Comparing proportions in overlapping samples

An unpublished paper by

J. Martin Bland
Professor of Medical Statistics
Department of Public Health Sciences,
St George's Hospital Medical School,
London SW17 0RE

Barbara K. Butland
Lecturer in Medical Statistics
Department of Public Health Sciences,
St George's Hospital Medical School,
London SW17 0RE

SUMMARY

Sometimes we want to estimate the difference in proportions between two groups where some subjects appear in both. We present an approach which does not require the assumption that the proportions in the overlapping and non-overlapping samples are the same, and which can be extended very easily to comparisons of means, odds, etc. The method has the disadvantage that we need each group to contain some subjects observed once only. We illustrate the method with an example from the UK National Child Development Study, and compare the results with other methods.

INTRODUCTION

Sometimes we want to compare two groups where some subjects appear in both. For example, we might follow a group of people over time, observing them at several points. People being what they are, we may not be able to observe everyone at each time. We may want to compare the subjects at two time points. Some subjects were observed at both times, some at the first time only, and some at the second only. The first and second samples, corresponding to the first and second time points, overlap.

For example, in the National Child Development Study,[1] children were studied at several ages, including 11 and 16. At these ages, parents were asked whether the child had experienced attacks of asthma or wheezy bronchitis in the past 12 months.[2] There were 9742 children with information on both occasions, a further 3952 children with information on the first occasion only and 1790 children with information on the second occasion only. We want to estimate the change in prevalence of reported disease.

There are several possible approaches to this problem. One described recently by Thomson [3] provides a confidence interval which is easy to calculate, but involves the assumption that the proportion in the overlapping and non-overlapping samples are the same. Here we present an alternative approach, which does not require this strong assumption, and can be extended very easily to comparisons of means, odds, etc. It has the disadvantage that we need each sample to contain some subjects observed once only.

THE PROPOSED METHOD

We can separate our subjects into two distinct comparisons: the paired comparison using subjects observed on both occasions, and the unpaired comparison of those seen on the first occasion only with those seen on the second occasion only.

We shall assume that the difference we wish to estimate is the same in the paired sample and between the two unpaired samples. However, we do not require the proportions themselves to be the same in the paired and unpaired samples. The method of Thomson [3] requires this because it forms combined estimates of the proportions for the first sample and for the second sample.

Methods for estimating the separate differences and their standard errors are familiar to most medical researchers: the McNemar test for paired data and the large sample comparison of two proportions for unpaired data. It is straightforward to get estimates for the paired difference, $\hat{d}_p$, and of the unpaired difference, $\hat{d}_u$, and their sampling variances.

Table 1. Paired and unpaired data in symbolic form

Paired sample			
First Sample	Second sample		Total at second
	Yes	No	
Yes	n_{11}	n_{10}	$n_{11} + n_{10}$
No	n_{01}	n_{00}	$n_{01} + n_{00}$
Total at first	$n_{11} + n_{01}$	$n_{10} + n_{00}$	n

Unpaired samples			
Sample	Yes	No	Total
First	n_x	$k - n_x$	k
Second	n_y	$m - n_y$	m

Using the notation of Table 1 we have

$$\hat{d}_p = \frac{n_{10}}{n} - \frac{n_{01}}{n} = \frac{n_{10} - n_{01}}{n}$$

$$\mathrm{Var}(\hat{d}_p) = \frac{n_{10} + n_{01}}{n^2} - \frac{(n_{10} - n_{01})^2}{n^3}$$

$$\hat{d}_u = \frac{n_x}{k} - \frac{n_y}{m}$$

$$\mathrm{Var}(\hat{d}_u) = \frac{n_x(k - n_x)}{k^3} - \frac{n_y(m - n_y)}{m^3}$$

To obtain the estimate $\hat{d}_u$ we must have both k and m greater than zero.

We can combine these two estimates by a weighted average. We find suitable weights, w_p and w_u, then get the combined estimate by

$$\hat{d} = \frac{w_p \hat{d}_p + w_u \hat{d}_u}{w_p + w_u}$$

We can find weights which make the variance of the weighted estimate a minimum using the inverses of the individual variances, $w_p = 1/\mathrm{Var}(\hat{d}_p)$ and $w_u = 1/\mathrm{Var}(\hat{d}_u)$.[4]

The variance of $\hat{d}$ is then

$$\mathrm{Var}(\hat{d}) = \frac{1}{\dfrac{1}{\mathrm{Var}(\hat{d}_p)} + \dfrac{1}{\mathrm{Var}(\hat{d}_u)}} = \frac{1}{w_p + w_u}$$

Using these weights ensures that Var($\hat{d}$) is always less than both Var($\hat{d}_p$) and Var($\hat{d}_u$), so it is better to use both paired and unpaired data than it is to drop either the paired or the unpaired samples.

In contrast, the method of Thomson [3] weights the paired and unpaired data by the number of subjects.

This general approach can be used for estimation of the difference between proportions, log odds, means, indeed anything for which a standard error can be calculated. If the proportions in the unpaired samples are clearly different from those in the paired sample, log odds may be preferred for dichotomous data.

It is straightforward to check the assumption that the differences in the paired and unpaired samples are the same. We have estimated separately the differences in proportion for the paired and unpaired samples. These are assumed to estimate the same quantity and so should be similar. We can test this formally, though as always non-significant results should be treated with caution. For the test, instead of the weighted average we find the difference $\hat{d}_p$ – $\hat{d}_u$, which has variance Var($\hat{d}_p$) + Var($\hat{d}_u$). Hence we can test the null hypothesis that the difference in the population is zero by

$$z = \frac{\hat{d}_p - \hat{d}_u}{\sqrt{\mathrm{Var}(\hat{d}_p) + \mathrm{Var}(\hat{d}_u)}}$$

which, under the usual large sample assumptions, would follow a Standard Normal distribution if the null hypothesis were true.

If we want a test of significance for $\hat{d}$, rather than a confidence interval, we can simply divide the estimated difference by its standard error and refer to the Standard Normal distribution. However, this does not take into account the presence in the paired and unpaired variances of the two proportions, which are the same under the null hypothesis. We can replace Var($\hat{d}_p$) and Var($\hat{d}_u$) by

$$\mathrm{Var}_{\mathrm{null}}(\hat{d}_p) = \frac{n_{10} + n_{01}}{n^2}$$

$$\mathrm{Var}_{\mathrm{null}}(\hat{d}_u) = \frac{(n_x + n_y)(k + m - n_x - n_y)}{(k+m)^2}\left(\frac{1}{k} + \frac{1}{m}\right)$$

APPLICATION TO THE EXAMPLE

For the National Child Development Study data, Table 2 shows the reported asthma or wheezy bronchitis for the paired sample and Table 3 shows data for the unpaired sample. The different structures of Tables 2 and 3 reflect the different structure of the paired and unpaired samples.

Table 2. Reported asthma or wheezy bronchitis for children with information on both occasions

Asthma/wheezy bronchitis at age 11	Asthma/wheezy bronchitis at age 16		
	Yes	No	Total at age 11
Yes	151	298	449 (4.74%)
No	203	8820	9023
Total at age 16	354 (3.74%)	9118	9472

Table 3. Reported asthma or wheezy bronchitis for children with information on one occasion only

Age at report	Asthma/wheezy bronchitis		
	Yes	No	Total
11	215 (5.44%)	3737	3952
16	73 (4.08%)	1717	9023

The proportions reporting asthma or wheeze are larger in the unpaired than in the paired samples at age 11 and at age 16, although the difference is not significant (P=0.08, Mantel Haenszel test). This difference could reflect a ‘healthy respondent’ effect, where people with problems are less likely to respond to research requests. Thus a method which does not assume that the proportions in the paired and unpaired samples are the same is desirable for these data.

For the paired data, the difference in proportions (age 11 minus age 16) was $\hat{d}_p$ = 0.0474 – 0.0374 = 0.0100, with standard error se_p = 0.00236. For the unpaired data, the difference was $\hat{d}_u$ = 0.0544 – 0.0408 = 0.0136, with standard error se_u = 0.00591. The weights are $w_p = 1/0.00236^2 = 179546$ and $w_u = 1/0.00591^2 = 28630$. The paired sample carries more weight than the unpaired, partly because it is larger but also because the paired data give us a more precise estimate of the difference. The weighted estimate is thus

$$\hat{d} = \frac{179546\times 0.0100 + 28630\times 0.0136}{179546 + 28630} = 0.0105$$

with variance

$$\mathrm{Var}(\hat{d}) = \frac{1}{179546 + 28630} = 0.0000048212$$

and standard error √0.0000048212 = 0.00219. Hence the 95% confidence interval for the difference is 0.0105 ± 1.96×0.00219 which gives 0.0062 to 0.0148. Thus we estimate that the prevalence of reported asthma or wheezy bronchitis in the past 12 months was between 0.6 and 1.5 percentage points lower at age 16 than at age 11.

To check the assumption that the differences in the paired and unpaired samples are the same, we have $\hat{d}_p - \hat{d}_u$ = 0.0100 – 0.0136 = –0.0036, fairly small compared to the magnitude of the difference between the first and second samples. For the test of significance,

$$z = \frac{-0.0036}{\sqrt{0.00236^2 + 0.00591^2}} = -0.57$$

which has P = 0.6. Hence there is no evidence that the change in prevalence differs between the paired and unpaired samples.

For the test of the null hypothesis that there is no difference in proportion between the first and second samples, we have

$$\mathrm{Var}_{\mathrm{null}}(\hat{d}_p)=\frac{298+203}{9472^2}=0.0000055841$$

$$\mathrm{Var}_{\mathrm{null}}(\hat{d}_u)=\frac{(215+73)(3952+1790-215-73)}{(3952+1790)^2}\left(\frac{1}{3952}+\frac{1}{1790}\right)=0.000038670$$

The new weights are the inverse of these, 1/0.0000055841 = 179080 and 1/0.000038670 = 25860. The weighted estimate using these null variances is

$$\hat{d}_{\mathrm{null}}=\frac{179080\times 0.0100+25860\times 0.0136}{179080+25860}=0.0105$$

with variance

$$\mathrm{Var}_{\mathrm{null}}(\hat{d}_{\mathrm{null}})=\frac{1}{79080+25860}=0.0000048795$$

and standard error √0.0000048795 = 0.0022090. The Standard Normal deviate is thus

$$z=\frac{0.0105}{0.0022090}=4.75$$

which is highly significant, P<0.0001.

COMPARISON WITH OTHER METHODS

Thomson [3] estimated the difference between the usual point estimates of the proportions, ignoring the pairing:

$$\hat{p}_x=\frac{n_x+n_{10}+n_{11}}{k+n}$$

$$\hat{p}_y=\frac{n_x+n_{01}+n_{11}}{m+n}$$

The variance of the difference then allows for the non-independence of $\hat{p}_x$ and $\hat{p}_y$ to give

$$\mathrm{Var}(\hat{p}_x - \hat{p}_y) = \frac{p_x(1-p_x)}{k+n} + \frac{p_y(1-p_y)}{m+n} - \frac{2n(p_{11} - p_x p_y)}{(k+n)(m+n)}$$

where p_{11} is the expected proportion of the *n* units which are 'yes' on both occasions. We do not know this, but we can estimate it by

$$\hat{p}_{11} = \frac{n_{11}}{n}$$

the observed proportion of the *n* units which are 'yes' on both occasions.

For our data this gives $\hat{p}_x - \hat{p}_y$ = 0.01155, Var($\hat{p}_x - \hat{p}_y$) = 0.00000497885, standard error = 0.0022313 and hence 95% confidence interval = 0.0072 to 0.0159. The Thomson method gives similar results to the proposed method, with a variance which is very slightly larger. The numerical example given by Thomson cannot be analysed by our method, because $m = 0$.

In general, the Thomson method gives more weight to the unpaired sample than does ours, and so may be expected to give a larger variance. How much larger depends on the correlation between the variable on the first and second occasion. On the other hand, the stronger assumption of the Thomson method may lead to a smaller variance when it is met.

For the significance test, Thomson gives a variance under the null hypothesis which is relatively complicated to find, involving iteration. We agree with him that estimation is far preferable anyway and so do not think the much simpler formulation here is a great advantage.

From the significance test point of view, Choi and Stablein [5] investigated seven methods of approaching this problem, as follows:

1. ignoring the paired sample
2. ignoring the unpaired sample
3. pairing unmatched observations randomly, discarding any excess
4. use of a weighted combination of Standard Normal deviates found for the
5. separate paired and unpaired comparisons, weighted by number of subjects
6. combining P values for the paired and unpaired comparisons
7. comparing combined weighted estimates of the first and second
8. proportions, a method similar to that of Thomson $ ^3 $
9. likelihood ratio test

Choi and Stablein [5] conclude that method 7 is best but computationally difficult, and the methods 4 and 6 are to be recommended. As might be expected, methods 1, 2 and 3, which sacrifice some of the data, lose power as a result. All these tests can lead to possibly biased results if the mechanisms which caused the incompleteness of the data are related to the factor under investigation. [5] In the NCDS example, subjects who provided data on only one occasion included those who refused to provide data on the other. This self-selection should make us reluctant to assume that proportions are the same in this group as in those who provided data on both occasions, if we can avoid it. Although not significant, the proportions are higher for those with partial information than for those with full information. Our proposed method does not require this assumption.

Shih [6, 7] has produced a maximum likelihood method, which also requires the assumption that the paired and unpaired samples are from the same population ('missing at random').

EXTENSION TO ODDS RATIOS

We can extend the method to odds ratios fairly easily, provided the samples are large enough. For the paired data, the log odds ratio is given by $\hat{l}_p = \log(n_{10} / n_{01})$ with approximate variance

$$\text{Var}(\hat{l}_p) = \frac{1}{n_{10}} + \frac{1}{n_{01}}$$

This approximation is reasonable provided the sample is large enough, discussed below. For the unpaired data, we have $\hat{l}_u = \log\left(n_x(m-n_{y)})/n_y(k-n_x)\right)$ with

$$\text{Var}(\hat{l}_u) = \frac{1}{n_x} + \frac{1}{m-n_y} + \frac{1}{n_y} + \frac{1}{k-n_x}$$

For the example, $\hat{l}_p$ = log(298/203) = 0.38389 with estimated variance

$$\text{Var}(\hat{l}_p) = \frac{1}{298} + \frac{1}{203} = 0.0082818$$

and for the unpaired data $\hat{l}_u = \log\left((215\times1717)/(73\times3737)\right)$ = 0.30247 with

$$\mathrm{Var}(\hat{l}_u) = \frac{1}{15} + \frac{1}{1717} + \frac{1}{73} + \frac{1}{3737} = 0.019200$$

The weights are w_p = 1/0.0082818 = 121 and w_u = 1/0.019200 = 52, giving combined estimate

$$\hat{l} = \frac{121 \times 0.38389 + 52 \times 0.30247}{121 + 52} = 0.35942$$

The variance of this estimate is

$$\mathrm{Var}(\hat{l}) = \frac{1}{121 + 52} = 0.0057803$$

and so the standard error is √0.0057803 = 0.076028 . The 95% confidence interval is thus 0.35942 ± 1.96×0.076028 which gives 0.21041 to 0.50843. Exponentiating we get estimated odds ratio = 1.43, 95% confidence interval 1.23 to 1.66.

We can check the assumption that the log odds ratios in the paired and unpaired samples are the same. The difference between them is 0.38389 – 0.30247 = 0.08142 with standard error √(0.0082818 + 0.019200) = 0.16578. As the standard error is larger than the estimate there is no evidence that the assumption of equal log odds ratios is not met.

ADEQUACY OF THE APPROXIMATION

Any approach using standard errors for proportions relies on the sample being large enough for the standard error to be a reasonable estimate. For matched samples, only the discordant pairs enter into the estimate and its standard error. The effect of this is that standard errors are not so well estimated for *n* matched pairs as they are for the comparison of two independent groups each of size *n*. The standard error for both the difference between two independent proportions and for the log odds ratio are quite well estimated provided the observed frequencies exceed 5. (The condition applies to observed rather than expected frequencies because we are dealing with estimation rather than testing.) For matched samples, simulations suggest that the standard error of the difference between two proportions is reasonable provided the off diagonal frequencies exceed 10, but for the log odds ratio these frequencies should exceed 20. Many authors do not give the standard error method for the paired log odds ratio, but quote more complex formulae which give better approximations when numbers are small.

DISCUSSION

The method proposed here seems to us a simple and attractive solution to the problem of comparing overlapping samples, which can be applied easily to several types of data. We have presented the extension to odds ratios, it would be equally straightforward to apply the method to the combination of paired and two-sample t tests. No doubt we are not the first to consider this approach, and others have either not thought it worth while publishing, or it has been published and we and others have failed to retrieve it. Only the presence in the literature of methods which are computationally more difficult [3, 5, 6, 7] or wasteful of data [5] made us think it worth presenting to the medical statistics community.

Compared to the method described by Thomson, [3] our method has two advantages: it is slightly computationally easier, particularly for the significance test, and it does not require the assumption of equal proportions in the paired and unpaired samples. In repeated surveys, where data may be

present on only one occasion because subjects refuse to answer on the other, it is quite possible that non-response is related to the outcome. For such data, the assumption that refusers are from the same population as acceptors is a strong one. In his example, Thomson [3] emphasised that the reason for data being present on only one occasion was administrative and unrelated to the variable observed. Our proposed method, which requires only that the difference be the same in the paired and unpaired observations, is therefore preferable for many applications. It has the disadvantage of requiring that there be unpaired data in both the first and second samples.

If unpaired data are present in only one sample, we cannot estimate the difference for the unpaired data. We can only use the unpaired data to improve the estimate of the proportion in the sample to which it applies. To do this, we must assume that the proportion in the paired and unpaired data are the same and a method which does not require this assumption cannot be used. If we cannot make this assumption, the only possible approach is to omit the unpaired data altogether.

Thus we conclude that our method is preferable to others on the grounds of ease of computation and less stringent assumptions, provided there are unpaired data in both samples. If not, the method of Thomson [3] would be the method of choice.

Acknowledgements

We thank the Economic and Social Science Research Council Data Archive, University of Essex, for supplying copies of the data. We are grateful to Janet Peacock, Jan Poloniecki, and Fiona Reid for helpful comments on an earlier draft. This paper was stimulated by an email query from Chris Stride.

References

1. City University Social Statistics Research Unit National Child Development Study, 1958 [computer file] Colchester: ESRC data archive, 1993.

2. Anderson, HR, Bland, JM, Patel, S, Peckham, C. (1986) The natural history of asthma in childhood. *Journal of Epidemiology and Community Health* **40**, 121-9.

3. Thomson, P.C. (1995) A hybrid paired and unpaired analysis for the comparison of proportions. *Statistics in Medicine* **14**, 1463-1470.

4. Hald, A. (1962) *Statistical Theory with Engineering Applications* New York: John Wiley and Sons, p243-4.

5. Choi, S.C. and Stablein, D.M. (1982) Practical tests for comparing two proportions in incomplete data. *Applied Statistics* **31**, 256-262.

6. Shih, W.J. (1985) Maximum likelihood estimation and likelihood ratio test with incomplete pairs. *Journal of Statistical Computation and Simulation* **21**, 187-194.

7. Shih, W.J. (1987) Maximum likelihood estimation and likelihood ratio test for square tables with missing data. *Statistics in Medicine* **6**, 91-97.