Análise de correlação entre os dados sobre defeitos de software, extraídos do repositório de defeitos do SINAPAD.

Duas hipóteses foram levantadas, considerando avaliar se o aumento da porcentagem de defeitos apresentada em um projeto influencia no aumento do tempo médio gasto, em dias, para resolver problemas neste mesmo projeto:

- H *nula*: não existe correlação forte entre a porcentagem de defeitos de cada um dos projetos com a média de tempo gasto em dias na resolução de defeitos;
- H *alternativa*: existe correlação forte entre a porcentagem de defeitos de cada um dos projetos e a média de tempo gasto, em dias, na resolução de defeitos em um mesmo projeto.

Para aceitar uma das hipóteses, é realizado o cálculo do Coeficiente da correlação de Pearson, tal como descrito em [Levin, et. al, 2012].

Coeficientes de Correlação indicam intensidade e direção da correlação dos dados, expressos da seguinte forma:

- Até -1,00: Correlação Negativa Perfeita.
- Até -0,60: Correlação Negativa Forte.
- Até -0,30: Correlação negativa moderada.
- Até -010: Correlação negativa fraca.
- Até 0,00: Nenhuma correlação
- Até 010: Correlação positiva Fraca.
- Até 0,30: Correlação positiva Moderada.
- Até 0,60: Correlação positiva Forte.
- Até 0,60: Correlação positiva Forte.
- Até 1,00: Correlação positiva Perfeita.

Em resumo, valores negativos indicam correlação negativa e valores positivos, indicam correlação positiva. Quando mais próximo de -1,00 ou 1,00, maior a intensidade da correlação entre os dados.

Neste sentido, será calculado o coeficiente da correlação de Pearson entre Porcentagem e Tempo Médio. Para isto, será considerada a Tabela 1:

Projetos	Porcentagem	Tempo Médio
Sinapad-framework	51	38,2

Portais SINAPAD	15	61,5
Profager	7	1,25
LuaPZ	7	1
SINAPAD-WEB	5	69,3
CAM GRID	5	5,6
OSC	5	31,6
ACES3	3	1
GT-mc²: Minha Cloud	2	58

Os dados da Tabela 1 foram obtidos da seguinte forma:
A porcentagem de defeitos por projeto:

Porcentagem = Quantidade Defeitos em um projeto / Quantidade total de defeitos.

O Tempo Médio para resolução de defeitos em um projeto é dado por:

Tempo Médio =(Σ (Data da última modificação do registro do defeito - Data de início da resolução do defeito)) / quantidade de registros de defeitos em um projeto).

Agora, considera-se a Tabela 2, complementar a Tabela 1:

Projetos	(P) Porcentagem	(TM) Tempo Médio	A = (P – P')	B = (TM – TM')	A * B
Sinapad-framework	51	38,2	39,8	8,42	31,38
Portais SINAPAD	15	61,5	3,89	31,79	-27,9
Profager	7	1,25	-4,11	-28,46	24,35
LuaPZ	7	1	-4,11	-28,71	24,6
SINAPAD-WEB	5	69,3	-6,11	39,59	-45,7
CAM GRID	5	5,6	-6,11	-24,11	18
OSC	5	31,6	-6,11	1,89	-8
ACES3	3	1	-8, 11	-28,71	20,60
GT-mc²: Minha Cloud	2	58	-9,11	28,29	-37,4
	Σ(P = 100) Média = P' = 11,11	Σ(TM = 267,45) Média = TM' = 29,71			SP = $\Sigma[A * B] =$ 562,73

Onde A, que é o valor calculado do desvio entre duas variáveis, é obtido por:

A = Porcentagem (P) – Valor da Média (P’), que é igual ao resultado do somatório das 9 porcentagens obtidas divididas pela quantidade de porcentagens obtidas, que é 9.

Já B, que é o valor calculado do desvio entre duas variáveis, é obtido por:

B= Tempo Médio (TM) – Valor da Média (TM’), que é igual ao resultado do somatório dos 9 tempos médios obtidos, divididos pela quantidade de tempos médios obtidos, que também é 9.

O valor calculado dos desvios é dado pela multiplicação dos desvios de A e B.

SP é obtido através do somatório de todos os valores calculados na coluna 6 da Tabela 2.

Dado que o valor de SP é um valor positivo, isto indica que há uma associação dita positiva, entre Porcentagem e Tempo Médio. Resta agora descobrir a intensidade da correlação entre os dados. Considera-se então a Tabela 3.

Tabela 3. Continuação da Tabela 2.

A = (P – P’)	B = (TM – TM’)	A^2	B^2
39,8	8,42	1584,04	70,8964
3,89	31,79	15,1321	1010,6041
-4,11	-28,46	16,8921	809,9716
-4,11	-28,71	16,8921	824,2641
-6,11	39,59	37,3321	1567,3681
-6,11	-24,11	37,3321	581,2921
-6,11	1,89	37,3321	3,5721
-8, 11	-28,71		824,2641
-9,11	28,29	82,9921	800,3241
		SQ(P) = 1893,71	SQ (TM)= 6492,55

A Tabela 3 apresenta os valores calculados de A e de B, elevados ao quadrado, como passos intermediários do cálculo do coeficiente de correlação, onde:

SQ(P) = Somatório dos valores de A^2.

SQ(TM) = Somatório dos valores de B^2.

Tendo os dados necessários, pode-se obter a Correlação de Pearson através da Fórmula:

$$r = \frac{\Sigma(A)(B)}{\sqrt{\Sigma(A^2)\Sigma(B^2)}}$$

$$= SP/\sqrt{SQ(P)SQ(TM)}$$

=

$$r = 562,73/ \sqrt{(1893,71)(6492,55)}$$
$$r = 562,73/\sqrt{3506,42}$$
$$r = +0,16.$$

Conclusão: a Correlação de Pearson indica que a Porcentagem (P) de defeitos por projeto se relaciona fracamente com o Tempo Médio gasto na resolução de um defeito. Na prática isto significa que não necessariamente sempre que a porcentagem de defeitos por projeto aumenta o tempo médio gasto para resolução de um defeito aumentou na mesma proporção. Sendo assim, a Hipótese Nula é aceita.

REFERÊNCIAS

Levin, Jack. Fox, James Alan. Forde, David R. Estatística para Ciências Humanas. (2012). 11ª ed. São Paulo: Pearson Education do Brasil.