

Tutorial # 4 - Calculating Significance by Monte Carlo

In the homework you should have found the correlation coefficients for some data. The Pearson Correlation Coefficient is

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_x^2 S_y^2}} \quad (1)$$

where

$$\bar{x} = \frac{1}{N} \sum_i x_i \quad S_x^2 = \frac{1}{(N-1)} \sum_i (x_i - \bar{x})^2 \quad (2)$$

A non-zero r_{xy} should indicate that there is a correlation between x and y . r_{xy} is an estimator of the populations true correlation coefficient

$$\rho_{xy} = \frac{C_{XY}}{\sqrt{\sigma_x^2 \sigma_y^2}} \quad (3)$$

But r_{xy} is a function of random variables so we would not expect it to be exactly 0 even if there were no correlations. We could try to calculate the probability distribution of r_{xy} analytically, but this might be difficult (more on this in lecture).

Monte Carlo

Instead lets find the significance of your measured r_{xy} by Monte Carlo given the hypothesis that x and y are not correlated.

- 1) Create two vectors (X and Y) of random normally distributed numbers with variance 1 and mean zero. Each vector should be 1000 elements long as was the data in the homework. Calculate r_{xy} for these.
- 2) Repeat step 1 a thousand times to get a distribution of r_{xy} .
- 3) Plot a histogram of your r_{xy} values.

4) What is the fraction of times $|r_{xy}|$ is larger than your measured value for the data in file homework_01_2d-datafile.csv ? Would you expect to find this if there were no correlation?

5) Do the exercise above over but this time use the measured S_x^2 , S_y^2 , \bar{x} and \bar{y} from homework_01_2d-datafile.csv to generate the X and Y variables. Plot the new histogram of r_{xy} over the old one. Is there any difference?

6) Order your sample of r_{xy} 's from smallest to largest. Take the i th value to be an estimate of the r_{xy} where $(N - i)/N$ of the probability distribution is larger than it. For example the 95% upper bound would be at $i/N = 0.95$. What is the 95% upper bound on r_{xy} if there is no correlation between X and Y ? We will call this $r_{0.95}$. In lecture we will find that the variance in this estimator is

$$\text{Var}[r_p] = \frac{2p(1-p)}{Nf(r_p)^2} \quad (4)$$

for large N where $f(r)$ is the pdf of r_{xy} . Assuming that r_{xy} is Gaussian distributed and its variance is the one you measure, what is the variance in your estimate of $r_{0.95}$?