A Simple Rule of Thumb for Statistically Significant Correlation

Dennis Walsh, Middle Tennessee State University

The correlation ρ between two normal random variables is considered statistically significant if the sample correlation coefficient *r* is sufficiently large or sufficiently small. We present an easy rule of thumb for detecting statistical significance at the .05 level of significance. Specifically, we will show that if $|\,r\,| > 2/\sqrt{n}$, then the correlation is statistically significant at approximately the .05 level of significance.

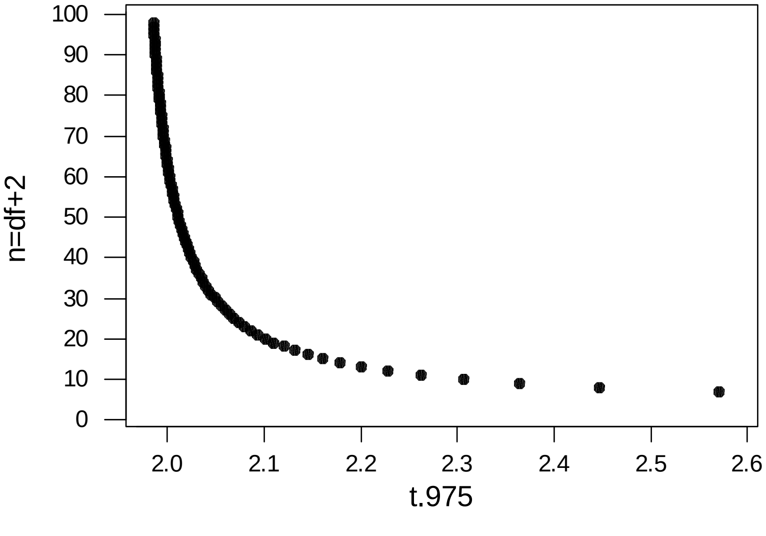
I. Approximating a Quantile from Student's T Distribution

Our result is based on the observation that the .975 quantile of a t distribution with $n-2$ degrees of freedom is closely approximated by a function of n. Specifically, for $n \geq 5$,

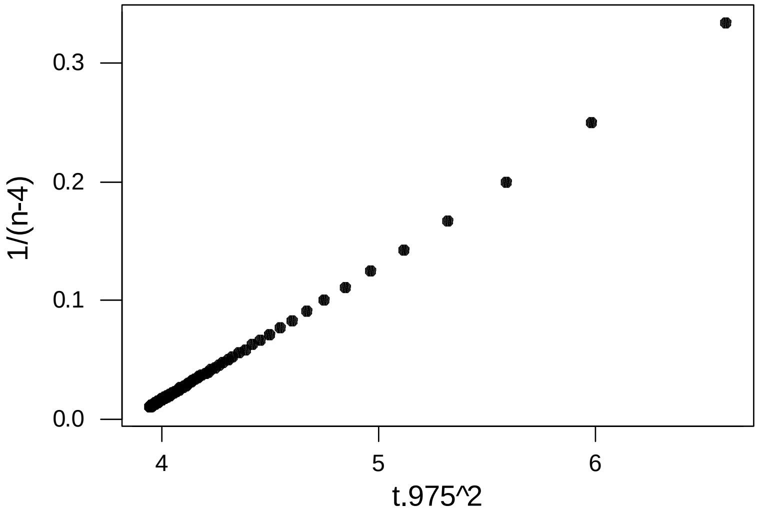
$$t \approx \sqrt{\tfrac{4(n-2)}{n-4}} \qquad (1)$$

where t satisfies $P(T \leq t) = .975$ when T has a $t(n-2)$ distribution. We illustate the derivation of this approximation below.

First observe the plot of n versus the $.975$ quantiles of the $t(n-2)$ distribution.



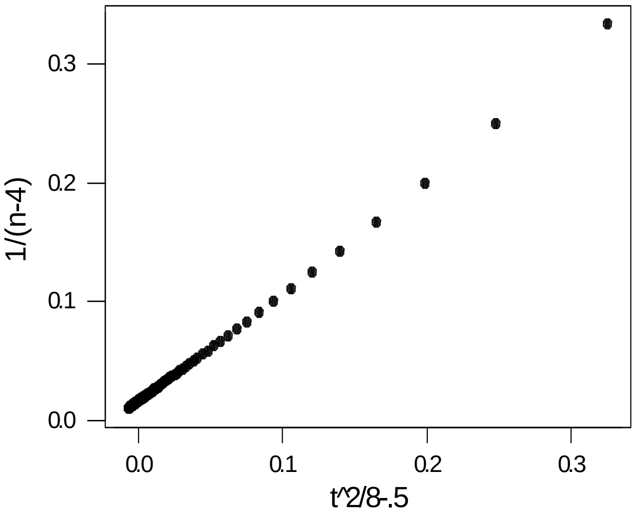
A transformation of both variables above will help linearize the plot. After some fiddling around, we let the dependent variable be $1/(n-4)$ and let the independent variable be $(t_{.975})^2$. A plot of these new variables is shown below.



We now regress $y = 1/(n-4)$ on the variable $x = (t_{.975})^2$ and obtain the fitted regression equation $y = -0.450 + 0.117\,x$. To simplify, we round -0.45 to $-.50$ and replace 0.117 with .125. Hence

$$\frac{1}{n-4} \approx -\frac{1}{2} + \frac{1}{8}t^2 \qquad (2)$$

as shown in the plot below.



Finally solving (2) for t gives us (1), that is, $t \approx \sqrt{\frac{4(n-2)}{n-4}}$.

II. Derivation of the Rule of Thumb

For a bivariate data set $\{(x_i, y_i) : i = 1, ..., n\}$, the sample correlation coefficient r is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} . \qquad (3)$$

If the data comes from a random sample of independent normal random variables X and Y, then the random variable T with observed value

$$t = r\left(\frac{n-2}{1-r^2}\right)^{1/2} \qquad (4)$$

has Student's t distribution with $n - 2$ degrees of freedom.

Therefore, when testing $H_o : \rho = 0$ versus $H_a : \rho \neq 0$ using level of significance $\alpha = .05$, we reject H_o if the test statistic $|t| = |r|\left(\frac{n-2}{1-r^2}\right)^{1/2}$ exceeds $t_{.975, n-2}$, the .975 quantile from the t distribution with $n - 2$ degrees of freedom.

To derive our rule of thumb, we proceed as follows. We reject H_o if

$$|r|\left(\frac{n-2}{1-r^2}\right)^{1/2} > t_{.975, n-2}$$

or, using the approximation in (1), if

$$|r|\left(\frac{n-2}{1-r^2}\right)^{1/2} > \sqrt{\frac{4(n-2)}{n-4}} \qquad (5).$$

Solving (5) for $|r|$ gives us $r^2(n-4) > 4(1-r^2)$ and, in turn, $r^2 > 4/n$. Thus, we reject the null hypothesis of zero correlation if $|r| > 2/\sqrt{n}$.

III. Accuracy of the Rule of Thumb

Below we compare the actual level of significance (when we use the rule of thumb) to the nominal level of $.05$. The actual level of significance is given by

$$\begin{aligned} \alpha &= 2P\left(|r| > 2/\sqrt{n}\right) \\ &= 2P\left(T > \sqrt{\frac{4(n-2)}{n-4}}\right). \end{aligned}$$

The table below gives the actual level of significance α, rounded to 3 decimal places, for the specified sample size n.

n	5	10	15	20	40	60	80	100
α	$.041$	$.050$	$.049$	$.048$	$.047$	$.046$	$.046$	$.046$

Note that the maximum error is $.50 - .041 = .009$ which occurs for $n = 5$. For $n > 100$, the level of significance remains at $\alpha = .046$. Hence, the rule of thumb is a bit conservative since the nominal level of .05 is actually slightly larger than the actual level.

IV. The Rule of Thumb Applied to Spearman's Coefficient

The correlation coefficient r is also known as Pearson's product-moment correlation coefficient. It is not the only correlation coefficient typically used. If the bivariate data consists of paired ranks or if the data is replaced by ranks because the populations sampled are not normally distributed, the most commonly used correlation coeffificient is Spearman's ρ, which is defined by

$$\rho = \frac{\sum_{i=1}^{n}(R(x_i)-n(n+1)/2)(R(y_i)-n(n+1)/2)}{\sqrt{\sum_{i=1}^{n}(R(x_i)-n(n+1)/2)^2\sum_{i=1}^{n}(R(y_i)-n(n+1)/2)^2}}.$$

We note that ρ is equivalent to r applied to the ranks of the data.

Not surprisingly, our rule of thumb works quite well for Spearman's ρ. The table below provides the critical values (for a two-sided test using level of significance $\alpha = .05$) based on our rule of thumb, on the $t-$distribution, and on Spearman's coefficient .

n	5	10	20	40	60	80	100
$2/\sqrt{n}$	$.894$	$.632$	$.447$	$.316$	$.258$	$.224$	$.200$
$\frac{t}{\sqrt{n-2+t^2}}$	$.878$	$.632$	$.444$	$.312$	$.254$	$.220$	$.197$
Spearman' cv	.900	.636	.445	$(.314)$	$(.255)$	$(.221)$	$(.197)$

V. The Rule of Thumb Applied to Paired Binary Data

The simplest kind of ranked or categorical data is binary data which takes on values 0 or 1. Suppose a bivariate random sample produced the data set $\{(x_i,\, y_i) : x_i$ and $y_i \in \{0,1\},\, i = 1,2,...,n\}$. Let a denote the number of (0,0) pairs, b the number of (0,1) pairs, c the number of (1,0) pairs, and d the numberof (1,1) pairs. Then, after a little algebra, Pearson's r can be expressed in the equivalent form given by

$$r = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}},$$

is sometimes called the phi coefficient and denoted by ϕ. Once again, if $|\,r\,| > 2/\sqrt{n}$, we can say that r is statistically significant at the approximate .05 level of significance, provided that n is sufficiently large.

Our justification is now based on a chi-squared test for independence. Let Q be defined by $Q \;= nr^2$. Then

$$Q = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

the familiar chi-squared test statistic used to test for independence when the associated 2×2 contingency table is of the form

$$\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\,.$$

Note that a is the number of $(0,0)$ ordered pairs in the random sample, b is the number of $(0,1)$ pairs, c is the number of $(1,0)$ pairs, and d is the number of $(1,1)$ pairs.

The null hypothesis of independent variables is rejected when Q exceeds the 3.841, which is the .95 quantile of the chi-square distribution with 1 degree of freedom. But $Q > 3.841$ implies $|\,r\,| > \sqrt{3.841/n} \approx 2/\sqrt{n}$. When n is large, the level of significance for our rule of thumb is in the ballpark of .046, based on the approximate chi-squared distribution of Q.

To get exact levels of significance requires extensive counting. For example, when n is 5, the contingency tables that make $r > 2/\sqrt{5}$ are the following:

4	0
0	1

1	0
0	4

0	4
1	0

0	1
4	0

3	0
0	2

2	0
0	3

0	3
2	0

0	2
3	0

.

Under the null hypothesis of independence, the probability of obtaining table $\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}$ is the multinomial probability $\frac{n!}{a!\,b!\,c!\,d!}\; p_{11}^a p_{12}^b p_{21}^c p_{22}^d$, where $p_{11} = P(X = 0,\, Y = 0)$,

$p_{12} = P(X = 0,\, Y = 1),\; p_{21} = P(X = 1,\, Y = 0)$, and $p_{22} = P(X = 1,\, Y = 1)$. Since the p_{ij} are unknown, we adopt the conservative approach and assign the value 1/4 to each, insuring that the true level of significance α is no greater than that calculated. Thus, for $n = 5,$ we have

$$\begin{aligned} \alpha &= P(r > 2/\sqrt{5}\,|\, r = 0) \\ &= P(\text{obtaining one of the tables above, assuming independence}) \\ &\le\; 4 \cdot \tfrac{5!}{4!\;1!}\left(\tfrac{1}{4}\right)^5 + 4 \cdot \tfrac{5!}{3!\;2!}\left(\tfrac{1}{4}\right)^5 \\ &=\; .05859375 \end{aligned}$$

Similar calculations to those above give us $\alpha \le .03027$ when $n = 6,\; \alpha \le .0364$ when $n = 7,$ and $\alpha \le .062439$ when $n = 8.$