Correlation coefficient and the fallacy of statistical hypothesis testing

Marcin Kozak

A correlation coefficient is likely the most well-known and commonly used statistical tool by researchers in any field and any country. Yet it does deserve some attention.

The issue I want to discuss here is the testing of the population correlation coefficient. In applied sciences it is done almost always when the correlation is estimated. Let us disregard here how the testing is reported since it has nothing to do with the main topic, and let us focus on the testing itself. I do believe that testing is most often done by convention only, and usually not only does it provide no useful information to the researcher, but also it may be deceptive. Why and how? Let us have a closer look at this issue.

A sample correlation coefficient r between two normal variables is determined to estimate a population correlation coefficient ρ between these variables. In addition, when one wants to check whether there is a linear relationship between the two variables, one tests the null hypothesis H_0: $\rho = 0$; its rejection means that the relationship is statistically significant, while its acceptance means there is no linear relationship.

All this can make sense, but the point is that in practice often the meaning of this testing is overestimated, and once the coefficient is significant at a particular significance level, it is reported as meaningful.

To see how far from correct this may be, let us note that the testing depends on a sample size. Provided that the sample is huge, anything will be significant. I remember seeing a correlation smaller than 0.025 reported as significant. Let me remind the reader about this well-known cliché, though an accurate one, that statistical significance does not necessarily mean the actual significance[1]. So how can a correlation of 0.025, and even that of 0.10, be significant? Of course, this does depend on a problem one applies correlation for, but 0.025?

One important thing to remember here is that the null hypothesis under consideration states that the population correlation coefficient between two variables equals zero, which is commonly understood as lack of a linear relationship between the variables. I would rather say misunderstood, because such a claim implies that if only the correlation is not zero, the variables are linearly related. This is not true. Note that if a population correlation coefficient equals 0.025, it does not equal zero, but would anyone claim that two such variables are linearly related? And with large sample sizes, it is quite possible that the null hypothesis will be rejected even for a small value of the coefficient's estimator. Just to present this, I generated two normal variables with means 0 and variance 1 and of size 10,000, with the correlation coefficient between them equal to –0.0235 (Figure 1). A locally weighted regression[2] (loess) curve is added to the scatterplot to help grasp the pattern of the relationship. Box 1 presents the results of the correlation analysis, including testing the two-sided null hypothesis that correlation is zero; this is a standard output of the default settings from the cor.test() function of *R* language[3]. If one chooses 0.05 type-I error probability level, the null hypothesis will be rejected, which some would understand as indicating the linear relation between the variables. Look at Figure 1 once more and ask yourself the question: are x and y linearly related?

Reporting a low correlation as significant is one thing. Now imagine that the population correlation coefficient one wants to estimate is $\rho = 0.50$. Based on a ten-element sample one has estimated ρ as $r = 0.50$, so the bias of this estimation is nil and the estimation is perfect. This value is, nonetheless, non-significant from a statistical point of view at the 0.05 significance level (and lower, of course). Thus there is no significant linear relationship. But had the estimated correla-

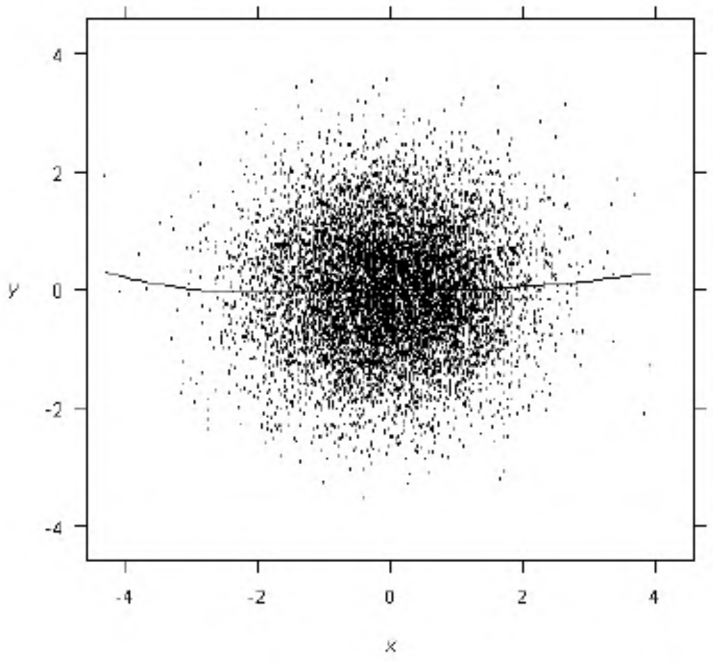

Figure 1. Scatterplot of two standard normal variables, x and y, of size 10,000 and with a correlation coefficient equal –0.0235. The locally weighted regression curve has been added to the plot to show the pattern of the relationship between the variables. Although it equals almost zero, the coefficient is significant at the probability level $P \leq 0.05$. Does it mean that the variables are negatively linearly dependent? The scatterplot and the loess curve show how far from true such a claim would be. So large a sample size makes the trivial correlation coefficient statistically significant, for which reason some might consider it 'important' or 'noticeable'.

Box 1. Standard output from *R*'s cor.test() function. The variables *x* and *y* from Figure 1 are significantly correlated at $P \leq 0.05$, although the estimator of the correlation coefficient equals –0.0235.

```
Pearson's product-moment correlation

data:  x and y
t = -2.3489, df = 9998, p-value = 0.01885
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
 -0.043065016  -0.003886488
sample estimates:
       cor
-0.02348477
```

tion equalled 0.20 and had it been based on a 100-element sample, it would be significant.

So, the point is that what one (usually) aims to is estimate the population correlation coefficient, and not necessarily test the null hypothesis on its zero value. The testing does depend on sample size. If one has a huge sample, anything will be significant; if the sample is small (five-element, say), then even $r = \pm 0.80$ will be non-significant at the 0.05 significance level. Testing usually does not provide any interesting information, although sometimes the testing might be of interest indeed – I suppose this is seldom the case. The estimator provides most of the information one needs. The rest of what one needs can be provided by the interval estimator to show how much confidence in the estimator one may have. If the sample is too small, the confidence interval will show it by its width. Note that for most estimators a sufficient way is to provide their standard errors; however, a confidence interval for the correlation coefficient is not distributed evenly on both sides of the estimate. So for this particular estimator it is best to provide the confidence interval. And usually this is all one needs when it is Pearson's population correlation coefficient between two normal variables one aims to estimate and interpret.

The above comments hold true also for statistical hypothesis testing in general[4,5]. Sample size strongly influences testing, and this needs to be remembered always when one verifies any statistical test. Indeed, statistical testing has been criticized for some time[6]; Quinn and Keough write, '(…) everything else being the same, larger sample sizes are more likely to produce a statistically significant result and with very large sample sizes, trivial effects can produce a significant result'[6]. For correlation coefficient, whether this is testing indeed what one is interested in should be decided on a case-by-case basis, but one should be cautious that testing is not as often of interest as it is usually thought so.

Fisher and Switzer[6] ask, 'Is a picture worth 100 tests?'. It can be.

1. Reese, R. A., *Significance*, 2004, **1**, 39–40.
2. Cleveland, W. S., *J. Am. Stat. Assoc.*, 1979, **74**, 829–836.
3. R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2007; ISBN 3-900051-07-0, http://www.R-project.org
4. Rigby, A. S., *Health Educ. Res.*, 1999, **14**, 713–715.
5. Quinn, G. P. and Keough, M. J., *Experimental Design and Data Analysis for Biologists*, Cambridge University Press, Cambridge, 2002.
6. Fisher, N. I. and Switzer, P., *Am. Stat.*, 2001, **55**, 233–239.

Marcin Kozak is in the Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences, Nowoursynowska 159, 02-787 Warsaw, Poland.
e-mail: m.kozak@omega.sggw.waw.pl