Test statistics for comparing two proportions with partially overlapping samples.

Abstract

Standard tests for comparing two sample proportions of a dichotomous dependent variable where there is a combination of paired and unpaired samples are considered. Four new tests are introduced and compared against standard tests and an alternative proposal by Choi and Stablein (1982). The Type I error robustness is considered for each of the test statistics. The results show that Type I error robust tests that make use of all the available data are more powerful than Type I error robust tests that do not. The Type I error robustness and the power among tests introduced in this paper using the phi correlation coefficient is comparable to that of Choi and Stablein (1982). The use of the test statistics to form confidence intervals is considered. A general recommendation of the best test statistic for practical use is made.

Key Words

Partially overlapping samples; Partially matched pairs; Partially correlated data; Equality of proportions.

1 Introduction

Tests for comparing two sample proportions of a dichotomous dependent variable with either two independent or two dependent samples are long established. Let π_1 and π_2 be the proportions of interest for two populations or distributions. The hypothesis being tested is $H_0 : \pi_1 = \pi_2$ against $H_1 : \pi_1 \neq \pi_2$. However, situations arise where a data set comprises a combination of both paired and unpaired observations. In these cases, within a sample there are, say a total of 'n_{12}' observations from both populations, a total of 'n_1' observations only from population one, and a total of 'n_2' observations only from population two. The hypothesis being tested is the same as when either two complete independent samples or two complete dependent samples are present. This situation with respect to comparing two means has been treated poorly in the literature (Martinez-Camblor et al, 2012). This situation with respect to comparing proportions has similarly been poorly treated.

Early literature in this area with respect to comparing proportions, refers to paired samples studies in the presence of incomplete data (Choi and Stablein, 1982; Ekbohlm, 1982), or missing data (Bhoj, 1978). These definitions have connotations suggesting that observations are missing only by accident. Recent literature for this scenario refers to partially matched pairs (Samawi and Vogel, 2011), however this terminology may be construed as the pairs themselves not being directly matched. Alternatively, the situation outlined can be referred to as part of the 'partially overlapping samples framework' (Martinez-Camblor et al, 2012). This terminology is more appropriate to cover scenarios where paired and independent samples may be present by accident or design. Illustrative scenarios where partially overlapping samples may arise by design include:

i) Where the samples are taken from two groups with some common element. For example, in education, when comparing the pass rate for two optional modules, where a student may take one or both modules.

ii) Where the samples are taken at two points in time. For example, an annual survey of employee satisfaction will include new employees that were not employed at time point one, employees that left after time point one and employees that remained in employment throughout.

iii) When some natural pairing occurs. For example, a survey taken comparing views of males and females, there may be some matched pairs ‘couples’ and some independent samples ‘single’.

Repeated measures designs can have compromised internal validity through familiarity (e.g. learning, memory or practise effects). Likewise, a matched design can have compromised internal validity through poor matching. However, if a dependent design can avoid extraneous systematic bias, then paired designs can be advantageous when contrasted with between subjects or independent designs. The advantages of paired designs arise by each pair acting as its own control helping to have a fair comparison. This allows differences or changes between the two samples to be directly examined (i.e. focusing directly on the phenomenon of interest). This has the result of removing systematic effects between pairs. This leads to increased power or a reduction in the sample size required to retain power compared with the alternative independent design. Accordingly, a method of analysis for partially overlapping samples that takes into account any pairing, but does not lose the unpaired information, would be beneficial.

Historically, when analysing partially overlapping samples, a practitioner will choose between discarding the paired observations or discarding the independent observations and proceeding to perform the corresponding ‘standard’ test. It is likely the decision will be based on the sample sizes of the independent and paired observations. Existing ‘standard’ approaches include:

Option 1: Discarding all paired observations and performing Pearson’s Chi square test of association on the unpaired data.

Option 2: Discarding all unpaired observations and performing McNemar’s test on the paired data.

> Option 3: Combining p-values of independent tests for paired and unpaired data. This can be done by applying Fisher's inverse Chi square method or Tippett's test. These approaches make use of all of the available data. These techniques were considered by Samawi and Vogel (2011) and are shown to be more powerful than techniques that discard data. However, it should be noted that the authors did not consider Type I error rates.

Other ad-hoc approaches for using all available data include randomly pairing any unpaired observations, or treating all observations as unpaired ignoring any pairing. These ad-hoc approaches are clearly incorrect practice and further emphasise the need for research into statistically valid approaches.

Choi and Stablein (1982) performed a small simulation study to consider standard approaches and ultimately recommended an alternative test making use of all the available data as the best practical approach. This alternative proposal uses one combined test statistic weighting the variance of the paired and independent samples, see Section 3.2 for definition. The authors additionally considered an approach using maximum likelihood estimators for the proportions. This approach was found to be of little practical benefit in terms of Type I error rate or power. Others have also considered maximum likelihood approaches. For example Thomson (1995) considered a similar procedure, using maximum likelihood estimators, and found the proposed procedure to perform similarly to that of Choi and Stablein (1982). It was noted by Choi and Stablein (1982) that given the additional computation, the maximum likelihood solution would not be a practical solution.

Tang and Tang (2004) proposed a test procedure which is a direct adaption of the best practical approach proposed by Choi and Stablein (1982). This adaption is found to be not Type I error robust in scenarios considered when $n_1 + n_2 + 2n_{12} = 20$. The test proposed by

Choi and Stablein (1982) is found to be Type I error robust in this scenario. The literature reviewed suggests that a solution to the partially overlapping samples case will have to outperform the best practical solution by Choi and Stablein (1982). Tang and Tang (2004, p.81) concluded that, ‘there may exist other test statistics which give better asymptotic or unconditional exact performance’.

In this paper, we introduce four test statistics for comparing the difference between two proportions with partially overlapping samples. These test statistics are formed so that no observations are discarded. The statistics represent the overall difference in proportions, divided by the combined standard error for the difference.

This paper will explore test statistics for testing H_0, in the presence of partially overlapping samples. In Section 2, existing ‘standard’ approaches and variants of are defined. In Section 3, our alternative proposals making use of all the available data are then introduced, followed by the most practical proposal of Choi and Stablein (1982).

In Section 4, a worked example applying all of the test statistics is given, followed by the simulation design in Section 5.

In Section 6.1, for all of the test statistics, the Type I error robustness is assessed when H_0 is true. This is measured using Bradley’s (1978) liberal criteria. This criteria states that the Type I error rate should be between $\alpha_{\text{nominal}} \pm 0.5\,\alpha_{\text{nominal}}$.

There is no standard criteria for quantifying when a statistical test can be deemed powerful. The objective is to maximise the power of the test subject to preserving the Type I error rate α_{nominal}. If Type I error rates are not equal it is not possible to correctly compare the power of tests. The preferred test where Type I error rates are not equal should be the one with the

Type I error rate closest to α_{nominal} (Penfield 1994). In Section 6.2, power will be considered under H_1 for the test statistics that meet Bradley's liberal criteria.

There is frequently too much focus on hypothesis testing. Confidence intervals may be of more practical interest (Gardner and Altman 1986). Confidence intervals allow insight into the estimation of a difference and the precision of the estimate. In Section 6.3, the coverage of the true difference under H_1 within 95% confidence intervals is considered. This is considered only for the most powerful test statistics that are Type I error robust.

2 Definition of standard test statistics

Assuming a dichotomous dependent variable, where a comparison in proportions between two samples is required, the layout of frequencies for the paired and the independent samples would be as per Table 1 and Table 2 respectively.

Table 1. Paired samples design for two samples and one dichotomous dependent variable.

		Response Sample 2	
Response Sample 1	Yes	No	Total
Yes	a	b	m
No	c	d	n_{12} - m
Total	k	n_{12} -k	n_{12}

Table 2. Independent samples design for two samples and one dichotomous dependent variable.

		Response	
	Yes	No	Total
Sample 1	e	f	n_1
Sample 2	g	h	n_2

2.1 Option 1: Discarding all paired observations

For two independent samples in terms of a dichotomous variable, as per Table 2, a Chi-square test of association is typically performed. This test will be displayed in standard textbooks in terms of χ_1^2. A chi square distribution on one degree of freedom is equivalent to the square of the z-distribution. Therefore under the null hypothesis an asymptotically N(0,1) equivalent statistic is defined as:

$$z_1 = \frac{\hat{\mathrm{p}}_1 - \hat{\mathrm{p}}_2}{\sqrt{\frac{\hat{\mathrm{p}}(1-\hat{\mathrm{p}})}{n_1} + \frac{\hat{\mathrm{p}}(1-\hat{\mathrm{p}})}{n_2}}} \text{ where } \hat{\mathrm{p}}_1 = \frac{e}{n_1},\ \hat{\mathrm{p}}_2 = \frac{g}{n_2} \text{ and } \hat{\mathrm{p}} = \frac{e+g}{n_1+n_2}.$$

For small samples, Yates's correction is often performed to reduce the error in approximation. Yate's correction is given by:

$$z_2 = \sqrt{\frac{(n_1+n_2)((|eh-fg|-0.5(n_1+n_2))^2}{(e+g)(f+h)n_1n_2}}.$$

The statistic z_2 is referenced against the upper tail of the standard normal distribution.

An alternative to the Chi square approach is Fisher's exact test. This is computationally more difficult. Furthermore, Fisher's exact test is shown to deviate from Type I error robustness (Berkson, 1978). Fisher's exact test will not be considered for the analysis of the partially overlapping samples design in this paper.

2.2 Option 2: Discarding all unpaired observations

For two dependent samples in terms of a dichotomous variable, as per Table 1, McNemar's test is typically performed. Under the null hypothesis, the asymptotically N(0,1) equivalent to McNemar's test is:

$$z_3 = \frac{b-c}{\sqrt{b+c}}.$$

When the number of discordant pairs is small, a continuity correction is often performed. McNemar's test with continuity correction is the equivalent to:

$$z_4 = \sqrt{\frac{(|b-c|-1)^2}{b+c}}.$$

The statistic z_4 is referenced against the upper tail of the standard normal distribution.

Test statistics based on Option 1 and Option 2 are likely to have relatively low power for small samples when the number of discarded observations is large. A method of analysis for partially overlapping samples that takes into account the paired design but does not lose the unpaired information could therefore be beneficial.

2.3 Option 3: Applying an appropriate combination of the independent and paired tests using all of the available data

Given that test statistics for the paired samples and dependent samples can be calculated independently, an extension to these techniques which makes use of all of the available data would be some combination of the two tests.

In terms of power, Fisher's test and Tippett's test are comparable to a weighted approach using sample size as the weights (Samawi and Vogel, 2011). Tippett's method and Fisher's method are not as effective as Stouffer's weighted z-score test (Kim et al, 2013). Stouffer's weighted z-score, for combining z_1 and z_3 is defined as:

$$\mathrm{z}_5 = \frac{w\mathrm{z}_1 + (1-w)\mathrm{z}_3}{\sqrt{w^2 + (1-w)^2}} \text{ where } w = \frac{n_1 + n_2}{2n_{12} + n_1 + n_2}.$$

Under the null hypothesis, the test statistic z_5 is asymptotically N(0,1).

Many other procedures for combining independent p-values are available, but these are less effective than Stouffer's test (Whitlock, 2005).

The drawbacks of Stouffer's test are that it has issues in the interpretation and confidence intervals for the true difference in population proportions cannot be easily formed.

3 Definition of alternative test statistics making use of all of the available data

The following proposals are designed to overcome the drawbacks identified of the standard tests. In these proposals observations are not discarded and the test statistics may be considered for the formation of confidence intervals.

3.1 Proposals using the phi correlation or the tetrachoric correlation coefficient.

It is proposed that a test statistic for comparing the difference in two proportions with two partially overlapping samples can be formed so that the overall estimated difference in proportions is divided by its combined standard error, i.e.

$$\frac{\bar{\mathrm{p}}_1 - \bar{\mathrm{p}}_2}{\sqrt{Var(\bar{\mathrm{p}}_1) + Var(\bar{\mathrm{p}}_2) - 2r_x Cov(\bar{\mathrm{p}}_1, \bar{\mathrm{p}}_2)}}$$

where $Var(\bar{\mathrm{p}}_1) = \frac{\bar{\mathrm{p}}_1(1-\bar{\mathrm{p}}_1)}{n_{12} + n_1}$, $Var(\bar{\mathrm{p}}_2) = \frac{\bar{\mathrm{p}}_2(1-\bar{\mathrm{p}}_2)}{n_{12} + n_2}$, $Cov(\bar{\mathrm{p}}_1, \bar{\mathrm{p}}_2) = \frac{\sqrt{\bar{\mathrm{p}}_1(1-\bar{\mathrm{p}}_1)}\sqrt{\bar{\mathrm{p}}_2(1-\bar{\mathrm{p}}_2)}n_{12}}{(n_{12} + n_1)(n_{12} + n_2)}$

and r_x is a correlation coefficient.

Test statistics constructed in this manner will facilitate the construction of confidence intervals, for example a 95% confidence interval θ would be equivalent to:

$$\theta = (\bar{\mathrm{p}}_1 - \bar{\mathrm{p}}_2) \pm 1.96 \times \sqrt{Var(\bar{\mathrm{p}}_1) + Var(\bar{\mathrm{p}}_2) - 2r_x Cov(\bar{\mathrm{p}}_1, \bar{\mathrm{p}}_2)}\,.$$

Pearson's phi correlation coefficient or Pearson's tetrachoric correlation coefficient are often used for measuring the correlation r_x between dichotomous variables.

Pearson's phi correlation coefficient is calculated as $r_1 = \dfrac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$.

The result of r_1 is numerically equivalent to Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient applied to Table 1, using binary outcomes '0' and '1' in the calculation. In this 2×2 case, r_1 is also numerically equivalent to Kendall's Tau-a and Kendall's Tau-b as well as Cramér's V and Somer's *d* (symmetrical). This suggests that r_1 would be an appropriate correlation coefficient to use.

Alternatively, assuming the underlying distribution is normal, a polychoric correlation coefficient may be considered. A special case of the polychoric correlation coefficient for two dichotomous samples is the tetrachoric correlation coefficient.

An approximation to the tetrachoric correlation coefficient as defined by Edwards and Edward (1984) is:

$$r_2 = \frac{s-1}{s+1} \text{ where } s = \left(\frac{ad}{bc}\right)^{0.7854}.$$

Other approximations are available, however there is no conclusive evidence which is the most appropriate (Digby, 1983). In any event, r_1 is likely to be more practical than r_2 because if any of the observed paired frequencies are equal to zero then the calculation of r_2 is not possible.

Constructing a test statistic using correlation coefficients r_1 and r_2 respectively, the following test statistics are proposed:

$$z_6 = \frac{\overline{\mathrm{p}}_1 - \overline{\mathrm{p}}_2}{\sqrt{\frac{\overline{\mathrm{p}}_1(1-\overline{\mathrm{p}}_1)}{n_{12}+n_1} + \frac{\overline{\mathrm{p}}_2(1-\overline{\mathrm{p}}_2)}{n_{12}+n_2} - 2r_1\left(\frac{\sqrt{\overline{\mathrm{p}}_1(1-\overline{\mathrm{p}}_1)}\sqrt{\overline{\mathrm{p}}_2(1-\overline{\mathrm{p}}_2)}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

$$z_7 = \frac{\overline{\mathrm{p}}_1 - \overline{\mathrm{p}}_2}{\sqrt{\frac{\overline{\mathrm{p}}_1(1-\overline{\mathrm{p}}_1)}{n_{12}+n_1} + \frac{\overline{\mathrm{p}}_2(1-\overline{\mathrm{p}}_2)}{n_{12}+n_2} - 2r_2\left(\frac{\sqrt{\overline{\mathrm{p}}_1(1-\overline{\mathrm{p}}_1)}\sqrt{\overline{\mathrm{p}}_2(1-\overline{\mathrm{p}}_2)}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

where: $\overline{\mathrm{p}}_1 = \frac{a+b+e}{n_{12}+n_1}$ and $\overline{\mathrm{p}}_2 = \frac{a+c+g}{n_{12}+n_2}$.

Under H_0, $\pi_1 = \pi_2 = \pi$, therefore two additional test statistics that may be considered are defined as:

$$z_8 = \frac{\overline{\mathrm{p}}_1 - \overline{\mathrm{p}}_2}{\sqrt{\frac{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}{n_{12}+n_1} + \frac{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}{n_{12}+n_2} - 2r_1\left(\frac{\sqrt{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}\sqrt{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

$$z_9 = \frac{\overline{\mathrm{p}}_1 - \overline{\mathrm{p}}_2}{\sqrt{\frac{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}{n_{12}+n_1} + \frac{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}{n_{12}+n_2} - 2r_2\left(\frac{\sqrt{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}\sqrt{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}n_{12}}{(n_{12}+n_1)(n_{12}+n_2)}\right)}}$$

where $\overline{\mathrm{p}} = \frac{(n_1+n_{12})\overline{\mathrm{p}}_1 + (n_2+n_{12})\overline{\mathrm{p}}_2}{2n_{12}+n_1+n_2}$.

The test statistics z_6, z_7, z_8 and z_9 are referenced against the standard normal distribution.

In the extreme scenario of $n_{12} = 0$, it is quickly verified that $z_8 = z_9 = z_1$. Under H_0, in the extreme scenario of $n_1 = n_2 = 0$, as $n_{12} \to \infty$ then $z_8 \to z_3$. This property is not observed for z_9. The properties of z_8 give support from a mathematical perspective as a valid test statistic to interpolate between the two established statistical tests where overlapping samples are not present.

3.2 Test statistic proposed by Choi and Stablein (1982).

Choi and Stablein (1982) proposed the following test statistic as the best practical solution for analysing partially overlapping sample:

$$z_{10} = \frac{\bar{\mathrm{p}}_1 - \bar{\mathrm{p}}_2}{\sqrt{\bar{\mathrm{p}}(1-\bar{\mathrm{p}})\left\{\frac{\psi_1^2}{n_1} + \frac{(1-\psi_1)^2}{n_{12}} + \frac{\psi_2^2}{n_2} + \frac{(1-\psi_2)^2}{n_{12}}\right\} - 2D}}$$

where $\psi_1 = \frac{n_1}{n_1 + n_{12}}$, $\psi_2 = \frac{n_2}{n_2 + n_{12}}$ and $D = \frac{(1-\psi_1)(1-\psi_2)(p_a - \bar{\mathrm{p}}^2)}{n_{12}}$.

The test statistic z_{10} is referenced against the standard normal distribution.

The authors additionally offer an extension of how optimization of w_1 and w_2 could be achieved, but suggest that the additional complication is unnecessary and the difference in results is negligible.

In common with the other statistics presented, z_{10} is computationally tractable but it may be less easy to interpret, particularly if $\psi_1 + \psi_2 \neq 1$.

4 Worked example

The objective of a Seasonal Affective Disorder (SAD) support group was to see if there is a difference in the quality of life for sufferers at two different times of the year. A binary response, 'Yes' or 'No' was required to the question whether they were satisfied with life. Membership of the group remains fairly stable, but there is some natural turnover of membership over time. Responses were obtained for $n_{12} = 15$ paired observations and a further $n_1 = 9$ and $n_2 = 6$ independent observations. The responses are given in Table 3.

Table 3. Responses to quality of life assessment.

		Response Time 2	
Response Time 1	Yes	No	Total
Yes	8	1	9
No	3	3	6
Total	11	4	15
		Response	
	Yes	No	Total
Time 1	5	4	9
Time 2	6	0	6

The elements of the test statistics (rounded to 3 decimal places for display purposes), are calculated as: $\hat{\mathrm{p}}_1 = 0.556$, $\hat{\mathrm{p}}_2 = 1.000$, $\hat{\mathrm{p}} = 0.733$, $\bar{\mathrm{p}}_1 = 0.583$, $\bar{\mathrm{p}}_2 = 0.810$, $\bar{\mathrm{p}} = 0.689$, $r_1 = 0.431$, $r_2 = 0.673$, $w = 0.333$, $\psi_1 = 0.375$, $\psi_2 = 0.286$, $D = 0.002$. The resulting test statistics are given in Table 4.

Table 4. Calculated value of test statistics (with corresponding p-values).

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
z-score	-1.907	1.311	-1.000	0.500	-1.747	-2.023	-2.295	-1.937	-2.202	-1.809
p-value	0.057	0.190	0.317	0.617	0.081	0.043	0.022	0.053	0.028	0.070

At the 5% significance level, whether H_0 is rejected depends on the test performed. It is of note that the significant differences arise only with tests introduced in this paper, z_6, z_7 and z_9.

Although the statistical conclusions differ for this particular example, the numeric difference between many of the tests is small. To consider further the situations where differences between the test statistics might arise, simulations are performed.

5 Simulation design

For the independent observations, a total of n_1 and n_2 unpaired standard normal deviates are generated. For the n_{12} paired observations, additional unpaired standard normal deviates X_{ij} are generated where *i* = (1,2) and *j* = (1,2,…., n_{12}). These are converted to correlated normal bivariates Y_{ij} so that:

$$Y_{1j} = \sqrt{\frac{1+\rho}{2}}X_{1j} + \sqrt{\frac{1-\rho}{2}}X_{2j} \text{ and } Y_{2j} = \sqrt{\frac{1+\rho}{2}}X_{2j} - \sqrt{\frac{1-\rho}{2}}X_{1j}$$

where ρ = correlation between population one and population two.

The normal deviates for both the unpaired and correlated paired observations are transformed into binary outcomes using critical values $C_{\pi i}$ of the normal distribution. If $X_{ij} < C_{\pi i}, Y_{ij} = 1$, otherwise $Y_{ij} = 0$

10,000 iterations of each scenario in Table 5 are performed in a 4×4×5×5×5×7=14000 factorial design.

Table 5. Values of parameters simulated for all test statistics.

Parameter	Values
π_1	0.15, 0.30, 0.45, 0.50
π_2	0.15, 0.30, 0.45, 0.50
n_1	10, 30, 50, 100, 500
n_2	10, 30, 50, 100, 500
n_{12}	10, 30, 50, 100, 500
ρ	-0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75

A range of values for n_1, n_2 and n_{12} likely to be encountered in practical applications are considered which offers an extension to the work done by Choi and Stablein (1982). Simulations are conducted over the range π from 0.15 to 0.5 both under H_0 and H_1. The values of π have been restricted to π <= 0.5 due to the proposed statistics being palindromic invariant with respect to π and $1-\pi$. Varying ρ is considered as it is known that ρ has an impact on paired samples tests. Negative ρ has been considered so as to provide a comprehensive overview and for theoretical interest, although ρ < 0 is less likely to occur in practical applications.

Two sided tests with $\alpha_{\text{nominal}} = 0.05$ is used in this study. For each combination of 10,000 iterations, the percentage of p-values below 0.05 is calculated to give the Type I error rate α. The Type I error rate under H_0, for each combination considered in the simulation design, should be between 0.025 and 0.075 to meet Bradley's liberal criteria and to be Type I error robust.

All simulations are performed in R.

6 Simulation Results

A comprehensive set of results with varying independent and paired sample sizes, correlation, and proportions was obtained as outlined in Section 5.

6.1 Type I error rates

Under H_0, 10,000 replicates were obtained for $4 \times 5 \times 5 \times 5 \times 7$=3500 scenarios. For assessment against Bradley's (1978) liberal criteria, Figure 1 shows the Type I error rates for all scenarios where $\pi_1 = \pi_2$ using $\alpha_{\text{nominal}} = 0.05$.

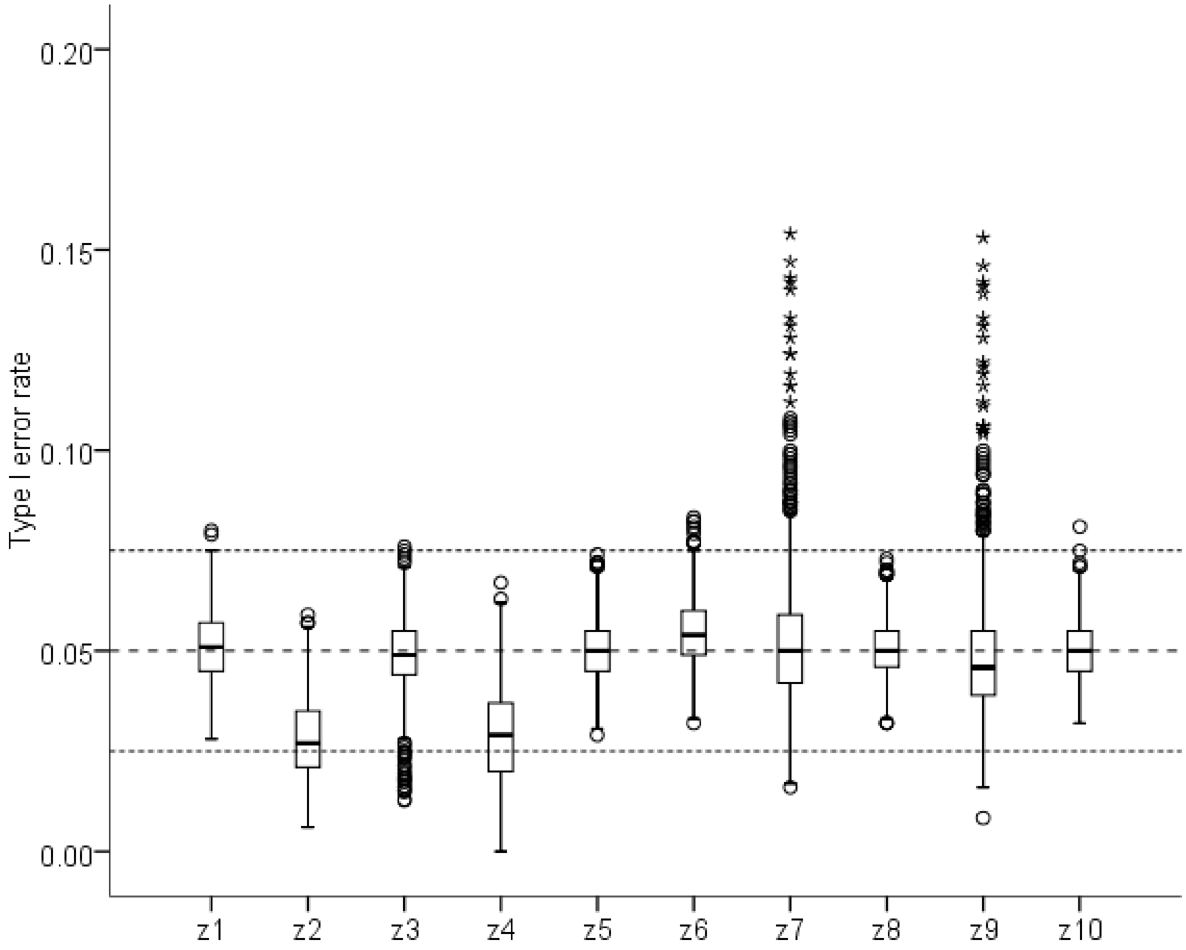

Figure 1: Type I error rates for each test statistic.

As may be anticipated, z_1 is Type I error robust because matched pairs are simply ignored. Similarly, z_3 performs as anticipated because the unpaired observations are ignored. Deviations from robustness for z_3 appear when n_{12} is small and ρ is large. Although deviations from stringent robustness are noted for z_3, this is not surprising since the cross product ratio is likely to be small when the proportion of success is low and the sample size is low. Crucially, the deviations from Type I error robustness of z_3 are conservative and will result in less false-positives, as such the tests statistic may not be considered unacceptable.

The corrected statistics, z_2 and z_4, generally give Type I error rates below the nominal alpha, particularly with small sample sizes. Ury and Fleiss (1980) found that z_1 is Type I error robust even with small samples, however applying Yate's correction is not Type I error robust and gives Type I error rates less than the nominal alpha. It is therefore concluded that z_2 and z_4 do not provide a Type I error robust solution.

The statistics using the phi correlation coefficient, z_6 and z_8, are generally liberal robust. For z_6 there is some deviation from the nominal Type I error rate. The deviations occur when min$\{n_1, n_2, n_{12}\}$ is small, max$\{n_1, n_2, n_{12}\}$ – min$\{n_1, n_2, n_{12}\}$ is large and $\rho < 0$. In these scenarios the effect of this is that z_6 is not liberal robust and results in a high likelihood of false-positives. It is therefore concluded that z_6 does not universally provide a Type I error robust solution to the partially overlapping samples situation.

The statistics using the tetrachoric correlation coefficient, z_7 and z_9, have more variability in Type I errors than the statistics that use the phi correlation coefficient. The statistics using the tetrachoric correlation coefficient inflate the Type I error when $\rho > 0.25$ and n_{12} is large. When min$\{n_1, n_2, n_{12}\}$ is small the test statistic is conservative. A test statistic that performs consistently would be favoured for practical use. It is therefore concluded that z_7 and z_9 do not provide a Type I error robust solution to the partially overlapping samples situation.

Three statistics making use of all of the available data, z_5, z_8 and z_{10}, demonstrate liberal robustness across all scenarios. Analysis of Type I error rates show near identical boxplots to Figure 1 when each of the parameters are considered separately. This means these statistics are Type I error robust across all combinations of sample sizes and correlation considered.

6.2 Power

The test statistics z_2, z_4, z_6, z_7 and z_9 are not Type I error robust. Therefore only z_1, z_3, z_5, z_8 and z_{10} are considered for their power properties (where H_1 is true). Table 6 summarises the power properties where $\pi_1 = 0.5$.

Table 6. Power averaged over all sample sizes.

π_1	π_2	ρ	z_1	z_3	z_5	z_8	z_{10}
0.5	0.45	>0	0.095	0.173	0.208	0.221	0.221
		0		0.133	0.168	0.186	0.186
		<0		0.112	0.150	0.166	0.166
0.5	0.3	>0	0.509	0.653	0.807	0.856	0.855
		0		0.569	0.772	0.828	0.827
		<0		0.508	0.746	0.801	0.801
0.5	0.15	>0	0.843	0.874	0.975	0.989	0.989
		0		0.834	0.970	0.985	0.986
		<0		0.795	0.966	0.980	0.982

For each of the test statistics, as the correlation increases from -0.75 through to 0.75 the power of the tests increase. Similarly, as sample sizes increase the power of the test increases.

Clearly, z_5 is more powerful than the other standard tests z_1 and z_3, but it is not as powerful as the alternative methods that make use of all the available data.

The power of z_8 and z_{10} are comparable. Separate comparisons of z_8 and z_{10} indicates that the two statistics are comparable across the factorial combinations in the simulation design. Either test statistic could reasonably be used for hypothesis testing in the partially overlapping samples case.

6.3 Confidence interval coverage

For z_8 and z_{10}, the coverage of the true difference of population proportions within 95% confidence intervals has been calculated as per the simulation design in Table 5 where $\pi_1 \neq \pi_2$. The results are summarised in Figure 2.

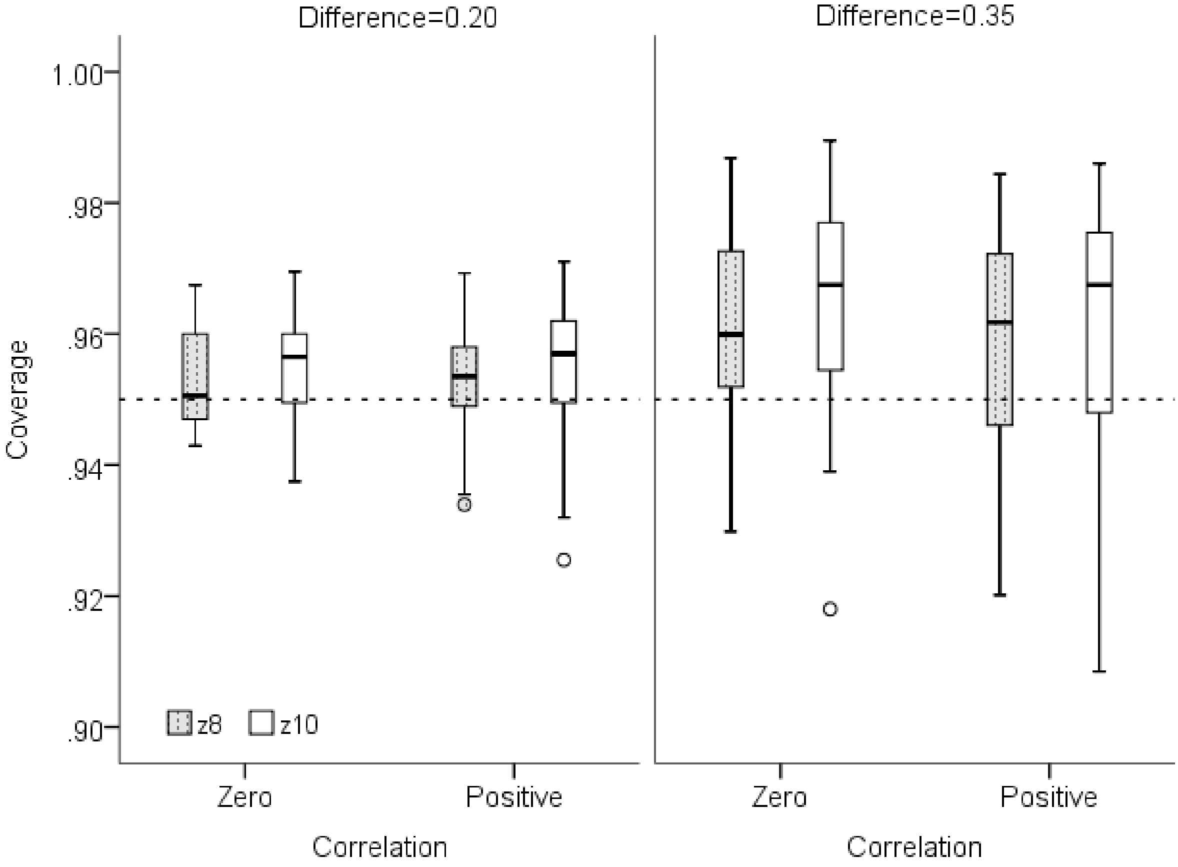

Figure 2: Percentage of iterations where the true difference is within the confidence interval.

Both z_8 and z_{10} demonstrate reasonable coverage of the true population difference $\pi_1 - \pi_2$. However, Figure 2 shows that z_8 more frequently performs closer to the desired 95% success rate. Taking this result into account, when the objective is to form a confidence interval, z_8 is recommended as the test statistic of choice in the partially overlapping samples case.

7 Conclusion

Partially overlapping samples may occur by accident or design. Standard approaches for analysing the difference in proportions for a dichotomous variable with partially overlapping samples often discard some available data. If there is a large paired sample or a large unpaired sample, it may be reasonable in a practical environment to use the corresponding standard test. For small samples, the test statistics which discard data have inferior power properties to tests statistics that make use of all the available data. These standard approaches and other ad-hoc approaches identified in this paper are less than desirable.

Combining the paired and independent samples z-scores using Stouffer's method is a more powerful standard approach, but leads to complications in interpretation, and does not readily extend to the creation of confidence intervals for differences in proportions. The tests introduced in this paper, as well as the test outlined by Choi and Stablein (1982) are more powerful than the test statistics in 'standard' use.

The alternative tests introduced in this paper, z_6, z_7, z_8 and z_9, overcome the interpretation barrier, in addition confidence intervals can readily be formed.

Tests introduced using the phi correlation coefficient, z_6 and z_8, are more robust than the equivalent tests introduced using the tetrachoric correlation coefficient, z_7 and z_9.

The most powerful tests that are Type I error robust are z_8 and z_{10}. The empirical evidence suggests that z_8 is better suited for forming confidence intervals for the true population difference than z_{10}. Additionally, z_8 has relative simplicity in calculation, strong mathematical properties and provides ease of interpretation. In conclusion, z_8 is recommended as the best practical solution to the partially overlapping samples framework when comparing two proportions.

References

Berkson J. In dispraise of the exact test. Journal of Statistic Planning and Inference. 1978;2:27–42.

Bhoj D. Testing equality of means of correlated variates with missing observations on both responses. Biometrika. 1978;65:225-228.

Bradley JV. Robustness?. British Journal of Mathematical and Statistical Psychology. 1978;31(2):144-152.

Choi SC, Stablein DM. Practical tests for comparing two proportions with incomplete data. Applied Statistics. 1982;31:256-262.

Digby PG. Approximating the tetrachoric correlation coefficient. Biometrics. 1983;753-757.

Edwards JH, Edwards AWF. Approximating the tetrachoric correlation coefficient. Biometrics. 1984;40(2):563.

Ekbohm G. On testing the equality of proportions in the paired case with incomplete data. Psychometrika. 1982;47(1):115-118.

Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypothesis testing. BMJ. 1986;292(6522):746-750.

Kim SC, Lee SJ, Lee WJ, Yum YN, Kim JH, Sohn S, Park JH, Jeongmi L, Johan Lim, Kwon SW. Stouffer's test in a large scale simultaneous hypothesis testing. Plos one. 2013;8(5):e63290.

Martinez-Camblor P, Corral N, De la Hera JM. Hypothesis test for paired samples in the presence of missing data. Journal of Applied Statistics. 2012;40(1):76-87.

Penfield DA. Choosing a two-sample location test. Journal of Experimental Education. 1994;62(4):343-360.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. 2014; version 3.1.2.

Samawi HM, Vogel R. Tests of homogeneity for partially matched-pairs data. Statistical Methodology. 2011;8(3):304-313.

Tang ML, Tang NS. Exact tests for comparing two paired proportions with incomplete data. Biometrical Journal. 2004;46(1),72-82.

Thomson PC. A hybrid paired and unpaired analysis for the comparison of proportions. Statistics in Medicine. 1995;14:1463-1470.

Ury HK, Fleiss JL. On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. Biometrics. 1980;347-351.

Whitlock MC. Combining probability from independent tests: the weighted z-method is superior to Fisher's approach. Journal of Evolutionary Biology. 2005;18(5):1368-1373.