Inference for correlations

Casper Albers*

Last update: June 19, 2015

Introduction. In the first year, it was taught how to interpret the correlation coefficient ρ and how to compute the sample correlation coefficient r. In short, ρ measures the strength and direction of the linear relation ('co-relation') between two variables (x and y, say). No relation corresponds to $\rho = 0$, a perfect positive relation to $\rho = 1$, and a perfect negative relation to $\rho = -1$. For a given sample of paired observations x_i and y_i ($i = 1, \ldots, n$), the sample correlation coefficient r is computed via

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} \right).$$

In case you need to refresh your memory on the computation and interpretation of correlations, please reread your first year material on these topics (e.g. Moore, McCabe & Craig, Section 2.3). You will need a good understanding of these basics in order to understand the following.

In this text, the basics of inference for correlation(s) is discussed. A few references to technical notes are placed (in the format $^{\text{TN}x}$). At the end of the document, these technical notes are listed. These notes serve as additional explanation and are not part of the exam material. Depending on your mathematical proficiency, you might find these notes either helpful or confusing. Feel free to skip these notes if they confuse you.

Each section closes with a few basic exercises. At the end of the document, the solutions to the exercises are presented.

Hypothesis testing for ρ. The computation of the correlation coefficient r tells something about the strength and direction of the linear relation between x and y in the sample. The next step is relating what this can tell us about the population correlation. Of special interest is the null hypothesis H_0: $\rho = 0$, versus either the two sided alternative H_A: $\rho \neq 0$, or a one sided alternative.

As you (should) know, there is a strong relation between the correlation coefficient ρ and the slope β_1 of the simple linear regression equation $y = \beta_0 + \beta_1 x + \varepsilon$; expression via

$$\beta_1 = \rho \frac{\sigma_y}{\sigma_x} \quad \Leftrightarrow \quad \rho = \beta_1 \frac{\sigma_x}{\sigma_y}.$$

From these equations, you can directly see that $\rho = 0$ if and only if $\beta_1 = 0$. Thus, testing H_0: $\rho = 0$ is equivalent to testing H_0: $\beta_1 = 0$ (against either a one- or two-sided alternative), performed through the test statistic

$$t = \frac{b_1}{SE_{b_1}}. \tag{1}$$

Under the null hypothesis, t follows the t-distribution with $n - 2$ degrees of freedom. An alternative formula$^{\text{TN1}}$ to compute the same value is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \tag{2}$$

Exercises Consider Table 1 which contains data on 8 patients suffering from agoraphobic problems. Here, x denotes the number of therapy sessions a patient has received, and y is the grade the patient gave himself for dealing with the problems after the last session (a higher grade indicates better capability of dealing with the problems). Use this data set throughout in the exercises.

1. Confirm (either by hand or by using software) that the regression equation $y = b_0 + b_1 x$ has

*University of Groningen; c.j.albers@rug.nl

x	1	2	3	4	5	6	7	8
y	4	2	6	5	6	8	9	6

Table 1: Data used in the exercises

$b_0 = 2.750$ and $b_1 = 0.667$; and that $SE_{b_1} = 0.243$. Interpret the value for b_1.

2. Compute (by hand) the correlation coefficient. Use $\bar{y} = 5.75$, $\bar{x} = 4.5$, $SD(y) = 2.188$, and $SD(x) = 2.450$.

3. Test, at $\alpha = 5\%$, H_0: $\rho = 0$ versus H_a: $\rho > 0$ using approaches (1) and (2). Explain why it is chosen to use the one-sided alternative. Interpret the outcome of the test.

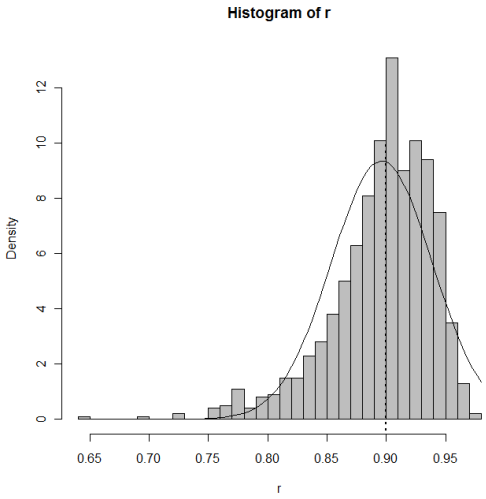

Figure 1: Histogram of the 1000 r-scores.

Confidence Intervals for ρ. For many parameters, such as the slope of a regression equation and the mean of a sample, confidence intervals (CIs) are constructed via

$$\text{(estimator)} \pm \text{(critical value)} \times \text{(standard error)}.$$

This construction of CIs relies, amongst others, on (approximate) normality of the sampling distribution of the estimator. In general, however, the distribution of r is not symmetric around ρ. This can best be made clear on basis of an example. From a given population with known correlation $\rho = 0.90$, 1000 random samples of size $n = 100$ have been drawn. For each of the 1000 samples, the correlation coefficient r has been computed. Figure 1 displays the histogram of the 1000 estimates for ρ. It is clear that this histogram is not symmetric: the 'tail' on the left hand side is heavier than that on the right.

This is understandable: By definition, a correlation coefficient can never be larger than 1. Thus, no estimate can deviate more than 0.10 to the right from $\rho = 0.90$. However, they can deviate more than 0.10 to the left from ρ and this indeed happens (in 36 of the 1000 cases).

To solve this problem and to be able to create a confidence interval after all, we use the so-called *Fisher z-transformation*, named after the 20th century statistician Sir R.A. Fisher. In short, this technique transforms all values to a 'parallel world' where the assumption of (approximate) normality of the sampling distribution *does* hold. In that parallel world, the standard technique to create a CI is employed, after which the transformation is applied 'the other way round' to obtain a CI in the 'real world'.

The Fisher z-transformation transforms an r-value into a new value, denoted by r' (sometimes it is also denoted by z or r_z) through[1]

$$r' = \tfrac{1}{2} \log\left(\tfrac{1+r}{1-r}\right).$$

A visualisation of this transformation is given in Figure 2. A few things can be observed: (i) whenever r is positive, r' will be positive (and similarly, negative r yield negative r'); (ii) there is a monotonic relation between r and r', i.e. every r-score corresponds to a unique r'-score, and vice versa; (iii) the r'-scores are not restricted to be inside the interval $[-1, 1]$.

Values of r closer to ± 1 will be affected more by the transformation than values close to 0. This ensures that, in our example, the right half of the plot, which was restricted to $[0.9, 1.0]$ now has a

[1]Recall that in statistics the convention is to always use the natural logarithm (ln) and denoting this as log, unless explicitly stated otherwise.

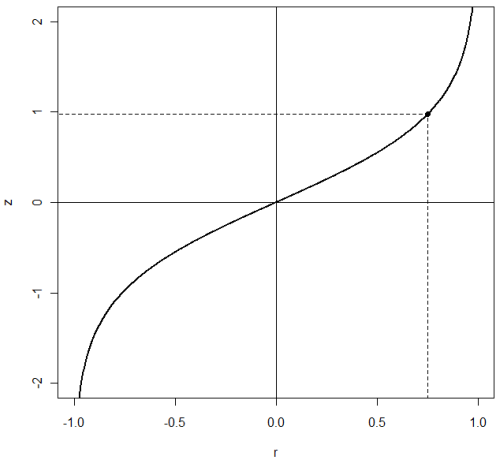

Figure 2: Fisher z-transformation.

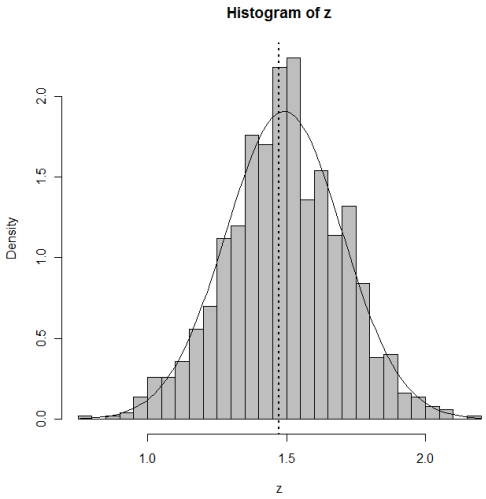

Figure 3: Histogram of the 1000 r'-scores.

wider range of values, whereas the left half of the plot isn't changed too much: It already had a wide range of values. In Figure 3, the same 1000 correlations from Figure 1 are shown again, but this time transformed into r'-scores. As you can see, now the histogram does look nicely symmetric. Indeed, Fisher showed that, after the z-transformation, the sampling distribution of the correlation coefficient is (approximately) normal[TN2], with mean ρ' (the Fisher z-transform of ρ) and variance $\frac{1}{n-3}$. This can be used to construct the confidence intervals.

The $(1-\alpha)$% CI for ρ_z is given by

$$\left(r' - z^*_{\alpha(2)}\frac{1}{\sqrt{n-3}}, r' + z^*_{\alpha(2)}\frac{1}{\sqrt{n-3}}\right)$$

where $z^*_{\alpha(2)}$ is the two-sided level-α critical value for the standard normal distribution, e.g. 1.96 when $\alpha = 0.05$. Now we have the CI for ρ', we are not finished yet: We are not interested in ρ', we only use that as a technical tool to obtain an interval for ρ. Thus, we need to transform the CI for ρ' to one for ρ. The inverse transformation of the Fisher z-distribution is

$$r = \frac{e^{2r'} - 1}{e^{2r'} + 1}.$$

By applying this inverse transformation to both the lower and the upper bound of the CI for ρ', one obtains the interval for ρ.

Exercises

4. Construct the 95% CI for ρ for the data of the previous exercises.

A test to compare two ρ's. Often in statistics, you want to compare different experiments and decide whether there are population differences or not. Suppose for instance, that you conduced an experiment to measure the correlation between the amount of time spent studying some material and the grade obtained for an exam based on the material. Suppose as well that there are two distinct groups (e.g. men/women, or those who visited training sessions/those who studied by themselves). Group A has sample size n_1 and correlation ρ_1, and group B has n_2 and r_2. It is of interest to test the hypothesis H_0: $\rho_1 = \rho_2$ versus H_a: $\rho_1 \neq \rho_2$ (or, perhaps, versus a one-sided alternative).

Also here, the Fisher z-transformation helps us out. Obviously, when $\rho_1 = \rho_2$ then also $\rho'_1 = \rho'_2$. Thus, testing H_0: $\rho_1 = \rho_2$ is essentially equivalent to testing H_0: $\rho'_1 = \rho'_2$. For this hypothesis the test statistic

$$Z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

can be used, and, using a table with critical values (or a computer), the z score can be converted into a p-value.

Exercises

5. Consider again the data in Table 1. For an adapted version of the therapy, data has been collected. In this study, $n = 12$ and $r = 0.500$. Test the null hypothesis of equality of correlations versus the two-sided alternative.

Conclusion. After working through this document, you should: (i) be able to construct hypothesis tests and confidence intervals for the correlation coefficient; (ii) be able to construct hypothesis tests for comparing correlation coefficients; (iii) understand why the Fisher z-transformation is needed.

Technical notes* Feel free to skip this section. It will not be examined. It can help, however, in gaining a better understanding of the material that is part of the exam material.

1. That both formulas give the same t-value can be seen from $b_1 = rs_y/s_x$ and

$$SE_{b_1} = \frac{1}{\sqrt{n-2}}\sqrt{\frac{s_y^2}{s_x^2} - b_1^2}.$$

Then,

$$\begin{aligned} t &= \frac{b_1}{SE_{b_1}} \\ &= \frac{r\frac{s_y}{s_x}}{\frac{1}{\sqrt{n-2}}\sqrt{\frac{s_y^2}{s_x^2} - b_1^2}} \\ &= \frac{\frac{s_y}{s_x}}{\frac{s_y}{s_x}}\frac{r\sqrt{n-2}}{\sqrt{1 - \frac{s_x^2}{s_y^2}b_1^2}} \\ &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \end{aligned}$$

2. The text provides a heuristic explanation as to why Fisher's transformation works. A full, detailed, and very technical explanation as to why exactly this transformation is the best one (and not any other transformation that affects high correlations more than near-zero ones), is not presented here. It can be found in Fisher, R.A. (1915), Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population, *Biometrika*, 10,
3. I haven't described the Bayesian approach to inference on correlations. The main reason is that I've restricted attention to approaches that can be computed manually which the Bayesian approach generally can't. In short, the Bayesian approach entails the following steps: (i) assume a bivariate normal distribution for x and y and impose (non-informative) priors on μ_x, μ_y, σ_x and σ_y; (ii) impose a prior on ρ. Common choices are the uniform $U(-1, 1)$ one and the symmetrized reference prior; (iii) through MCMC (and this is the step you don't want to do manually) estimate the posterior density and (iv) use this to obtain a credible interval for ρ or as a step towards computing a Bayes Factor.

Answers to exercises

1. These values are indeed correct. The slope, 0.667, can be interpreted as: 'for each additional session, the self-reported grade increases by two-thirds of a unit'.
2. $r = 0.746$.
3. The test is one sided, because one is only interested in therapies with a positive effect (mathematically it would be equivalent to test H_0: $\rho \leq 0$ vs H_a: $\rho > 0$). The approach based on the regression context provides

$$t = \frac{b}{SE_{b_1}} = \frac{0.667}{0.243} = 2.748,$$

the approach based on correlation provides

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.746\sqrt{6}}{\sqrt{1-0.746^2}} = 2.748.$$

They are indeed both the same. To obtain the p-value, look at a table with critical values (and df $= n-2 = 6$), to find $0.01 < p < 0.02$. When using software, the p-value can be computed more precisely as $p = 0.0167$. This is smaller than $\alpha = 0.05$ thus the null hypothesis is rejected. We conclude that there is a significant effect of therapy.

4. From $r = 0.746$ we have

$$r' = \tfrac{1}{2}\log\tfrac{1.746}{0.254} = 0.964.$$

Since we need the 95% interval, $z^*_{\alpha(2)} = 1.96$, thus the interval is

$$0.964 \pm 1.96/\sqrt{5} = 0.964 \pm 0.877 = (0.087, 1.841).$$

The inverse transformation of the lower bound is

$$r = \frac{e^{2 \cdot 0.087} - 1}{e^{2 \cdot 0.087} + 1} = \frac{1.190 - 1}{1.190 + 1} = 0.087$$

(note: for values r close to zero, it holds that $r \approx r'$, this explains why both r and r' coincide up to three decimal places). Similarly, the upper bound is computed via

$$r = \frac{e^{2 \cdot 1.841} - 1}{e^{2 \cdot 1.841} + 1} = \frac{39.726 - 1}{39.726 + 1} = 0.951.$$

Thus, the 95% CI is $(0.087, 0.951)$. This shows that, even though the correlation differs significantly from zero, the interval covers almost fully all positive values: We can not accurately tell what the value of the correlation is (which is obviously due to the small sample size).

5. For both studies, r' needs to be computed:

Study	n	r	r'
A	8	0.746	0.964
B	12	0.500	0.549

Next, compute

$$Z = \frac{0.964 - 0.549}{\sqrt{\frac{1}{5} + \frac{1}{9}}} = \frac{0.415}{0.558} = 0.746.$$

Next, use a table to find the corresponding p-value: $p = 0.46$. Thus, H_0 cannot be rejected.