

Distribution of Sample Correlation in the Presence of Censoring

Erica Cross, Youngstown State University
Faculty Mentor: Dr. Scott Linder, Ohio Wesleyan University

Background: In many applications, data are subjected to censoring.

Let $\begin{pmatrix} x \\ y \end{pmatrix} \sim$ bivariate normal $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$

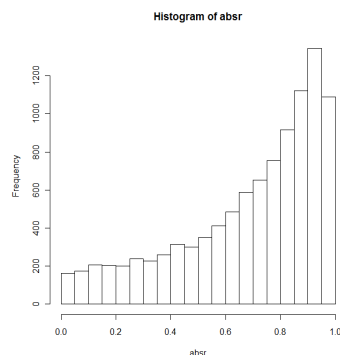
We start with n bivariate normal random vectors, but observe only those associated with the smallest p values of X .

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}, \dots, \begin{pmatrix} x_p \\ y_p \end{pmatrix}, \begin{pmatrix} x_{p+1} \\ y_{p+1} \end{pmatrix}, \begin{pmatrix} x_{p+2} \\ y_{p+2} \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

In this context, the sampling distributions of many ordinary, commonly used statistics are mathematically intractable.

Goal: Approximate the sampling distribution of the sample correlation, r .

Example: Simulation of 10,000 $|r|$'s with $\rho=0.9$, $n=20$, and $p=5$



Possible to approximate the sampling distribution of $|r|$ with a beta distribution. A beta distribution is bound between 0 and 1. A beta distribution is highly versatile through suitable selection of parameters.

To fit a $beta(\alpha, \beta)$ distribution to the sampling distribution of $|r|$, we need to do two things:

- Determine parameters α and β (as functions of experimental conditions n , p and ρ)
- Determine goodness of fit of the approximation.

Determining Parameters

First, we simulate 10,000 $|r|$'s under fixed experimental conditions (n , p , and ρ). Given this large collection of $|r|$'s, we then estimate the parameters α and β , using two methods.

Two Methods for Parameter Estimation:

- Method of moments
- Maximum likelihood estimation

Method of Moments (moment matching)

Find $\hat{\alpha}$ and $\hat{\beta}$ by matching the following:

$$E(x) = \frac{\alpha}{\alpha + \beta} \leftarrow \overline{|r|}$$

$$Var(x) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \leftarrow S_{|r|}^2$$

Maximum Likelihood Estimators

More consistent and generally has better properties than method of moments. Values of parameters were obtained numerically through Matlab.

Determining Goodness of Fit

Having determined parameters of α and β , we now hypothesize that $|r|$ has a $beta(\alpha, \beta)$ distribution. How well does this approximation fit?

We use the Chi-Square goodness of fit test statistic to measure lack of fit. First, we partition the interval (0,1) into 10 equi-probable "bins", determined by the hypothesized $beta(\alpha, \beta)$ distribution. Next, we simulate 2,000 new $|r|$'s. Under the hypothesis that $|r|$ has a $beta(\alpha, \beta)$ distribution, we would expect to see about 200 in each "bin".

$$\text{Let } \chi^2 = \sum \frac{(O - E)^2}{E}$$

Under the hypothesis, χ^2 should be about 9.

Example: Test statistic and p-value using estimated parameters when $n=20$ and $\rho=0.5$.

	MLE		MOM	
	5	12.17 (0.2038943)	16.67 (0.0541416)	
	7	22.64 (0.0070575)	9.84 (0.3635929)	
	9	38.45 (0.0000145)	33.59 (0.0001054)	
	11	67.71 (0.0000000)	33.00 (0.0001336)	
P	13	36.42 (0.0000334)	57.90 (0.0000000)	
	15	92.62 (0.0000000)	84.98 (0.0000000)	
	17	75.45 (0.0000000)	96.68 (0.0000000)	
	19	93.81 (0.0000000)	79.82 (0.0000000)	

test statistic
(p-value)

As evidenced by the table, both method of moments and maximum likelihood estimators provide poor overall goodness of fit.

We can also observe the percentile estimates to evaluate the goodness of fit at the lower and upper tails of the distribution using $n=20$ and $\rho=0.5$.

	T	0.01			0.05			0.95		
		B-MLE	E		T	B-MLE	E	T	B-MLE	E
7	0.00586	0.01124	0.918206	0.03375	0.04523	0.340113	0.81596	0.83596	0.024521	
11	0.00933	0.01526	0.634467	0.04998	0.05169	0.034164	0.74572	0.77695	0.041868	
15	0.01227	0.02871	1.339161	0.05606	0.07728	0.378584	0.70911	0.76923	0.084787	
19	0.02552	0.07609	1.982038	0.13108	0.14804	0.129443	0.7281	0.77816	0.068753	

	T	0.01			0.05			0.95		
		B-MOM	E		T	B-MOM	E	T	B-MOM	E
7	0.009495	0.012667	0.333994	0.036573	0.04892	0.337589	0.808306	0.837904	0.036617	
11	0.013622	0.01927	0.414652	0.039793	0.059803	0.50285	0.745551	0.771806	0.035215	
15	0.013628	0.036269	1.66129	0.062791	0.08931	0.422336	0.727788	0.760129	0.044437	
19	0.033005	0.094143	1.852419	0.121824	0.169259	0.389377	0.72877	0.7669	0.05232	

Results

Given that $|r|$ is distributed as $beta(\alpha, \beta)$ we can obtain a confidence interval for ρ by inverting this distribution.

Observed confidence levels for nominally 95% confidence intervals for ρ formed by inverting the approximating sampling distribution.

n=20	p	p		
		0.01	0.05	0.09
	6	0.9499	0.9352	0.8779
	12	0.9505	0.9195	0.8285
	18	0.9485	0.9434	0.9292

Using the approximate sampling distribution $|r| \sim beta(\alpha, \beta)$, we can determine combinations of n and p (experimental conditions) in order to test for independence of X and Y with specified power.

		True value of ρ			
		0.3	0.5	0.7	0.9
Power	0.4	N=20 P>=18	N=20 P>=12	N=20 P>=6	N=20 P>=6
	0.6	N=60 P>=52	N=60 P>=30	N=60 P>=16	N=60 P>=8
Power	0.6	N=20 P>=20	N=20 P>=14	N=20 P>=8	N=20 P>=8
	0.8	N=60 P>=60	N=60 P>=38	N=60 P>=22	N=60 P>=10
			N=20 P>=18	N=20 P>=10	N=20 P>=10
		N=60 P>=48	N=60 P>=30	N=60 P>=14	N=60 P>=14