The Equivalence of Two Proportions Revisited

A. MARTÍN ANDRÉS* & I. HERRANZ TEJEDOR**
*Bioestadística, Facultad de Medicina, Universidad de Granada, Spain, **Bioestadística, Facultad de Medicina, Universidad Complutense, Madrid, Spain

(Received September 2002; revised February 2003; accepted July 2003)

ABSTRACT The classic conditional test for checking that the difference between two independent proportions is not null may not be appropriate in many circumstances. Dunnett & Gent (1977) showed that in clinical trials, in studies of drugs, etc, the aim is to prove the practical equality (equivalence) of both proportions. On other occasions the aim may be the opposite: i.e. to prove that the two proportions are substantially different (biologically significant). Both cases are usually solved by two one-sided tests (TOST test). In this article, this procedure is shown to be conservative and two true two-sided tests for each case are proposed.

KEY WORDS: Confidence intervals, comparisons of two proportions, conditional test, equivalence test, Z test, $\chi^2$ test, $2 \times 2$ tables

Introduction

In the field of hypothesis tests, it is usual to contrast the null hypothesis that the two treatments are equal with the alternative one that they are different. Dunnett & Gent (1977) showed that in certain situations (studies of bioavailability, clinical trials, the evaluation of drugs or vaccines, etc), the aim of the experiment is to prove the equivalence of the two treatments (the new one and the classic one), which means having to restate the test. On other occasions (for example, when there are two competing treatments), the aim is to prove that both treatments are substantially different (biologically significant). Traditionally, both aims have been dealt with by a two-sided confidence interval for the difference (or for the ratio) between the average of the answers to each treatment, but today it is increasingly usual to approach the problem from the standpoint of the hypothesis tests. To this end (and with regard to the case of the difference $d$, that is the current aim), the researcher must propose a value $\Delta$ (positive) and contrast $H_{PE}$: $|d| \geq \Delta$ versus $K_{PE}$: $|d| < \Delta$ (in order to show that both treatments are practically equal; case PE), or contrast $H_{SD}$: $|d| \leq \Delta$ versus $K_{SD}$: $|d| > \Delta$ (in order to show that both treatments are substantially different; case SD). Generally, $d$ refers to the difference between two means ($\mu_2 - \mu_1$) or to the difference between two
proportions \((p_2 - p_1)\). In this paper, we shall focus on the second case, where \(p_i\) refers to the true proportions of successes.

The solutions of Dunnett & Gent (1977, 1988) and Johnson (1988) to the problem PE were found by using two one-sided tests (known today as the TOST of Schuirmann, 1987): \(H_{SG}: d \leq -\Delta \) versus \(K_{SG}: d > -\Delta \) and \(H_{SN}: d \geq +\Delta \) versus \(K_{SN}: d < +\Delta \). The first test is known as the non-inferiority test (case NI). The second test (case SS) is the opposite case to the substantial superiority test \(H_{SS}: d \leq +\Delta \) versus \(K_{SS}: d > +\Delta \) or case SS). All of these were illustrated by the authors mentioned above. In the following, it is advisable to generalize the above and define the following cases of \((generalized) \) superiority (SG) and of \((generalized) \) inferiority (IG):

\[
H_{SG}: d \leq \delta \text{ versus } K_{SG}: d > \delta \quad \text{and} \quad H_{IG}: d \geq \delta \text{ versus } K_{IG}: d < \delta
\]  

(1)

where \(\delta\) may be positive or negative. Figure 1 summarizes the various \(H\) and \(K\) hypotheses for the different acronyms used above (and others which appear below).

Since Dunnett & Gent first proposed the said tests, a variety of papers have been published about them (Miettinen & Nurminen, 1985; Mau, 1988; Farrington & Manning, 1990; Roeburgh & Kühn, 1995; Chan, 1998, 1999; Falk & Koch, 1998; Röhml & Mannsmann, 1999; Kang & Chen, 2000; Frick, 2000). Some authors develop the tests from the conditional point of view (the total successes of both treatments remain constant throughout the inference); others, from the unconditional point of view. However, very few authors deal with cases PE and SD. The usual values for \(\Delta\) are 0.10, 0.15 and 0.20 (Röhmel, 2001). The usefulness of these tests in industry and in evaluating agencies (conformity testing) can be seen in Holst et al. (2001). An analysis of cases PE and SD for \(\mu_2 - \mu_1\) can be found in Mehring (1993).

Dunnett & Gent approached the case PE from the conditional standpoint, but their solution was not very satisfactory. Johnson (1988) showed what the error was and Dunnett & Gent (1988) agreed with him (although with a proviso that will be shown below). The aim of this article is to prove that the solution of both is conservative, and to propose two new solutions, extending them to the case SD. At the same time the authors bring the different tests up to date in the light of other recently published articles.
The Three Basic Tests

Let \( X_i \sim B(n_i, p) \), where \( i = 1(2) \), and let \( x_i \) be the observed value of \( X_i \). D&G proposed three types of test for contrasting \( H_{SG} \) or \( H_{IG} \): the tests \( \chi^2 \) and \( Z \) (asymptotic) and Gart's test (semi-exact, since from the conditioned point of view an exact test does not exist). If \( H_{IG} : d = \delta \) refers to the two null hypotheses \( H_{SG} \) and \( H_{IG} \), considered only as an equality, then \( H_{IG} \) implies that \( p_2 = p_1 + \delta \), and the value \( p_1 = p \) will be a nuisance parameter (unknown) which must be evaluated. D&G proposed the estimator that is obtained by making the expected quantities add up to the observed: \( \hat{p} = (a_1 - n_2 \delta)/n \) where \( a_i = x_1 + x_2 \) and \( n = n_1 + n_2 \). Farrington & Manning (1990) observed that this estimation can furnish illicit values (for example: \( \hat{p} < 0 \)), and proposed using the estimator \( \hat{p} \) of maximum likelihood. The value \( \hat{p} \) is the solution of a cubic equation. This solution is given in F&M and Miettinen & Nurminen (1985).

From the perspective of a normal approximation, the base statistic is:

\[
Z = \frac{\hat{d} - \delta}{\left( \frac{p(1-p)}{n_1} + \frac{(p+\delta)(1-p-\delta)}{n_2} \right)^{1/2}}
\]

(2)

where \( \hat{d} = \hat{p}_2 - \hat{p}_1 \) and \( \hat{p}_i = x_i/n_i \). When \( p \) is substituted by \( \bar{p} \) (\( \hat{p} \)) the statistic \( Z_1 \) (Z2) is obtained.

From the perspective of chi-squared test, the base statistic is \( \chi^2 = \Sigma (O - E)^2/E \), where \( O = x_1, n_1 - x_1, x_2 \) and \( n_2 - x_2 \) the observed quantities, and \( E = n_1 p, n_1 (1-p), n_2 (p+\delta), n_2 (1-p-\delta) \) the expected quantities. Nam (1995) proved that \( Z_1^2 \) is equal to the statistic \( \chi^2 \) where \( p \) is substituted by \( \bar{p} \) (for this reason this last version is not considered in the following). When \( p \) is substituted by \( \hat{p} \) in the statistic \( \chi^2 \), D&G's statistic \( Z_3^2 \) is obtained. An alternative expression for this is:

\[
Z_3^2 = (n_1 \hat{p} - x_1)^2 \left\{ \frac{1}{n_1 \bar{p}(1-\bar{p})} + \frac{1}{n_2(\bar{p} + \delta)(1-\bar{p} - \delta)} \right\}
\]

(3)

D&G experimentally showed that \( Z_2^2 \neq Z_3^2 \).

In order to place \( Z_1, Z_2 \) and \( Z_3 \) statistic in a similar format, bear in mind that \( n_1 \hat{p} - x_1 = n_1 n_2 (\bar{d} - \delta)/n \); so, the three asymptotic statistics are:

\[
Z = \frac{\bar{d} - \delta}{s} \left\{ s = (\hat{p}(1-\hat{p})n_1 + (\bar{p} + \delta)(1-\bar{p} - \delta)n_2 \right\}^{1/2} \quad \text{for } Z_1
\]

\[
Z = \frac{\bar{d} - \delta}{s} \left\{ s = (\hat{p}(1-\hat{p})n_1 + (\bar{p} + \delta)(1-\bar{p} - \delta)n_2 \right\}^{1/2} \quad \text{for } Z_2
\]

\[
Z = \frac{\bar{d} - \delta}{s} \left\{ s = (n/n_1 n_2) \left\{ 1/[n_1 \hat{p}(1-\hat{p})] + 1/[n_2(\bar{p} + \delta)(1-\bar{p} - \delta)] \right\}^{-1/2} \quad \text{for } Z_3
\]

(4)

A continuity correction (c.c. in the following) can be performed on statistic \( Z \): one need only subtract from the numerator the quantity \( c = n/2 n_1 n_2 \) in the case SG (or to add it, in the case IG).
From the perspective of the method of Gart (1971), the conditional probability of the value \(i\) (under \(H_p\)) is:

\[
G(i) = \left( \begin{array}{c} n_1 \\ i \\ a_1 - i \\
\end{array} \right) \left( \begin{array}{c} n_2 \\ a_2 \end{array} \right) \theta_d^{i} / \sum_{h=r}^{s} \left( \begin{array}{c} n_1 \\ h \\ a_1 - h \\
\end{array} \right) \theta_h^{h} \tag{5}
\]

where \(r = \max\{0; a_1 - n_2\}\), \(s = \min\{a_1; n_3\}\) and \(\theta_h = p(1 - p - \delta)/(1 - p)(p + \delta)\) the value of the odds-ratio \(\theta = p_1(1 - p_2)/p_2(1 - p_1)\) under \(H_p\). As \(p\) is unknown, expression (5) cannot be applied. One approximate way of working with it is to substitute \(p\) by \(\hat{p}\), so giving rise to the method G1 (G2) proposed by D&G.

For the SG case, the \(p\)-value will be obtained in the classic way using the statistic \(Z\):

\[
P = P(z \geq Z),
\]

where \(z\) is a typical normal. For the method of Gart, because \(K_{SG}\) is equivalent to \(\theta < \theta_s\), the \(p\)-value will be \(P = \sum_{i=1}^{n} G(i)\). Similarly for the IG case. Roebruck & Kühn (1995) proved that the statistics based on \(\hat{p}\) are preferable to those based on \(\hat{\hat{p}}\).

**Classic Solution using Two One-sided Tests: TOST Method**

Johnson (1988) and Dunnett & Gent (1988) agreed that, to perform the tests for the case PE, it is necessary to perform the two tests implied in expression (1) for \(\delta = -\Delta\) and \(\delta = +\Delta\) respectively. When the two \(p\)-values \(-P_{SG}(-\Delta)\) and \(P_{IG}(+\Delta)\) are obtained, the \(p\)-value for the test PE is \(P_1 = \max\{P_{SG}(-\Delta), P_{IG}(+\Delta)\}\) for Johnson, but \(P_{DG} = P_{SG}(-\Delta) + P_{IG}(+\Delta)\) for D&G. In reality, this way of proceeding is now a classic in the field of (bio)equivalence tests, and is known as the TOST (two one-sided tests) test of Schuirmann (1987). D&G defend their solution on the grounds that it is the way to guarantee that the approach of the test is equivalent to the approach of the confidence intervals (for \(d\)). But this is not the case. As Munk & Pfüger (1999) indicated, if one wants to perform a hypothesis test to the error \(x\) using a confidence interval, the error \(z_{CI}\) of the interval depends on the convexity of the alternative hypothesis: \(z_{CI} = 2x(z)\) for the case PE (SD). Therefore, the correct solution is the one given by Johnson (see also Senn, 2001), as justified in the following paragraph.

Generally, the procedure for effecting test PE by a confidence interval (CI in the following) consists of calculating a two-sided \((1-2x)\)-CI for \(d = [\delta_L, \delta_U]\) and deciding \(K_{PE}\) (to the error \(x\)) when \(I \subseteq (-\Delta, +\Delta)\). In the case of SD, the interval \(I\) is a two-sided \((1-x)\)-CI and \(K_{SD}\) is decided (to the error \(x\)) when \(I \subseteq (-\Delta, +\Delta)\) = \(\phi\). The interval \(I\) to the error \(\gamma\) is usually obtained by inverting the two one-sided tests SG and IG to the error \(\gamma/2\) (Chan & Zhang, 1999), so that:

\[
\begin{cases}
\delta_L = \min\{\delta | P_{SG}(\delta) > \gamma/2\} \\
\delta_U = \max\{\delta | P_{IG}(\delta) > \gamma/2\}
\end{cases} \Rightarrow \delta_L \leq d \leq \delta_U \text{ (with error } \leq \gamma),
\]

yields a CI obtained by the TOST method. If this CI is used, the first significance for the test PE is obtained when \(\delta_L = -\Delta\) or when \(\delta_U = +\Delta\), and hence Johnson’s solution will be the correct one:

\[
P_{PE(TOST)} = \max\{P_{SG}(-\Delta), P_{IG}(+\Delta)\} \text{ if } |d| < \Delta \tag{7}
\]
Similarly, in case SD, the first significance is obtained when \( \delta_u = -\Delta \) (if \( \hat{d} < -\Delta \)) or when \( \delta_u = +\Delta \) (if \( \hat{d} > +\Delta \)) and hence the \( p \)-value will be:

\[
P_{SD(TOST)} = \begin{cases} 
2P_{SG} (+\Delta) & \text{if } \hat{d} > +\Delta \\
2P_{IG} (-\Delta) & \text{if } \hat{d} < -\Delta 
\end{cases} = 2 \times \min \{P_{SG} (+\Delta), P_{IG} (-\Delta)\} \quad \text{if } |\hat{d}| > \Delta
\]

It is usually said that expression (7) is valid even where \( |\hat{d}| \geq \Delta \), but, if this were so, the desired concordance between the conclusions of the test and the CI would be lost. In effect, if, for example, \( \hat{d} < -\Delta \), the CI to the error \( P_{PE(TOST)} \) is of the type \( \{d < \delta_l\} \cup \{d > \delta_u\} \), which leads one to conclude \( H_{PE} \) (and not \( K_{PE} \)). The downside of expression (7) is that it does not allow a \( p \)-value to be obtained when \( |\hat{d}| \geq \Delta \). In fact what happens is that the \( p \)-value in that case is unity and one must always conclude \( H_{PE} \). For the definition to be valid it is necessary for the \( p \)-value \( p_v \) to verify \( P(p_v \leq \alpha) \leq \alpha \) (Berger & Boos, 1994), and this is what happens. In effect, if \( \beta = \max_{\delta \leq \Delta} P_{\delta} (-\Delta < d < +\Delta) \) - where \( P_{\delta} \) is the probability under \( d = \delta \) -- then, when \( \alpha \leq \beta \) (\( \beta < x < 1 \)) the point is such that \( |\hat{d}| < \Delta(\hat{d}| \geq \Delta) \) and the condition is verified since the \( p \)-value is licit in that zone \( P(p \leq \alpha) = P(p_v < \delta) = \beta < \alpha \). Similarly for the case SD and for the other methods which follow.

To illustrate the most practical way of acting in the case PE, let us consider the data in D&G referring to whether the health care provided by nurse-practitioners (group 2) could be considered equivalent to conventional care (group 1). The values \( p_i \) refer here to a proportion of adequate care. The data were \( \hat{p}_1 = x_1/n_1 = 148/225 = 0.6578 \) and \( \hat{p}_2 = x_2/n_2 = 115/167 = 0.6886 \), where the aim was to prove the equivalence of the two proportions for \( \Delta = 0.10 \). Johnson’s solution for the Z3 without c.c. (which was the one he used) are obtained thus:

1. As \( \hat{d} = \hat{p}_2 - \hat{p}_1 = 0.0308 \), then \( |\hat{d}| < 0.1 \) and the test PE may be significant.
2. Under \( \delta = -0.1 \), \( \hat{p} = (263 + 167 \times 0.1)/392 = 0.7135 \) and \( s = 0.0481 \), and so \( Z3 = 2.719 \) and the \( p \)-value will be \( P_{SG} (-0.1) = P(z \geq 2.719) = 0.0033 \).
3. By proceeding in the same way under \( \delta = +0.1 \) one obtains \( \hat{p} = 0.6283 \), \( s = 0.0470 \), \( Z3 = -1.472 \) and \( P_{IG} (+0.1) = P(z \leq -1.472) = 0.0705 \).
4. Thus \( P_{PE(TOST)} = 7.05% \). D&G point out (correctly) that it is always advisable to perform a c.c. of \( c = 392/(2 \times 225 \times 167) = 0.0052 \), which results in \( P_{PE(TOST)} = 8.7% \).

When both the compared treatments are either new ones or classic ones, the researcher must ask him/herself if there is a biologically significant difference between them. If a difference \( |\hat{d}| \leq \Delta \) is understood that ‘it is not biologically significant’, the test to be performed will be SD for the said \( \Delta \). For example, Irwin (1935) who, together with Fisher (1935) and Yates (1934) proposed the well-known Fisher’s exact test, refers to the case of comparing the number of cases of measles prevented and not prevented by the use of convalescent serum in each of two different schools. The results were \( \hat{p}_1 = x_1/n_1 = 67 \text{ (prevented)}/69 = 0.9710 \) in School I and \( \hat{p}_2 = x_2/n_2 = 76/88 = 0.8636 \) in School II. The Fisher’s exact test gives a \( p \)-value with two sides of 0.0232: the proportions prevented in the two schools are significantly different. But are they biologically significant?
If a difference $|d| \leq 5\% = \Delta$ is not relevant, then it is advisable to modify the result using test SD. The steps now are:

1. As $\hat{d} = -0.1074$, then $|\hat{d}| > 0.05$ and the test SD may be significant.
2. As $\hat{d} < -0.05$, it is enough to calculate $P_{IG}(-\Delta)$ according to expression (8). For $\hat{d} = -0.05$, one obtains $\hat{p} = 0.9389$, $s = 0.0426$ and $P_{IG}(-0.05) = P(Z \leq (-0.1074 + 0.05 + 0.0118)/0.0426) = 0.1422$, since $c = 143/(2 \times 69 \times 88) = 0.0118$. Therefore $P_{SD(TOST)} = 28.4\%$.

### The Present Solution Using a True Two-sided Test: NEW Method

The well-known conservatism of the previous TOST method is due to the fact that a two-sided test (PE or SD) is carried out using two one-sided tests. In order to carry out a true two-sided test one must act as follows (and this will give the NEW method for this section).

When performing a hypothesis test to target error $\alpha$, the researcher must propose a critical region (CR in the following) and calculate the probability for it under the null hypothesis: this is the real error $x^* \leq \alpha$. As in the cases PE and SD the null hypothesis is double ($d = -\Delta$ or $d = +\Delta$), one must calculate the errors $x^*$ under each model $-x^*(-\Delta)$ and $x^*(+\Delta)$—and calculate the final error as $x^* = \max \{x^*(-\Delta), x^*(+\Delta)\}$. Under the conditional model, the only random variable of the problem is $X_1$, and the CR for the case PE (SD) will have to be in the form $x_1 \leq X_1 \leq x'_1$ ($\{X_1 \leq x'_1\} \cup \{X_1 \geq x_1\}$). For the results of the inference to be coherent, and concordant with expressions (7) and (8), it is also necessary that the values $d$ and $\hat{d}$ which yield $x_1$ and $x'_1$ verify $-\Delta < \hat{d} < -\Delta$, $\hat{d}(-\Delta)$ and $\hat{d} > +\Delta$ for PE (SD). Once the CR has been obtained, the value $x^*$ for each case will be:

$$x^*_{PE(NEW)} = \max_{\delta = -\Delta, +\Delta} P_{\delta}(x_1 \leq X_1 \leq x'_1)$$

$$x^*_{SD(NEW)} = \max_{\delta = -\Delta, +\Delta} \{P_{\delta}(X_1 \leq x'_1) + P_{\delta}(X_1 \geq x_1)\}$$

(9)

which defines the present NEW method.

Obtaining the CR depends on the statistic used ($Z$ or $G$) and the test to be carried out (PE or SD). For example, the statistic $Z$ in the case PE orders the values of $X_1$ (that are compatible with $-\Delta < \hat{d} < +\Delta$) from the largest to the smallest value of $|Z|$ at the point in question. As $H_{PE}$ refers to $d = -\Delta$ or $d = +\Delta$, then the order of entry into the CR is from the largest to the smallest value of $\min_{\delta = -\Delta, +\Delta} |Z_{\delta}(X_1)|$, where $Z_{\delta}(X_1)$ refers to the value of the expression (4) in $\delta$ and $X_1$. If the aim is to obtain the $p$-value of the experimental data, the value of $X_1$ for these will have to be $x_1$ or $x'_1$. Let us assume that it is $x_1$, which implies that $|Z_{+\Delta}(x_1)| = |Z_0|$, $|Z_{-\Delta}(x'_1)| = |Z_0|$. If this is not so, one need only permute the order of the two samples. Following from what has been shown above, $x'_1$ will be the licit value of $X_1$ furthest from $r$ which verifies the conditions $x'_1 \geq x_1$ (equivalent to $\hat{d} \leq d$) and $|Z_{+\Delta}(x'_1)| \geq |Z_0|$. In the case of the statistic $Z^2$, the licit values are those in which $\hat{d} = (n_1 a_1 - n x'_1)/n_1 n_2 > -\Delta$, so that now $x'_1$ will have to verify $|Z_{+\Delta}(x'_1)| \geq |Z_0|$ and $x_1 \leq x'_1 < E = n_1 a_1 + n_2 \Delta/n_1 n_2$ (where $\hat{p}_{+\Delta}$ is the value of $\hat{p}$ in $\delta = +\Delta$). Acting similarly for the order based on $G$, and
for the test SD, the following rule of behaviour is obtained: if the samples are ordered so that \( x_1 \) verifies:

\[
|Z_{+\delta}(x_i)| = |Z_0| \leq |Z_{-\delta}(x_i)| \text{ or } G_{+\delta}(x_i) = G_0 \geq G_{-\delta}(x_i)
\]  

(10)

then the value \( x'_1 \) for PE (SD) is the largest (smallest) value of \( X_1 \), where \( x_1 \leq X_1 < E \) (\( E < X_1 \leq s = \min\{a_1; n_1\} \)) and \( E = n_1(a_1 + n_2\Delta)/n \), in which:

\[
|Z_{-\delta}(x'_1)| \geq |Z_0| \text{ or } G_{-\delta}(x'_1) \leq G_0
\]  

(11)

When \( x_1 \) and \( x'_1 \) are known, expression (9) will yield the required \( p \)-value.

In order to obtain the probabilities \( P_\delta(\cdot) \) of expression (9), it is advisable to use the same statistic used to determine the value \( x'_1 \). For example, for the case PE and the statistic \( Z \), \( P_{+\delta}(x_1 \leq X_1 \leq x'_1) = P_{+\delta}(d \leq \hat{d} \leq \hat{d}) = P_{+\delta}(d \leq \hat{d}) - P_{+\delta}(d < \hat{d}) \). The first quantity is the \( p \)-value of the original table for test IG in \( \delta = +\Delta \), which implies that the c.c. has to be + \( c \). The second quantity is the \( p \)-value for the table \( x'_1 \) in the previous test, which implies that the c.c. has to be \(-c \) (because < rather than \( \leq \) appears). Moreover, this \( p \)-value will have to be calculated for the value \( \hat{p} \) that was obtained in the \( x_1 \) original table (in the case of using \( \hat{p} \) it makes no difference: it has the same value in both tables). By proceeding in the same way in the remaining cases the following rules are obtained:

\[
P_\delta(x_1 \leq X_1 \leq x'_1) = \begin{cases} 
F\left(\frac{\hat{d} - \delta + c}{s_\delta}\right) + F\left(\frac{\hat{d} - \delta - c}{s_\delta}\right) & \text{for } Z \\
\sum_{i=x_1}^{x'_1} G_\delta(i; x_1) & \text{for } G 
\end{cases}
\]

(12)

\[
P_\delta(X_1 \leq x_1) + P_\delta(X_1 \geq x'_1) = \begin{cases} 
F\left(\frac{\hat{d} + \delta + c}{s_\delta}\right) - F\left(\frac{\hat{d} - \delta + c}{s_\delta}\right) & \text{for } Z \\
\sum_{i=x_1}^{\min(a_1; n_1)} G_\delta(i; x_1) + \sum_{i=x'_1} G_\delta(i; x_1) & \text{for } G 
\end{cases}
\]

(13)

where \( F(\cdot) \) refers to the distribution function of a \( N(0,1) \) random variable, \( \hat{d} (\hat{d}) \) is the estimation of \( d = p_2 - p_1 \) in \( X_i = x_1 \) (\( X_i = x'_1 \)), \( x'_1 \) is obtained as in equation (11), \( c = n/2n_1 n_2 \) is the c.c. of Yates (if c.c. is not wanted, one need just make \( c = 0 \)), \( s_\delta \) is obtained as in equation (4) and \( G_\delta(i; x_1) \) as in equation (5) (the last two values with \( p \) calculated for \( \delta \) and for table \( x_1 \)). Two special cases may arise. In the case PE, one may get \( x'_1 = x_1 \) and so \( \hat{d} = d \). In the case SD, it is possible that \( x'_1 \) is not a licit value (\( x'_1 < r = \max\{0; a_1 - n_2\} \) or \( x'_1 > s = \min\{a_1; n_1\} \)) does not exist—and hence the second expression (13) for \( Z \) will be \( F\{(-d + \delta + c)/s_\delta\} \).

The procedure is especially useful when the statistics are \( Z_2 \) or \( Z_3 \). For the case PE, because this should be \( (\hat{d} + \Delta)/s_{-\Delta} \leq -Z_0 \) (since \( Z_0 \) is negative), with \( s_{-\Delta} \) the value of \( s \) for \( \delta = -\Delta \) (which is constant in \( X_1 \))
Tests PE and SD using the NEW method and the statistics Z2 and Z3. Given an experimental table of key values, x1, x2, a1 = x1 + x2, n1 and n2:

1. Check that [d|] < Α (|d| > Α), where  \( \hat{d} = \hat{p}_2 - \hat{p}_1 \) and  \( \hat{p}_1 = x_1/n_1 \), in the case PE (SD). If this is not so, accept  \( H_{PE} \) ( \( H_{SD} \)). Otherwise:
2. Calculate  \( \hat{\rho}_2(a_1 - n_2\delta)/n \) for  \( \delta = -\Delta \) and  \( \delta = +\Delta \), and, using expression (4), the values of  \( s_\delta \) and  \( Z_\delta \). Let  \( |Z_\delta| = \min_{\delta = -\Delta, +\Delta}|Z_\delta| \).
3. The value \( x_1' \) for the other side is determined as follows:

   \[
   x_1' = \left[ \frac{n_1 \{ a_1 + n_2 [s_{\mathrm{sgn}(Z_\delta)\Delta}Z_0 - \mathrm{sgn}(Z_\delta)\Delta]\}^{|\mathrm{sgn}(Z_\delta)|}}{n} \right] \]  

   for PE  

   \[
   x_1' = \left[ \frac{n_1 \{ a_1 + n_2 [s_{-\mathrm{sgn}(Z_\delta)\Delta}Z_0 + \mathrm{sgn}(Z_\delta)\Delta]\}^{|\mathrm{sgn}(Z_\delta)|}}{n} \right] \]  

   for SD

   where  \( \mathrm{sgn}(Z_\delta) = +(-) \) when  \( Z_0 > 0 \) ( \( Z_0 < 0 \)) and  \( [x]^+ \) (\( [x]^− \)) referring to the first integer larger than or equal (smaller than or equal) to  \( x \).

4. Let \( \hat{d} \) the value of  \( \hat{d} \) for  \( x_1' \) and  \( x_2' = a_1 - x_1' \), and let  \( \hat{d}_{\text{max}} \) and  \( \hat{d}_{\text{min}} \) be the maximum and the minimum for  \( d \) and  \( \hat{d} \). Then the  \( p \)-value  \( P_{\text{PE(NEW)}} \) or  \( P_{\text{SD(NEW)}} \) is  \( \max_{\delta = -\Delta, +\Delta} P_{\delta} \), where:

   \[
   P_{\delta} = F\left\{ \frac{\hat{d}_{\text{max}} - \delta + c}{s_\delta} \right\} - F\left\{ \frac{\hat{d}_{\text{min}} - \delta - c}{s_\delta} \right\} \]  

   for PE

   \[
   P_{\delta} = F\left\{ \frac{-\hat{d}_{\text{max}} - \delta + c}{s_\delta} \right\} + F\left\{ \frac{\hat{d}_{\text{min}} - \delta + c}{s_\delta} \right\} \]  

   for SD

   In the case SD, if  \( x_1' \) is an illicit value, that is if  \( x_1' < r = \max\{0; a_1 - n_2\} \) or  \( x_1' > s = \min\{a_1; n_1\} \), then  \( P_\delta = F\{(-|\hat{d} - \delta| + c)/s_\delta\} \).

   In particular, when  \( \Delta = 0 \) the test SD is the classic chi-squared test for comparing two proportions,  \( Z_1^2 = Z_2^2 = Z_3^2 \) and  \( x_1' \) is the value for the other side advised by Mantel (1974).

   In order to illustrate case PE, let us return to the data of D&G. In the following it is assumed that the statistic  \( Z_3 \) will be used (in order to be able to compare the present results with those of Johnson). The way to act is as follows:

   1. As  \( \hat{d} = 0.0308 \), then  \( |\hat{d}| < 0.1 \) and it is right to carry out the PE test.
   2. In the third section,  \( \hat{\rho}_{-0.1} = 0.7135 \),  \( \hat{\rho}_{+0.1} = 0.6283 \),  \( s_{-0.1} = 0.0481 \),  \( Z_{3-0.1} = 2.719 \),  \( s_{+0.1} = 0.0470 \) and  \( Z_{3+0.1} = -1.472 \) was obtained. With this,  \( Z_0 = -1.472 \) and  \( \mathrm{sgn}(Z_0) = - \).
(3) Using expression (14), \( x'_1 = \frac{[225(263 + 167(-0.0481 \times 1.472 + 0.1))]/392]}{\sqrt{153}} 

(4) In order to apply the first expression of (15) it must be remembered that 
\( \hat{d} = 0.0308 = d_{\text{max}} \) (in \( x_1 = 148 \), \( \hat{d} = 110/167 - 153/225 = -0.0213 = d_{\text{min}} \) (in \( x'_1 = 153 \)), \( c = 0.0052, s_{-0.1} = 0.0481 \) and \( s_{+0.1} = 0.0470 \). With this \( P_{0.05} = F(2.827) - F(1.528) = 6.1\% \), \( P_{+0.1} = F(-1.362) - F(-2.691) = 8.3\% \), and the \( p \)-value will be \( P_{\text{PE(NEW)}} = 8.3\% \).

The present \( p \)-value \( P_{\text{PE(NEW)}} = 8.3\% \) is smaller than the \( p \)-value of Johnson, and this in spite of the samples being quite large. The reason is that the \( p \)-value is \( P_{+0.1}(148 \leq x_1 \leq 153) \) here, while for Johnson it is \( P_{+0.1}(x_1 \geq 148) \).

This generally occurs for any example—the NEW test calculates \( P_{+0.1}(x_1 \leq x_1) \), for example, while the TOST test calculates \( P_{+0.1}(x_1 \geq x_1) \)—and so it is that the present test is always more powerful. This increase in power is particularly relevant when the samples are small.

In order to illustrate case SD, let us return to the data of Irwin. The way to act is as follows:

(1) As \( \hat{d} = -0.1074 \), then \( |\hat{d}| > 0.05 \) and it is right to carry out the SD test.

(2) In the third section it was already seen that \( \hat{p}_{-0.05} = 0.9389, s_{-0.05} = 0.0426 \), \( Z_{3-0.05} = -1.347, \hat{p}_{+0.05} = 0.8828, s_{+0.05} = 0.0456 \) and \( Z_{3+0.05} = -3.452 \), and therefore \( Z_0 = -1.347 \) and \( \text{sgn}(Z_0) = - \).

(3) Using expression (14), \( x'_1 = [69(143 + 88(-0.0456 \times 1.374 - 0.05))/157]^{-1/2} \).

(4) To apply the second expression of (15) it must be remembered that 
\( \hat{d} = -0.1074 = d_{\text{min}} \) (in \( x_1 = 67 \), \( \hat{d} = 85/88 - 58/69 = 0.1253 = d_{\text{max}} \) (in \( x'_1 = 58 \)) and \( c = 143/(2 \times 69 \times 88) = 0.0118 \). As a result \( P_{-0.05} = F(-3.838) + F(-1.070) = 14.2\% \), \( P_{+0.05} = F(-1.393) + F(-3.193) = 8.3\% \), and the \( p \)-value will be \( P_{\text{SD(NEW)}} = 14.2\% \): there is no evidence that \( |p_2 - p_1| > 5\% \).

Observe that again the \( p \)-value for the NEW method \( P_{\text{SD(NEW)}} = 14.2\% \) is quite a bit smaller than that of the TOST method \( P_{\text{SD(TOST)}} = 28.4\% \).

The Present Solution Using Two Two-sided Tests: TTST Method

In the third section we have already referred to the fact that a customary way of performing a test PE (SD) to the error \( x \) is by a CI to the error \( 2x \) (or \( x \)). This has the advantage of the conclusions of the test and the interval being compatible. But there are many ways of constructing a CI. If the \((1 - 2x)\)-CI is obtained by inverting two one-sided tests, as in expression (6), it yields the test PE(TOST) to the error \( x \). However, Agresti & Min (2001) proved that inverting the two-sided test SG2 \( H_{SG2}: d = \delta \) versus \( K_{SG2}: d \neq \delta \) yielded a licit CI

\[
\text{If } \delta_L(\delta_U) = \min(\max) \{ \delta | P_{SG2}(\delta) > \gamma \} \Rightarrow \delta_L \leq d \leq \delta_U \text{ (with error } \leq \gamma) \quad (16)
\]

which is generally narrower than that of the TOST method. As a consequence, if the CI is used to effect the PE test, the new procedure will produce the test
PE(TTST) which will generally be more powerful than the test PE(TOST). This is also valid for case SD, and thus, based on what has been indicated earlier:

$$P_{PE(TTST)} = \frac{1}{2} \max \{ P_{SG2}(-\Delta), P_{SG2}(+\Delta) \} \text{ if } |\hat{d}| < \Delta$$

$$P_{SD(TTST)} = \begin{cases} P_{SG2}(+\Delta) & \text{if } \hat{d} > +\Delta \\ P_{SG2}(-\Delta) & \text{if } \hat{d} < -\Delta \end{cases} = \max \{ P_{SG2}(-\Delta), P_{SG2}(+\Delta) \} \text{ if } |\hat{d}| > \Delta,$$

since in case SD the CI must be to the error $\alpha$ (and not $2\alpha$, as Mehring, 1993, thought).

The test SG2 was illustrated by D&G (although they—wrongly—believed they were performing the test PE). Unexpectedly, at the end the test SG2 is useful. In order to see this, let us suppose (without loss of generality) that $d > \hat{d}_\alpha$, where $\hat{d}_\alpha$, is the value $\hat{d}$ referred in the fourth section; therefore $P_{PE(NEW)} = P_N = P(d_\alpha \leq d \leq \hat{d}) = P(d \leq \hat{d}) - P(d < d_\alpha)$. If $\hat{d}_T$, is the value $\hat{d}$ obtained from (19), then $P_{PE(TTST)} = P_T = \frac{P(d \geq \hat{d}_T) + P(d < \hat{d})}{2}$, because in this case $P_{SG2}(+\Delta) > P_{SG2}(-\Delta)$.

For example, and for the data of D&G (test PE): (a) in $\delta = +0.1$ one obtains $x_1' = 134$ and $P_{SG2}(+0.1) = 15.02\%$; (b) in $\delta = -0.1$ one obtains $x_1' = 174$ and $P_{SG2}(-0.1) = 0.7\%$; (c) as a consequence, $P_{PE(TTST)} = 7.51\%$, a value which is even smaller to that of the NEW test in the fourth section. As the NEW method always performs better than the TOST method, which generally (although not always) performs worse than the TTST method, one can deduce that the methods TTST and NEW compete with each other. In order to carry out a comparative study on them, let us suppose (without loss of generality) that $d > \hat{d}_T$, where $\hat{d}_T$, is the value $\hat{d}$ referred in the fourth section; therefore $P_{PE(NEW)} = P_N = P(d_\alpha \leq d \leq \hat{d}_T) = P(d \leq \hat{d}_T) - P(d < d_\alpha)$. If $\hat{d}_T$, is the value $\hat{d}$ obtained from (19), then $P_{PE(TTST)} = P_T = \frac{P(d \geq \hat{d}_T) + P(d < \hat{d})}{2}$, because in this case $P_{SG2}(+\Delta) > P_{SG2}(-\Delta)$.

For Irwin’s data (test SD), as $d = -0.1074 < -0.05$, one only has to perform the test SG2 for $\delta = -0.05$. For this, $x_1' = 62$ and $P_{SD(TTST)} = P_{SG2}(-0.05) = 22.1\%$, a value that is larger than that of the NEW test in the fourth section. In fact, the NEW test is always more powerful than the TTST test. In order to see this, let us assume that $d > +\Delta$. If $\hat{d}_T$ and $\hat{d}_\alpha$ are the values $\hat{d}$ obtained from expressions (19) and (14) respectively, then $\hat{d}_\alpha < \hat{d}_T$. As $P_{SD(NEW)} = P_N = P(d \geq \hat{d}_T) + P(d \leq \hat{d}_\alpha)$ and $P_{SD(TTST)} = P_T = P(d \geq \hat{d}_T) + P(d < \hat{d}_\alpha)$, then generally $P_T > P_N$.

Conclusions

The classic test for comparing two proportions (H: $p_1 = p_2$ versus K: $p_1 \neq p_2$) is frequently used. However, in the field of health sciences it is frequently necessary
to prove that two proportions are practically the same—$K_{PE}(|p_2 - p_1| < \Delta)$—or that both are substantially different—$K_{SD}(|p_2 - p_1| > \Delta)$—where $\Delta$ is a previously given positive number. The traditional solution to both problems is to carry out two one-sided tests (Dunnett & Gent, 1988; Johnson, 1988) or the TOST method.

In this paper two new procedures are given: a true two-sided test (NEW method) and the TTST method (based on the CI of Agresti & Min, 2001, which is obtained when both two-sided tests are performed). On the other hand, the $p$-value for each test may be obtained in an approximate (statistics $Z_1$, $Z_2$ and $Z_3$) or semi-exact (statistics G1 and G2) fashion.

From what is known in published articles, the statistic G1 performs better than G2, and the statistic $Z_1$ performs better than $Z_2$ and $Z_3$, but $Z_2$ is the most widely used (as it is the most convenient). In all the statistics, $Z$, a correction for continuity $c = n/2n_1n_2$, should be performed. The article offers explicit expressions for calculating the $p$-value for the method $Z_2$ and for the NEW and TTST tests.

In the paper it is proved that the TOST method always (almost always) is less powerful than the NEW (TTST) method. It is also shown that the TTST method always (in large samples) is less powerful than the NEW method in the case of test SD (PE).

Finally, the above methods yield tests for proving that $-\Delta < d < +\Delta$ ($d < -\Delta$ or $d > +\Delta$) or cases PE (SD). These methods can immediately be extended to the case if one wants to show that $\delta_1 < d < \delta_2$ ($d < \delta_1$ or $d > \delta_2$), or cases PEG (SDG). For example $P_{PEG(TTST)} = \max \{P_{SG2}(\delta_1), P_{SG2}(\delta_2)\}/2$.

Acknowledgement

This research was supported by the Dirección General de Investigación, Spain, Grant BFM2000–1472.

References
