Qual Life Res (2008) 17:651–652
DOI 10.1007/s11136-008-9351-4

The scales were highly correlated: $P = 0.0001$

Peter M. Fayers

Accepted: 16 April 2008 / Published online: 7 May 2008

This is the first of a series of brief notes, describing some of the statistical issues encountered when reviewing submissions to *Quality of Life Research*—or, for that matter, to any other journal. The review process aims to catch infelicitous applications of statistics before publication, and of course it is unreasonable to expect that authors should be aware of all the issues themselves. Unfortunately, some do slip through. That stimulated the writing of this present note.

A cornerstone of traditional instrument development and the evaluation of multi-item scales is the use of correlations. For example, correlations underpin factor analysis and structural equation modelling, and are used for multi-trait scaling, for assessing convergent and discriminant validity, and even appear in disguise as Cronbach's α. Yet there are several traps for the unwary when reporting correlations.

As a statistician involved in randomized clinical trials as well as a variety of non-randomized studies, I am usually pleased when I find differences that are significant with $P < 0.01$, and I am delighted by $P < 0.001$. Thus, alarm bells start ringing whenever I see a table with a column of P-values that are nearly all more extreme than 0.0001.[1] And when I see that the table is reporting correlation coefficients, it is obvious exactly what is wrong.

Most dimensions or scales in health-related quality of life (HRQoL) research are bound to be correlated. We know, beyond all reasonable doubt, that responses to questions about depression will be correlated with pain scores, with fatigue, with physical functioning, with emotional well-being, etc. A significance test as to whether r (correlation coefficient) $= 0$ is not only uninteresting, it is pointless. We already know that the "null hypothesis" of $r = 0.00$ is bound to be false. In a small-sized study, a not-significant result would therefore be ascribed to inadequate sample size. However, many studies are designed with many other analyses in mind, such as factor analysis or other analyses, and will have a suitably large sample size and thus more than ample power for confirming that r is not zero. Hence the occurrence of a plethora of highly significant correlation coefficients, $P < 0.0001$.

Why do researchers so frequently report this uninteresting and self-evident information? Perhaps partly because so many journals and reviewers request that all results be tested for significance. Partly, no doubt, because the majority of statistical packages automatically display the P-values, even when they are irrelevant and uninformative. Unfortunately, by reporting these extreme values in a publication, readers who are not statistically astute may be misled into believing that the results indicate more than they do.

How should correlations be reported, then? Some possibilities, depending on the objectives of the study, include the following:

- Perhaps the simplest approach is to regard the correlation as an estimate, in which case the natural information to supply is the confidence interval (CI). This is very easy to calculate, as shown below. The CI indicates to the reader the precision of the estimated r-coefficient, and provides an idea as to whether correlations differ significantly from each other. A small sample size results in a wide a

P. M. Fayers (✉)
Department of Public Health, University of Aberdeen Medical School, Foresterhill, Aberdeen AB25 2ZD, UK
e-mail: P.Fayers@abdn.ac.uk

P. M. Fayers
Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Trondheim, Norway

[1] An even worse statistical sin is to report $P = 0.0000$. P-values are never zero, although they can be, say, <0.0001.

CI, reflecting the lack of information for estimating the true value of r. If the CI crosses zero then there is insufficient evidence that the correlation is significantly different from zero—either r is small or zero, or the sample size is inadequate.

- Another possibility is indeed to use a significance test, but with a different target threshold. Choice of threshold will depend on the objective, and should be specified a priori (that is, in advance and recorded in the study protocol). For example, we may have stated in the study protocol that we expect certain scales to be correlated with $r > 0.7$. This is a testable hypothesis. When carrying out multitrait analysis, on the other hand, it has been suggested that during initial scale development item convergence is supported for items that correlate at least moderately ($r > 0.3$) with the scale that they are hypothesised to belong to, but that for subsequent evaluation the values of r should be at least 0.4 [1]. Of course, the CI provides equivalent information; it is usually more informative, too, and many investigators prefer to use CIs rather than significance tests with P-values. In any case, whenever values of r are important, details of sample size and/or CI should accompany significance tests.
- A third possibility is to test whether one correlation is significantly greater than another—this effectively being what is also done in multitrait-scaling analysis, when we claim item discrimination for an item if it correlates (significantly) more strongly with the scale that it is contained in rather than with other scales. If, on the other hand, an item correlates significantly more strongly with other scales than with its own, a "scaling error" is indicated and, unless there are overriding clinical or other practical or theoretical reasons, the item (or the scales involved) should be considered for modification.
- In addition, correlation coefficients such as Pearson's r assume a linear relation between two variables. Even if plots are not reproduced in a publication, readers should be told how the investigators examined and confirmed this assumption—or, when appropriate, what allowance was made for non-linearity.

Calculating the confidence interval of *r*

The most common form of correlation is called Pearson's r, or the product–moment correlation coefficient. Although r itself does not have a normal distribution, there is a simple transformation that can convert r to a variable Z that does. This transformation is

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right).$$

Furthermore, it can be shown that, for a sample size of n, the standard error (SE) of Z is given by

$$SE(Z) = \frac{1}{\sqrt{n-3}}.$$

These equations assume that n is reasonably large—in practice, more than 50 observations. It is also assumed that the two variables being analysed are plausibly normally distributed, and that they have a linear relationship.

A CI for Z can then be calculated as for any data that follow a normal distribution. For example, a 95% CI is

$$Z_{lower} = Z - 1.96 \times \frac{1}{\sqrt{n-3}} \quad \text{to}$$
$$Z_{upper} = Z + 1.96 \times \frac{1}{\sqrt{n-3}}.$$

Finally, Z_{lower} and Z_{upper} can be converted back to obtain the CI for r itself, using

$$r_{lower} = \left(e^{2Z_{lower}} - 1\right)/\left(e^{2Z_{lower}} + 1\right),$$

with a similar expression for r_{upper}.

A simple example of the calculation is taken from Fayers and Machin (2007), who also show how to test for a difference between two correlations [1].

> Suppose a correlation of $r = 0.58$ has been observed for the association between Current QoL and Pain using the FLIC, with a sample size of 98 patients. Then:

$$Z = \frac{1}{2}\log_e\left(\frac{1+0.58}{1-0.58}\right) = 0.6625 \ , \quad \text{and}$$
$$SE(Z) = \frac{1}{\sqrt{98-3}} = 0.1026 \ .$$

The *CI* for Z is

$$Z_{lower} = 0.6625 - (1.96 \times 0.1026) = 0.4614 \quad \text{to}$$
$$Z_{upper} = 0.6625 + (1.96 \times 0.1026) = 0.8636$$

Therefore the 95% *CI* for r itself is

$$r_{lower} = \left(e^{2\times0.4614} - 1\right)/\left(e^{2\times0.4614} + 1\right) = 0.43$$

to

$$r_{upper} = \left(e^{2\times0.8636} - 1\right)/\left(e^{2\times0.8636} + 1\right) = 0.70.$$

Hence we would expect that the true value of r is likely to lie between 0.43 and 0.70.

Reference

1. Fayers P. M., & Machin D. (2007). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes* (2nd ed.). Chichester: John Wiley & Sons. ISBN 13: 978-0-470-02452-2.