Confidence in Correlation

doi:10.13140/RG.2.2.23673.49769

Gunnar Taraldsen

Department of Mathematical Sciences
Norwegian University of Science and Technology

November 11, 2020

Abstract

In 1895 Karl Pearson published his definition of the empirical correlation coefficient, but the idea of statistical correlation was anticipated substantially before this. Linear regression, and the associated correlation, is the principal statistical methodology in many applied sciences. We derive an explicit formula for the exact confidence density of the correlation. This can be used to replace the approximations currently in use.

Keywords: **confidence distributions; fiducial inference; correlation coefficient; binormal distribution; Gaussian law**

1 Introduction

The result of an experiment is given by four points with (x, y) coordinates $(773, 727)$, $(777, 735)$, $(284, 286)$, and $(519, 573)$. There are reasons a priori for assuming a linear relationship. This is further supported by Figure 1, and a high value for the coefficient of determination $R^2 = 97.00\%$. The R^2 equals the square of the empirical correlation $r = 98.49\%$. An approximate 95% one sided confidence interval for the correlation ρ based on the Fisher (1921) z-transformation is $[66.08, 100]\%$. Linear interpolation in the table presented by Fisher (1930, p.434) gives an exact 95% confidence interval $[67.42, 100]\%$. [1] Our Theorem 1, without linear interpolation, gives the true exact 95% confidence interval $[67.39, 100]\%$.

The previous analysis is probably familiar to many readers with the possible exception of the exact solution. Unfortunately, the exact solution by Fisher (1930)

[1] 66.4037 + (71.6298 - 66.4037)*(98.4893 - 98.4298)/(98.7371-98.4298) = 67.4156

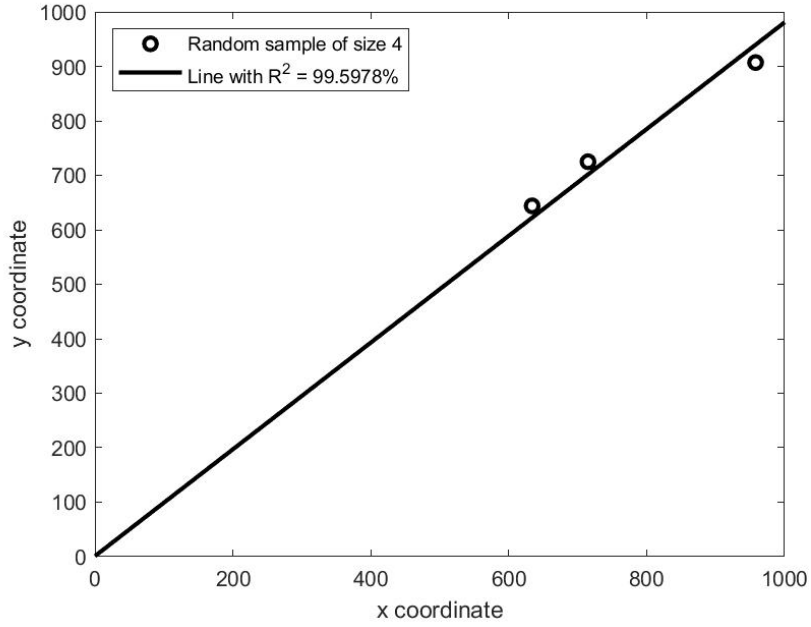

Figure 1: A sample of size 4 with a regression line.

seems to be essentially forgotten. It can, and should, be implemented in standard statistical software and practice. The purpose of this paper is to explain the necessary theory, and to expand on the analysis given by Fisher (1930). The main result is an explicit formula for the exact confidence density for the correlation. This can be seen as adding an important example to the theory of confidence distributions as presented by Schweder and Hjort (2016). It can also be seen as an important example of fiducial inference as formulated by Hannig et al. (2016) and others.

Much has been written on the correlation coefficient. A major source of inspiration for the presented proof is the above mentioned references and the work of Hotelling (1953). In stead of giving a more thorough introduction we refer to Rodgers and Nicewander (1988) and Rovine and von Eye (1997) which give further references and several different interpretations of the correlation.

2 Theory

The correlation ρ between two random variables X and Y equals the cosine of the angle β between $X - \mu_X$ and $Y - \mu_Y$ in the Hilbert space of finite variance random

variables. It is given by

$$\rho = \cos(\beta) = \frac{\langle X - \mu_X, Y - \mu_Y \rangle}{\sigma_X \sigma_Y} \tag{1}$$

exactly as for the calculus definition for vectors in $\mathbb{R}^2$. The inner product $\langle X, Y \rangle = \mathrm{E}(XY) = \int X(\omega)Y(\omega)\ \mathrm{P}(d\omega)$ defines $\|X\|^2 = \langle X, X \rangle$ and orthogonality $X \perp Y$ by $\langle X, Y \rangle = 0$. The mean μ_X and standard deviation σ_X of the random variable X equals the projection $\mu_X = \langle 1, X \rangle$ and the norm $\sigma_X = \|X - \mu_X\|$.

The reader may feel that the Hilbert space approach is unnecessarily abstract. It is, in fact, rather useful. The problem has now been reformulated into the problem of making inference regarding an angle between two vectors based on a random sample. The empirical correlation r for a random sample of size n is then, naturally, given by the cosine of the angle between the vectors $(x_i - \overline{x})$, $(y_i - \overline{y})$ in $\mathbb{R}^n$. This was the key geometrical idea when Fisher (1915) derived his explicit formula for the probability density of the empirical correlation. The Hilbert space approach is also very well described and motivated by Brockwell and Davis (1991) in their text on time series where the correlation function of a process is the main tool.

The best linear predictor $\hat{Y}$ of Y given X is the projection

$$\hat{Y} = \mu_Y + \rho\sigma_Y \frac{X - \mu_X}{\sigma_X} \tag{2}$$

of Y onto the subspace spanned by the orthonormal basis $\{1, (X - \mu_X)/\sigma_X\}$. Alternatively, equation (2) can be seen to correspond to the elementary definition of the cosine from a triangle in a plane. The angle in the triangle is given by the angle β between the vectors $X - \mu_X$ and $Y - \mu_Y$ spanning a two dimensional subspace. Equation (2) gives that the correlation can be interpreted as the slope of the best predictor line for standardized variables. This is possibly the most direct natural interpretation in applications. Furthermore, it can be generalized to give a similar interpretation for partial correlation.

If X and Y are jointly Gaussian, then $\hat{Y} = \mathrm{E}(Y \mid X)$, and the conditional law of Y given $X = x$ is Gaussian. This gives the link between equation (2) and ordinary regression

$$y_i = a + bx_i + \sigma v_i \tag{3}$$

where $v_1, \ldots, v_n$ is a random sample from the standard normal law. Comparison of equation (3) with equation (2) gives the constant term $a = \mu_Y - \rho\mu_X\sigma_Y/\sigma_X$, the slope $b = \rho\sigma_Y/\sigma_X$, and the conditional variance $\sigma^2 = (1 - \rho^2)\sigma_Y^2$. The binormal is hence parameterized alternatively by $(\mu_X, \sigma_X, a, b, \sigma^2)$. A random sample $((x_1, y_1), \ldots, (x_n, y_n))$ of size n from the binormal can be generated by equation (3)

where $x_i = \mu_X + \sigma_X u_i$ and $u_1, \ldots, u_n$ is a random sample from the standard normal law. Combining gives

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} + \begin{bmatrix} \sigma_X & 0 \\ \rho\sigma_Y & \sqrt{1-\rho^2}\sigma_Y \end{bmatrix} \cdot \begin{bmatrix} u_i \\ v_i \end{bmatrix} \tag{4}$$

The data generating equation (4), and multivariate generalizations beyond Gaussian, is treated by Fraser (1968, 1979). This involves group actions, maximal invariants, and leads to optimal inference methods as demonstrated by Taraldsen and Lindqvist (2013). A particular consequence of equation (4), proved by Fraser (1964, p.853), is

$$\sqrt{u}\frac{\rho}{\sqrt{1-\rho^2}} - \sqrt{v}\frac{r}{\sqrt{1-r^2}} = z \tag{5}$$

where $u \sim \chi^2(\nu), v \sim \chi^2(\nu-1), z \sim \mathrm{N}(0,1)$ are independent.

Equation (5) gives the law of ρ when r is known. The degrees of freedom $\nu = n-1$ for sample size n. With known mean the $\nu = n$, and r is the cosine of the angle between the vectors $(x_i - \mu_X)$, $(y_i - \mu_Y)$ in $\mathbb{R}^n$. The following result holds for any real $\nu > 1$ as a consequence of equation (5).

Theorem 1. *Let r be the empirical correlation of a random sample of size n from the binormal. The confidence density for the correlation ρ is*

$$\pi(\rho \,|\, r, \nu) = \frac{\nu(\nu-1)\Gamma(\nu-1)}{\sqrt{2\pi}\Gamma(\nu+\frac{1}{2})}(1-r^2)^{\frac{\nu-1}{2}} \cdot (1-\rho^2)^{\frac{\nu-2}{2}} \cdot (1-r\rho)^{\frac{1-2\nu}{2}} F(\frac{3}{2}, -\frac{1}{2}; \nu+\frac{1}{2}; \frac{1+r\rho}{2})$$

where F is the Gaussian hypergeometric function and $\nu = n-1 > 1$.

Proof. The idea is that equation (5) gives the conditional densities, and hence the marginal densities after integration over u, v. This integration is done by a change of variables resulting in a gamma integral and the above density. The details are as follows.

The conditional density of s given u, v is normal by equation (5) with $(s \,|\, u, v) \sim \mathrm{N}(\sqrt{\frac{v}{u}}t, 1/u)$. Using this, the law of u, v, and $ds = (1-\rho^2)^{-3/2}d\rho$ give the joint density of ρ, u, v as

$$(1-\rho^2)^{-3/2} \cdot \frac{u^{\frac{\nu}{2}-1}e^{-\frac{u}{2}}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} \cdot \frac{v^{\frac{\nu-1}{2}-1}e^{-\frac{v}{2}}}{2^{\frac{\nu-1}{2}}\Gamma(\frac{\nu-1}{2})} \cdot \sqrt{\frac{u}{2\pi}} e^{-\frac{u}{2}(s-\sqrt{\frac{v}{u}}t)^2} \tag{6}$$

The terms in the exponential are

$$-\frac{1}{2}\left[\frac{u}{1-\rho^2} - \frac{2\sqrt{uv}\rho r}{\sqrt{(1-\rho^2)(1-r^2)}} + \frac{v}{1-r^2}\right] = -\frac{\nu(s_1^2 - 2s_1s_2r\rho + s_2^2)}{2(1-r^2)} \tag{7}$$

using new coordinates (s_1, s_2) defined by $\nu s_1^2 = u(1-r^2)/(1-\rho^2)$ and $\nu s_2^2 = v$. Let $s_1 = \sqrt{\alpha}\exp(-\beta/2)$ and $s_2 = \sqrt{\alpha}\exp(\beta/2)$. The density for ρ, α, β from equation (6) is

$$\frac{2^{1-\nu}\nu^\nu}{\sqrt{\pi}\Gamma(\frac{\nu}{2})\Gamma(\frac{\nu-1}{2})}(1-r^2)^{-\frac{\nu+1}{2}}(1-\rho^2)^{\frac{\nu-2}{2}}e^{-\beta}\alpha^{\nu-1}e^{\frac{-\nu\alpha(\cosh(\beta)-\rho r)}{1-r^2}} \tag{8}$$

Integration over α gives $\pi(\rho\,|\,r,\nu)$ using the identity $\pi(\nu-2)! = \sqrt{\pi}2^{\nu-2}\Gamma(\frac{\nu}{2})\Gamma(\frac{\nu-1}{2})$ and adjusting an integral representation of F (Olver et al., 2010, 14.3.9,14.12.4). □

The Fisher (1921) z-transformation argument implies

$$\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}) - \frac{1}{2}\ln(\frac{1+r}{1-r}) \approx z/\sqrt{\nu-2} \tag{9}$$

Replacing equation (5) with this gives the z-transform approximate confidence density

$$\tilde{\pi}(\rho\,|\,r,\nu) = \sqrt{\frac{\nu-2}{2\pi}}(1-\rho^2)^{-1}e^{\frac{2-\nu}{8}\left[\ln(\frac{(1+\rho)(1-r)}{(1-\rho)(1+r)})\right]^2} \tag{10}$$

for $\nu > 2$.

3 Examples

Fisher (1930, p.534) considers the case with an observed correlation $r = 99\%$ from a sample of size $n = 4$. Fisher, relying on calculations of Miss F. E. Allan, states that the corresponding 5% ρ is equal to about 76.5%. Using equation (5) on these ρ and r values confirms this.

Consider next the data presented in Figure 1. Theorem 1 applied on these data confirms the calculations giving the stated confidence intervals. More complete information is given by the confidence densities shown in Figure 2. The empirical correlation is $r = 0.9849$. The exact confidence density in Figure 2 illustrates the corresponding uncertainty corresponding to all possible confidence intervals with all possible confidence levels.

Figure 3 shows the cd4 counts for 20 HIV-positive subjects (Efron, 1998, p.101). The x-axis gives the baseline count and the y-axis gives the count after one year of treatment with an experimental antiviral drug. The empirical correlation is $r = 0.7232$, and the equitail z-transform 90% approximate confidence interval is $[47.41, 86.51]\%$. Figure 4 shows the closeness of the confidence density and the z-transform density. The exact equitail 90% confidence interval from Theorem 1 is $[46.54, 85.74]\%$. It is shifted to the left as also can be inferred from Figure 4.

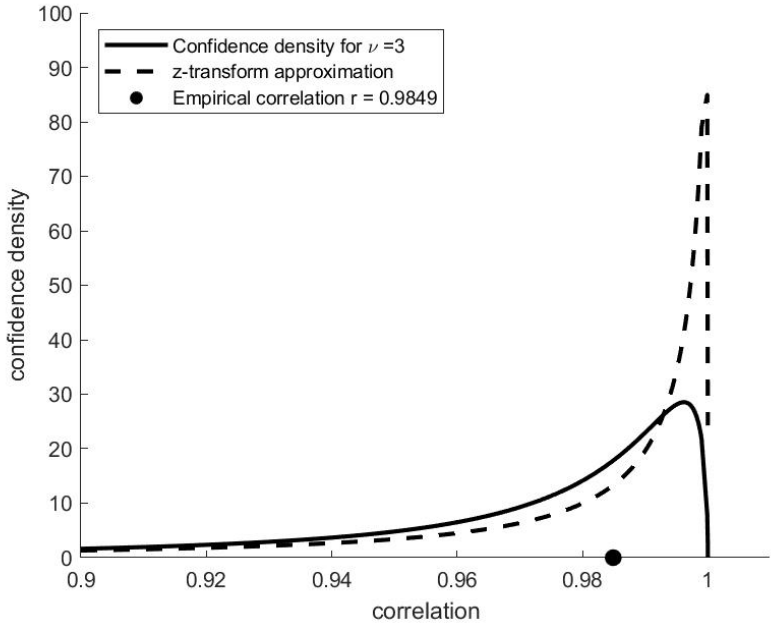

Figure 2: The confidence density and the z-transform density.

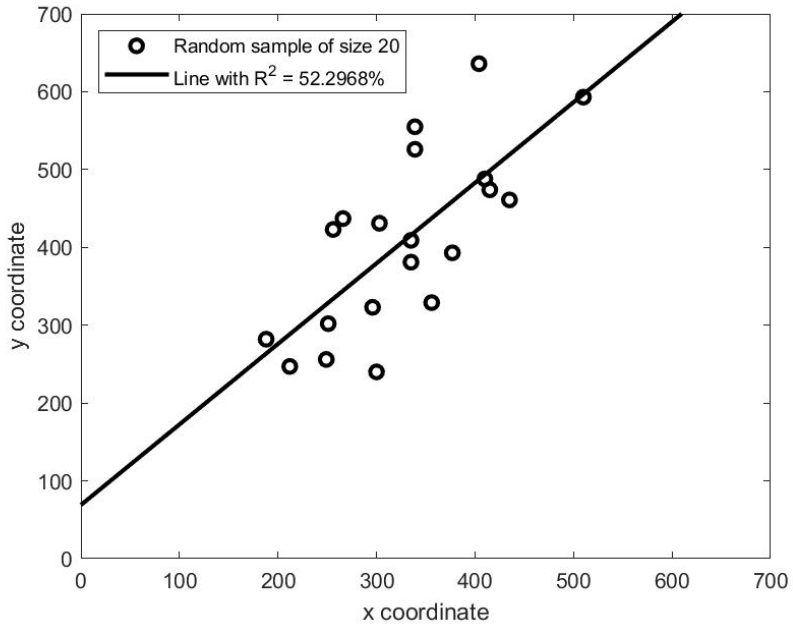

Figure 3: The cd4 data of DiCiccio and Efron (1996, Table 1).

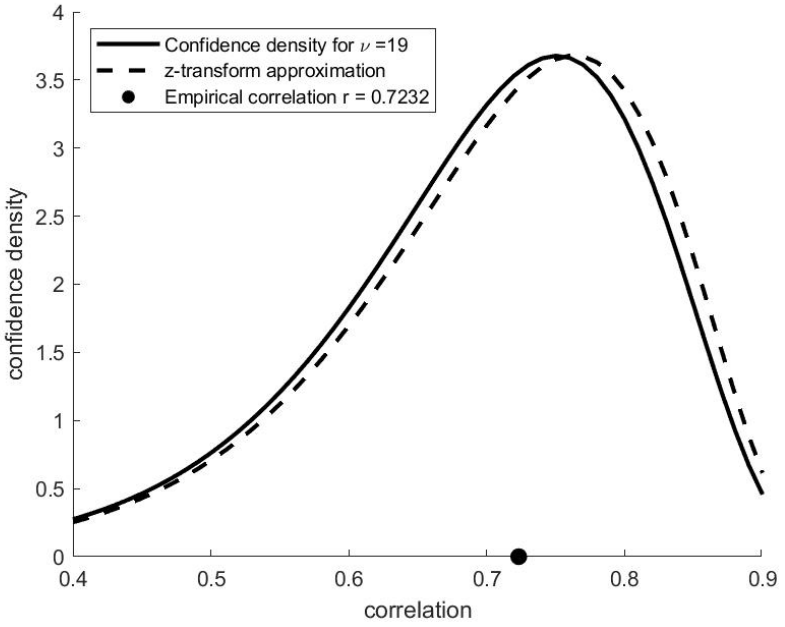

Figure 4: The confidence density and the z-transform density.

4 Conclusion

The z-transform has been historically convenient since confidence intervals can be calculated directly from tables of the standard normal distribution. Today, there seems to be little reason for using this in stead of the exact result in Theorem 1 since the hypergeometric function is implemented in standard numerical libraries.

References

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (Second ed.). Springer Series in Statistics. New York: Springer-Verlag.

DiCiccio, T. J. and B. Efron (1996). Bootstrap Confidence Intervals. *Statistical Science 11*(3), 189–212.

Efron, B. (1998). R. A. Fisher in the 21st century (Invited paper presented at the 1996 R. A. Fisher Lecture). *Statistical Science 13*(2), 95–122.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficent in samples from an indefinitely large population. *Biometrika 10*, 507–21.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron 1*(4), 1–32.

Fisher, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc. 26*, 528–535.

Fraser, D. A. S. (1964). On the definition of fiducial probability. *Bull. Int. Statist.Inst 40*, 842–856.

Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley.

Fraser, D. A. S. (1979). *Inference and Linear Models*. McGraw-Hill.

Hannig, J., H. Iyer, R. C. S. Lai, and T. C. M. Lee (2016). Generalized Fiducial Inference: A Review and New Results. *Journal of the American Statistical Association 111*(515), 1346–1361.

Hotelling, H. (1953). New Light on the Correlation Coefficient and its Transforms. *Journal of the Royal Statistical Society. Series B (Methodological) 15*(2), 193–232.

Olver, F. W. J., D. W. Lozier, R. F. Boisvert, and C. W. Clark (Eds.) (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press.

Rodgers, J. L. and W. A. Nicewander (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician 42*(1), 59–66.

Rovine, M. J. and A. von Eye (1997). A 14th Way to Look at a Correlation Coefficient: Correlation as the Proportion of Matches. *The American Statistician 51*(1), 42–46.

Schweder, T. and N. L. Hjort (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press.

Taraldsen, G. and B. H. Lindqvist (2013). Fiducial theory and optimal inference. *Annals of Statistics 41*(1), 323–341.