

Comparing discrete distributions when the sample space is small

Comparando las distribuciones discretas cuando el espacio muestral es pequeño

Recibido: junio 1 de 2012 | Revisado: agosto 1 de 2012 | Aceptado: agosto 20 de 2012

RAND R. WILCOX^{*}
CHERYL VIGEN^{**}
FLORENCE CLARK
MICHAEL CARLSON

University of Southern California, United States

ABSTRACT

This paper describes two new methods for comparing two independent, discrete distributions, when the sample space is small, using an extension of the Storer–Kim method for comparing independent binomials. These methods are relevant, for example, when comparing groups based on a Likert scale, which was the motivation for the paper. In essence, the goal is to test the hypothesis that the cell probabilities associated with two independent multinomial distributions are equal. Both a global test and a multiple comparison procedure are proposed. The small-sample properties of both methods are compared to four other techniques via simulations: Cliff's generalization of the Wilcoxon–Mann–Whitney test that effectively deals with heteroscedasticity and tied values, Yuen's test based on trimmed means, Welch's test and Student's *t* test. For the simulations, data were generated from beta-binomial distributions. Both symmetric and skewed distributions were used. The sample space consisted of the integers 0(1)4 or 0(1)10. For the global test that is proposed, when testing at the 0.05 level, simulation estimates of the actual Type I error probability ranged between 0.043 and 0.059. For the new multiple comparison procedure, the estimated family wise error rate ranged between 0.031 and 0.054 for the sample space 0(1)4. But for 0(1)10, the estimates dropped as low as 0.016 in some situations. Given the goal of comparing means, Student's *t* is well known to have practical problems when distributions differ. Similar problems are found here among the situations considered. No single method dominates in terms of power, as would be expected, because different methods are sensitive to different features of the distributions being compared. But in general, one of the new methods tends to have relatively good power based on both simulations and experience with data from actual studies. If, however, there is explicit interest in comparing means, rather than comparing the cell probabilities, Welch's test was found to perform well. The new methods are illustrated using data from the Well-Elderly Study where the goal is to compare groups in terms of depression and the strategies used for dealing with stress.

Keywords authors

Multinomial distribution; Likert scales, Storer–Kim method; Hochberg's method; Yuen's method.

Keywords plus

Welch's Test, Student's Test, Well Elderly Study, Stress, Depression, Methodology.

doi:10.11144/Javeriana.UPSY12-5.cdds

Para citar este artículo: Wilcox, R. R., Vigen, C., Clark, F., & Carlson, M. (2013). Comparing discrete distributions when the sample space is small. *Universitas Psychologica*, 12 (5), 1583-1595. doi:10.11144/Javeriana.UPSY12-5.cdds

^{*} University of Southern California. Dept of Psychology. E-mail: rwilcox@usc.edu

^{**} University of Southern California. Occupational Science and Occupational Therapy

RESUMEN

En este artículo se describen dos nuevos métodos para comparar dos distribuciones discretas independientes, cuando el espacio muestral es pequeño, usando una extensión del método Storer-Kim para comparar binomios independientes. Estos métodos son relevantes, por ejemplo, cuando se comparan grupos basados en una escala Likert, la cual motivó la escritura del artículo. En esencia, el objetivo es evaluar la hipótesis de que las probabilidades de células asociadas con dos distribuciones multinomiales independientes son iguales. Se propone una prueba global y un procedimiento de comparación múltiple. Las propiedades de las muestras pequeñas de ambos métodos fueron comparadas con otras cuatro técnicas a través de simulaciones: generalización de Cliff de la prueba de Wilcoxon-Mann-Whitney que trata eficazmente con heteroscedasticidad y valores vinculados, la prueba de Yuen basada en medias truncadas, la prueba de Welch y la prueba t de Student. Para las simulaciones, los datos se generaron a partir de distribuciones beta-binomiales. Se utilizaron distribuciones tanto simétricas como asimétricas. El espacio muestral consistió en los enteros 0(1)4 o 0(1)10. Para la prueba global que se propone, cuando se evaluó al nivel de 0.05, la simulación estimó la probabilidad del error tipo I osciló entre 0.043 y 0.059. Para el nuevo procedimiento de comparación múltiple, la tasa de error estimada oscilaba entre 0.031 y 0.054 para el espacio de la muestra 0(1)4. Pero para 0(1)10, las estimaciones fueron tan bajas como 0.016 en algunas situaciones. Teniendo en cuenta el objetivo de la comparación de medias, la prueba t de Student es bien conocida por tener problemas prácticos cuando distribuciones difieren. Problemas similares se encontraron entre las situaciones consideradas. No existe un único método que domina en términos de poder, como sería de esperar, debido a que los diferentes métodos son sensibles a las diferentes características de las distribuciones que son comparadas. Pero en general, uno de los nuevos métodos tiende a tener relativamente buen poder basado tanto en simulaciones y la experiencia con los datos de estudios reales. Si, sin embargo, existe un interés explícito en comparar medias, en lugar de comparar las probabilidades de celda, la prueba de Welch se encuentra que tiene un buen desempeño. Los nuevos métodos se ilustran usando datos del estudio Well-Elderly donde el objetivo es comparar los grupos en cuanto a la depresión y las estrategias utilizadas para hacer frente al estrés.

Palabras clave autores

Distribución multinomial, escalas likert, método Storer-Kim, método de Hochberg, método de Yuen.

Palabras clave descriptores

Prueba de Welch, prueba t de student, estudio *Well-Elderly*, estrés, depresión, metodología.

Introduction

Consider two independent, discrete random variables, X and Y, each having a sample space with N elements. Of particular interest here are situations

where Likert scales are being compared, in which case N is relatively small and typically has a value between 3 and 10, but the results reported here are relevant to a broader range of situations, as will become evident. The means are denoted by μ_1 and μ_2 , respectively. Denote the elements of the sample space by x_1, \dots, x_N , and let $p_j = P(X = x_j)$ and $q_j = P(Y = x_j)$, $j = 1, \dots, N$. There are, of course, various ways these two distributions might be compared. Six methods are considered here, two of which are new. One of the new methods compares multinomial distributions based on a simple extension of the Storer and Kim (1990) method for comparing two independent binomial distributions. The other is a multiple comparison procedure for all N cell probabilities based on the Storer-Kim method combined with Hochberg's (1998) method for controlling the probability of one or more Type I errors. (Hochberg's method represents an improvement on the Bonferroni procedure.) These two new methods are compared to four alternative strategies in terms of power and the probability of Type I error.

One of the more obvious strategies for comparing the two distributions is to test

$$H_0 : \mu_1 = \mu_2 \quad (1)$$

with Student's t test or Welch's (1938) test, which is designed to allow heteroscedasticity. Note that for the special case where $N = 2$, testing Equation 1 corresponds to testing the hypothesis that two independent binomial probability functions have the same probability of success. That is, if $x_1 = 0$, $x_2 = 1$, testing 1 corresponds to testing $H_0: p_2 = q_2$, for which numerous methods have been proposed (e.g. Wilcox, 2012b, Section 5.8). The method derived by Storer and Kim (1990) appears to be one of the better methods in terms of both Type I errors and power. The point here is that their method is readily generalized to situations where the goal is to test

$$H_0 : p_j = q_j, \text{ for all } j = 1, \dots, N. \quad (2)$$

That is, the goal is to test the global hypothesis that all N cell probabilities of two independent multinomials are equal. In contrast, there might

be interest in testing N hypotheses corresponding to each of the N cell probabilities. That is, for each j , test $H_0: p_j = q_j$ with the goal of determining which cell probabilities differ. Moreover, it is desirable to perform these individual tests in a manner that controls the probability of one or more Type I errors. There is a rather obvious approach based in part on the Storer–Kim method, but here it is found that with small sample sizes, a slight adjustment is needed to avoid having the actual probability of one or more Type I errors well below the nominal level.

$$\text{Let } p = P(X < Y) + 0.5P(X = Y). \quad (3)$$

yet another approach to comparing two distributions is testing $H_0: p = 0.5$, which has been advocated by Cliff (1996), Acion et al. (2006), and Vargha and Delaney (2000), among others. Moreover, compared to several other methods aimed at testing (3), Cliff’s method has been found to compete well in terms of Type I errors (Neuhäuser, Lösch & Jöckel, 2007). Note that the Wilcoxon–Mann–Whitney test is based in part on a direct estimate of p . However, under general conditions, when distributions differ in shape, it uses the wrong standard error, which can affect power, Type I error probabilities, and the accuracy of the confidence interval for p . Cliff’s method deals effectively with these problems, including situations where tied values are likely to occur. In particular, it uses a correct estimate of the standard error even when the distributions differ.

It is evident that the methods just outlined are sensitive to different features of the distributions being compared (e.g., Roberson et al., 1995). If, for example, there is explicit interest in comparing means, testing (1) is more reasonable than testing (2). However, even when the means are equal, the distributions can differ in other ways that might be detected when testing (4) or (2). Nevertheless, it is informative to consider how these methods compare in terms of power. This paper reports simulation results aimed at addressing this issue, as well as new results on controlling the probability of a Type I error. Data stemming from a study dealing with the

mental health of older adults are used to illustrate that there can be practical advantages to testing (2).

Details of the Methods To Be Compared

This section describes the methods to be compared. Three are based on measures of location: the usual Student’s t test, Welch’s (1938) test, and a generalization of Welch’s method derived by Yuen (1974) designed to test the hypothesis that two independent groups have equal (population) trimmed means. For brevity, the computational details of Student’s test are not described simply because it is so well known. Yuen’s test is included because it has been examined extensively in recent years and found to have advantages when dealing with distributions that differ in skewness or when outliers are commonly encountered (e.g., Guo & Luh, 2000; Lix & Keselman, 1998; Wilcox & Keselman, 2003; Wilcox, 2012a, 2012b). However, for the types of distributions considered here, evidently there are few if any published results on how well it performs. The fourth approach considered here is Cliff’s method. The final two methods are aimed at comparing independent multinomial distributions as indicated in the introduction.

The Welch and Yuen methods

Here, Yuen’s method is described based on 20% trimming. When there is no trimming, it reduces to Welch’s test. Under normality, Yuen’s method yields about the same amount of power as methods based on means, but it helps guard against the deleterious effects of skewness and outliers. For the situation at hand, it will be seen that skewness plays a role when comparing means.

For notational convenience, momentarily consider a single group and let X_1, \dots, X_n be a random sample of n observations from the j th group and let $g_j = [0.2n]$, where $[0.2n]$ is the value of $0.2n$ rounded down to the nearest integer. Let $X_{(1)} \leq \dots \leq X_{(n)}$ be

the n observations written in ascending order. Let $h = n - 2g$. That is, h is the number of observations left after trimming by 20%. The 20% trimmed mean is

$$\bar{X}_t = \frac{X_{(g+1)} + \dots + X_{(n-g)}}{n = 2g}$$

Winsorizing the observations by 20% means that the smallest 20% of the observations that were trimmed, when computing the 20% trimmed mean, are instead set equal to the smallest value not trimmed. Simultaneously, the largest values that were trimmed are set equal to the largest value not trimmed. The *Winsorized mean* is the average of the Winsorized values, which is labeled. In symbols, the Winsorized mean is

$$\bar{X}_w = \frac{1}{n} \{ (g+1)X_{(g+1)} + X_{(g+2)} + \dots + X_{(n-g-1)} + (g+1)X_{(n-g)} \}$$

The Winsorized variance is s_w^2 , the sample variance based on the Winsorized values.

For two independent groups, let

$$d_j = \frac{(n_j - 1)s_{wj}^2}{h_j(h_j - 1)},$$

where s_{wj}^2 is the Winsorized variance for the j th group. Yuen's test statistic is

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}}$$

The degrees of freedom are

$$\hat{\nu} = \frac{(d_1 + d_2)^2}{\frac{d_1^2}{h_1 - 1} + \frac{d_2^2}{h_2 - 1}}$$

The hypothesis of equal trimmed means is rejected if $|T_y| \geq t$, where t is the $1 - \alpha/2$ quantile of Student's t distribution with $\hat{\nu}$ degrees of freedom. If there is no trimming, Yuen's method reduces to Welch's method for means. For convenience, Student's t , Welch's test, and Yuen's test are labeled methods T, W and Y, respectively.

Cliff's Method

Let

$$p_1 = P(X_{i1} > X_{i2}),$$

$$p_2 = P(X_{i1} = X_{i2}),$$

and

$$p_3 = P(X_{i1} < X_{i2}).$$

Cliff (1996) focuses on testing

$$H_0 : \delta = p_1 - p_3 = 0,$$

his method is readily adapted to making inferences about p , as will be indicated. (Note that $p_3 + 0.5p_2$ corresponds to the left side of equation 3.) For the i th observation in group 1 and the h th observation in group 2, let $d_{ih} = -1$ if $P(X_{i1} < X_{h2})$, $d_{ih} = 0$ if $X_{i1} = X_{h2}$ and $d_{ih} = 1$ if $X_{i1} > X_{h2}$. An estimate of $\delta = P(X_{i1} > X_{i2}) - P(X_{i1} < X_{i2})$ is

$$\hat{\delta} = \frac{1}{n_1 n_2} \sum \sum d_{ih},$$

the average of the d_{ih} values. Let

$$\bar{d}_{.i} = \frac{1}{n_2} \sum_h d_{ih},$$

$$\bar{d}_{.h} = \frac{1}{n_1} \sum_i d_{ih},$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\bar{d}_{.i} - \hat{\delta})^2$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{h=1}^{n_2} (\bar{d}_{.h} - \hat{\delta})^2,$$

$$\tilde{\sigma}^2 = \frac{1}{n_1 n_2} \sum \sum (d_{ih} - \hat{\delta})^2$$

Then

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \tilde{\sigma}^2}{n_1 n_2}$$

estimates the squared standard error of \bar{X} . Let z be the $1 - \alpha/2$ quantile of a standard normal distribu-

tion. Rather than use the more obvious confidence interval for δ , Cliff (1996, p. 140) recommends

$$\frac{\hat{\delta} - \hat{\delta}^3 \pm z\hat{\sigma} \sqrt{(1 - \hat{\delta}^2)^2 + z^2 \hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2 \hat{\sigma}^2}$$

The parameter δ is related to p in the following manner:

$$\delta = 1 - 2p,$$

so
$$p = \frac{1 - \delta}{2}.$$

Letting

$$C_\ell = \frac{\hat{\delta} - \hat{\delta}^3 - z\hat{\sigma} \sqrt{(1 - \hat{\delta}^2)^2 + z^2 \hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2 \hat{\sigma}^2}$$

and

$$C_u = \frac{\hat{\delta} - \hat{\delta}^3 + z\hat{\sigma} \sqrt{(1 - \hat{\delta}^2)^2 + z^2 \hat{\sigma}^2}}{1 - \hat{\delta}^2 + z^2 \hat{\sigma}^2}$$

a $1 - \alpha$ confidence interval for p is

$$\left(\frac{1 - C_u}{2}, \frac{1 - C_\ell}{2} \right).$$

As noted in the introduction, this confidence interval has been found to perform well compared to other confidence intervals that have been proposed. This will be called method C.

An Extension of the Storer–Kim Method

The Storer and Kim method for comparing two independent binomials can be extended to comparing independent multinomials in a straightforward manner. The basic strategy has been studied in a broader context by Alba-Fernández and Jiménez-Gamero (2009). In essence, a bootstrap method is used to determine a p-value assuming (2) is true. (See Liu & Singh, 1997, for general results on bootstrap methods for computing a p-value.) First, we review the Storer–Kim method.

Note that the possible number of successes in the first group is any integer, x , between 0 and n_1 , and for the second group it is any integer, y , between 0 and n_2 . Let r_j be the number of successes observed in the j th group and let

$$\hat{p} = \frac{r_1 + r_2}{n_1 + n_2}$$

be the estimate of the common probability of success assuming the null hypothesis is true. For any x between 0 and n_1 and any y between 0 and n_2 , set

$$a_{xy} = \begin{cases} 1 & \text{if } \left| \frac{x}{n_1} - \frac{y}{n_2} \right| \geq \left| \frac{r_1}{n_1} - \frac{r_2}{n_2} \right|, \\ 0 & \text{otherwise} \end{cases}$$

otherwise $a_{xy} = 0$.

The test statistic is

$$T = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} a_{xy} b(x, n_1, \hat{p}) b(y, n_2, \hat{p}),$$

where

$$b(x, n_1, \hat{p}) = \frac{n_1!}{x!(n_1 - x)!} \hat{p}^x (1 - \hat{p})^{n_1 - x}$$

and $b(y, n_2, \hat{p})$ is defined in an analogous fashion. The null hypothesis is rejected at the α level if

$$T \leq \alpha.$$

That is, T is the p-value.

Generalizing, let c_{kj} be the number of times the value x_k ($k=1, \dots, N$) is observed among the n_j participants in the j th group, let

$$\hat{p}_k = \frac{c_{k1} + c_{k2}}{n_1 + n_2}$$

be the estimate of the assumed common cell probability, and let

$$S = \sum \left(\frac{c_{k1}}{n_1} - \frac{c_{k2}}{n_2} \right)^2.$$

The goal is to determine whether S is sufficiently large to reject (2). But as N gets large, computing the probabilities associated with every element in the sample space of a multinomial distribution becomes impractical. To deal with this, randomly sample n_1 observations from a multinomial distribution having cell probabilities $\hat{p}_1, \dots, \hat{p}_N$. Repeat this process only now generating n_2 observations. Based on the resulting cell counts, compute S and label the result S^* . Repeat this process B times yielding S_1^*, \dots, S_B^* . Then a p-value is given by

$$1 - \frac{1}{B} \left(\sum_{b=1}^B I_{S > S_b^*} + .5 \sum_{b=1}^B I_{S = S_b^*} \right),$$

where the indicator function $I_{S > S_b^*} = 1$ if $S > S_b^*$; otherwise $I_{S > S_b^*} = 0$. Here, $B = 500$ is used, which generally seems to suffice in terms of controlling the Type I error probability (e.g., Wilcox, 2012b), but a possibility is that a larger choice for B will result in improved power (e.g., Wilcox, 2012a, p. 277). This will be called method M.

Multiple Comparisons

The final method tests $H_0: p_j = q_j$ for each j with the goal of controlling the probability of one or more Type I errors. This is done using a modification of Hochberg's (1988) sequentially rejective method. (A slight modification of Hochberg's method, derived by Rom, 1990, might provide a slight advantage in terms of power.) To describe Hochberg's method, let p_1, \dots, p_N be the p-values associated with the N tests as calculated by the extended Storer-Kim method, and put these p-values in descending order yielding $p_{[1]} \geq p_{[2]} \geq \dots \geq p_{[C]}$. Beginning with $k = 1$, reject all hypotheses if

$$p_{[k]} \leq \alpha/k.$$

If $p_{[k]} > \alpha$, proceed as follows:

1. Increment k by 1. If
2. $p_{[k]} \leq \alpha/k$,

stop and reject all hypotheses having a p-value less than or equal $p_{[k]}$

If $p_{[k]} > \alpha/k$, repeat step 1.

3. Repeat steps 1 and 2 until a significant result is found or all N hypotheses have been tested.

Based on preliminary simulations, a criticism of the method just described is that the actual probability of one or more Type I errors can drop well below the nominal level when the sample size is small. The adjustment used here is as follows. If the p-value (i.e., $p_{[k]}$) is less than or equal to twice the nominal level, the minimum sample size is less than 20, $N > 4$, and the sample sizes are not equal, then divide the p-value by 2. This will be called method B.

3 Simulation Results

Using the software R, simulations were used to study the small-sample size properties of the six methods described in section 2. To get a reasonably broad range of situations, observations were generated from a beta-binomial distribution that has probability function

$$p(x) = \frac{N!}{x!(N-x)!} \frac{B(x+r, N-x+s)}{B(r, s)}$$

($x = 0, \dots, N$), where the parameters r and s determine the first two moments and B is the beta function. The mean of the beta-binomial distribution is

$$\frac{Nr}{r+s}$$

and the variance is

$$\frac{Nrs(N+r+s)}{(r+s)^2(1+r+s)}$$

when $r = s$, the distribution is symmetric. Note that if both r and s are multiplied by some constant, d , the mean remains the same but the variance generally, but not always, decreases. If $r < s$, the distribution is skewed to the right. When $r = s$ and $r < 1$, the distribution is U-shaped. Here $N = 4$ and 10 are used with sample sizes 20 and 40. Figure 1 shows

TABLE 1: Estimated Type I error probabilities, $\alpha = 0.05$

N	n1	n2	r	s	Methods					
					B	C	M	Y	W	T
4	20	20	0.5	0.5	0.054	0.044	0.052	0.054	0.047	0.047
4	20	20	2	2	0.053	0.046	0.050	0.043	0.048	0.048
4	20	20	1	3	0.046	0.055	0.059	0.043	0.054	0.056
4	20	40	0.5	0.5	0.033	0.048	0.046	0.062	0.049	0.048
4	20	40	2	2	0.048	0.048	0.054	0.041	0.051	0.05
4	20	40	1	3	0.031	0.041	0.045	0.036	0.048	0.042
10	20	20	0.5	0.5	0.016	0.040	0.047	0.052	0.048	0.048
10	20	20	2	2	0.019	0.043	0.044	0.047	0.05	0.05
10	20	20	1	3	0.019	0.040	0.046	0.044	0.045	0.046
10	20	40	0.5	0.5	0.056	0.050	0.046	0.058	0.053	0.052
10	20	40	2	2	0.053	0.047	0.043	0.049	0.052	0.05
10	20	40	1	3	0.048	0.048	0.049	0.053	0.049	0.048

the probability function for $N = 10$, $(r, s)=(0.5,0.5)$ (indicated by lines connecting the points labeled o), $(r, s) = (2, 2)$ (indicated by the points corresponding to *) and $(r, s) = (1, 3)$ (indicated by a +).

Table 1 shows the estimated probability of a Type I error, based on 4000 replications, when X and Y have identical distributions and the nominal Type I error probability is 0.05. (If the method in Wilcox, 2012, p. 166, is used to test the hypothesis that the actual Type I error probability is 0.05, based on a 0.95 confidence interval, then power exceeds 0.99 when the actual Type I error probability is 0.075. Also, with 4000 replications, the resulting half-lengths of the 0.95 confidence intervals for the actual Type I error probability were approximately 0.007 or less.) In terms of avoiding Type I error probabilities larger than the nominal level, all of the methods perform well. Only method Y has an estimate larger than 0.06, which occurred when $N = 4$ and $r = s = 0.5$. In terms of avoiding Type I error probabilities well below the nominal level, say less than 0.04, all of the methods performed well except method B when $N = 10$ and $n_1 = n_2 = 20$, where

the estimate drops below 0.02. This suggests that method B might have less power than method M, but situations are found where the reverse is true.

From basic principles, when $N = 2$, meaning that the goal is to test the hypothesis that two independent binomials have the same probability of success, reliance on the central limit theorem can be problematic when the probability of success is close to zero or one. Also, it is known that when distributions differ in terms of skewness, this can adversely affect the control over the probability of a Type I error when using Student's *t* test, particularly when the sample sizes differ. Of course, a related concern is that confidence intervals can be inaccurate in terms of probability coverage. It is noted that these issues remain a concern for the situation at hand.

Consider, for example, the situation where $N = 10$, $(r, s) = (1, 2)$ for the first group, and $(r, s) = (10, 1)$ for the second group. These two probability functions are shown in Figure 2, with o corresponding to $(r, s) = (1, 2)$. So the first group is skewed to the right and the second is skewed to the left with the events $X = 0$ and 1 having probabilities close to

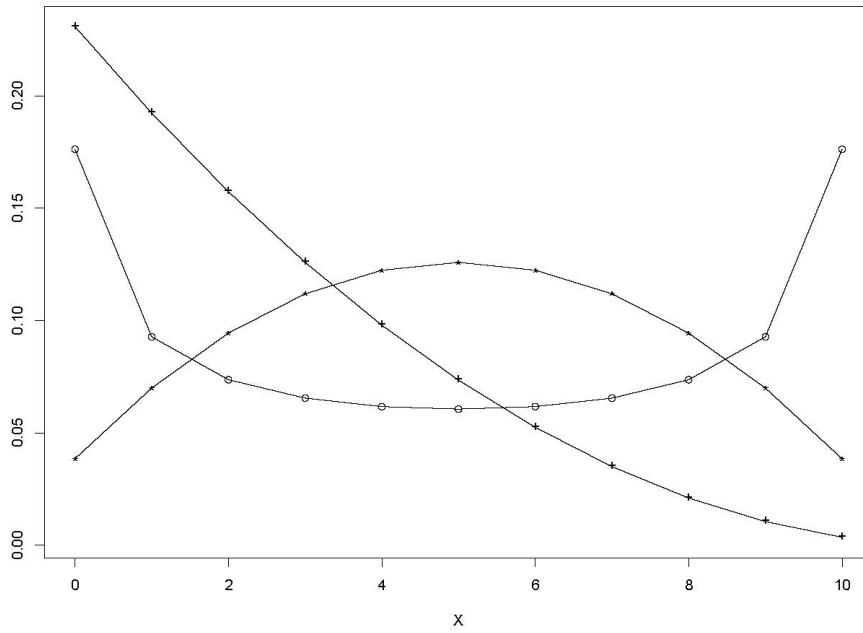


Figure 1: Beta-binomial probability functions used in the simulations

$N = 10$, $(r, s) = (0.5, 0.5)$ (indicated by lines connecting the points labeled o), $(r, s) = (2, 2)$ (indicated the points corresponding to +) and $(r, s) = (1, 3)$ (indicated by a *).

Source: Own work.

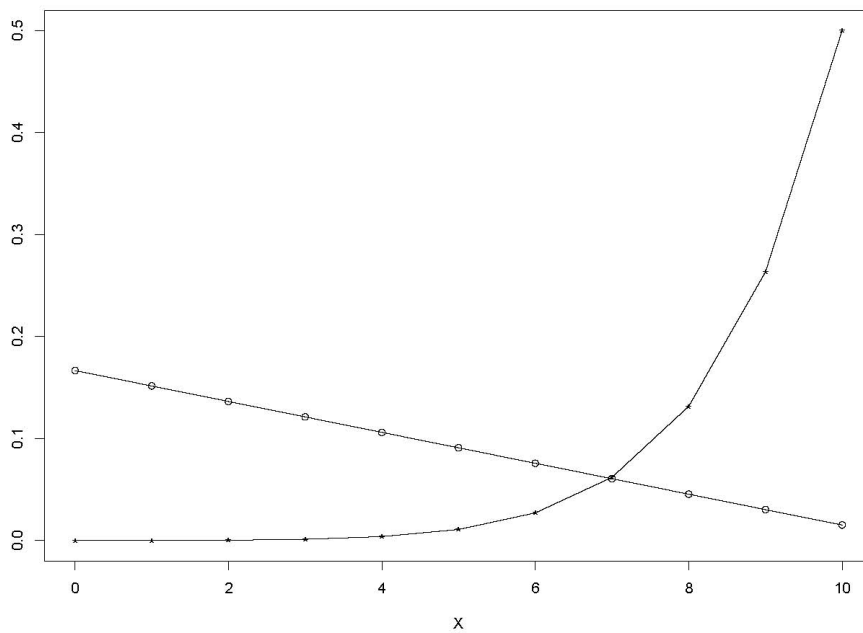


Figure 2: An example of two distributions where Student's t can be unsatisfactory

Source: Own work.

TABLE 2: Estimated power

N	n1	n2	Group 1		Group 2		Methods					
			r	s	r	s	B	C	M	Y	W	T
4	20	20	1	2	3	6	0.078	0.044	0.085	0.058	0.043	0.044
4	20	20	1	2	4	8	0.082	0.051	0.095	0.059	0.052	0.053
4	20	20	1	2	2	1	0.672	0.836	0.696	0.804	0.847	0.847
4	20	20	1	2	6	3	0.716	0.868	0.729	0.813	0.882	0.883
4	20	20	1	2	8	4	0.732	0.877	0.728	0.808	0.892	0.892
4	20	40	1	2	3	6	0.149	0.053	0.086	0.074	0.054	0.064
4	20	40	1	2	4	8	0.171	0.056	0.114	0.073	0.056	0.064
4	20	40	1	2	2	1	0.843	0.922	0.815	0.904	0.924	0.932
4	20	40	1	2	6	3	0.900	0.935	0.844	0.889	0.949	0.960
4	20	40	1	2	8	4	0.903	0.928	0.847	0.884	0.941	0.955
10	20	20	1	2	3	6	0.051	0.054	0.120	0.061	0.054	0.056
10	20	20	1	2	4	8	0.056	0.052	0.135	0.06	0.047	0.047
10	20	20	1	2	2	1	0.146	0.942	0.614	0.921	0.949	0.949
10	20	20	1	2	6	3	0.177	0.973	0.722	0.958	0.983	0.983
10	20	20	1	2	8	4	0.185	0.973	0.73	0.962	0.983	0.983
10	20	40	1	2	3	6	0.133	0.051	0.139	0.067	0.047	0.07
10	20	40	1	2	4	8	0.192	0.062	0.178	0.070	0.054	0.08
10	20	40	1	2	2	1	0.609	0.987	0.773	0.978	0.989	0.989
10	20	40	1	2	6	3	0.802	0.985	0.852	0.978	0.993	0.996
10	20	40	1	2	8	4	0.818	0.985	0.872	0.977	0.994	0.998
4	20	20	2	2	3	6	0.184	0.319	0.18	0.285	0.344	0.345
4	20	20	2	2	4	8	0.200	0.331	0.20	0.289	0.363	0.364
4	20	20	2	2	2	1	0.228	0.312	0.222	0.313	0.319	0.32
4	20	20	2	2	6	3	0.190	0.338	0.184	0.297	0.358	0.36
4	20	20	2	2	8	4	0.202	0.323	0.196	0.279	0.356	0.357
4	20	40	2	2	3	6	0.40	0.419	0.259	0.382	0.43	0.479
4	20	40	2	2	4	8	0.422	0.407	0.264	0.375	0.427	0.484
4	20	40	2	2	2	1	0.347	0.415	0.284	0.418	0.405	0.423
4	20	40	2	2	6	3	0.423	0.421	0.269	0.386	0.441	0.48
4	20	40	2	2	8	4	0.430	0.411	0.266	0.365	0.43	0.483
10	20	20	2	2	3	6	0.046	0.498	0.177	0.466	0.56	0.561
10	20	20	2	2	4	8	0.053	0.505	0.196	0.479	0.574	0.577
10	20	20	2	2	2	1	0.052	0.458	0.183	0.432	0.468	0.469
10	20	20	2	2	6	3	0.054	0.502	0.182	0.464	0.559	0.562
10	20	20	2	2	8	4	0.053	0.511	0.183	0.487	0.579	0.582
10	20	40	2	2	3	6	0.198	0.581	0.227	0.517	0.634	0.705
10	20	40	2	2	4	8	0.228	0.589	0.263	0.538	0.661	0.735
10	20	40	2	2	2	1	0.168	0.616	0.267	0.557	0.593	0.601

TABLE 3: The mean and variance of the distributions used in Table 2.

N	r	s	Mean	Variance	N	r	s	Mean	Variance
4	1	2	1.33	1.56	10	1	2	3.33	7.22
4	3	6	1.33	1.16	10	3	6	3.33	4.22
4	4	8	1.33	1.09	10	4	8	3.33	3.76
4	2	1	2.67	1.56	10	2	1	6.67	7.22
4	6	3	2.67	1.56	10	6	3	6.67	4.22
4	8	4	2.67	1.09	10	8	4	6.67	3.76
4	2	2	2.00	1.6	10	2	2	5.0	7.0

Source: Own work.

zero. If the first distribution is shifted to have a mean equal to the mean of the second group, and if $n_1 = n_2 = 20$, the actual Type I error probability is approximately 0.064 when testing at the 0.05 level with Student's t . But if $n_2 = 50$, the actual Type I error probability is approximately 0.12. In contrast, when using Welch's method, the Type I error probability is 0.063. Put another way, if the first groups is not shifted, the actual difference between the means is 1.8. But the actual probability coverage of the nominal 0.95 confidence interval based on Student's t is only 0.88 rather than 0.95, as intended. Increasing the sample sizes to 50 and 80, Student's t now has a Type I error probability of 0.085.

Numerous papers have demonstrated that with unequal sample sizes, violating the homoscedasticity assumption underlying Student's t test can result in poor control over the probability of a Type I error. Adding to this concern is the fact that under general conditions, Student's t uses the wrong standard error (Cressie & Whitford, 1986). It is noted that these concerns are relevant here. For example, if $N = 10$, $n_1 = 20$, $n_2 = 50$, $r = s = 0.5$ for the first group, and $r = s = 4$ for the second, in which case the distributions are symmetric with $\mu_1 = \mu_2 = 4.5$, the actual level of Student's t is approximately 0.112. Increasing the sample sizes to $n_1 = 50$ and $n_2 = 80$, the actual level is approximately 0.079. In contrast, when using method W, the actual levels for these two situations are 0.057 and 0.052, respectively.

As seems evident, the method that will have the most power depends on how the distributions

differ, simply because different methods are sensitive to different features of the distributions. Nevertheless, in terms of maximizing power, it is worth noting that the choice of method can make a substantial difference with each of the methods offering the most power in certain situations. This is illustrated in Table 2, which shows the estimated power when $(r, s) = (1, 2)$ or $(2, 2)$ for the first group and the second group is based on different choices for r and s , which are indicated in columns 6 and 7. The mean and variance associated with each of these distributions is shown in Table 3. Note that when $(r, s) = (1, 2)$ for group 1 and for group 2 $(r, s) = (3, 6)$, and $(4, 8)$, the population means are equal, so Welch's methods and Student's t are designed to have rejection rates equal to 0.05.

Also, the null hypothesis associated with Cliff's method is nearly true, the value of p being approximately 0.052. Now, with $N=4$, there is an instance where Student's t (method T) has an estimated rejection rate of 0.064, a bit higher than any situation where the distributions are identical. Methods B and M have a higher probability of rejecting, as would be expected, simply because the corresponding null hypothesis is false. Also, as expected, the method with the highest power is a function of how the distributions differ. Generally, method M performs well, but it is evident that it does not dominate. Note, for example, that when $(r, s) = (2, 2)$ for group 1, $N=4$ and the sample sizes are 20 and 40, method B can have a substantial power advantage compared to method M.

Also, there are instances where methods Y and W offer a substantial gain in power. This is the case, for example, when $N=10$ and when the first group has $r=s=2$.

Some Illustrations

A practical issue is whether the choice of method can make a difference in terms of the number of significant results when a relatively large number of tests are performed. Data from the Well Elderly 2 study (Jackson et al., 1997) are used to illustrate that this is indeed the case, with method M tending to report more significant results. The participants were 460 men and women aged 60 to 95 years (mean age 74.9) who were recruited from 21 sites in the greater Los Angeles area.

One portion of the study dealt with the processes older adults use to cope with stress, which was measured by the COPE scale developed by Carver, Scheier and Weintraub (1989). The scale consists of 36 items where each participant gave one of four responses: I usually didn't do this at all, I usually did this a little bit, I usually did this a medium amount, I usually did this a lot. Method M was applied to all 36 items with the goal of comparing males to females. (The frequencies associated with each item are available on the first author's web page in the files `cope_freq_males` and `cope_freq_females`.) Controlling for the probability of one or more Type I errors at the 0.05 level, among the 36 hypotheses tested, was done via Hochberg's improvement on Bonferroni method. Five significant differences were found. COPE contains four items related to a scale named "turning to religion," which constituted four of the five significant results. The fifth significant result was for the single scale item alcohol-drug disengagement.

Method B provides more detail about how the individual items differ. For the scale named turning to religion, the fourth possible response (I usually did this a lot) was the most significant. (The estimated difference between the probability a female responds 4, minus the same probability for males, ranged between 0.20 and 0.23.) If Cliff's method is used instead, three of the five items that are found

to be significant via method M are again significant, with no differences found otherwise. (All three significant results are associated with religion). Using method W gives the same result as method C.

Another portion of the Well Elderly study dealt with feelings of depression as measured by the Center for Epidemiologic Studies Depression Scale (CESD). The CESD (Radloff, 1977) consists of twenty items, each scored on a four-point scale. The CESD is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Foley et al., 2002; Lewinsohn et al., 1988). Here, we compare three ethnic groups: white, black/African American, and Hispanic or Latino. Applying method M to each of the twenty items, no difference is found between the first two groups. However, comparing group 1 to 3, as well as group 2 to group 3, again controlling for the probability of one or more Type I errors at the 0.05 level via the Hochberg's method, four significant differences are found (items 3, 4, 15 and 17). Using Cliff's method, three significant differences are found (items 4, 15 and 17), Welch's method and Yuen's method find two significant differences (items 4 and 17). Method B rejects for the same four items as method M and provides more detail about how the groups differ.

Comparing groups 2 and 3, again method M yields four significant results (items 4, 15, 17 and 18). Method C rejects in three instances (items 4, 15 and 18), while the Welch and Yuen methods reject twice (items 4 and 15).

As a final example, males and females are compared based on nine items aimed at measuring mental health and vitality. This time Cliff's methods C and Welch's method rejected for four of the items (namely, items 1, 2, 4 and 7), while methods M and Yuen returned three significant results (items 1, 2 and 7). The only point is that the choice of method can make a practical difference, as would be expected because each method is sensitive to different features of the data and designed to answer different questions. Limited results suggest that method M often has the most power, but the only certainty is that exceptions occur.

Concluding Remarks

In summary, methods were proposed for testing the hypothesis that two independent multinomial distributions have identical cell probabilities. All indications are that the probability of a Type I error can be controlled reasonably well. It is not suggested that these methods be used to the exclusion of all other techniques. As illustrated, the most powerful method depends on the nature of the distributions being compared. Also, if there is explicit interest in comparing measures of location, the Welch and Yuen methods are preferable. But simultaneously, methods based on measures of location can miss true differences that are detected by methods B and M. Also, method B has the potential of providing a more detailed understanding of how distributions differ, as was illustrated.

If Student's t is viewed as a method for testing the hypothesis that two distributions are identical, it appears to control the probability of a Type I error reasonably well. But when distributions differ, generally it seems to offer little or no advantage in terms of power. Also, consistent with past studies, it was found that Student's t can be unsatisfactory as a method for comparing means. Consequently, it is not recommended.

A criticism of using multiple methods to compare groups is that this increases the probability of one or more Type I errors. Again, one might use the Bonferroni method to deal with this issue or some recent improvement on the Bonferroni method, such as Hochberg's (1988) method. Yet another possibility is to control the false discovery rate using the technique derived by Benjamini and Hochberg (1995). But this comes at the cost of reducing the probability of detecting all true differences. Consequently, if one of the methods compared here tests a hypothesis that is of direct interest, an argument can be made to use it to the exclusion of all other techniques. In general, however, it seems prudent to keep in mind that the choice of method can make a practical difference in terms of detecting differences between two groups.

Given that method M controls the probability of a Type I error reasonably well, this suggests a

modification of method B. If method M rejects, compute p-values using method B and at a minimum, reject the hypothesis of equal cell probabilities corresponding to the situations having the two smallest p-values. That is, if the distributions differ, it must be the case that at least two cell probabilities differ. So even if the modified Hochberg method rejects for one cell probability only, at least one other pair of cell probabilities would be rejected as well.

Finally, R software for applying methods M and B are available from the first author upon request. The R functions `disc2com` and `binband` perform methods M and B, respectively. These functions are contained in the file `Rallfun-v20`, which can be downloaded at <http://college.usc.edu/labs/rwilcox/home>. (Use the source command to get access to these functions.) Alternatively, install the R package WRS with the R command `install.packages("WRS", repos="http://R-Forge.R-project.org")`. This requires the most recent version of R.

References

- Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, *25*, 591-602.
- Alba-Fernández, V., & Jiménez-Gamero, M. D. (2009). Bootstrapping divergence statistics for testing homogeneity in multinomial populations. *Mathematics and Computers in Simulation*, *79*, 3375-3384.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *B*, *57*, 289-300.
- Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology*, *56*, 267-283.
- Cliff, N. (1996). *Ordinal Methods for Behavioral Data Analysis*. Mahwah, NJ: Erlbaum.
- Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample t -test. *Biometrical Journal*, *28*, 131-148.

- Foley K., Reed P., Mutran E., et al., (2002). Measurement adequacy of the CES-D among a sample of older African Americans. *Psychiatric Research*, 109, 61-9.
- Guo, J. H., & Luh, W. M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, 49, 1-7.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800-802.
- Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., Chou, C.-P., Jordan-Marsh, M., Forman, T., White, B., Granger, D., Knight, B., & Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials*, 6 90-101.
- Lewinsohn, P.M., Hoberman, H. M., & Rosenbaum M., 1988. A prospective study of risk factors for unipolar depression. *Journal of Abnormal Psychology*, 97, 251-64.
- Liu, R. G., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92, 266-277.
- Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, 58, 409-429.
- Neuhäuser M., Lösch C., & Jöckel K-H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics & Data Analysis*, 51, 5055-5060.
- Radloff L., (1977). The CES-D scale: a self report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385-401.
- Roberson, P. K., Shema, S. J., Mundfrom, D. J., & Holmes, T. J. (1995). Analysis of paired Likert data: How to evaluate change and preference questions. *Family Medicine*, 27, 671-675.
- Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77, 663-666.
- Storer, B. E., & Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association*, 85, 146-155.
- Vargha A., & Delaney H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25, 101-132.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.
- Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8, 254-274.
- Wilcox, R. R. (2012a). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. New York: Chapman & Hall/CRC press
- Wilcox, R. R. (2012b). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd Ed. San Diego: Academic Press.

