Robustness of Pearson correlation

28 May 2009

PHILLIP GOOD

Information Research, 205 W. Utica Ave., Huntington Beach CA 92648 USA
drgood@statcourse.com

The standard Pearson Correlation test provides exact significance levels regardless of the distributions from which the data are drawn. Its power is equivalent to that of the corresponding permutation test.

Keywords and Phrases: correlation, ordered dose response, parametric test, permutation test, distribution-free test, robust test.

Lehmann [1986; p248] has shown that if the variable pair \{X,Y\} is bivariate normal \(N(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)\), then a UMP test of the hypothesis \(\rho=0\) against the alternative \(\rho\neq 0\) is given by rejecting if \(\hat{\rho}\) is large as determined by tables of a related t-distribution. Note that this test is also invariant under linear transformations of the form \(X'=aX+b; Y'=cY+d\).

Still, in many practical situations, distribution-free permutation tests offer the advantage of being both exact (because they are distribution free) and more powerful than comparable parametric methods; for example, in the multivariate two-sample comparison of means using Hotelling’s-\(T^2\) [Good, 2005, Section 9.2.2], the k-sample univariate comparison of means [Good and Lunneborg, 2005] and the analysis of contingency tables [Mehta and Patel, 1980, 1983]. In others, the parametric test is both powerful and sufficiently robust against non-normality such that the permutation test offers little or no advantage. We sought to determine if this was also true in the case of tests for non-zero correlation.
After elimination of factors that are invariant under permutation, a distribution-free permutation test based on \( \hat{\rho} \) reduces to a test based on \( \sum x_i y_i \) the inner product of \( \hat{x} \) and \( \hat{y} \). This test is also invariant with respect to \( T \). Let \( Y \) denote a random variable whose distribution is the conditional distribution of \( Y \) given \( x \). This test is unbiased against all alternatives for which, given \( x > x' \), \( Y_x \) is stochastically larger than \( Y_{x'} \) [Lehmann, 1986; p252].

In Frank, Trzos, and Good [1978], we proposed the use of permutation methods based on \( \sum x_i y_i \) to test for an ordered dose response. We report here that the parametric version of the Pearson Correlation used when the data are known to be normal would be as effective regardless of the distributions of \( X \) and \( Y \).

To establish this, we conducted a series of simulations in which the data for the first variable \( X \) were generated from one of the following distributions:

1. Fixed (1, 2, ..., N)
2. Uniform (0,1)
3. Normal (0,1).
4. Contaminated normal, both because such mixtures of distributions are common in practice and because they cannot be readily transformed to normal distributions.
5. Weibull, because such distributions arise in reliability and survival analysis and cannot be readily transformed to normal distributions. A shape parameter of 1.5 was specified.

A temporary variable \( V \) was then generated from one of these same five distributions. For verifying that significance levels were exact for testing \( \rho = 0 \) against
\(|\rho|>0\), we set \(Y = V\). For comparing the power of the permutation and parametric test, we set \(Y = \rho X + (1 - \rho)V\), where \(0 < \rho < 1\). Note we can test \(\rho \leq \rho_0\) against \(\rho > \rho_0 > 0\), by first forming the set \(\{y'_i = y_i - \rho_0 x_i\}\).

In every instance, even with sample sizes as small as 5, the p-values associated with the permutation and parametric tests were identical to within the precision of the Monte Carlo simulation. That is, both tests were exact and their power was equivalent. For the benefit of those who may wish to replicate or extend our results, the R code we use is appended.

These results are not surprising. Chance [1986] has shown by geometric means that the distribution of the linear correlation coefficient 

\[
\hat{\rho} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]

\(\rho = \frac{\text{cov}(XY)}{\sqrt{\text{var}(X) \text{var}(Y)}} = 0\) is independent of the distributions of \(X\) and \(Y\) (providing that the corresponding variances exist and are finite).

In contrasting our findings with those of Kowalski [8] and Edgell and Noon [9], note that advances in computer technology meant we were able to run more simulations with more precision.

Phillip Good

REFERENCES

Chance, W.A., 1986, A geometric derivation of the distribution of the correlation coefficient \(|r|\) when \(\rho = 0\). Amer. Math. Monthly. 93, 94-98.


APPENDIX: R Computer Code

```r
simcor=function(sample_size,N,MC,p, rho){
  # set up counters for number of rejections at the p significance level
  cntA=0
  cntP2=0
  # generate N samples of the two variables
  for(i in 1:N){
    X=gen1(sample_size)
    Y= X*rho + (1-rho)*gen2(sample_size)
    # compute Pearson Correlation and check to see if it rejects
    ...
  }
}
```

```r
# example usage
simcor(100,1000,10,0.05,0.5)
```
d=cor.test(Y, X, method = "p", alternative = "g")
if (d[[3]]<=p)cntA=cntA+1
#compute Pitman Correlation and check to see if it rejects
if (permc(MC,Y,X)<=p)cntP2=cntP2+1
}
list(cntP2=cntP2, cntA=cntA)
}

#Compute p-value for permutation test
permc=function(MC,x,y){
  rho0 = sum(x*y)
  cnt= 0
  #select MC random rearrangements of the first sample
  for (i in 1:MC){
    pX = sample (x)
    rho = sum(pX*y)
    if (rho0 <= rho ) cnt=cnt+1
  }
  return (cnt/MC)
}

#sample code for generating data
gen1=function(sample_size){
  #generate sample from a mixture of normals
  data=rnorm(sample_size,1+2*rbinom(sample_size,1,0.3))
  return (data)
}
gen2=function(sample_size){
  #generate sample from a uniform distribution
  data=runif(sample_size,0,1)
  return (data)