0.1 Sample correlation

Consider two Gaussian random variables x and y distributed with the density

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_x\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_x)^2}{\sigma_1^2} - 2\frac{\rho(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right]\right) \tag{1}$$

where m_x and m_y are the expected values of x and y, respectively, σ_x and σ_y their standard deviations and ρ is the correlation coefficient between x and y. Now the statistics of a sample of N independent couples (x_i, y_i) extracted from the density of Eq. (1) are

$$\hat{m}_x = \frac{1}{N}\sum_{k=1}^{N} x_k \qquad \hat{m}_y = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{2}$$

$$\hat{\sigma}_x = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{m}_x)^2} \qquad \hat{\sigma}_y = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k-\hat{m}_y)^2} \tag{3}$$

$$\hat{\rho} = \frac{\sum_{k=1}^{N}(x_k-\hat{m}_x)(y_k-\hat{m}_y)}{N\hat{\sigma}_x\hat{\sigma}_y} \tag{4}$$

Fisher found the density function of the vector $(\hat{m}_x, \hat{m}_y, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho})$ describing the sample statistics of N variables. The density factorizes in the joint pdf $u(\hat{m}_x, \hat{m}_y)$ for the sample means and the joint pdf $v(\hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho})$ for the elements of the covariance matrix. After integrating v over the two standard deviations one finally obtains the density for the sample correlation coefficient

$$p(\hat{\rho}) = \frac{N-2}{\pi}(1-\rho^2)^{(N-1)/2}(1-\hat{\rho})^{(N-4)/2}\int_0^1 \frac{x^{N-2}}{(1-\rho\hat{\rho}x)^{N-1}\sqrt{1-x^2}}\,dx \tag{5}$$

This integral cannot be computed analytically but the table for the distribution of $\hat{\rho}$ as a function of N and ρ are given in many books for hypothesis testing. From Eq. (5) one can compute the approximate value of the mean and the variance of $\hat{\rho}$ which are

$$E\{\hat{\rho}\} \cong \rho, \qquad D^2\{\hat{\rho}\} \cong \frac{(1-\rho^2)^2}{N} \tag{6}$$

Unfortunately these expressions are valid only when N is large (say of the order of 500) and this is because the distribution for $\hat{\rho}$ is highly asymmetric.

Fisher discovered also that the random variable

$$U = \frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} \tag{7}$$

has, even for small N, approximately the normal distribution

$$p(U) = N\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(N-1)}; \frac{1}{\sqrt{N-3}}\right) \tag{8}$$

This result shows that (i) the mean sample correlation coefficient tends to overestimate the correlation coefficient, even if this bias decreases as N^{-1}, and (ii) the uncertainty on the sample correlation coefficient decreases as $1/\sqrt{N}$ as usual for other statistical samples (e.g. the mean).