

\$P\$ Values Maximized Over a Confidence Set for the Nuisance Parameter



Roger L. Berger; Dennis D. Boos

Journal of the American Statistical Association, Vol. 89, No. 427 (Sep., 1994),
1012-1016.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28199409%2989%3A427%3C1012%3AVMOACS%3E2.0.CO%3B2-S>

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

P Values Maximized Over a Confidence Set for the Nuisance Parameter

Roger L. BERGER and Dennis D. BOOS*

For testing problems of the form $H_0: \nu = \nu_0$ with unknown nuisance parameter θ , various methods are used to deal with θ . The simplest approach is exemplified by the t test where the unknown variance is replaced by the sample variance and the t distribution accounts for estimation of the variance. In other problems, such as the 2×2 contingency table, one conditions on a sufficient statistic for θ and proceeds as in Fisher's exact test. Because neither of these standard methods is appropriate for all situations, this article suggests a new method for handling the unknown θ . This new method is a simple modification of the formal definition of a p value that involves taking a maximum over the nuisance parameter space of a p value obtained for the case when θ is known. The suggested modification is to restrict the maximization to a confidence set for the nuisance parameter. After giving a brief justification, we give various examples to show how this new method gives improved results for 2×2 tables and solves previously intractable semiparametric problems.

KEY WORDS: Behrens-Fisher problem; Confidence set; Contingency table; Permutation test; Pivotal quantity; Scale differences.

1. INTRODUCTION

Testing problems are often complicated by the presence of a nuisance parameter vector θ . Consider first a model in which there is no nuisance parameter. Suppose that the data X have a probability distribution P_ν defined in terms of a parameter ν , and that we wish to test the simple hypothesis $H_0: \nu = \nu_0$. If the test statistic T is used to test H_0 and if large values of T give evidence against H_0 , then for an observed value $T = t$, the p value is $p = P_{\nu_0}(T \geq t)$.

Now consider a model with a nuisance parameter θ . The distribution of X has two parameters, ν and θ . We still wish to test $H_0: \nu = \nu_0$, but this hypothesis is no longer simple, because the value of θ is unspecified. Using a test statistic as before, the p value is now $p = \sup_\theta P_{\nu_0, \theta}(T \geq t)$ (see, for example, Bickel and Doksum 1977, pp. 171-172). Unfortunately, the need to calculate the \sup_θ has complicated the problem.

This complication is usually handled in one of three ways. First, in some problems it can be shown that for all values of t , the \sup_θ is always attained at a particular value θ_0 . In this case the p value is simply $p = P_{\nu_0, \theta_0}(T \geq t)$, and the parameter (ν_0, θ_0) is called the least favorable configuration. For example, in common one-sided testing problems, the boundary of the null hypothesis space is least favorable.

A second way to handle the unknown θ is to choose judiciously a test statistic T (that usually depends on estimated values of θ) whose distribution under H_0 does not depend on θ . That is, T is ancillary under H_0 . Then $P_{\nu_0, \theta}(T \geq t)$ is the same for all θ , so calculation of the \sup_θ is avoided. For example, in normal means problems we replace unknown variances with sample variances and use t or F distributions to account for the estimated variances.

A third method to handle the unknown θ is to condition on the value of a statistic S that is sufficient for θ under H_0 . Then the conditional distribution of any statistic, given S , does not depend on θ (under H_0), and the p value is taken to be $p = P_{\nu_0}(T \geq t | S = s)$. For example, in a 2×2 contingency table with common "success" probability θ under

H_0 , one can condition on the marginals (a sufficient statistic for θ under H_0) and use Fisher's exact test.

All three methods replace the calculation of the \sup_θ by the calculation of a single probability, and each method can result in a valid p value; that is, a statistic p such that under the null hypothesis,

$$P(p \leq \alpha) \leq \alpha, \quad \text{for each } \alpha \in [0, 1]. \quad (1)$$

We call a statistic that satisfies (1) a valid p value because it can be used in the standard way to define a level- α test. Consider the test that rejects the null hypothesis if and only if $p \leq \alpha$. Then under the null hypothesis, $P(\text{reject null}) = P(p \leq \alpha) \leq \alpha$; that is, the test so defined is a level- α test.

In many situations, however, none of the three methods is satisfactory. For example, the value of θ at which the \sup_θ occurs may depend on the value t in a complicated way. Also, exact distributional results are often not available for statistics with estimated parameters. And, finally, it may not be possible to find an appropriate sufficient statistic on which to condition.

In this article we consider a different approach for obtaining valid p values. Suppose that a valid p value $p(\theta_0)$ may be calculated when the true value θ_0 of the nuisance parameter vector θ is known. Here it should be noted that the calculation of $p(\theta_0)$ does not have to be based on the same test statistic for different values of θ_0 . Indeed the test statistic may depend directly on the assumed known value of θ_0 . All that is needed is that for each value of θ_0 , $p(\theta_0)$ must be a statistic that satisfies (1). If θ_0 is not known, then a valid p value may be obtained by maximizing $p(\theta)$ over the parameter space of θ ; that is, $p_{\text{sup}} = \sup_\theta p(\theta)$ clearly satisfies (1).

The use of p_{sup} has two potential difficulties, one computational and the other statistical. If the parameter space for θ is unbounded and if the \sup_θ is calculated numerically (as it often will be), then it may be uncertain whether the numerical method did indeed find the overall maximum. Of course, there is always uncertainty about the result of a numerical maximization, but this uncertainty is worse if the set being maximized over is unbounded. Statistically, it seems

* Roger L. Berger is Professor and Dennis D. Boos is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203.

a waste of information in the data to take the sup over all values of θ . Having observed the data, we should be able to estimate θ and should not need to consider values of θ that are completely unsupported by the data. Storer and Kim (1990) and others have used this idea to propose as a *p* value $p(\hat{\theta})$, where $\hat{\theta}$ is an estimate of θ (usually the maximum likelihood estimate). But *p* values defined in this way may not be valid; see the computations of Storer and Kim (1990).

A valid *p* value that addresses both of the aforementioned concerns is defined as follows. Let C_β be a $1 - \beta$ confidence set for the nuisance parameter when the null hypothesis is true. Intuition suggests that we might be able to restrict the maximization to the set C_β . Indeed we show in section 2 that

$$p_\beta = \sup_{\theta \in C_\beta} p(\theta) + \beta \tag{2}$$

is an alternative valid *p* value. This *p* value may be preferred to p_{sup} on computational grounds (due to maximizing over bounded sets) and on statistical principles (restricting interest to likely regions of θ). The value of β and the confidence set C_β should of course be specified before looking at the data. Note that p_β is never smaller than β . So in practice, β will be chosen rather small, such as .001 or .0001. If p_β is to be used to define a level- α test, then β must be less than α to obtain a useful test. Because the largest possible value of p_β is $1 + \beta$, p_β could be replaced by $\min\{p_\beta, 1\}$. It is also a valid *p* value and is always between 0 and 1.

We give the theoretical justification for p_β in the following lemma. The rest of the article is a series of illustrative examples. The first example, a pedagogical example, concerns tests about a normal mean when the variance is unknown. The remaining, more realistic examples concern 2×2 contingency tables, the Behrens–Fisher problem, and nonparametric testing for scale differences.

2. VALIDITY OF p_β

Lemma. Suppose that $p(\theta)$ satisfies (1) for any assumed known value θ . Let C_β satisfy $P(\theta \in C_\beta) \geq 1 - \beta$, if the null hypothesis is true. Let p_β be given by (2). Then p_β is a valid *p* value.

Proof. Suppose that the null hypothesis is true. Denote the true but unknown θ by θ_0 . If $\beta > \alpha$, then, because p_β is never smaller than β , $P(p_\beta \leq \alpha) = 0 \leq \alpha$. If $\beta \leq \alpha$, then

$$\begin{aligned} P(p_\beta \leq \alpha) &= P(p_\beta \leq \alpha, \theta_0 \in C_\beta) + P(p_\beta \leq \alpha, \theta_0 \in \bar{C}_\beta) \\ &\leq P(p(\theta_0) + \beta \leq \alpha, \theta_0 \in C_\beta) + P(\theta_0 \in \bar{C}_\beta) \\ &\leq P(p(\theta_0) \leq \alpha - \beta) + \beta \\ &\leq \alpha - \beta + \beta = \alpha. \end{aligned}$$

The first inequality follows because $\sup_{\theta \in C_\beta} p(\theta) \geq p(\theta_0)$ when $\theta_0 \in C_\beta$.

3. EXAMPLES

Example 1: Pedagogical Example About a Normal Mean. Let X_1, \dots, X_n be a random sample from a normal population with mean μ and variance σ^2 . We consider testing

$H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$, where μ_0 is a fixed value and σ^2 is the nuisance parameter. We consider this familiar example to illustrate our method, not to offer a serious contender to the usual *t* test.

If σ^2 were known, then we could use the test statistic $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$, where \bar{X} is the sample mean. Then the two-sided *p* value would be

$$p(\sigma^2) = 2\Phi(-|z_{\text{obs}}|),$$

where z_{obs} is the value of the test statistic calculated from the data and $\Phi(z)$ is the standard normal cumulative distribution function. As a confidence interval for σ^2 , we will use the upper confidence bound given by

$$C_\beta = \left\{ \sigma^2: 0 \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_\beta^2} \right\},$$

where s^2 is the sample variance and χ_β^2 is the 100β percentile of a chi-squared distribution with $n - 1$ degrees of freedom. The valid *p* value we propose is

$$p_\beta = \sup_{\sigma^2 \in C_\beta} p(\sigma^2) + \beta = \sup_{\sigma^2 \in C_\beta} 2\Phi(-|z_{\text{obs}}|) + \beta.$$

Because $|z_{\text{obs}}|$ is a decreasing function of σ , the \sup_{C_β} occurs at the upper endpoint. (This is why we chose to use an upper confidence bound.) Thus $p_\beta = 2\Phi(-|z_{\text{max}}|) + \beta$, where z_{max} is the test statistic calculated with $\sigma^2 = (n - 1)s^2/\chi_\beta^2$.

In this example the test statistic Z depends on the value of the nuisance parameter, a possibility mentioned in Section 1. Also, in this example the *p* value p_{sup} , although valid, is useless because it always has the value 1, because $|z_{\text{obs}}| \rightarrow 0$ as $\sigma \rightarrow \infty$. So the fact that maximization is restricted to C_β when calculating p_β is of critical importance in getting a reasonable answer.

This example is a bit unusual in that the \sup_{C_β} can be calculated exactly. In many cases this will need to be calculated numerically.

This example is also unusual in that the exact size of the test based on p_β can be calculated. Suppose that we reject H_0 if $p_\beta \leq \alpha$. Then the actual size of the test is

$$\begin{aligned} P(p_\beta \leq \alpha) &= P(2\Phi(-|Z_{\text{max}}|) + \beta \leq \alpha) \\ &= P(\Phi(-|Z_{\text{max}}|) \leq (\alpha - \beta)/2) \\ &= P(-|Z_{\text{max}}| \leq z_{(\alpha - \beta)/2}) \\ &= 2P(T \leq \sqrt{(n-1)/\chi_\beta^2} z_{(\alpha - \beta)/2}), \end{aligned}$$

where T has a Student's *t* distribution with $n - 1$ degrees of freedom and z_α is the 100α percentile of a standard normal distribution. It can be shown that $\sqrt{(n-1)/\chi_\beta^2}$ converges to 1 as n goes to infinity. So the actual size of the test, which is at most α because the *p* value is valid, converges to $\alpha - \beta$.

Example 2: 2×2 Contingency Table with Independent Binomial Sampling. Consider a 2×2 contingency table consisting of two independent binomial samples: 14 “successes” out of 47 trials for group 1 and 48 “successes” out of 283 trials for group 2. This data appeared in table 1 of Emerson and Moses (1985), who obtained it from Taylor et al. (1982). We consider here the usual 2×2 table chi-squared statistic

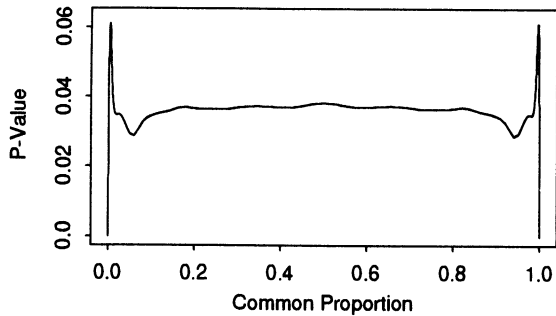


Figure 1. Exact p Values for the 2×2 Table Chi-Squared Statistic. Calculations are from independent binomial distributions with common proportion π .

$$Z^2 = \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where $\hat{\pi} = (n_1\hat{\pi}_1 + n_2\hat{\pi}_2)/(n_1 + n_2)$ and $\hat{\pi}_1$ and $\hat{\pi}_2$ are the sample proportions in the two groups. Figure 1 shows the p value $p(\pi)$ for detecting the difference between the two binomial proportions π_1 and π_2 as a function of the unknown common π under the null hypothesis $H_0: \pi_1 = \pi_2 = \pi$. The p value $p(\pi)$ for a fixed value of π is computed from the binomial distribution as

$$p(\pi) = \sum b(x; 47, \pi)b(y; 283, \pi),$$

where $b(x; n, \pi)$ is the binomial probability of x successes in n trials with success probability π and the sum is over all pairs (x, y) of x successes from group 1 and y successes from group 2 that give a Z^2 value bigger than or equal to the $Z^2 = 4.346$ value calculated from this data. The usual unconditional p value for this problem is $p_{\text{sup}} = \sup_{\pi \in [0,1]} p(\pi) = .061$. Suissa and Shuster (1985) discussed this p value and recommended it as an appropriate p value for this problem.

Looking at Figure 1, however, it would seem natural to restrict the region over which the maximization takes place to a region around the null maximum likelihood estimate $\hat{\pi} = (48 + 14)/(283 + 47) = .188$. A .999 confidence interval for π under the null hypothesis is given by [.123, .267] (see, for example, Casella and Berger 1990, p. 499). Numerically calculating the sup of $p(\pi)$ over this interval yields the value .036. Thus the new p value is $p_{.001} = .036 + .001 = .037$. This improvement in the p value is not unusual. We have found similar improvement in numerous 2×2 contingency table examples.

This example illustrates two important points. First, the supremum of $p(\pi)$ may occur at a π value far from the null maximum likelihood estimate $\hat{\pi}$. One might question the relevance of $p(.003) = .061$, because the data indicate that π is near .188 and not .003. Our method defines a set of π values close to $\hat{\pi}$ that should be examined in defining a valid p value. Storer and Kim (1990) and others have taken this notion to the extreme and have evaluated only $p(\hat{\pi})$. But $p(\hat{\pi})$ is typically not a valid p value.

Second, the function $p(\pi)$ may be quite variable. Unless it is performed carefully, numeric maximization can fail to

find the spikes in Figure 1. The function $p(\pi)$ may be much more stable on the confidence set C_β , and maximization on this restricted set is then much easier. In this example $p(\pi)$ is nearly constant between .123 and .267.

Example 3: Behrens-Fisher Problem. The classical Behrens-Fisher problem has two independent samples, X_1, \dots, X_m and Y_1, \dots, Y_n , from normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 . The null hypothesis is $H_0: \mu_1 = \mu_2$, where σ_1^2 is not assumed equal to σ_2^2 .

Best and Rayner (1987) reaffirmed the practical value of the Welch solution based on

$$t_w = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}},$$

where \bar{X} , \bar{Y} , s_1^2 , and s_2^2 are the usual sample means and sample variances and critical values are obtained from a t distribution with estimated degrees of freedom. But numerous studies have shown that the Welch solution can be slightly liberal. In other words, the corresponding p value does not satisfy (1) for certain combinations of m and n and $\theta = \rho = \sigma_2^2/\sigma_1^2$.

Here we can use our approach along with t_w to get a valid p value, because, under $H_0: \mu_1 = \mu_2$, the distribution of t_w depends only on the ratio of variances $\rho = \sigma_2^2/\sigma_1^2$. Although the distribution of t_w is not simple, we can easily simulate from normal distributions to get a p value for each value of ρ . Figure 2 shows the results for a data set with sample sizes $m = 9$ and $n = 13$, sample means 0 and 6.225, and sample variances 18 and 78 (an example taken from Barnard 1984). A .999 confidence interval for ρ obtained from the F distribution of s_1^2/s_2^2 is (.32, 38.72). On this interval the maximum two-sided p value is .048, so that $p_{.001} = .048 + .001 = .049$. Because the p value was obtained from 1,000,000 Monte Carlo replications, the standard error of the estimate .049 is around .0002. For comparison purposes, note that the Welch solution p value is .041, the pooled t p value is .065, and the Behrens-Fisher p value is .050.

Another way to use our approach in this problem follows from the quantity

$$t(\rho) = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{m} + \frac{\rho}{n}\right) \frac{(m-1)s_1^2 + (n-1)s_2^2/\rho}{(m+n-2)}}},$$

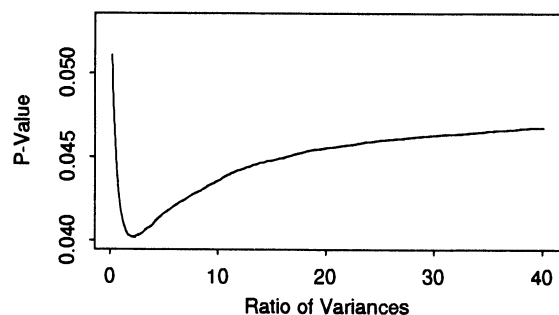


Figure 2. Estimated p Values for Welch's t as a Function of the Ratio of Variances ρ . Number of Monte Carlo replications = 1,000,000.

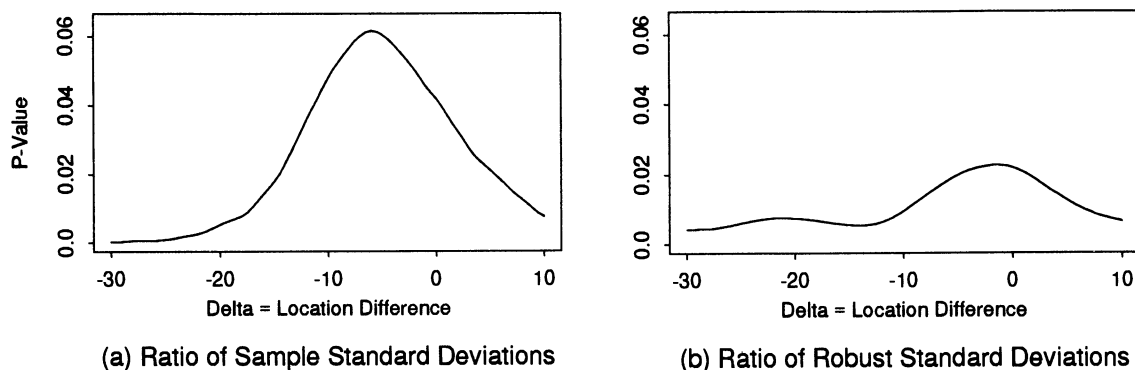


Figure 3. Estimated *p* Values for Tests of Scale for Weight Gain Data: (a) Ratio of Sample Standard Deviations; (b) Ratio of Robust Standard Deviations. Number of random permutations = 10,000.

given by Fisher (1939, p. 176). For a given value of ρ , $t(\rho)$ has a t distribution with $m + n - 2$ degrees of freedom under H_0 . Thus we might consider using our approach with $t(\rho)$ and this latter t distribution. The appropriate $p(\rho)$ is easy to calculate and has intuitive appeal. Unfortunately, this $p(\rho)$ is much more sensitive to changes in ρ than the simulation $p(\rho)$ based on t_w . We do not display the results for $p(\rho)$ but note that $p_{.001} = .233 + .001 = .234$ and $p_{.01} = .15 + .01 = .16$. Clearly, the method based on t_w is superior.

In fact, we believe that there is a general principle here concerning our method to the effect that one should use statistics such as t_w whose null distribution depends on the nuisance parameter rather than use pivotal quantities such as $t(\rho)$ that are functions of the nuisance parameter but whose null distributions do not depend on the nuisance parameter.

Our p value based on t_w is a valid p value for the Behrens-Fisher problem. Barnard (1984, sec. 6), Robinson (1976), and Tsui and Weerahandi (1989) claimed that the Behrens-Fisher solution p value is also valid. It would be interesting to compare the power properties of these two competing procedures.

The three previous examples were parametric problems where the nuisance parameter θ was confined to $(0, \infty)$, $[0, 1]$, and $(0, \infty)$. Now we turn to a more ambitious semi-parametric problem where θ is a location parameter belonging to $(-\infty, \infty)$ but there also is a second infinite dimensional nuisance parameter corresponding to an unknown distribution function. This is really not much more difficult than the previous examples, however, because we can handle this latter nuisance parameter using classical permutation test methods. That is, for each given value of θ , we will obtain a permutation p value and then carry on as in Examples 2 and 3.

Example 4: Testing for Scale Differences in Two Populations with Unknown Locations. Consider two iid samples X_1, \dots, X_m and Y_1, \dots, Y_n with distribution functions $F((x - \mu_1)/\sigma_1)$ and $F((x - \mu_2)/\sigma_2)$. The null hypothesis is $H_0: \sigma_1 = \sigma_2$; F, μ_1 , and μ_2 are unknown. This model is not identifiable, but an equivalent description in which all parameters are identifiable is for the X 's and Y 's to have distribution functions $F(x)$ and $F((x - \Delta)/\rho)$. The null hypothesis is then $H_0: \rho = 1$; F and Δ are unknown nuisance parameters.

The literature contains numerous good test statistics for this problem but none accompanied by valid finite sample p values. Actually, one can randomly pair the data in each sample and create differences $X_i - X_j$ and $Y_i - Y_j$, thereby eliminating the unknown locations. Rank and permutation tests on the differences then yield valid tests, but the loss in power due to the random pairing makes this approach unsuitable. A good review of test statistics and practical methods was provided by Conover et al. (1981).

If the difference in locations Δ was known, then we could subtract Δ from each of the Y 's, pool the X 's and the transformed Y 's, and carry out the standard permutation approach. That is, we compute a statistic T for each of the $\binom{m+n}{m}$ distinct permutation data sets $(X_1^*, \dots, X_m^*; Y_1^*, \dots, Y_n^*)$ drawn without replacement from the set $(X_1, \dots, X_m, Y_1 - \Delta, \dots, Y_n - \Delta)$. The permutation p value is then the proportion of these values that are greater than or equal to the statistic calculated from the original data.

For illustration, we consider the weight gain of a group of $m = 30$ control rats and of a second group of $n = 20$ rats whose diet included calcium EDTA. The data are from Brownie et al. (1986). The observed values for the control group are

34, 22, 51, 33, 20, 32, 35, 24, 13, 22, 26, 38, 34, 30, 20,

30, 25, 32, 36, 22, 26, 28, 31, 28, 32, 31, 28, 28, 31, 31;

those for the treated group are

9, 23, 16, 13, -13, 32, 10, 26, 14,

-24, 8, 29, 24, 27, 22, 2, 19, 21, 27, -1.

Figure 3 shows the estimated p values for $|\log(s_1/s_2)|$ and $|\log(g_1/g_2)|$, where s_1^2 and s_2^2 are the sample variances and the g_i are robust scale estimators with the form

$$g_1 = \frac{1}{M - [M(.25)]} \sum_{i=1}^{M - [M(.25)]} Z_{(i)},$$

where the $Z_{(i)}$ are the $M = m(m - 1)/2$ ordered values of $|X_j - X_k|$. These trimmed versions of Gini's mean difference were studied by Janssen, Serfling, and Veraverbeke (1987), and subsequently found to have good efficiency and robustness properties.

An exact $1 - \beta$ confidence interval for Δ under H_0 may be obtained by inverting any two-sample rank test for location differences. Here we use the interval based on the Wilcoxon rank sum statistic with the form $[D_{(k)}, D_{(l)}]$, where $D_{(1)}, \dots, D_{(mn)}$ are the ordered differences of the form $Y_j - X_i$ (see Randles and Wolfe 1979, p. 180). The .999 confidence interval for these data is $[-24, -3]$. This leads to $p_{.001} = .062 + .001 = .063$ for the variance-based statistic of Figure 3a and to $p_{.001} = .022 + .001 = .023$ for the robust statistic of Figure 3b. The standard errors of these p values are about .002 due to using 10,000 random permutations.

Asymptotic arguments given by Boos, Janssen, and Veraverbeke (1989) justify the use of $p(\hat{\Delta})$ in large samples, where $\hat{\Delta}$ is estimated from the data. For example, $\bar{Y} - \bar{X} = 14.2 - 29.1 = -14.9$, leading to $p(-14.9) = .018$ and .006 from Figures 3a and 3b. Taking the ratios $.063/.018$ and $.023/.006$ suggests a "cost" factor around 3 to 4 for getting a valid p value for these data in place of an asymptotic approximate p value.

We also note that the p value for the nonrobust statistic based on sample variances is much more sensitive to changes in Δ ; it ranges from .0012 to .062 over $\Delta \in [-24, -3]$, whereas the robust statistic based on g_1 and g_2 ranges from .005 to .022.

4. SUMMARY

Nuisance parameters may be handled in various ways in testing problems. In this article we have introduced a new method for modifying the standard definition of a p value given by $p = \sup_{\theta} P_{v_0, \theta}(T \geq t)$ to allow for taking the supremum over a confidence interval for θ instead of over the whole parameter space.

The new method is not intended to supplant standard methods for handling nuisance parameters when those methods give tractable answers. But our examples suggest that the new method can indeed give improved procedures, as in the case of the 2×2 contingency table using the Z^2 statistic. In other situations the new method can give finite-sample level- α tests where none previously existed.

Finally, we should like to reemphasize the principle mentioned at the end of Example 3: It is preferable to take suprema over the *distribution* of statistics such as t_w rather

than take suprema over pivotal quantities like $t(\rho)$ whose distribution does not depend on the nuisance parameter. In a heuristic sense, the supremum after averaging (or making probability calculations) tends to be smaller than averaging after taking the supremum.

[Received August 1992. Revised August 1993.]

REFERENCES

- Barnard, G. (1984), "Comparing the Means of Two Independent Samples," *Applied Statistics*, 33, 266-271.
- Best, D. J., and Rayner, J. C. W. (1987), "Welch's Approximate Solution for the Behrens-Fisher Problem," *Technometrics*, 29, 205-210.
- Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.
- Boos, D., Janssen, P., and Veraverbeke, N. (1989), "Resampling from Centered Data in the Two-Sample Problem," *Journal of Statistical Planning and Inference*, 21, 327-345.
- Brownie, C. F., Brownie, C., Noden, D. S., Krook, L., Haluska, M., and Aronson, A. L. (1986), "Teratogenic Effect of Calcium Edetate (CaEDTA) in Rats and the Protective Effect of Zinc," *Toxicology and Applied Pharmacology*, 82, 426-443.
- Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Pacific Grove, CA: Wadsworth.
- Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981), "A Comparative Study of Tests for Homogeneity of Variances, With Applications to the Outer Continental Shelf Bidding Data," *Technometrics*, 23, 351-361.
- Emerson, J. D., and Moses, E. M. (1985), "A Note on the Wilcoxon-Mann-Whitney Test for $2 \times k$ Ordered Tables," *Biometrics*, 41, 303-309.
- Fisher, R. A. (1939), "The Comparison of Samples With Possibly Unequal Variances," *Annals of Eugenics*, 9, 174-180.
- Janssen, P., Serfling, R., and Veraverbeke, N. (1987), "Asymptotic Normality of U -Statistics Based on Trimmed Samples," *Journal of Statistical Planning and Inference*, 16, 63-74.
- Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: John Wiley.
- Robinson, G. K. (1976), "Properties of Student's t and of the Behrens-Fisher Solution to the Two Means Problem," *The Annals of Statistics*, 4, 963-971.
- Storer, B. E., and Kim, C. (1990), "Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions," *Journal of the American Statistical Association*, 85, 146-155.
- Suissa, S., and Shuster, J. (1985), "Exact Unconditional Sample Sizes for the 2×2 Binomial Trial," *Journal of the Royal Statistical Society, Ser. A*, 148, 317-327.
- Tsui, K., and Weerahandi, S. (1989), "Generalized p -Values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters," *Journal of the American Statistical Association*, 84, 602-607.
- Taylor, D. N., Wachsmuth, I. K., Shangkuang, Y., Schmidt, E. V., Barrett, T. J., Schrader, J. S., Scherach, C. S., McGee, H. B., Feldman, R. A., and Brenner, D. J. (1982), "Salmonellosis Associated With Marijuana: A Multistate Outbreak Traced by Plasmid Fingerprinting," *New England Journal of Medicine*, 306, 1249-1253.