What is Strong Correlation?

KEYWORDS:
Teaching;
Interpretation;
Rule of thumb.

Marcin Kozak
Warsaw University of Life Sciences, Poland.
e-mail: nyggus@gmail.com

Summary
Interpretation of correlation is often based on rules of thumb in which some boundary values are given to help decide whether correlation is non-important, weak, strong or very strong. This article shows that such rules of thumb may do more harm than good, and instead of supporting interpretation of correlation – which is their aim – they teach a schematic approach to statistics. Therefore they should be avoided in a statistics course.

◆ INTRODUCTION ◆

What correlation is strong? Is it $|\rho| = 0.70$? Or maybe $|\rho| = 0.50$? And $|\rho| > 0.70$ indicates very strong correlation, doesn't it? And there is weak correlation, when $|\rho|$ lies in an interval from 0.20 to 0.50, and non-important correlation, that lower than 0.20. Is that right?

The world is not that easy. Correlation cannot be contained within such frames. Despite this, people – including some teachers of statistics – sometimes try to *facilitate* the world by such frames, being nothing more than subjective rules of thumb, to support understanding correlation. The aim of this article is to show that this is wrong and that correlation should not be thought of as such a simple general interpretation tool: its interpretation should always be linked to a problem it describes.

We will discuss population correlation and take no notice of sample correlation – interpretation of the latter aims to be as close to the interpretation of the former as possible. Thereby we will omit the problems of sample size (Kozak 2009) and significance (Reese 2004) of a correlation coefficient.

◆ CORRELATION *FRAMES* ◆

To use strict frames for the correlation coefficient is thought of by some as facilitation, especially in teaching. To show it is a wrong approach is easy, although experience shows that convincing *frame-users* that this is so is rather difficult. The example of a general *theory* of strict frames (rules) for correlation may be presented as follows.

$0 \le |\rho| \le 0.20$: the correlation is non-important;
$0.20 < |\rho| \le 0.50$: the correlation is weak;
$0.50 < |\rho| \le 0.70$: the correlation is strong;
$|\rho| > 0.70$: the correlation is very strong.

The boundary points of these frames may be changed subjectively, and indeed they often are. Let us take no notice here of the obvious problem – commonly found in many statistical situations, not only in correlation – of values being close to a boundary point (i.e. $\rho = 0.49$ would be considered weak whereas $\rho = 0.50$ strong): this likely has no easy solution. The aim here is to convince a reader that the frame-approach to interpretation of correlation is in general wrong (even though it may have sense in particular correlation problems).

The main point of this article is that correlation has always to be linked to a problem it is applied to. Thus, in one situation, $\rho = 0.50$ may be thought of as very strong whereas, in another, as very weak. Consider the two following situations.

1. If a correlation between two variables should be non-existent, any correlation different from zero may be thought of as unexpected and thus important.
2. If a correlation between two variables should be very strong, say near 1, ρ below it, even close to 0.90 (which is usually considered very strong), may be thought of as weak.

By the *should be*, in these situations I mean that knowledge of the actual process suggests the value that should be expected for the correlation. To say that a correlation is indeed very strong, or that it is indeed non-important, is a researcher's (who should be an expert) task. There is no way to point out the general boundary points for a particular problem, which could be used any time.

The two situations show that there is no generality in interpretation of correlation. Let us consider some particular examples that illustrate these situations.

Example 1 (situation 1). Consider correlation between the results of drawing a random number (say from 0 to 100) by two persons using the same computer programme.

Example 2 (situation 1). Consider correlation between IQ of a person and average volume of tea this person drinks a day; let us limit the population to Poles. A positive correlation would mean that the more tea one drinks, the higher IQ one has, and vice versa: the higher IQ one has, the more tea one drinks.

Example 3 (situation 2). Consider two bathroom scales and the correlation between weights of people obtained from them.

Example 4 (situation 2). Consider a set of objects, say leaves, of length from 10 cm to 100 cm. Consider correlation between the objects' measurements with two rulers: one that has the 1-mm scale and the other the 1-cm scale.

In examples 1 and 2 we would expect no correlation, so any correlation different from zero (either positive or negative) would be unexpected and hence should not be called non-important. In examples 3 and 4, on the other hand, we would expect the correlation to be as high as possible, close to 1, and any correlation noticeably lower than 1, even such *strong* (as would usually be called) correlation as 0.90, might be thought of as weak.

These examples are, of course, just a drop in the ocean of similar examples we could use, but they should show clearly that correlation should not be simply contained within some artificial frames. Such frames, or rules, may sometimes be useful, but in some situations, like those given above, a rule of thumb used thoughtlessly may do more harm than good.

◆ CONCLUSION ◆

Rules are good in teaching, but only when they are correct. Rules for correlation, those *frames* this article discusses, that are often taught are not correct – there are so many situations to which the rules do not apply that we cannot simply call them exceptions.

The main point and conclusion of this article is that students should not be taught any limits that are supposedly to help interpret correlation. Statistics is, among others, a tool of interpretation, in which logic plays an important role, so let students think a little about a correlation problem when it comes to interpreting it. Had they simply applied any general rule of thumb, their interpretation could have become totally senseless.

Acknowledgements

I wish to thank Dr Anna Rajfura, Dr Dariusz Gozdowski and Mr Jakub Paderewski as well as my other colleagues from the Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences, and Dr. Stan Lipovetsky from GFK Custom Research North America for discussions on the topic of this article.

References

Kozak, M. (2009). How to show that sample size matters. *Teaching Statistics*, **31**(2), 52–54.

Reese, R.A. (2004). Does significance matter? *Significance*, **1**(1), 39–40.