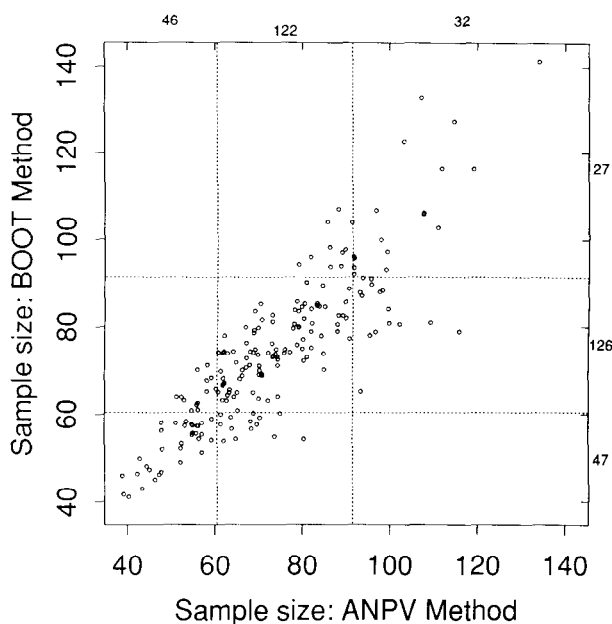


Figure 4. Sample Sizes Estimated by the ANPV and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 20$ Observations per Treatment Generated From a Normal Distribution. Target power = .90 at alternative $\delta = .5$. The exact sample size is $n_0 = 73$. See Figure 2 for an explanation of the dashed lines.

Some representative examples of the plots are shown in Figures 2–11 (pp. 331–334). To show all individual points, the sample sizes were jittered slightly before

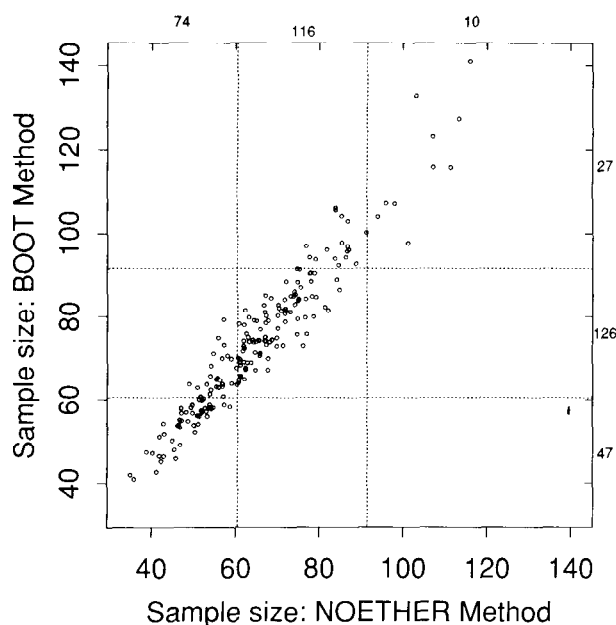


Figure 5. Sample Sizes Estimated by the NOETHER and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 20$ Observations per Treatment Generated From a Normal Distribution. Target power = .90 at alternative $\delta = .5$. The exact sample size is $n_0 = 73$. See Figure 2 for an explanation of the dashed lines.

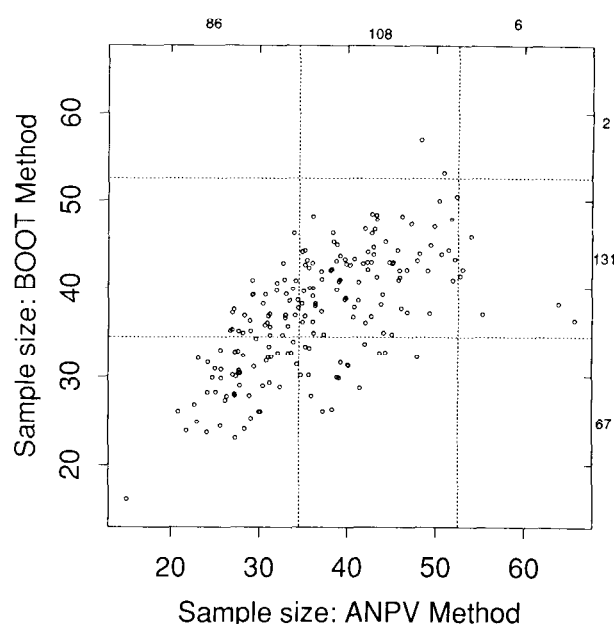


Figure 6. Sample Sizes Estimated by the ANPV and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From a Uniform Distribution. Target power = .90 at alternative $\delta = .2$. The exact sample size is $n_0 = 42$. See Figure 2 for an explanation of the dashed lines.

being plotted. Jittering was accomplished by adding to each estimated sample size a random value chosen from a uniform $(-.4, .4)$ distribution. Each plot region is square with the same scale on both axes. If the two methods produced identical estimated sample sizes, the points would lie on a line at a 45-degree angle through the origin. Superimposed on each scatterplot are vertical and horizontal bands enclosing those sample sizes for which the power achieved at $\hat{n}_0$ is within .05 of the target power κ_0 . If both methods had always produced sample sizes for which the true power values were within .05 of the target power, then all points would be inside the small square region formed by the intersection of the vertical and horizontal bands. The bands partition the plot into three vertical strips and into three horizontal strips. The number of points in each strip are printed in the top and right margins of each plot. Graphics were created using the S statistical programming language (Becker, Chambers, and Wilks 1988) on a Micro Vax II minicomputer and laser printer at Montana State University.

In some cases (viz., for the Cauchy distribution), the ANPV method produced extremely large sample-size estimates. This deficiency of the ANPV method was anticipated because the sample variance is a nonrobust estimator of scale. In plotting the estimated sample sizes for these cases, it was not feasible to scale the plot region to show all exact points. Thus in Figure 10 the plot region was truncated and

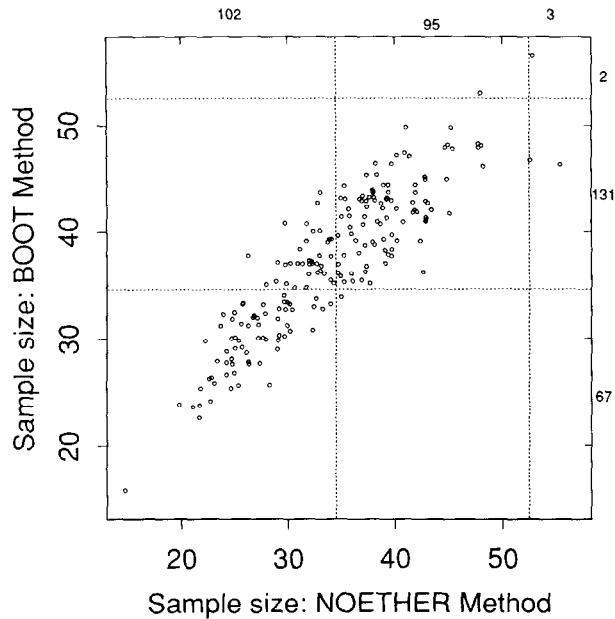


Figure 7. Sample Sizes Estimated by the NOETHER and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From a Uniform Distribution. Target power = .90 at alternative $\delta = .2$. The exact sample size is $n_0 = 42$. See Figure 2 for an explanation of the dashed lines.

the plot symbol $>$ was used to mark points where the ANPV sample size was larger than the plot region.

We also measured the error of each method by

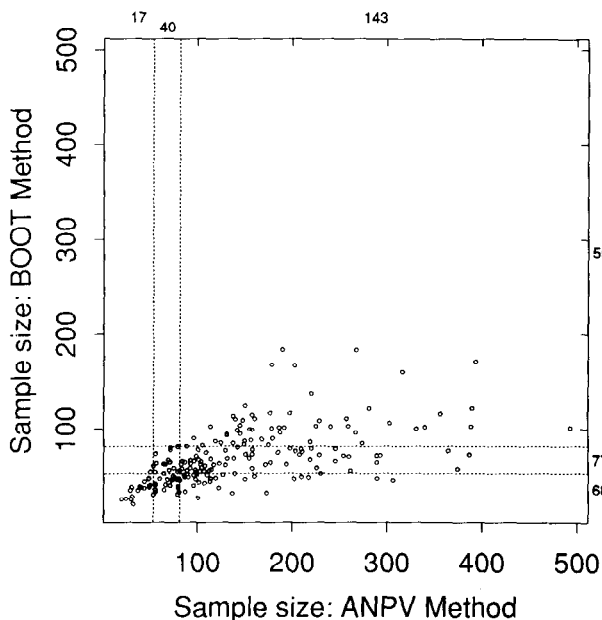


Figure 8. Sample Sizes Estimated by the ANPV and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From an Exponential Distribution. Target power = .90 at alternative $\delta = .35$. The exact sample size is $n_0 = 65$. See Figure 2 for an explanation of the dashed lines.

analyzing the discrepancy between the estimated sample size and the exact sample size and the discrepancy between the power achieved at the estimated sample size and the target power. Some of the quantitative summaries for BOOT and ANPV are shown in Table 2. Analyses of the estimated sample sizes and achieved power included calculation of the mean, median, variance, mean squared error, 10th percentile, and 90th percentile. In addition, we calculated (a) the mean absolute error (MAE) between the power realized at the estimated sample size and the target power, and (b) the mean relative squared error (MRSE) between realized power and target power. Specifically,

$$\text{MAE} = \Sigma |\hat{\kappa}(F; \delta_0, \alpha, \hat{n}_0) - \kappa_0| / 200,$$

and

$$\text{MRSE} = \Sigma \{[\hat{\kappa}(F; \delta_0, \alpha, \hat{n}_0) - \kappa_0]^2 / R(\hat{n}_0)\} / 200,$$

where $R(\hat{n}_0) = \kappa_0$, if $\hat{\kappa}(F; \delta_0, \alpha, \hat{n}_0) < \kappa_0$, and $R(\hat{n}_0) = 1 - \kappa_0$, if $\hat{\kappa}(F; \delta_0, \alpha, \hat{n}_0) > \kappa_0$.

3.4 Results of the Simulation Study

The results indicate that (a) estimators of sample size are plagued by high variability; (b) the BOOT and NOETHER methods give very similar answers, but NOETHER is prone to underestimation; (c) when F is normal or uniform, the BOOT method and the ANPV method are equally reliable; and (d) when F

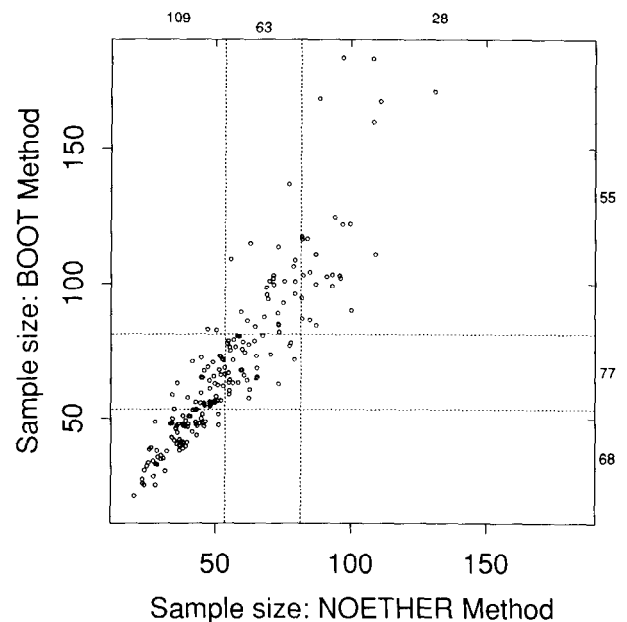


Figure 9. Sample Sizes Estimated by the NOETHER and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From an Exponential Distribution. Target power = .90 at alternative $\delta = .35$. The exact sample size is $n_0 = 65$. See Figure 2 for an explanation of the dashed lines.

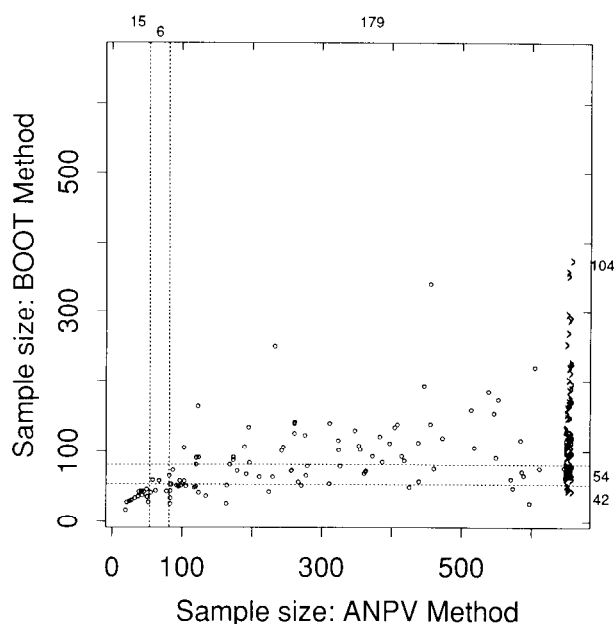


Figure 10. Sample Sizes Estimated by the ANPV and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From a Cauchy Distribution. Target power = .90 at alternative $\delta = 1.0$. The exact sample size is $n_0 = 65$. See Figure 2 for an explanation of the dashed lines.

is exponential or Cauchy, the BOOT and NOETHER methods are better than the ANPV method, primarily because the ANPV method is extremely conservative and produces many substantial overestimates. We conclude that the BOOT method is strongly preferable to ANPV and is weakly preferable to NOETHER.

For F fixed, the results were quite consistent across cases. It suffices to discuss a specific case typical of the results for that F . First, consider the normal distribution and the specific case $m = 10$, $\delta_0 = .5$, and $\kappa_0 = .9$ presented in Table 2 and Figures 2 and 3, for which the exact sample size is $n_0 = 73$. The median BOOT estimate of n_0 is 66 and the middle 80% of the BOOT estimates are between 43 and 99. These BOOT sample-size estimates correspond to a median achieved power of .87, and 80% of the achieved power values are between .73 and .96. The median NOETHER estimate of n_0 is 53 and the middle 80% of the NOETHER estimates are between 35 and 79. These NOETHER sample-size estimates correspond to a median achieved power of .80, and 80% of the achieved power values are between .65 and .92. The median ANPV estimate is 69 and the middle 80% of the ANPV estimates are between 43 and 102. These ANPV sample-size estimates correspond to a median achieved power of .89, and 80% of the achieved power values are between .73 and .97. Figure 2 shows that about half of the pilot experiments

led to estimated sample sizes for which the associated power achieved is between .85 and .95—specifically, 97 of 200 for BOOT and 91 of 200 for ANPV. Figure 3 shows that only 50 of the 200 pilot experiments led to similarly accurate NOETHER estimates.

Table 2 and Figures 4 and 5 show results for the same situation when $m = 20$. Doubling the size of the pilot experiment does improve the reliability of the estimates. Substantially more of the pilot data sets produced sample sizes for which the achieved power values are between .85 and .95; specifically, 126 of 200 for BOOT, 116 of 200 for NOETHER, and 122 of 200 for ANPV.

Summary statistics and scatterplots indicate that BOOT performs as well as ANPV, and NOETHER performs nearly as well, even though ANPV uses the correct assumption that F is normal. All methods exhibit a high degree of variability; none do a terrific job when based on small pilot studies. Nevertheless, we believe that BOOT is sufficiently accurate for most practical situations when F is normal and $m \geq 10$.

Next consider Table 2 and Figures 6 and 7, where F is uniform and $m = 10$. At $\delta_0 = .2$ and $\kappa_0 = .9$, the exact sample size is 42. All three methods tend to underestimate; no one method seems to be significantly superior to the other. Overall, the BOOT method is at least as good as the other two.

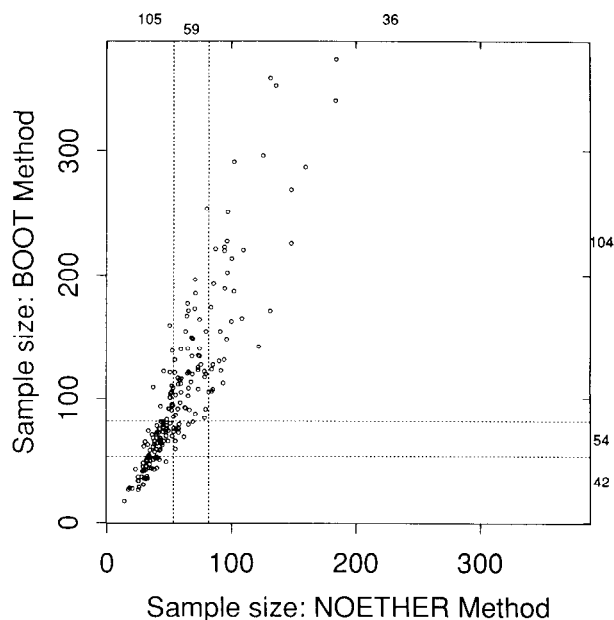


Figure 11. Sample Sizes Estimated by the NOETHER and BOOT Methods for Each of 200 Simulated Pilot Experiments of $m = 10$ Observations per Treatment Generated From a Cauchy Distribution. Target power = .90 at alternative $\delta = 1.0$. The exact sample size is $n_0 = 65$. See Figure 2 for an explanation of the dashed lines.

Table 2. Comparison of the BOOT, ANPV, and NOETHER Methods Based on 200 Simulated Pilot Experiments of m Observations per Treatment When the Underlying Distribution Is F , the Target Power $\kappa_0 = .9$, and the Shift of Interest Is δ_0 (see Sec. 3)

Method	Distribution of 200 sample-size estimates						Distribution of power achieved at sample-size estimates							
	Mean	Variance	MSE	Percentiles			Mean	Variance	MSE	MRSE	MAE	Percentiles		
				10	50	90						10	50	90
(a)														
BOOT	69.7	635.8	647.0	43	66	99	.856	.009	.011	.107	.075	.726	.873	.962
ANPV	72.4	578.1	578.5	43	69	102	.866	.010	.011	.136	.075	.726	.886	.966
NOETHER	56.3	404.8	685.0	35	53	79	.793	.011	.022	.065	.119	.646	.803	.921
(b)														
BOOT	74.1	294.9	296.2	54	74	96	.890	.003	.004	.094	.046	.809	.905	.957
ANPV	73.4	305.1	305.3	52	70	97	.886	.004	.004	.099	.051	.794	.890	.959
NOETHER	65.4	227.5	284.8	47	64	85	.855	.004	.006	.046	.061	.757	.864	.936
(c)														
BOOT	37.2	44.5	67.9	28	37	45	.858	.004	.005	.020	.052	.773	.867	.919
ANPV	36.6	70.8	99.5	27	36	48	.849	.005	.008	.039	.067	.760	.858	.933
NOETHER	34.0	50.3	113.9	25	34	43	.828	.005	.010	.021	.078	.729	.840	.909
(d)														
BOOT	69.9	898.9	923.1	38	64	106	.876	.010	.011	.206	.079	.728	.900	.983
ANPV	143.8	7703.8	13913.0	55	117	267	.953	.007	.010	.632	.088	.857	.989	1.000
NOETHER	55.5	440.7	530.6	31	51	85	.821	.013	.019	.104	.106	.650	.833	.958
(e)														
BOOT	65.8	351.9	352.6	43	63	93	.883	.006	.006	.132	.060	.775	.896	.970
ANPV	141.8	3555.0	9453.2	77	131	207	.980	.001	.008	.741	.084	.941	.994	1.000
NOETHER	60.3	307.3	329.8	39	58	87	.860	.007	.008	.091	.069	.737	.874	.961
(f)														
BOOT	105.1	4239.2	5848.8	43	87	189	.917	.011	.011	.449	.086	.771	.961	1.000
ANPV	>10 ⁶	>10 ⁶	>10 ⁶	77	549	25977	.973	.007	.013	.875	.104	.941	1.000	1.000
NOETHER	59.2	853.8	887.8	30	52	94	.818	.017	.024	.149	.117	.632	.835	.971

NOTE: (a) Exact $n_0 = 73$; $F = \text{normal}$; $\delta_0 = .5$; $m = 10$. (b) Exact $n_0 = 73$; $F = \text{normal}$; $\delta_0 = .5$; $m = 20$. (c) Exact $n_0 = 42$; $F = \text{uniform}$; $\delta_0 = .2$; $m = 10$. (d) Exact $n_0 = 65$; $F = \text{exponential}$; $\delta_0 = .35$; $m = 10$. (e) Exact $n_0 = 65$; $F = \text{exponential}$; $\delta_0 = .35$; $m = 20$. (f) Exact $n_0 = 65$; $F = \text{Cauchy}$; $\delta_0 = .5$; $m = 10$.

Now consider Table 2 and Figures 8 and 9, where F is exponential, $m = 10$, $\delta_0 = .35$, and $\kappa_0 = .9$, a case for which the exact sample size is $n_0 = 65$. These results indicate that ANPV is too conservative, typically producing a sample size much larger than necessary. The BOOT and NOETHER estimators, however, perform nearly as well as they did for the normal distribution. The figures show that the BOOT and NOETHER sample-size estimates are much more closely clustered around $n_0 = 65$ than are the ANPV estimates, and BOOT is slightly better than NOETHER, which tends to underestimation. Although only one case is shown here, this superiority of the BOOT procedure holds uniformly across all summary statistics and plots for exponential data.

Finally, consider Table 2 and Figures 10 and 11, where F is Cauchy, $m = 10$, $\delta_0 = 1.0$, and $\kappa_0 = .9$, a case for which the exact sample size is $n_0 = 65$. The ANPV method substantially overestimates the exact sample size, and the BOOT and NOETHER estimators, although also somewhat conservative, perform nearly as well as when the underlying distribution is normal or exponential. This superiority

of the BOOT and NOETHER procedures holds uniformly across all summary statistics and plots for Cauchy data.

There are two main reasons why the ANPV does not work well for the Cauchy and exponential distributions. One is that it relies on the strong assumption of normality; the other is that it uses a nonrobust estimate of the scale parameter. To eliminate the distribution assumption, something like bootstrapping is required. Although we have not explored the possibility, we guess that the BOOT and NOETHER methods would still compare favorably to the assume normal method, even if a robust estimate of scale were used.

The ANPV method is clearly inferior. Between BOOT and NOETHER, we prefer BOOT. It performed at least as well as NOETHER in all cases. It is a versatile technique, applicable to tests other than the Wilcoxon. The main disadvantage of BOOT is its computational requirements. The NOETHER procedure is computationally easy and produces reasonably reliable, robust estimates of the required sample size. The NOETHER method is applicable, however, only to the Wilcoxon test; for a different

test, theoretical analysis similar to that of Noether (1987) is required.

4. DISCUSSION

4.1 Related Literature

Bickel and Freedman (1981) proved the consistency and asymptotic normality of the bootstrap estimator for the power of the Wilcoxon one-sample test, and they stated that their proof can be extended to show the asymptotic properties of bootstrap power estimates in the case of the Wilcoxon two-sample test. Beran (1986) studied the asymptotic behavior of the one-sample t test under the assumption that the underlying distribution is normal. He mentioned the possibility of using estimated power curves as a way to choose among competing test procedures and as a way to choose sample sizes for the next experiment. In contrast to the previously cited works, we have not pursued asymptotic considerations here. Our interest is in the small-sample situation in which the pilot experiment is of size $m = 10$ or 20 .

Sayn and Merkle (1989) and Goldstein (1989) reviewed computer software available for calculating power and determining sample sizes for various experimental designs and statistical tests, including the two-sample t test. The books by Cohen (1988) and Kraemer and Thiemann (1987) reviewed power analysis and sample-size determination. These articles and books describe traditional methods; bootstrapping is not used.

4.2 Extensions of the Bootstrap Method

If the goal of the research effort is to estimate the magnitude of the difference between treatments, rather than to decide whether a difference is statistically significant, one could base sample-size decisions on confidence interval considerations rather than on power. Kupper and Hafner (1989) and Beal (1989) discussed this approach for the two-sample t test. In the Wilcoxon-test setting, one could use bootstrapping to estimate the sample size required to achieve a specified expected width for the nonparametric confidence interval estimate of δ . Note that Kupper and Hafner described situations in which sample-size estimates based on the power of the t test perform quite well, but sample-size estimates based on the confidence interval formulas are typically too small.

The bootstrap power and sample-size estimation method can be easily extended to apply to two-sample tests other than the Wilcoxon test. The extension is accomplished by using the desired test statistic and critical value in place of T and T_α in step 2 of the algorithm described in Section 1. The BOOT method has a straightforward extension to the k -sample ex-

periment in which, for example, it could be used to estimate the sample size for a Kruskal-Wallis test.

One can also use bootstrapping to create an adaptive nonparametric test—that is, to produce an algorithm for selecting a powerful test procedure and associated sample size to use for the main experiment. For example, one could estimate the power curve for each of several alternative test statistics and then base the sample size calculations on the test statistic having the largest estimated power. Collings and Hamilton (1988) calculated power curves via the bootstrap method for the t test and the median test as well as the Wilcoxon test for a pilot experiment of $m = 12$ observations. The median test and the t test had approximately the same power, and the Wilcoxon test had greater power than either. For those data, we believe that the Wilcoxon test is preferable to the t test and the median test and that the investigator should use the Wilcoxon test and the associated sample-size determinations. We are presently studying a formal adaptive method for selecting a sum-of-scores test and associated sample size.

The bootstrap estimator of this article should be applied only in situations in which the alternatives to the null hypothesis are shifts. A transformation of the measurement scale may be required to achieve compatibility with the shift model—for example, the logarithmic transformation used for the time-to-failure example. The bootstrap method however, could be adapted to alternatives other than shift alternatives. For example, consider the Lehmann alternative $F_Y = F_X^\nu$ for some $\nu \geq 1$. Then power and the required sample size are functions of ν instead of δ . In this situation, one could alter step 1 of simulation procedure in Section 1.1 as follows: Draw a random sample of n X 's from G_X ; then draw a sample of n Y 's from G_X^ν . We have investigated only shift alternatives thus far.

We have focused on the value of the BOOT method as a *point estimator* of sample size. It is possible to construct an associated *interval estimator*. Davidson, Hinkley, and Schechtman (1986) and Hall (1986a, b) provided some guidance. Efron (1982) provided a good overview of bootstrap methodology.

The simulation results show that sample-size choice is far from an exact science. In practice, one could use the BOOT-determined sample size as a suggestion and then, taking account of the uncertainty, adjust the BOOT sample size up or down as dictated by the circumstances of the experiment at hand.

ACKNOWLEDGMENT

This article was prepared for the Rupert G. Miller, Jr., Memorial Session of the 1989 Joint Statistical Meetings, August 6, 1989, Washington, D.C. We

thank the *Technometrics* reviewers for their helpful comments, particularly the suggestion to include the NOETHER method in the simulation study.

[Received March 1990. Revised November 1990.]

REFERENCES

- Bain, L. J., and Engelhardt, M. (1987), *Introduction to Probability and Mathematical Statistics*. Boston: Duxbury Press.
- Beal, S. L. (1989), "Sample Size Determination for Confidence Intervals on the Population Mean and on the Difference Between Two Population Means," *Biometrics*, 45, 969–977.
- Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth and Brooks/Cole.
- Beran, R. (1986), "Simulated Power Functions," *The Annals of Statistics*, 14, 151–173.
- Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 14, 1196–1217.
- Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum.
- Collings, B. J. (1987), "Compound Random Number Generators," *Journal of the American Statistical Association*, 82, 525–527.
- Collings, B. J., and Hamilton, M. A. (1988), "Estimating the Power of the Two-Sample Wilcoxon Test for Location Shift," *Biometrics*, 44, 847–860.
- Davidson, A. C., Hinkley, D. V., and Schechtman, E. (1986), "Efficient Bootstrap Simulation," *Biometrika*, 73, 555–566.
- Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans* (National Science Foundation–Conference Board of the Mathematical Sciences Monograph 38). Philadelphia: Society for Industrial and Applied Mathematics.
- Goldstein, R. (1989), "Power and Sample Size via MS/PC-DOS Computers," *The American Statistician*, 43, 253–260.
- Hall, P. (1986a), "On the Bootstrap and Confidence Intervals," *The Annals of Statistics*, 14, 1431–1452.
- (1986b), "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *The Annals of Statistics*, 14, 1453–1462.
- Kraemer, H. C., and Thieman, S. (1987), *How Many Subjects? Statistical Power Analysis in Research*, Newbury Park, CA: Sage Publications.
- Kupper, L. L., and Hafner, K. B. (1989), "How Appropriate Are Popular Sample Size Formulas?" *The American Statistician*, 43, 101–105.
- Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.
- Noether, G. E. (1987), "Sample Size Determination for Some Common Nonparametric Tests," *Journal of the American Statistical Association*, 82, 645–647.
- Prochan, F. (1963), "Theoretical Explanation of Observed Decrease Failure Rate," *Technometrics*, 5, 375–384.
- Sayn, H., and Merkle, W. (1989), "Statistical Software for Sample Size Estimation, Power, Design-Power, and IFNS," *Statistical Software Newsletter*, 15, 56–59.