Skeletal Radiol (2011) 40:1–4
DOI 10.1007/s00256-010-1006-1

PERSPECTIVE

Assessment of paired binary data

Kaspar Rufibach

Received: 8 July 2010 / Revised: 8 July 2010 / Accepted: 8 July 2010 / Published online: 30 July 2010

Introduction

A typical statistical problem in studies published in imaging journals such as Skeletal Radiology is the comparison of (1) two paired proportions, (see e.g., [1]) and (2) the comparison of sensitivity and specificity of two diagnostic methods that are both compared to the gold standard on the basis of the same experimental units, e.g., images or patients (see e.g., [2–4]). In general, "paired", "correlated", or "clustered" binary data does not only arise when comparing two different methods on the same experimental units but also when measuring a binary response twice, e.g., before and after an intervention.

In [5], it was reported that in the first 6 months of 2007 approximately 25% of Skeletal Radiology papers the latter authors reviewed clustering in data was not properly accounted for in analysis. The intention of this solicited perspective is to briefly explain, from a statistician's point of view, the methods that can be used to analyze clustered binary data, illustrate those methods on an example published in Skeletal Radiology, discuss some common pitfalls, and provide some recommendations on how to report the results of such a comparison.

What we present here are standard statistical methods elaborated on in more detail e.g., in [6, Chap. 6], [7, Sect. 10.1], [8, Sect. 13.1], [9, Chap. 21], or [10]. However, since not all of the methods presented here are part of all standard software packages, we deem it appropriate to also provide some basic, admittedly involved, formulas.

K. Rufibach (✉)
Institute of Social and Preventive Medicine, Biostatistics Unit,
University of Zurich,
Hirschengraben 84,
8001 Zurich, Switzerland
e-mail: kaspar.rufibach@ifspm.uzh.ch

Comparison of paired proportions

A typical example is presented in [4]. From the comparison of an imaging method (conventional magnetic resonance, MR) in $n=92$ patients to the gold standard (arthroscopy, AR) in detection of supraspinatus tendon tears, the numbers in Table 1 resulted. To discuss the approaches of analysis, we introduce some notation in Table 2. The letters *a, b, c, d* and their sums provide frequencies we observe in a situation as displayed in Table 1, whereas population quantities, i.e., the underlying probabilities we want to assess, are provided in parentheses. When reporting results of this type, we recommend to provide absolute instead of relative empirical frequencies since, unlike in the case of unpaired proportions, absolute frequencies are needed to compute a confidence interval for the difference of paired proportions (see below).

The typical null hypothesis researchers want to assess in Table 1 is whether the proportion of samples judged positive by arthroscopy (Method 1 in Table 2), p_{1+} estimated by $\widehat{p}_{1+} = (a+b)/n$, is equal to those termed positive by MR (Method 2), p_{+1} estimated by $\widehat{p}_{+1} = (a+b)/n$:

$$H_{0,\text{McNemar}} : p_{1+} = p_{+1}. \quad (1)$$

This hypothesis is equivalent to testing whether the difference $d = p_{1+} - p_{+1}$ is equal to 0. Looking at tables similar to Table 1, researchers are often tempted to compute a χ^2 (or Fisher's exact) test and report the corresponding

Table 1 MR vs. gold standard

	MR +	MR −	Total
Arthroscopy +	16	7	23
Arthroscopy −	3	66	69
Total	19	73	92

results. However, an χ^2 test assesses a null hypothesis different from (1). Namely, in Table 1, whether the proportion of MR-positive results is the same in the AR positive group, p_{11}/p_{1+}, compared to the AR negative group, p_{11}/p_{+1}:

$$H_{0,\chi^2} : p_{11}/p_{1+} = p_{21}/p_{2+}. \tag{2}$$

These two quantities are estimated by $a/(a + b)$ and $c/(c + d)$. Comparing the null hypotheses $H_{0,\text{McNemar}}$ and H_{0,χ^2} reveals that, unlike for two-group comparisons for continuous data, the choice of the analysis method for a binary response does not only depend on the structure of the data ("dependent" vs. "independent") but also on the hypothesis one would like to assess.

A standard error of $\widehat{d} = \widehat{p}_{1+} - \widehat{p}_{+1} = (b - c)/n$, assuming the null hypothesis d=0, amounts to $se_0\left(\widehat{d}\right) = \sqrt{b + c/n}$. Note that the estimated difference of proportions $\widehat{d}$ and the corresponding standard error se0($\widehat{d}$) only depend on the off-diagonal elements of the underlying contingency table, b and c. Having available the standard error of the quantity of interest allows construction of a Wald-type test statistic $\omega^2 = \widehat{d}/\text{se}_0\left(\widehat{d}\right)$. The continuity-corrected version

$$\omega^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{3}$$

is due to McNemar (the original reference is [11]) and follows in large samples under the null hypothesis $H_{0,\text{McNemar}}$ an χ^2 distribution with 1 degree of freedom. To perform a statistical test at significance level $1 - \alpha$ for $H_{0,\text{McNemar}}$ we therefore compare ω^2 to the $(1 - \alpha)$-quantile of the χ^2 distribution with one degree of freedom. Alternatively, to quantify the evidence against $H_{0,\text{McNemar}}$, one can compute a p value (here, p=0.34).

A common rule-of-thumb for the validity of the asymptotic χ^2 McNemar test is that the number of discordant pairs is larger than $10 : b + c \geq 10$. If less discordant pairs are present, use of an exact binomial test is recommended. To this end, one conditions on $b + c$ and derives an exact test for the odds ratio, see [12], Chap. 5] for details. For the MR data, the exact p value amounts to 0.34, so is identical to the continuity-corrected p value from the asymptotic χ^2 McNemar test.

As is the case for all statistical tests, by performing one we do not get any information about the size of a possible effect. We therefore recommend to complement the result of a McNemar test with a corresponding $(1 - \alpha)$ (typically 95%) confidence interval. To get a Wald-type confidence interval, the standard error of $\widehat{d}$ needs to be generalized to the case where the underlying proportions are not hypothesized to be equal (see [8, Sect. 13.1]), namely

$$\text{se}\left(\widehat{d}\right) = \frac{1}{n}\sqrt{b + c - \frac{(b - c)^2}{n}} = \sqrt{\frac{\text{se}_0\left(\widehat{d}\right)^2 - \widehat{d}^2}{n}}. \tag{4}$$

A $(1 - \alpha)$ Wald-type confidence interval for the underlying difference of paired proportions can then be computed according to

$$\left[\widehat{d} - q_{1-\alpha/2}\text{se}\left(\widehat{d}\right) - 1/n, \widehat{d} + q_{1-\alpha/2}\text{se}\left(\widehat{d}\right) + 1/n\right] \tag{5}$$

where q_α is the α-quantile of the standard normal distribution and $1/n$ is a continuity correction. The reason for providing these formulas here is that a confidence interval for a paired proportion is, to the best of our knowledge, not part of all standard software packages. Having said that, we would like to point out that the Wald-type confidence interval typically exhibits poor coverage performance (see [13]) if the number of discordant pairs is small. Instead, [6] recommend using Newcombe's score-based interval (introduced in [10]). Corresponding closed formulas and a detailed worked out example can be found in [2, Chap. 6]. Further alternatives are the exact interval (see [4, Sect. 5.2]) or the interval proposed in [13]. Although closed formulas are available, computation of all these intervals is more involved than (5).

For the data in Table 1 we get an estimated difference of paired proportions of $\widehat{d} = \widehat{p}_{1+} - \widehat{p}_{+1} = 0.25 - 0.21 = 0.04$. Using the standard error computed under the assumption that d=0 yields a value of the test statistic $\omega^2 = \widehat{d}/\text{se}_0\left(\widehat{d}\right) = 1.33$ which is smaller than the 95% quantile 3.84 of the χ^2 distribution, so that we can not conclude that the proportion of positive results is different for MR compared to AR. The corresponding 95% confidence interval computed according to (5) amounts to [–0.030, 0.110]. So we can conclude that this interval covers the true underlying difference between

Table 2 General notation: empirical frequencies and underlying probabilities

	Method 2 +	Method 2 −	Total
Method 1 +	a (p_{11})	b (p_{12})	$a + b$ (p_{1+})
Method 1 −	c ($p21$)	d ($p22$)	$c + d$ ($p2+$)
Total	$a + c$ (p_{+1})	$b + d$ (p_{+2})	n (1)

p_{1+}, the proportion of samples judged positive by Method 1 (AR), and p_{+1}, the proportion of samples judged positive by Method 2 (MR), with a probability not less than 95%. Note that the confidence interval contains 0, the value of no effect.

However, as can be inferred from Table 1, the number of discordant pairs amounts to only 7+3=10, and reporting of Newcombe's interval should be considered. For the MR data, this interval is [–0.028, 0.116]. In this case, differences between the Wald and Newcombe confidence interval turn out to be negligible.

Comparison of sensitivity and specificity between groups

The main scientific question in [4] was to evaluate sensitivity and specificity of MR and a second experimental method, abduction external rotation (ABER), see Table 3 for the corresponding number of patients. The two experimental methods, MR and ABER, differ in the positioning of the patient's arm. To illustrate the procedures discussed here, we extracted the numbers from the AR (gold standard) positive patients in Tables 1 and 3 to generate Table 4, or rather to determine its row and column totals. Sensitivities for MR and ABER, respectively, are estimated as $\text{Sens}_{\text{MR}}=\widehat{p}_{1+}= 16/23 = 0.696$ and $\text{Sens}_{\text{ABER}}=\widehat{p}_{+1}= 13/23 = 0.565$. Obviously, comparing $\text{Sens}_{\text{ABER}}$ to Sens_{MR} is precisely assessing hypothesis (1) for Table 4 and therefore McNemar's test statistic ω^2 can be used to evaluate whether the two sensitivities coincide. Since in Table 4 the number of discordant pairs is with 4+1=5 small, we rely on the exact McNemar test, yielding a p value of 0.38. Thus we can not conclude that sensitivities are different between ABER and MR. Specificities between methods can be compared similarly.

However, unlike in Sect. 2, when comparing sensitivities or specificities between methods, we are often not so much interested in providing a confidence interval for the difference, but rather for sensitivity (specificity) in each group (here MR and ABER). This implies that we should provide a confidence interval for a single proportion, i.e., for the sensitivity (or specificity) in each group separately. In accordance with [6], we recommend to use the confidence interval due to Wilson (see e.g., [14] for a performance comparison of confidence intervals for a single proportion). In our example, we get 95% Wilson confidence intervals for Sens_{MR} and $\text{Sens}_{\text{ABER}}$ of [0.491, 0.844] and [0.368, 0.744], respectively.

Table 3 ABER vs. gold standard

	ABER +	ABER −	Total
Arthroscopy +	13	10	23
Arthroscopy −	4	65	69
Total	17	75	92

Table 4 Computation of sensitivity: MR and ABER status in AR-positive patients

	ABER +	ABER −	Total
MR +	12	4	16
MR −	1	6	7
Total	13	10	23

Further points

When prospectively planning a study for a primary endpoint that is binary and consists of paired observations, a decision must be made about the sample size to obtain. The classical way of planning a sample size for McNemar's test is to a priori specify the probability of discordance, $p12+p21$, and an odds ratio to be detected. However, often the investigator is hardly able to provide the probability of discordance, but can state, at least approximately, the marginal probabilities $p1+$ and $p+1$. How to plan a sample size in that scenario is elaborated in [15].

It is sometimes argued that assessing sensitivity and specificity separately, as proposed in Sect. 3, is problematic, since one method may have higher sensitivity but lower specificity. How to combine the two tests to get a single one is discussed in [16].

Finally, let us mention that more-involved methods, such as the ones described in [17], generalized linear mixed models, or generalized estimating equations, are necessary once we have more than two observations on each experimental unit, i.e., clustered data.

As mentioned in the introduction, general comments on the analysis of clustered data can also be found in [5].

Concluding remarks

To conclude, we briefly summarize those three points we tried to emphasize in this brief note. First, when reporting the result of a comparison of proportions, carefully think about whether the observations are paired, i.e., whether the same experimental units have been measured twice and what hypotheses you want to assess. Second, in case of paired binary observations, when the number $b + c$ of discordant observations is small, consider application of an exact test. For computation of a confidence interval, use of (5) is only recommended once one has a large sample.

Otherwise, computation of Newcombe's interval is advocated. And finally, always complement the result of a statistical test with the corresponding confidence interval.

Acknowledgements I thank the editors for helpful comments and Prof. Burkhardt Seifert for discussions and proofreading the article.

References

1. Blankenbaker DG, Ullrick SR, Davis KW, De Smet AA, Haaland B, Fine JP. Correlation of MRI findings with clinical findings of trochanteric pain syndrome. Skeletal Radiol. 2008;37(10):903–9. doi:10.1007/s00256-008-0514-8.
2. Dinauer PA, Flemming DJ, Murphy KP, Doukas WC. Diagnosis of superior labral lesions: comparison of noncontrast MRI with indirect MR arthrography in unexercised shoulders. Skeletal Radiol. 2007;36(3):195–202. doi:10.1007/s00256-006-0237-7.
3. Kijowski R, Blankenbaker D, Stanton P, Fine J, De Smet A. Correlation between radiographic findings of osteoarthritis and arthroscopic findings of articular cartilage degeneration within the patellofemoral joint. Skeletal Radiol. 2006;35(12):895–902. doi:10.1007/s00256-006-0111-7.
4. Schreinemachers SA, van der Hulst VPM, Willems WJ, Bipat S, van der Woude HJ. Is a single direct MR arthrography series in ABER position as accurate in detecting anteroinferior labroligamentous lesions as conventional MR arthography? Skelet Radiol. 2009;38(7):675–83. doi:10.1007/s00256-009-0692-z.
5. Obuchowski NA, Lieber ML. Statistics and methodology. Skeletal Radiol. 2008;37(5):393–6. doi:10.1007/s00256-008-0448-1.
6. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence. 2 edn. University Press Belfast; 2000.
7. Agresti A. Categorical data analysis. Wiley series in probability and mathematical statistics: applied probability and statistics. 2nd edn. New York: John Wiley & Sons Inc; 2002. A Wiley-Interscience Publication.
8. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. Wiley Series in Probability and Statistics, 3rd edn. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ; 2003. doi:10.1002/0471445428.
9. Kirkwood BR, Sterne JAC. Essential medical statistics. Malden: Blackwell Science; 2003.
10. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med. 1998; 17(22):2635–50. URL http://www.hubmed.org/display.cgi?uids=9839354.
11. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12 (2):153–7.
12. Breslow NE, Day NE. Statistical methods in cancer research. volume I - the analysis of case-control studies. IARC Sci Publ. 1980;32:5–338.
13. Zhou XH, Qin G. A new confidence interval for the difference between two binomial proportions of paired data. J Statist Plann Inference 2005;128(2):527–542. doi:10.1016/j.jspi.2003.11.005.
14. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Stat Sci. 2001;16(2):101–33. With comments and a rejoinder by the authors.
15. Lachenbruch P. On the sample size for studies based upon McNemar's test. Stat Med. 1992;11:1521–5.
16. Hamdan M, Pirie W, Arnold J. Simultaneous testing of McNemar's problem for several populations. Psychometrika. 1975;40(2):153–61.
17. Obuchowski N. On the comparison of correlated proportions for clustered data. Stat Med. 1998;17(13):1495–507.