Statistical Briefing Note No 1

Why use relative risks?

David Gordon, University of Bristol
December 2012

Significant Differences?

One of the key questions that readers of the perception of necessities graphs and reports want answered is: are the differences *between* groups (e.g. Men vs Women, older people vs young adults, etc) 'significant', that is whether the differences observed are likely to have occurred by chance or not.

To help answer this question, the PSE research team have used "relative risk" rather than "hypothesis testing". This statistical note sets out the differences between the two approaches and the reasons for the choice.

Hypothesis testing

The need to find a way of testing whether observed differences were significant or not was one of the key drivers behind the development of statistics during the 20th Century, particularly the development of hypothesis testing.

The origins of hypothesis testing can be traced back to Student's[1] invention of the *t*-test in 1908 for monitoring the quality of Guinness while working as the Brewer-in-Charge of the Experimental Brewery in Dublin (Fisher Box, 1987). Fisher built on Student's work and eventually published the extremely influential book, *Statistical Methods for Research Workers* in 1925, in which he explained the idea of hypothesis testing. He argued for the use of the < 5% level for testing significance that is this would have occurred by chance less than 5 in 100 times (and the 1% level for more stringent testing) and, equally importantly, he provided statistical look-up tables in the book (for X^2, *t*, *F*) at the 5% and 1% level (Kendall, 1963; Cowles & Davies, 1982; Lehmann, 1993)

The Problem of Hypothesis Testing

The major problem with hypothesis testing is that what readers want to know is 'given the survey data that have been collected what is the probability that the differences are significant?' Unfortunately, what a significance test actually tells us is: 'Given that there are no differences (that Ho – the 'null hypothesis' is correct), what is the probability of these data?' Unfortunately these are not the same questions and hypothesis testing is logically flawed. Pollard and Richardson (1987, p161) provide the following example to illustrate the logical problem, which is elaborated on by Cohen (1994, p999)

[1] Student was a pseudonym used by William Sealy Gosset (Fisher Box, 1987)

'*If a person is American, then he is probably not a member of Congress.*
This person is a member of Congress
Therefore, he is probably not an American'

Equally problematic, is the fact that most significance tests are affected by both the magnitude of the differences and the sample size. Thus, if you have a large enough sample of data (survey) then virtually everything you test will be 'significantly' different at the 5% level. In this case, all a Null Hypothesis Significance Test tells you is that you have a large sample of data and you already know that without performing any test! Hypothesis testing may tell you that the results are 'statistically significant' but this does not mean that the differences are 'important' or meaningful.

Thus, Rozeboom (1960, p417) argued, over 50 years ago, that Null Hypothesis Significance Testing was "*The statistical folkways of a more primitive past*".

Confidence intervals

In general, confidence intervals contain all the information found by significance testing and much more (Cohen, 1994)

Confidence intervals provide information about the range in which the true value lies with a certain degree of probability. There is of course *only one* true value, and the confidence interval defines the range where it's most likely to be. Thus a 95% confidence interval will include the' true' value 95% of the time i.e. the confidence interval includes the true value 95 times out of 100.

Relative Risk

During the past 50 years, statisticians have developed a wide range of techniques for measuring the size/magnitude of differences between two or more groups (effect sizes). Relative Risk (RR) ratios[2] are both simple to understand and their confidence intervals are easy to calculate (Morris & Gardner, 1988).

Relative risk tells us the **risk** or **probability** of one group (e.g. men) thinking an item is a necessity compared with the other group (e.g. women). Thus, a relative risk of 2.0 means twice the risk, a score of 0.5 means half the risk, a score of 3.0 is three times the risk and a score of 0.33 is a third of the risk, etc. A relative risk of 1 would mean that there are no differences between the two groups. We can use this approach to compare the differences in attitudes to necessities between groups (e.g. men v women, etc) by calculating their relative risk. A relative risk greater than 1.0 means men are more likely than women to think that a particular item is a necessity, by contrast a relative risk less than 1.0 means that women are more likely than men to think that the item is a necessity. The nearer the relative risk is to 1, the smaller the difference between the two groups.

If the 95% confidence Intervals of a relative risk ratio span 1.0 then you cannot be confident at the 5% level that the 'true' risk is different from 50:50, i.e. the difference is unlikely to be 'significant'. If the 95% confidence Intervals of a relative risk ratio do not span 1.0 then the differences between the two groups are likely to be statistically significant.

---

[2] Where a, b, c & d are the four cells of a two by two table then *Relative Risk* = (a / (a+b)) / (c / (c+d))

Thus relative risk ratios and their 95% Confidence Intervals (CI) provide intuitive and useful estimates about whether the differences between two groups are likely to be significant (or not) and also the size and direction of these differences Relative risk and their 95% confidence Intervals (CI) therefore tells you not only if the differences between two groups are 'statistically significant' it also tells you if these differences are likely to be important or meaningful (i.e. practically significant).

References

Cohen, J. (1994) The Earth is Round ($p < .05$). *American Psychologist*, 49, 997-1003.

Cowles, M. and Davis, C. (1982) On the Origins of the .05 Level of Statistical Significance. *American Psychologist*, 37, 553-558.

Fisher, R. A. (1925) Statistical Methods for Research Workers. Edinburgh, Oliver & Boyd

Fisher Box, J. (1987). "Guinness, Gosset, Fisher, and Small Samples". *Statistical Science* **2** (1): 45–52

Lehmann, E.L. (1993) The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?. *Journal of the American Statistical Association*, 88, 1242-1249

Kendall, M. G. (1963) Ronald Aylmer Fisher, 1890-1962, Biometrica, 50, 1-15.

Morris, J.A. and Gardner, M.J. (1988) Calculating confidence intervals for relative risk (odds ratios) and standardised ratios and rates. *British Medical Journal*, 296, 1313-1316.

Pollard, P. and Richardson, J. (1987) On the probability of making Type I errors. *Psychological Bulletin*, 102, 159-163.

Rozeboom, W. W. (1960) The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428.

Student (1908) The Probable Error of the Mean. *Biometrica*, 6, 1-25.