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 the data X have a probability distribution P_ν defined in terms of a parameter ν, and 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 above, 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 $\sup_\theta$ 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 $\sup_\theta$ 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 $\sup_\theta$ 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 two by two 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 $\sup_\theta$ by the calculation of a single probability, and each method can result in a *valid* p-value, i.e., a statistic p such that, under the null hypothesis,

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

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. That is, 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 above three methods is satisfactory. For example, the value of θ at which the $\sup_\theta$ 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 to condition upon.

In this paper we want to 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, $p(\theta_0)$ is 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 $\sup_\theta$ is calculated numerically (as it often will be), then it may be uncertain whether the numerical method indeed found the overall maximum. Of course, there is always uncertainty about the result of a numerical maximization, but the uncertainty is worse if the set being maximized over is unbounded. Statistically, it seems 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 it should be unnecessary to consider values of θ that are completely unsupported by the data. Authors such as Storer and Kim (1990) 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 above 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 below 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.

We will first give the theoretical justification for p_β in the following lemma. The rest of the paper 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 are about two by two contingency tables, the Behrens-Fisher problem, nonparametric testing for skewness, 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 the null hypothesis is true. Denote the true but unknown θ by θ_0. If $\beta > \alpha$, then since p_β is never smaller than β, $P(p_\beta \leq \alpha) = 0 \leq \alpha$. If $\beta \leq \alpha$, then

$$P(p_\beta \leq \alpha) \quad = \quad P(p_\beta \leq \alpha, \theta_0 \in C_\beta) + P(p_\beta \leq \alpha, \theta_0 \in \bar{C}_\beta)$$

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

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 $X_1, \ldots, X_n$ 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, 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

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

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

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

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

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, since $|z_{\text{obs}}| \to 0$ as $\sigma \to \infty$. 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 $\sup_{C_\beta}$ 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 we reject H_0 if $p_\beta \le \alpha$. Then the actual size of the test is

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

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 $\sqrt{(n-1)/\chi^2_\beta}$ converges to 1 as n goes to infinity. So the actual size of the test, which is at most α since the p-value is valid, converges to $\alpha - \beta$.

Example 2. *Two by two contingency table with independent binomial sampling.* Consider a two by two 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 two by two table chi-squared statistic

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

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

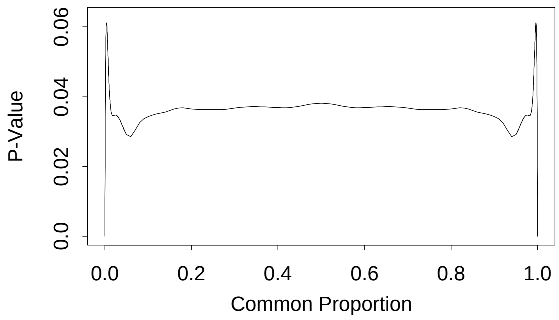


Figure 1: Exact p-values for the two by two table chi-squared statistic. Calculations are from independent binomial distributions with common proportion π.

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) discuss this p-value and recommend 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] (e.g., 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 since the maximum over $[0, 1]$ often occurs near 0 or 1, far from the estimated value of π (as we have found in numerous examples).

In fact, the program EXACTB by Shuster (1988) will compute the maximum of $p(\pi)$ over a .999 confidence interval, and report it as a p-value. But the program documentation does not provide any theory to justify this approach. Also, the value reported is just the maximum, not the maximum plus β as in (2). So the reported p-value may not be valid. As an aside, we note that EXACTB will also compute p_{sup}. But for this data, the value computed by EXACTB was $p_{\text{sup}} = .038$. Apparently, the maximization routine failed to detect the spikes in $p(\pi)$ near 0 and 1. But the spikes are real and the correct value is .061, as we reported above.

Example 3. *Behrens-Fisher problem.* The classical Behrens-Fisher problem has two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ 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) recently 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. Numerous studies have shown, however, 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 since, 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 means 0.0 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 $p_{.001} = .048 + .001 = .049$. Since the p-value was obtained from 1,000,000 Monte Carlo replications, the standard error of the estimate .049 is around .0002. For

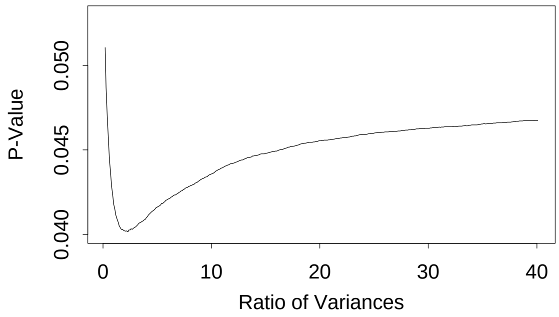


Figure 2: Estimated p-values for Welch's t as a function of the ratio of variances ρ. Number of Monte Carlo replications = 1,000,000.

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

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

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 methods 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 $t(\rho)$ which are functions of the nuisance parameter but whose null distributions do not depend on the nuisance parameter.

Our p-value based on t_w may be the first nontrivial valid p-value for the Behrens-

Fisher problem, although Barnard (1984, Sec. 6) claims that Robinson (1976) has shown that the Behrens-Fisher solution p-value is valid. It is not clear to us that Robinson (1976) has actually proved such a result. But from a practical view we must point out that neither our solution nor the Behrens-Fisher solution are likely to be robust to nonnormality since they both use the F distribution of s_1^2/s_2^2.

The three previous examples were parametric problems where the nuisance parameter θ was confined to $(0,\infty)$, [0,1] and $(0,\infty)$, respectively. Now we turn to more ambitious semi-parametric problems where θ is a location parameter belonging to $(-\infty,\infty)$, but in addition, there is a second infinite dimensional nuisance parameter corresponding to an unknown distribution function. This is really not much harder 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 skewness with unknown location.* We have a single iid sample $X_1,\ldots,X_n$ and wish to test whether the X's are symmetrically distributed about some unknown θ. Formally the null hypothesis is $H_0 : F(\theta + x) = 1 - F((\theta - x)^-)$ all $x \in (-\infty,\infty)$, F and θ unknown. A variety of good test statistics have been proposed for this problem, but no finite sample valid p-values have been given. In fact Schuster and Barker (1987) have proposed bootstrap methods because even approximate validity has been so elusive.

For illustration purposes we shall consider two simple statistics. The first is the sample standardized third moment

$$\sqrt{b_1} = m_3/m_2^{3/2},$$

where $m_k = \sum(X_i - \bar{X})^k/n$. D'Agostino, Belanger, and D'Agostino (1990) discuss the use of $\sqrt{b_1}$ as a test for normality, but there has been no valid method for using it as a test of asymmetry. For example, Randles, et al. (1980) show that the appropriate standardization of $\sqrt{b_1}$ by an estimate of its asymptotic standard deviation results in a normal test which is extremely liberal for a variety of symmetric distributions.

The second statistic we consider is the triples statistic

$$T = \hat{\eta}/\hat{\sigma},$$

where $\hat{\eta}$ is the U-statistic estimate of $\eta = P(X_1 + X_2 > 2X_3) - P(X_1 + X_2 < 2X_3)$ and $\hat{\sigma}$ is the corresponding estimate of standard deviation. T is asymptotically normal for any symmetric F, but Table 2 of Randles et al. (1980) shows that using normal or t critical values results in a test which is approximately valid in small samples but which can be liberal for certain symmetric Fs.

Our method in this situation is as follows. For given θ, a valid p-value can be obtained by calculating the statistic of interest for each of the 2^n possible samples of the form $\pm|X_1 - \theta|, \ldots, \pm|X_n - \theta|$. The permutation p-value is just the proportion of these values which are greater than or equal to the statistic calculated from the original n observations. Since 2^n is often a very large number, we typically randomly sample from the set of possible permutations.

The second ingredient of our method is a confidence interval for θ. The simplest interval is the exact confidence interval for the median (which equals θ under H_0) given by $(X_{(l)}, X_{(n-l+1)})$, where $X_{(1)} \le \cdots \le X_{(n)}$ are the order statistics and $\beta = (1/2)^{n-1} \sum_{i=0}^{l-1} \binom{n}{i}$ (see David, 1981, Sec. 2.5).

To illustrate, we consider the sample of $n = 62$ cholesterol values given in D'Agostino, Belanger, and D'Agostino (1990). Figure 3 shows estimated right-tailed p-values versus θ for $\sqrt{b_1}$ and T based on 10,000 random permutations. Using l=18 in the above confidence procedure, we get $\beta = .000497$, and so a .9995 confidence interval for θ under H_0 is $(X_{(18)}, X_{(45)}) = (220, 267)$ resulting in $p_{.0005} = .0370 + .0005 = .038$ for $\sqrt{b_1}$ and $p_{.0005} = .0177 + .0005 = .018$ for T. Use of the random permutations (instead of all 2^{62} permutations) introduces a standard error of about $((0.03)(0.97)/10,000)^{1/2} = 0.0017$.

The triples statistic for this data is T=2.501, and using a t distribution with $n = 62$ degrees of freedom as suggested by Randles et al. (1980), we get an approximate right-tailed p-value of .0075. Since $.018/.0075 \doteq 2.4$, we might say that there is a 2.4 "cost" factor in this case to obtain the valid p-value of .018, rather than an approximate value. Using $\sqrt{b_1}$ and the approximation given by D'Agostino, Belanger, and D'Agostino (1990), we obtain .00085 for a one-tailed p-value for normality versus a right-skewed

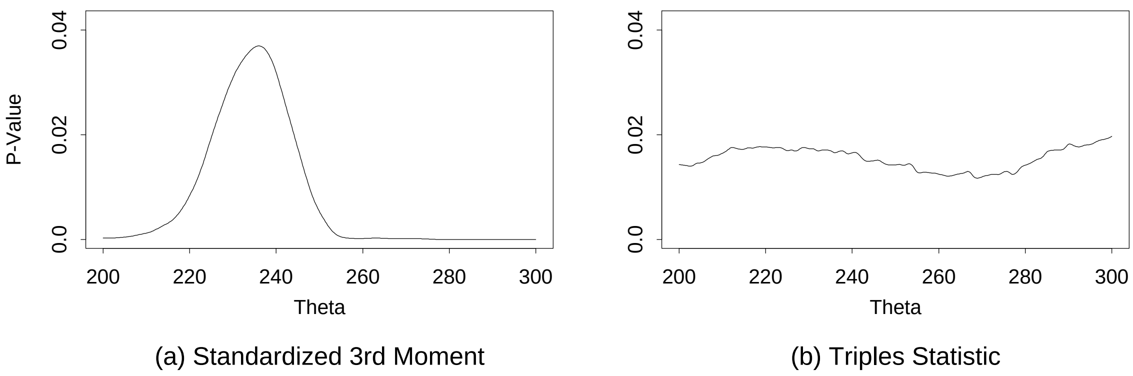


Figure 3: Estimated p-values for tests of skewness for cholesterol data. Number of random permutations = 10,000.

distribution. Here though,it is unfair to compare with the valid p-value of .038 since the null class of all symmetric distributions is much bigger than that of the set of normal distributions.

The range of variation of the p-values in Figure 3 is much larger for $\sqrt{b_1}$ than for T. We believe that the difference is due to the robustness of T compared to that of $\sqrt{b_1}$. That is, the sample third moment is very sensitive to outliers and to small changes in distributional shape. The triples T is an average of indicator functions and insensitive on a large scale to such changes although wiggles do occur because of the discontinuities caused by the indicator function. A second difference between $\sqrt{b_1}$ and T is the fact that T is studentized and $\sqrt{b_1}$ is not. A plot of p-values for $\hat{\eta}$ in place of T (not displayed) is qualitatively similar to Figure 3b and suggests that studentization is not a major cause of differences between Figures 3a and 3b.

Example 5. *Testing for scale differences in two populations with unknown locations.* Consider two iid samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ with distribution functions $F((x-\mu_1)/\sigma_1)$ and $F((x-\mu_2)/\sigma_2)$, respectively. The null hypothesis is $H_0 : \sigma_1 = \sigma_2$; F, μ_1 and μ_1 are unknown. This model is not identifiable, but an equivalent description in which all parameters are identifiable is for the Xs and Ys to have distribution functions $F(x)$ and $F((x-\Delta)/\rho)$, respectively. The null hypothesis is then $H_0 : \rho = 1$; F and Δ are unknown nuisance parameters.

As in Example 4, there are numerous good test statistics in the literature 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 is found in Conover et al. (1981).

If the difference in locations Δ were known, we could subtract Δ from each of the Ys, pool the Xs and transformed Ys, 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 $(X_1^*, \ldots, X_m^*; Y_1^*, \ldots, Y_n^*)$ drawn without replacement from the set $(X_1, \ldots, X_m, Y_1 - \Delta, \ldots, Y_n - \Delta)$. The permutation p-value is then the proportion of these values which 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 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,

and for the treated group are

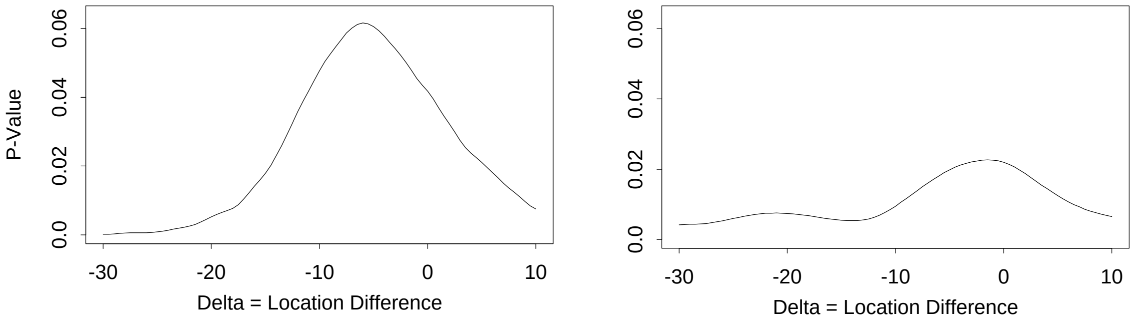
9, 23, 16, 13,-13, 32, 10, 26, 14,-24, 8, 29, 24, 27, 22, 2, 19, 21, 27, -1.

Figure 4 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

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

and 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 in 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



(a) Ratio of Sample Standard Deviations

(b) Ratio of Robust Standard Deviations

Figure 4: Estimated p-values for tests of scale for weight gain data. Number of random permutations = 10,000.

Wilcoxon rank sum statistic which has the form $[D_{(k)}, D_{(l)})$ where $D_{(1)}, \ldots, D_{(mn)}$ are the ordered differences of the form $Y_j - X_i$ (see Randles and Wolfe, 1979, p. 180). The .999 confidence interval for the above data is [-24,-3]. This leads to $p_{.001} = .062 + .001 = .063$ for the variance-based statistic of Figure 4a and to $p_{.001} = .022 + .001 = .023$ for the robust statistic of Figure 4b. The standard errors of these p-values are about .002 due to using 10,000 random permutations.

Asymptotic arguments are given in Boos, Janssen, and Veraverbeke (1989) which justify the use in large samples of $p(\hat{\Delta})$, 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 4a and 4b, respectively. Taking the ratios .063/.018 and .023/.006 suggests a "cost" factor around 3 to 4 for getting a valid p-value for this data in place of an asymptotic approximate p-value.

We also note that, as in Example 4, the p-value for the nonrobust statistic based on sample variances is much more sensitive to changes in Δ, ranging from .0012 to .062 over $\Delta \in [-24, -3]$, while the robust statistic based on g_1 and g_2 ranges from .005 to .022.

4 Summary

Nuisance parameters may be handled in a variety of ways in testing problems. In this paper we have introduced a new method for modifying the standard definition of a p-value given by $p = \sup_\theta P_{\nu_0,\theta}(T \geq t)$ to allow for taking the supremum over a confidence interval for θ instead of over the whole parameter space of θ.

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 two by two contingency table using the Z^2 statistic. In other situations the new method can give finite-sample level - α tests where none previously existed.

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, Inc.

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.

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.

D'Agostino, R. B., Belanger, A., and D'Agostino, R. B. Jr. (1990), "A Suggestion for

Using Powerful and Informative Tests of Normality," *The American Statistician*, 44, 316-321.

David, H. A. (1981), *Order Statistics*, 2nd Ed., New York: John Wiley.

Emerson, J. D., and Moses, E. M. (1985), "A Note on the Wilcoxon-Mann-Whitney Test for 2 x 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.

Randles, R. H., Fligner, M. A., Policello, G. E., II, and Wolfe, D. A. (1980), "An Asymptotically Distribution-Free Test for Symmetry versus Asymmetry," *Journal of the American Statistical Association*, 75, 168-172.

Robinson, G. K. (1976), "Properties of Student's t and of the Behrens-Fisher Solution to the Two Means Problem," *Annals of Statistics*, 4, 963-971.

Schuster, E. F., and Barker, R. C. (1987), "Using the Bootstrap in Testing Symmetry versus Asymmetry," *Communications in Statistics - Simulation & Computation*, 16, 69-84.

Shuster, J. (1988), "EXACTB: Exact Unconditional P-Values for the 2x2 Binomial Trial," Research Assistance Corp, Gainesville, Fl.

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 2x2 Binomial Trial," *Journal of the Royal Statistical Society*, Ser. A, 148, 317-327.

Taylor, D. N., Wachsmuth, I. K., Shangkuan, 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.