Exemplo de um experimento

Prof. Paulo Cesar Masiero

Wholin Cap. 11

Definição

- O PSP – Personal Software Process foi criado por Watts Humphrey, do SEI.
- É um processo sistemático para apoiar engenheiros de software a entender e melhorar o desempenho.
- Publicado no livro do SEI: PSP: A Self-Improvement Process for Software Engineers, Watts Humphrey.
- O experimento foi realizado no contexto de um curso de PSP.

Definição dos objetivos

- Objeto do estudo: participantes de um curso de PSP e suas habilidades em termos de desempenho com base na formação e na experiência.
- Objetivo: Avaliar o desempenho individual baseado na formação.
- Perspectiva: dos pesquisadores e professores do curso.

Definição dos objetivos (cont.)

- Foco de qualidade: Desempenho individual no curso de PSP.
  - O desempenho será medido pela Produtividade (KLOC/tempo-desenvolvimento) e por Densidade de defeitos (falhas/KLOC)
- Contexto do experimento: Contexto de PSP, em um curso sobre PSP, no Depto de Sistemas de Comunicação, da Un. de Lund, Suécia.

Definição dos objetivos - Resumo

Analisar *os resultados de PSP*

Com o objetivo de *avaliação*

Com respeito *à formação das pessoas*

Do ponto de vista dos *pesquisadores e professores*

No contexto *de um curso de PSP*

Planejamento - introdução

- É um experimento off line (fora do contexto de uma empresa)
- Estudantes de graduação (4º. Ano)
- Trata de um problema real, que são as diferenças individuais e como entender essas diferenças
- Tem o objetivo de generalizar os resultados e pode ser replicado.
- Os formulários de coleta serão os próprios recomendados no livro e o contexto experimental é fácil de ser preparado.

Formulação das hipóteses Informal

- Os estudantes são de CC and E e EE. Os primeiros têm mais disciplinas de computação e há uma crença que seriam mais produtivos.
- Os estudantes responderão a questões sobre o curso. Por exemplo, conhecimento de C. A hipótese é que estudantes com experiência geram menos falhas por linha de código.

Formulação das hipóteses Formal

1. Null hypothesis, H_0: There is no difference in productivity (measured as lines of code per total development time) between students from the Computer Science and Engineering program (CSE) and the Electrical Engineering program (EE).

   H_0: Prod(CSE) = Prod(EE)

   Alternative hypothesis, H_1: Prod(CSE) ≠ Prod(EE)

   Measures needed: student program (CSE or EE) and productivity (LOC/hour).

2. Null hypothesis, H_0: There is no difference between the students in terms of number of faults per KLOC (1000 lines of code) based on the prior knowledge in C.

   H_0: Number of faults per KLOC is independent of C experience.

   Alternative hypothesis, H_1: Number of faults per KLOC changes with C experience.

   Measures needed: C experience and Faults/KLOC.

   The hypotheses mean that we have to collect the following data:

Variáveis

- (I)Formação do estudante: CCE ou EE (escala nominal)
- (D) Produtividade (KLOC/TD), a ser medida para os 10 programas que constam do livro.
  - KLOC (escala quociente), medida para linhas novas e modificadas
  - Tempo de Desenvolvimento, medidos em minutos (escala quociente) e depois convertidos em horas.

Variáveis

- (I) Experiência em C, escala ordinal com quatro classificações:
  - 1. Sem experiência anterior
  - 2. Leu um livro ou fez um curso
  - 3. Alguma experiência profissional (menos que 6m)
  - 4. Experiência profissional (mais que 6m)
- (D) Falhas/KLOC, medida pelo número de falhas durante os testes dos programas dividido pelo número de linhas.

Seleção dos sujeitos

- Baseada em conveniência - os estudantes matriculados no curso. Eles são uma amostra dos estudantes do curso, mas não aleatória.
- É um quase-experimento.

Projeto do experimento

- Aleatorização
  - O experimento não está querendo comparar PSP com outro método.
  - Todos os sujeitos fazem os 10 programas.
  - A ordem não é importante
- Blocagem
  - Não há, não interesse por exemplo em estudar os programas em grupos.
- Balanceamento
  - O ideal seria balancear, mas isso depende dos estudantes matriculados. Provavelmente não será balanceado.

Projeto do experimento

- Tipos básicos de projeto
  - H-1: Um fator com dois tratamentos
    - fator: cursos
    - tratamentos: CCE e EE
    - Escala quociente
    - Teste paramétrico é possível → t-test
  - H-2: Um fator com mais de dois tratamentos
    - fator: experiência em C
    - Tratamento: os quatro níveis de experiência.

Ameaças à validade

- Validade interna (relação causal entre tratamento e resultado)
  - Não é problema, haverá um grande número de testes
- Validade de construção
  - Uma ameaça é que as medidas definidas não sejam apropriadas para o que queremos medir
  - O experimento é parte de um curso em que os alunos recebem notas. Como mitigação, no início do curso será dito aos alunos que as notas não dependem do resultados dos dados, mas sim da entrega dos dados no prazo e do entendimento expressado no relatórios a serem entreguem durante o curso.

Ameaças à validade

- Validade externa (generalização)
  - Alta probabilidade que os resultados se repitam para cursos similares na mesma universidade
  - É mais difícil generalizar para estudantes que não fazem o curso (não estão interessados em desenvolvimento de software)
  - Os resultados da análise podem provavelmente ser generalizados para outros cursos de PSP em que se possível comparar o desempenho de alunos de CSE e EE

Ameaças à validade

- Validade de conclusão (tratamento e resultados)
  - A qualidade dos dados pode ser uma ameaça, pois os estudantes devem produzir muitos resultados e podem cometer erros ou falsificar os dados
  - Esse problema, entretanto, não parece depender da formação do aluno.
  - Considera-se que as ameaças à conclusão não são críticas

Instrumentação

- Formação e experiência: pesquisa a ser realizada na primeira aula.
- Objetos experimentais: os programas desenvolvidos no curso de PSP, diretrizes e medidas constam do livro. Os formulários de coleta são os mesmos do PSP.

Table 37. *Student characterization*

Area	Description	Answer
Study program (denoted Line)	Answer: Computer Science and Engineering or Electrical Engineering	
General knowledge in computer science and software engineering (denoted SE)	1. Little, but curious about the new course 2. Not my speciality (focus on other subjects) 3. Rather good, but not my main focus (one of a couple of areas) 4. Main focus of my studies	
General knowledge in programming (denoted Prog.)	1. Only 1-2 courses 2. 3 or more courses, no industrial experience 3. A few courses and some industrial experience 4. More than 3 courses and more than 1 year industrial experience	
Knowledge about the PSP (denoted PSP)	1. What is it? 2. I have heard about it 3. A general understanding of what it is 4. I have read some material	
Knowledge in C (denoted C)	1. No prior knowledge 2. Read a book or followed a course 3. Some industrial experience (less than 6 months) 4. Industrial experience	
Knowledge in C++ (denoted C++)	1. No prior knowledge 2. Read a book or followed a course 3. Some industrial experience (less than 6 months) 4. Industrial experience	
Number of courses (denoted Courses)	A list of courses was provided and the students were asked to put down a yes or no whether they had taken the course or not. Moreover, they were asked to complement the list of courses if they had read something else they thought was a particularly relevant course.	

Operação

- Preparação
  - Os sujeitos não sabiam dos aspectos do que se pretendia estudar no curso (as hipóteses).
  - Foram informados de que se pretendia estudar investigar os resultados do curso de PSP em comparação com a formação dos alunos
  - Do ponto de vista dos alunos, eles estavam participando de um curso.
  - Foi garantido o anonimato dos estudantes
  - O Material foi preparado antes e os formulários eram os mesmos do curso.

Operação

- Execução
  - O experimento foi executado durante 14 semanas, durante as quais os 10 exercícios de programação foram atribuídos normalmente.
  - Os dados foram coletados por intermédio dos formulários
  - O experimento não atrapalhou o curso. A única diferença foi a pesquisa de formação feita no início.

Operação

- Foram coletados dados de 65 estudantes
- Validação dos dados
  - Dados de 6 estudantes foram removidos pelos pesquisadores e professores por terem sido considerados inválidos ou pelo menos questionáveis. As razões foram:
    - Dois com preenchimento incorreto
    - Um estudante entregou os dados com muito atraso
    - Dois entregaram com atraso e precisaram de muito auxílio para completar as tarefas (o auxílio extra foi considerado que poderia viesar os resultados)
    - Um estudante foi removido porque tinha formação diferente

Análise e interpretação dos resultados Estatística descritiva – produtividade

- 32 alunos de CSE e 27 de EE
- A produtividade foi agrupada em oito classes
- CSE: Média = 23 e DP= 8,7
- EE: Média = 16,4 e DP= 6,3
- A figura 26 mostra que
  - estudantes de EE parecem ter tido produtividade pior.
  - Estudantes de CSE parece ter tido maior variabilidade (dispersão) nos resultados

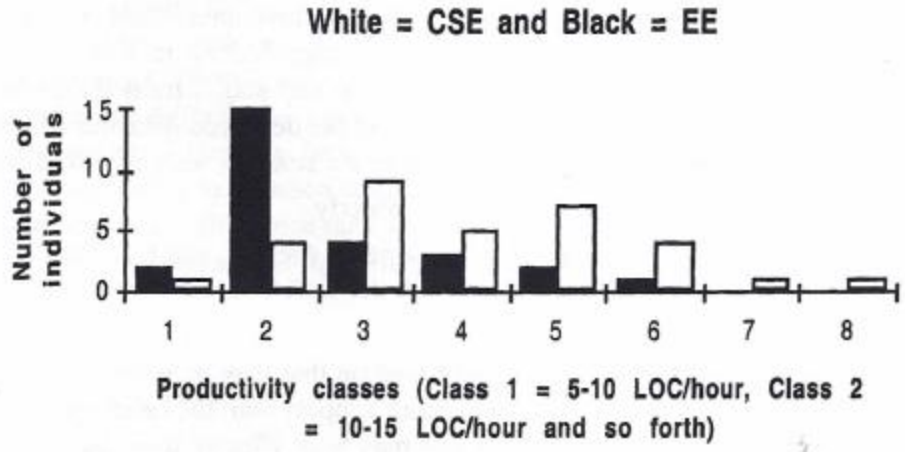

Figure 26. *Frequency distribution for the productivity (in classes).*

Análise e interpretação dos resultados Estatística descritiva – produtividade Box- Plot

- CSE:
  - Mediana = 22.7
  - tamanho da caixa = 29.4 – 17.6 = 11.8
  - Bigode superior: 29.4 + 11.8*1.5=47.1 → 42.5 (maior amostra)
  - ...e assim por diante.
- Esta análise também mostra que CSE teve maior produtividade
- Existe um outlier para EE, que não foi removido.

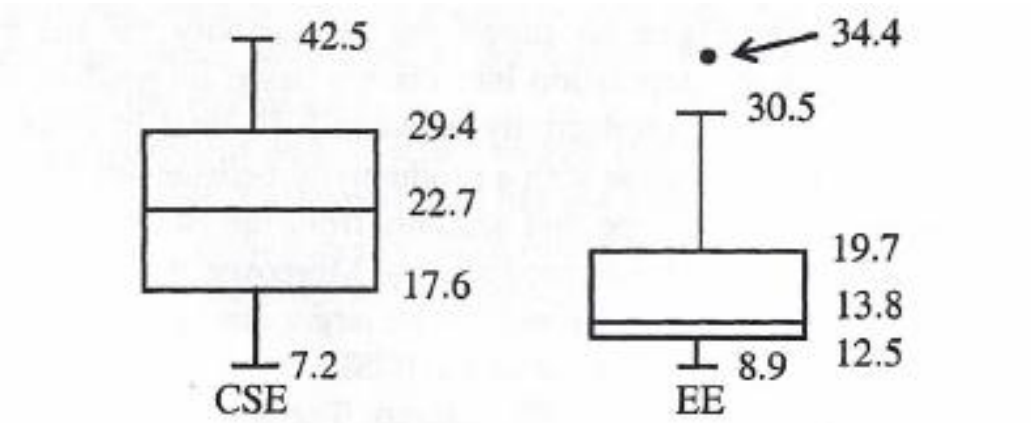

Figure 27. *Box plot of productivity for the two study programs.*

Análise e interpretação dos resultados Estatística descritiva – Exp-CvsFalhas

- A tabela 28 mostra que a distribuição é inclinada para “pouca experiência em C”.
- As médias mostram que os mais experientes geram menos falhas
- O desvio padrão é muito alto
- A mediana varia bastante em relação à média
- O desvio padrão para a primeira classe é muito alto e exige melhor investigação

Table 28. *Faults/KLOC for the different C experience classes.*

Class[a]	Number of students	Median value of faults/ KLOC	Mean value of faults/KLOC	Standard deviation of faults/KLOC
1	32	66.8	82.9	64.2
2	19	69.7	68.0	22.9
3	6	63.6	67.6	20.6
4	2	63	63.0	17.3

a. The different classes are explained in Section 11.2.2.

Análise e interpretação dos resultados Estatística descritiva – Exp-CvsFalhas

- Box-plots foram construídas para as 4 classes
- As box-plots não levam a informações adicionais para as classes 2, 3 e 4 → todos os valores estão entre os limites superiores e inferiores.

Análise e interpretação dos resultados Estatística descritiva – Exp-CvsFalhas

- A fig. 28 mostra o box-plot da classe 1
  - O bigode inferior é igual ao menor valor de falha/KLOC
  - O bigode inferior é menor que a maior amostra e isso implica em outliers
  - O outlier 398.1 é mais de 10 vezes superior ao menor valor e isso explica o alto desvio padrão.

A análise descritiva ajudou a entender melhor os dados coletados: o que se pode esperar do teste de hipótese e problemas potenciais que podem ser causados pelos outliers

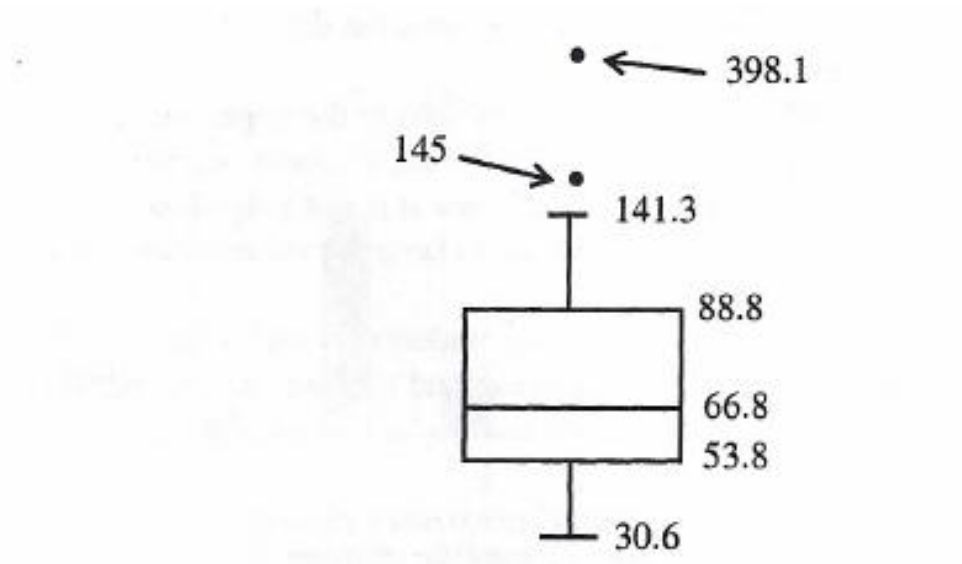

Figure 28. *Box plot for faults/KLOC for class 1.*

Redução de dados

- A redução de dados causa perda de informação, e por isso é delicada. Ela pode ocorrer de duas formas:
  - Remover pontos (de dados) isolados. Ex: outliers
  - Quando há alta correlação entre as variáveis, elas podem ser combinadas em um medida mais abstrata (reduzir o número de variáveis)
- Pontos de dados não devem ser removidos apenas por que eles não se encaixam em nossa hipótese!
- Por outro lado, deve-se remover pontos de dados que podem causar inconsistências nos resultados (ex. não remover um outlier extremo)

Redução de dados (cont.)

- Para remover e combinar variáveis, é preciso usar outras técnicas estatísticas.
- Neste caso, decidiu-se pela remoção do outlier extremo para o número de falhas/KLOC
- Isso diminuiu o valor da média e do desvio padrão, mas a diferença em relação às outras classes não ficou muito grande (fig. 29)

Table 29. *Faults/KLOC for C experience class 1.*

Class	Number of students	Median value of faults/ KLOC	Mean value of faults/KLOC	Standard deviation of faults/KLOC
1	31	66	72.7	29.0

Teste de Hipótese - produtividade

- Foi usado t-test (unpaired, two-tailed)
- Resultado na tabela 30.
- Na tabela A1 (mostrada na aula 1), o valor do t-value para 30 graus de liberdade é 2.042 e para 40 graus é 2.021
- Podemos concluir que Ho pode ser rejeitada. O valor de Po é muito baixo e, portanto, o resultado é altamente significante, isto é, há diferença significante entre o desempenho de estudantes com formação em diferentes cursos (neste caso EE e CSE).

Table 30. *Results from the t-test.*

Factor	Mean diff.	Degrees of freedom (DF)	t-value	p-value
CSE vs. EE	6.1617	57	3.283	0.0018

Teste de Hipótese - Exp-CvsFalhas

- Foi usado o teste ANOVA (tabela 31)
- Os resultados da análise não são significativos
- Então, não foi possível mostrar que há uma diferença significativa entre o número de falhas/KLOC com base na experiência em C.

Table 31. *Results from the ANOVA-test.*

Factor: C vs. Faults/KLOC	Degrees of freedom (DF)	Sum of squares	Mean square	F-value	p-value
Between treatments	3	3483	1160.9	0.442	0.7236
Error	55	144304	2623.7		

Analises adicionais

- Como o número de estudantes nas classes 3 e 4 é muito baixo, foi feita uma outra análise comparando as classes 2, 3 e 4 juntas, com a classe 1.
- Nenhum resultado significativo foi obtido.