Correlation analysis in automated testing

FOSDEM 2020

Łukasz Wcisło

3MDEB

Agenda

- Introduction
- Purpose
- Function definition & deviations
- Covariance matrix
- Pearson correlation coefficient
- Correlation Matrix
- Use-case

Introduction

Science may be described as the art of systematic over-simplification — the art of discerning what we may with advantage omit.

Karl Popper

Purpose

- Simplicity
- Time saving
- Logic
- Elegance

Function definition

Test result as a Boolean function, a relation between a release version and a result of a test.

Test/Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Probability
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0.5
4	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0.55
5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0.4
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	1	1	0	0	0.5
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0.4
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	1	1	0	0	0.5
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0.45
21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0.3
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0.3

Red - FAIL

Green - PASS

Function deviations

Instead of using expected value, we can use the probability.

Test/Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
4	0.45	0.45	-0.55	-0.55	-0.55	-0.55	-0.55	-0.55	-0.55	-0.55	-0.55	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
5	0.6	0.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	0.6	0.6	0.6	0.6	0.6
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0.5	0.5	0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	-0.5	-0.5	-0.5	0.5	0.5	-0.5	-0.5	0.5	0.5	-0.5	-0.5
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0.6	0.6	0.6	0.6	-0.4	-0.4	-0.4	-0.4	0.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	0.6	0.6
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0.5	0.5	0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	-0.5	-0.5	-0.5	0.5	0.5	-0.5	-0.5	0.5	0.5	-0.5	-0.5
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0.55	0.55	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	0.55	0.55	0.55	0.55	0.55	0.55	0.55
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0.7	0.7	0.7	0.7	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	0.7	0.7
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0.7	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	0.7	0.7	0.7	0.7	0.7

Covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Where

$\sigma_i^2 = D^2 X_i$

is a variance of variable X, and

$\sigma_{ij} = \mathbf{cov}(X_i, X_j)$

is a covariance between two standardized random variables.

(In our case - between two tests)

Covariance matrix 2

Covariance Matrix	3	4	5	9	12	16	20	23	25
3	0.0625	0.175	-0.05	-0.05	-0.05	-0.05	0.125	-0.05	0.01
4	0.175	0.0612562	0.18	0.025	0.03	0.025	0.2025	0.035	0.135
5	0.1	0.18	0.0576	0	0.09	0	0.22	0.08	0.18
9	-0.05	0.025	0	0.0625	0.05	0.25	0.025	0	0
12	-0.05	0.03	0.09	0.05	0.0576	0.05	0.07	0.18	0.08
16	-0.05	0.025	0	0.25	0.05	0.0625	0.025	0	0
20	0.125	0.2025	0.22	0.025	0.07	0.025	0.0612562	0.065	0.165
23	-0.05	0.035	0.08	0	0.18	0	0.065	0.0441	0.06
25	0.1	0.135	0.18	0	0.08	0	0.165	0.06	0.0441

We can extract meaningful tests for better performance. Diagonal contains variance of each test, covariance matrix is symmetric. Also, every covariance matrix is positive semi-definite.

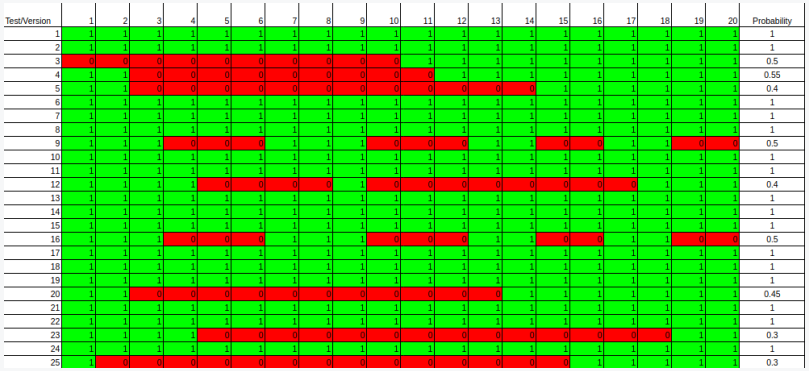
Pearson correlation coefficient

What brings us to Pearson correlation coefficient.

It is a covariance of two variables divided by the product of their standard deviations:

$$r_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m} P(X=x_i, Y=y_j) x_i y_j\right) - \overline{X}\,\overline{Y}}{\sqrt{\left(\sum_{i=1}^{n} P(X=x_i) x_i^2\right) - \overline{X}^2}\sqrt{\left(\sum_{i=1}^{m} P(Y=y_i) y_i^2\right) - \overline{Y}^2}}$$

Correlation Matrix

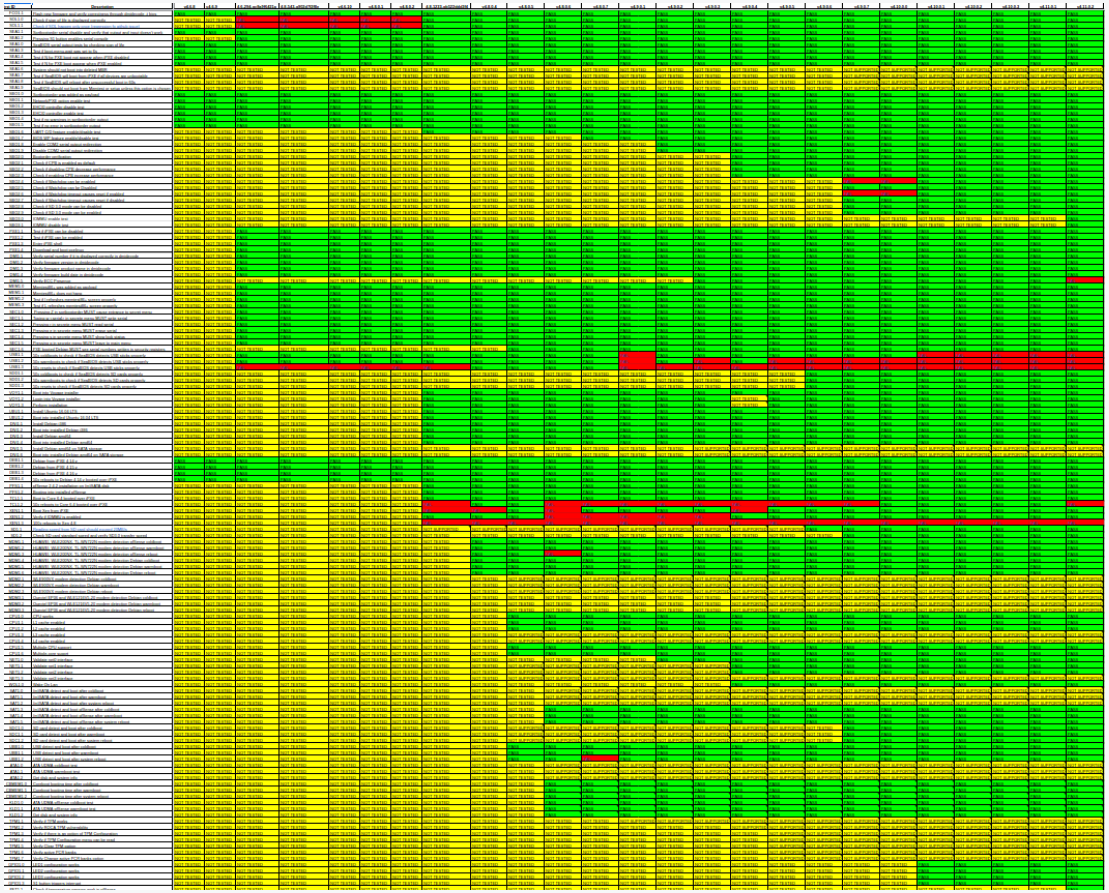


Test/Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Probability
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0.5
4	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0.55
5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0.4
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	1	1	0	0	0.5
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0.4
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	1	1	0	0	0.5
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0.45
21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0.3
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0.3

Correlation	3	4	5	9	12	16	20	23
4	0.70							
5	0.41	0.74						
9	-0.20	0.10	0.00					
12	-0.20	0.12	0.38	0.20				
16	-0.20	0.10	0.00	1.00	0.20			
20	0.50	0.82	0.90	0.10	0.29	0.10		
23	-0.22	0.15	0.36	0.00	0.80	0.00	0.29	
25	0.44	0.59	0.80	0.00	0.36	0.00	0.72	0.29

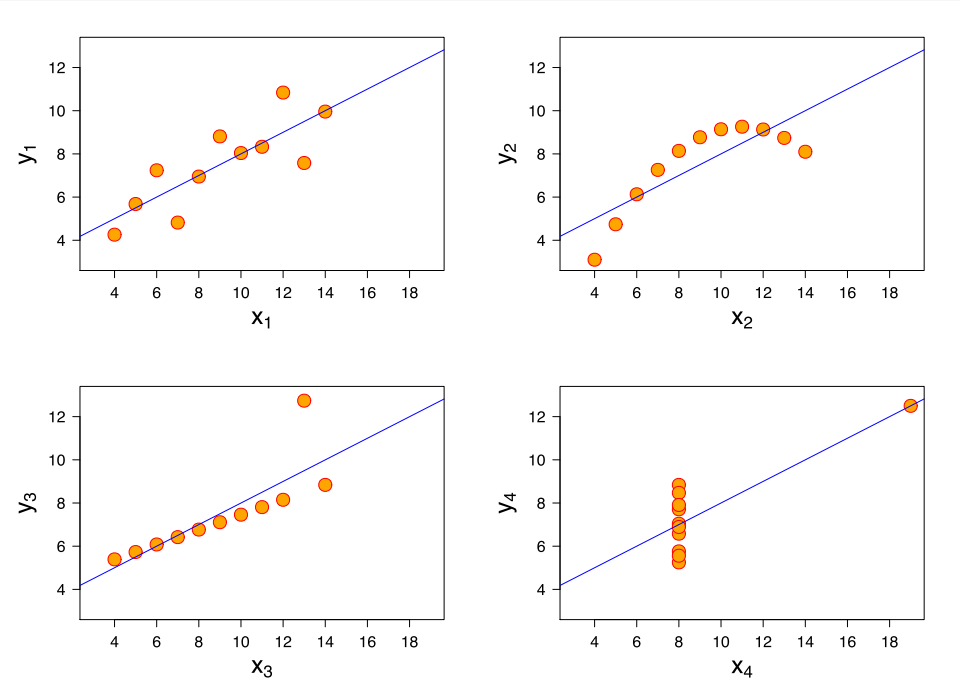
Where correlation is normalized and always stays between -1 and 1.

Actual use-case



Source

Anscombe's quartet



Mean of x, of y, variance of x, of y, correlation between x and y, linear regression and coefficient of determination of the linear regression are the same for each data set.

Bibliography

1. A. Buda and A.Jarynowski (2010) Life-time of correlations and its applications vol.1, Wydawnictwo Niezależne: 5–21, December 2010, ISBN 978-83-915272-9-0
2. W.J. Krzanowski: Principles of Multivariate Analysis. Nowy Jork: Oxford University Press, 2003, seria: Oxford Statistical Science. ISBN 0-19-850708-9.
3. Cox, D.R., Hinkley, D.V. (1974) Theoretical Statistics, Chapman & Hall (Appendix 3) ISBN 0-412-12420-3
4. Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966

3MDEB

Q & A

Thank you for your attention

"There are three kinds of lies: lies, damned lies, and statistics."

Benjamin Disraeli