



Experimentação em Engenharia de Software: Fundamentos e Conceitos

SBQS 2011

Guilherme Horta Travassos
 www.cos.ufrj.br/~ght

Grupo de Engenharia de Software Experimental
ese.cos.ufrj.br



Experimentação em Engenharia de Software: Fundamentos e Conceitos

- O material deste mini-curso representa a agregação de trabalhos anteriores de cientistas em Engenharia de Software, os quais autorizaram seu reuso no contexto da ESELAN (Experimental Software Engineering Latin American Network) e ESELAW por este pesquisador :

Victor R. Basili (**VRB**), UMD/USA

“The Role of Experimentation in Software Engineering: Past, Present, Future” - Keynote speaker at ICSE 18

Shari Lawrence Pfleeger (**SLP**), Rand Inc.

“Evaluating Software Technology” - Tutorial at SBES’2002
and Software Engineering Thesis Workshop talk at SBES’2002

Barbara Kitchenham (**BK**), Keele University/UK – Research Protocols and Systematic Literature Reviews, 3rd International Advanced School on Empirical Software Engineering, ISESE 2005

Tore Dybå (**TD**), SINTEF, Norway - Research Protocols and Systematic Literature Reviews, 3rd International Advanced School on Empirical Software Engineering, ISESE 2005

Travassos, G.H., Fabbri, S., Maldonado, J.C.. Experimental Software Engineering: an Introduction, Tutorial at the Brazilian Symposium on Software Engineering. 2005. Brazil.

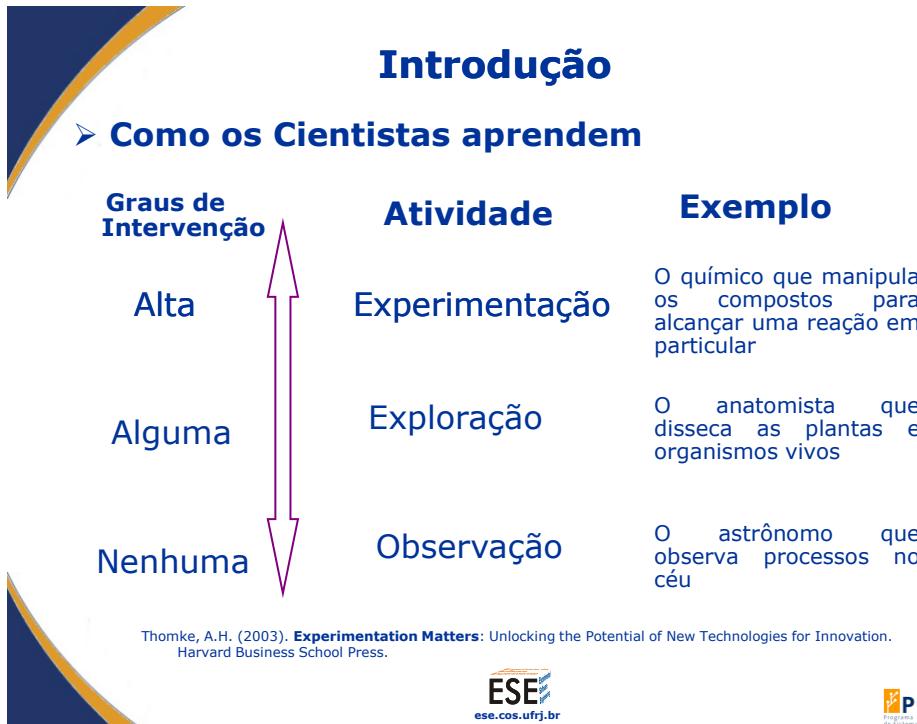
Biolchini, J. C.; Travassos, G.H.. Systematic Reviews Applied to Software Engineering. Tutorial at the Brazilian Symposium on Software Engineering, Brazil, 2007.

Araujo, M. A. ; Barros, M. O.; Murta, L. G. P.; Travassos, G.H. (**ABMT**). Statistical Methods Applied to Experimental Software Engineering. Tutorial at Brazilian Symposium on Software Engineering. Brazil, 2006, ESELAW 2009 e ClbSE 2011



Introdução

➤ Como os Cientistas aprendem



Introdução

➤ Evolução do Conhecimento: Construção de Modelos, Experimentação e Aprendizado

- O entendimento de uma disciplina envolve a construção de modelos
 - Ex.: Domínio da aplicação, processos para solução de problemas
- Para verificar se nosso entendimento está correto, precisamos:
 - Testar nossos modelos
 - Experimentá-los!

- Analisar os resultados de um estudo experimental requer aprendizado, encapsulamento de conhecimento e habilidade para modificar e refinar nossos modelos ao longo do tempo
 - O entendimento de uma disciplina evolui com o tempo
 - Encapsulamento de conhecimento nos permite lidar com níveis mais altos de abstração
 - Este paradigma vem sendo utilizado por diferentes áreas:
 - Ex.: Física, Medicina, Manufatura



Introdução

➤ Evolução do Conhecimento: Construção de Modelos, Experimentação e Aprendizado

- O que estas áreas tem em comum?
 - o Elas evoluíram como disciplina quando começaram a aplicar o ciclo de construção de modelos, experimentação e aprendizado
 - Começaram observando e registrando o que vinha sendo observado
 - Evoluíram manipulando as variáveis e estudando os efeitos de modificação destas variáveis
- Quais são as diferenças entre estas áreas?
 - o As diferenças são os objetos de estudo, as propriedades destes objetos, as propriedades do sistema que os contêm, o relacionamento do objeto com o sistema e a cultura da disciplina
 - o Isto afeta:
 - Como os modelos são construídos
 - Como a experimentação deve ser realizada



Introdução

➤ Evolução do Conhecimento: Construção de Modelos, Experimentação e Aprendizado

- **Física**
 - o Entende e prediz o comportamento do universo físico
 - o Pesquisadores: teóricos e experimentalistas
 - o Tem progredido devido ao forte relacionamento entre os grupos
- o **Teóricos** constroem modelos para descrever o universo
 - Predizem o resultado de eventos que podem ser medidos
 - Modelos são baseados na teoria sobre as variáveis essenciais e sua interação, dados de estudos experimentais anteriores
- o **Experimentalistas** observam, medem, experimentam para
 - Colocar a prova ou refutar uma hipótese ou teoria
 - Explorar um novo domínio
- o Independente do ponto de entrada no ciclo de aprendizado existe um padrão de modelagem, experimentação, aprendizado e remodelagem (método científico)
- o Os primeiros experimentalistas apenas observaram, não manipularam objetos. A Física moderna aprendeu a manipular o universo físico, como por exemplo, a física de partículas.



Introdução

➤ **Evolução do Conhecimento: Construção de Modelos, Experimentação e Aprendizado**

- **Medicina**

- o Pesquisadores e profissionais da prática (médicos)
- o Relacionamento claro entre os dois
- o Construção do conhecimento através da realimentação do profissional da prática para o pesquisador
- o **Pesquisador** objetiva entender as funcionalidades do corpo humano de forma a predizer os efeitos de vários procedimentos e medicamentos
- o **Profissional da Prática** aplica o conhecimento manipulando os processos e medicamentos no corpo para o propósito de cura
- o Começou como uma forma de arte (empirismo). Evoluiu como uma área científica quando começou a observar e construir modelos (experimentalismo)
- o **Experimentação**
 - De experimentos controlados a estudos de caso
 - As diferenças da raça humana causam problemas na interpretação de resultados
 - Dados podem ser difíceis de serem conseguidos
- Entretanto, nosso conhecimento sobre o corpo humano tem evoluído ao longo do tempo



Introdução

➤ **Evolução do Conhecimento: Construção de Modelos, Experimentação e Aprendizado**

- **Manufatura**

- o Domínios do pesquisador e de manufatura
- o Entende as características do processo e do produto
- o Constrói um produto para satisfazer a um conjunto de especificações
- o Evoluiu como uma disciplina quando começou a aprimorar seus processos
- o Relacionamento entre as características do processo e do produto é bem compreendido
- o Melhoria do Processo baseado em modelos de
 - Domínio do problema e espaço de solução
 - Paradigma evolutivo de construção de modelos, experimentação e aprendizado
 - Relacionamento entre os três
- o Modelos são construídos com boas capacidades de predição
 - Mesmo produto gerado diversas vezes e baseado no mesmo conjunto de processos
 - Compreensão dos relacionamentos entre processo e produto



Introdução

- Outras áreas de engenharia exploram modelos (experimentação) como uma ferramenta fundamental para desenvolvimento de suas tecnologias: Automóveis

Contrasting Experimentation Strategies in Car Safety Design	
Traditional Prototype Crashes	Crash Modeling and Simulation
Primarily design verification because of late and slow feedback	Learning and experimentation through early and fast feedback
Costly and time-consuming construction of each crash prototype	Cheap, fast generation of virtual prototypes after initial model is "built"
Destructive testing difficult to analyze	Simulation testing allows for careful and deep analysis
Experiments minimized because of high cost	Experiments maximized with diverse concepts
Experimental conditions limited to standard laboratory set up	Experimental conditions can be changed easily
Prototypes are closer to reality for some crashes	Models are approximations of reality and may be too complex for some tests
Interaction with humans more natural because it is physical	Requires better human-machine interfaces to maximize acceptance and learning

Thomke, A.H. (2003). **Experimentation Matters: Unlocking the Potential of New Technologies for Innovation.** Harvard Business School Press.



Introdução

- Outras áreas de engenharia exploram modelos (experimentação) como uma ferramenta fundamental para desenvolvimento de suas tecnologias: Medicamentos

Contrasting Experimentation Strategies in Drug Discovery	
Traditional (Medicinal) Chemistry and Testing	Combinatorial Chemistry and High-Throughput Screening
Primarily sequential experimentation with few compounds per round	Primarily parallel experimentation with many compounds per round
Costly, slow generation of "prototype" compounds	Cheap, fast generation of "prototype" compounds
Small testing capacity	Large testing capacity through automated screening
High "solution" diversity through proven compounds in chemical libraries	Limited "solution" diversity but expected to grow in the future
Nearly perfect chemical purity and thus high confidence in results	Lower chemical purity and thus less reliable results
Focus on the craft of perfection	Focus on the science of production
Bottleneck is generating and testing new compounds	Bottleneck is ability to process and understand massive amounts of data

Thomke, A.H. (2003). **Experimentation Matters: Unlocking the Potential of New Technologies for Innovation.** Harvard Business School Press.



Engenharia de Software

- O desenvolvimento de software depende de diferentes tecnologias.
 - Apesar de todo o esforço de pesquisa, ainda não existem evidências suficientes relativas aos benefícios, limitações, custos e riscos destas tecnologias
- Entretanto, engenheiros de software precisam responder a questões relacionadas as tecnologias para aumentar as chances de sucesso dos projetos:
 - Que tecnologia de software devo considerar para meu projeto?
 - Quanto treinamento e/ou investimento é necessário para inserir a tecnologia de software no meu processo de trabalho?
 - Quando e como será observado o Retorno do Investimento realizado?
 - Sob que condições uma tecnologia de software apresenta melhor desempenho?
- Para complicar, ainda existe baixa interação entre academia e indústria

Grande necessidade de soluções rápidas/imediatas

+

Necessidade para justificar investimentos

=

Tecnologia imatura transferida para a indústria



Engenharia de Software

➤ Natureza de uma Disciplina

- Assim como outras disciplinas de base científica, a Engenharia de Software necessita explorar o ciclo de construção de modelos, experimentação e aprendizado (método científico)
- **Engenharia de Software é ciência de laboratório**
- O papel do pesquisador é compreender a natureza dos processos, produtos e o relacionamento entre eles no contexto do sistema
- O papel do profissional da prática (engenheiro de software) é construir sistemas cada vez melhores, utilizando o conhecimento disponível e produzido pelos pesquisadores
- Mais que em outras disciplinas, estes papéis são simbióticos
 - o O pesquisador precisa dos laboratórios para observar e manipular as variáveis
 - Seus efeitos somente se manifestam quando os engenheiros de software constroem sistemas de software
 - o O engenheiro de software precisa compreender melhor como construir sistemas melhores
 - O pesquisador pode (e deve) produzir modelos para ajudar



Engenharia de Software

➤ **Natureza de uma Disciplina**

- Engenharia de Software é desenvolvimento e não produção
- As tecnologias da disciplina são baseadas no elemento humano
- O Software não é o mesmo todo o tempo
 - Existe um enorme número de variáveis que provocam diferenças
 - Quais são e seus efeitos precisam ser entendidos
- Atualmente,
 - Conjunto de modelos insuficientes que nos permita pensar sobre a disciplina
 - Falta de conhecimento sobre os limites das tecnologias para certos contextos de desenvolvimento
 - **Análise e experimentação insuficientes: empirismo x experimentação; experiência x experimentação**



Engenharia de Software

➤ A definição de processos de software é importante, entretanto deve nos permitir:

- Observar, medir e melhorar
- Disseminar/Divulgar facilmente as boas práticas de trabalho
- Ajudar a tornar o desenvolvimento de software mais “engenheirado” e menos artístico

➤ A definição de processos de software necessita iteração no campo

- Muitos aspectos de um processo não podem ser avaliados apenas em laboratório
- Alto risco ao introduzir processos imaturos na indústria de software
- Adaptação é necessária
- Gerenciar lições aprendidas e conhecimento adquirido é fundamental





Engenharia de Software

Precisamos aumentar nosso interesse em experimentação e entender como aplicá-la como ferramenta para caracterizar, compreender, avaliar, predizer, controlar e melhorar as tecnologias de software!



Experimentação em ES

- Engenharia de Software precisa ter sólidos fundamentos científicos tais como uma disciplina científica de engenharia
- Compartilhamento de conhecimento científico representa um desafio fundamental em ES
- O desenvolvimento de novas tecnologias de software demanda experimentação em larga escala

➤ **Software Engineering Is big science!** (Vic Basili)

- Entretanto, executar um estudo experimental em Engenharia de Software é difícil, consome tempo e gera um grande volume de dados, informação e conhecimento científico



Engenharia de Software Experimental

➤ Paradigmas de Pesquisa Disponíveis

- **Analítico**

- o Propõe uma teoria formal ou um conjunto de axiomas
- o Desenvolve uma teoria
- o Deriva os resultados e
- o Se possível, verifica os resultados com observações *empíricas*

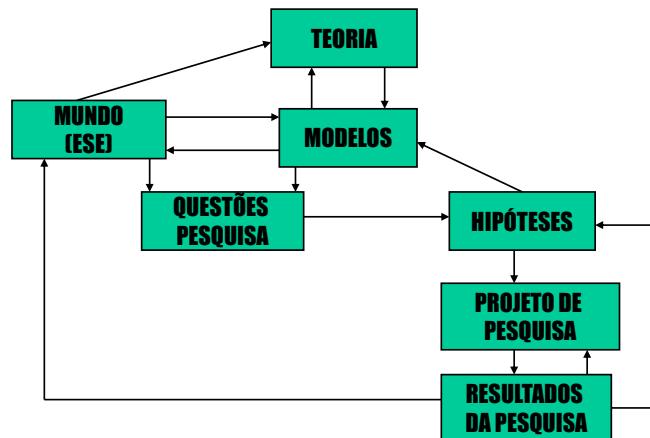
- **Experimental**

- o Observa o mundo (ou soluções existentes)
- o Propõe um modelo ou uma teoria de comportamento (ou melhores soluções)
- o Mede e analisa
- o Valida (coloca à prova) hipóteses do modelo ou da teoria (ou invalida)
- o Repete o procedimento para evoluir a base de conhecimento
- o O paradigma experimental envolve:
 - Projeto do estudo experimental
 - Observação
 - Análise quantitativa e qualitativa
 - Coleta de dados e validação do processo ou produto sendo estudado



Engenharia de Software Experimental

➤ Relacionamento da ESE para teoria, modelos e questões de pesquisa



Engenharia de Software Experimental

➤ Paradigmas de Pesquisa Experimental

o Análise Quantitativa

- Medição controlada e não intrusiva
- Objetiva
- Orientada a verificação

o Análise Qualitativa

- Observação natural e não controlada
- Subjetiva
- Orientada a descoberta

o Análise Semi-Quantitativa

- Associa observação e alguma medição (tendência)
- Semi-objetiva
- Orientada a observação



Engenharia de Software Experimental

➤ Paradigmas de Pesquisa Experimental

• Estudos Experimentais

- o Ato de descobrir algo desconhecido ou colocar a prova uma hipótese
- o Pode incluir todas as formas de análise quantitativa, qualitativa e semi-quantitativa
- o Estudos podem ser:

o Primários

➤ Experimentais

- Direcionado pelas hipóteses; análise quantitativa
- Experimentos controlados
- *quasi*-experimentos ou projetos pré-experimentais

➤ De Observação

- Direcionados ao entendimento; análise qualitativa predomina
- Estudo quantitativo, qualitativo ou semi-quantitativo
- Estudo qualitativo puro

o Secundários

- Revisões Sistemáticas
- Meta Análise



Classificação de Estudos em Engenharia de Software Experimental

Estudos Primários

- Representam os estudos regulares, usualmente incluindo estudos de observação (*survey*, estudos de caso, pesquisa-ação) e experimentais (*quasi* experimentos, experimentos controlados)
- Utilizados para observar algum comportamento no campo ou colocar uma hipótese a prova

Estudos In vivo

- envolvem os indivíduos em seus próprios ambientes de trabalho
 - Estudos experimentais executados nas organizações que desenvolvem software durante o processo de desenvolvimento e sob condições reais de trabalho podem ser classificados como *in vivo*

Estudos In vitro

- são executados em ambientes controlados, como um laboratório ou comunidade controlada
 - A maioria dos estudos experimentais *in vitro* são executados em universidades ou entre grupos selecionados em uma organização de desenvolvimento de software



Classificação de Estudos em Engenharia de Software Experimental

Estudos Primários

- Três exemplos de estudos primários *in vivo/in vitro*:
 - o **Pesquisa de Opinião (*survey*)**
 - É realizado depois do fato ocorrido, para tentar identificar alguma evidência. Não permite controle.
 - o **Estudos de Caso**
 - É realizado enquanto o desenvolvimento ou manutenção do software está sendo realizado, entretanto comportamento relevante não pode ser manipulado.
 - o **Experimentos Controlados (*quasi* experimentos)**
 - Um experimento permite controle e manipulação de comportamento.
- A diferença está no nível de controle:
 - **Usar um destes é mais útil e melhor que não usar nenhum!**

Mafra, S. N. e Travassos, G.H. (2006) Estudos Primários e Secundários apoiando a busca por Evidência em Engenharia de Software. Relatório Técnico ES-687/06 - PESC/COPPE/UFRJ. Disponível em <http://www.cos.ufrj.br/uploadfiles/1149103120.pdf>



Classificação de Estudos em Engenharia de Software Experimental

Estudos in virtuo

- Participantes interagem com um modelo computacional simulando a realidade
 - O comportamento do ambiente nos quais os indivíduos interagem é descrito através de um modelo representado por um programa de computador (simulador)
 - Em ES estes estudos são normalmente executados em universidades e centros de pesquisa e podem ser caracterizados por um pequeno grupo de indivíduos manipulando simuladores

Estudos in silico

- Participantes e mundo real (ambiente) descritos e simulados através de modelos computacionais
 - O ambiente é totalmente composto por modelos computacionais onde a interação humana é reduzida a um mínimo
 - Podemos encontrar, por exemplo, estudos *in silico* aplicados a experimentação relacionada a usabilidade de software (e.g. caracterização de desempenho)

TRAVASSOS, G. H. ; BARROS, M. O.. Contributions of *In Virtuo* and *In Silico* Experiments for the Future of Empirical Studies in Software Engineering. In: 2nd Workshop in Workshop Series on Empirical Software Engineering: The Future of Empirical Studies in Software Engineering, 2003, Roma. Proceedings of the WSESE03, 2003.



Tipos de Estudos Primários e Modelos em ESE

in vivo



Nenhum Modelo Necessário

in vitro



Modelo do Ambiente

in virtuo



Modelos Computacionais
do Objeto e Ambiente

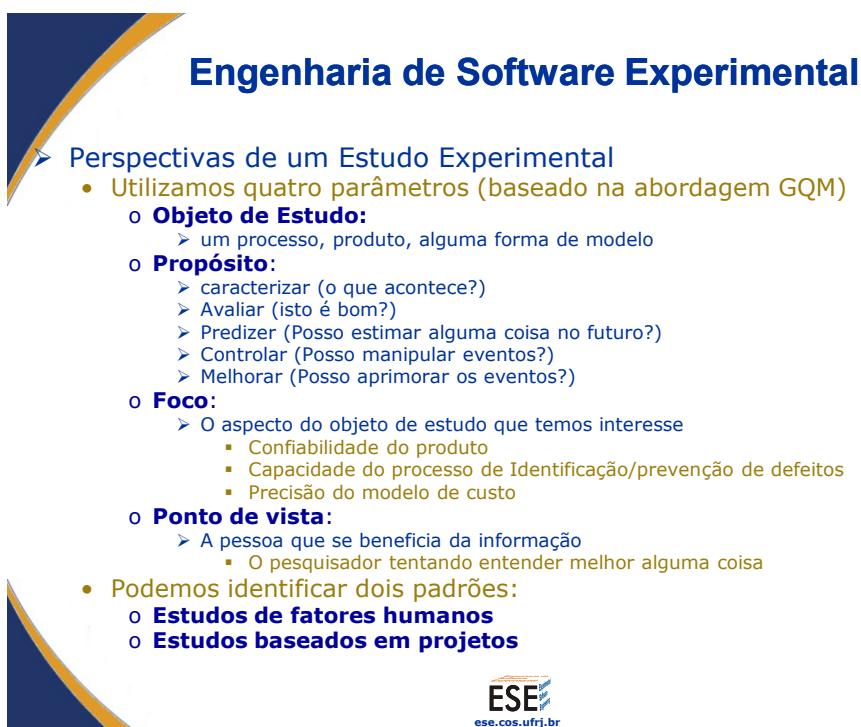
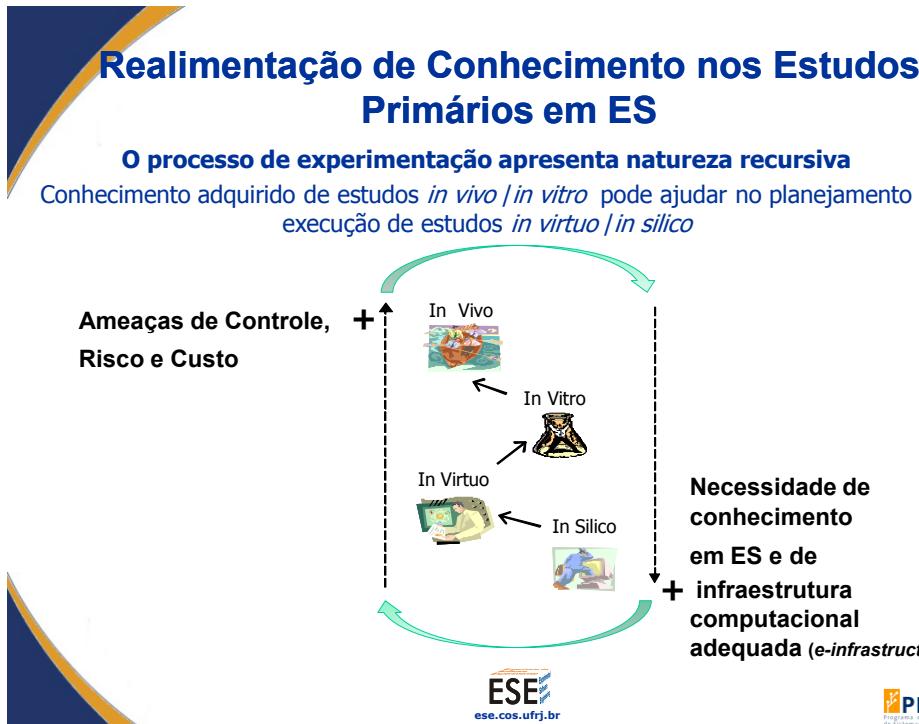
in silico



Modelos computacionais do
Comportamento do Participante,
Objeto e Ambiente

TRAVASSOS, G. H. ; BARROS, M. O.. Contributions of *In Virtuo* and *In Silico* Experiments for the Future of Empirical Studies in Software Engineering. In: 2nd Workshop in Workshop Series on Empirical Software Engineering: The Future of Empirical Studies in Software Engineering, 2003, Roma.





Engenharia de Software Experimental

➤ Maturidade da área ESE

- Quais são os componentes de Estudo

- **Estudos de Fatores Humanos**

- Objeto de estudo: uma pequena tarefa cognitiva
- Foco: alguma medida de desempenho
- Propósito: avaliação
- Ponto de vista: pesquisador

➤ Feito por/com especialistas em psicologia cognitiva confortáveis com experimentação

➤ Tem permitido estudos em pequena escala

- **Estudos Baseado em Projetos**

- Objeto de estudo: processo de software, produto, ...
- Foco: variação de confiabilidade do produto e custo até efeito do processo
- Propósito: caracterização/entendimento, avaliação, alguma predição
- Ponto de vista: o pesquisador (geralmente uma visão do profissional da prática)

➤ Feito normalmente por Engenheiros de Software, menos adeptos à experimentação

➤ Tem evoluído de itens pequenos e específicos,

- como características particulares de linguagens de programação,
- Para incluir processos de desenvolvimento como um todo, como tradicionais e baseados em agilidade



Engenharia de Software Experimental

➤ Tipos de estudo primário que tem sido executados

- 1) Os resultados dos estudos podem ser descritivos, correlacionais ou causa-efeito

- Descritivo: podem existir padrões nos dados mas o relacionamento entre a variáveis não foi examinado

- Correlacional: a variação das variáveis dependentes é relacionada a variação das variáveis independentes

- Causa-efeito: o tratamento das variáveis é a única causa possível de variação das variáveis dependentes

- Fatores humanos: maioria dos casos causa-efeito

- Sinal de maturidade dos experimentalistas, tamanho natural do problema

- Baseado em Projetos

- Evoluindo (?) de estudos correlacionais para descritivos

➤ Reflete crenças iniciais que o problema seria simples e que alguma combinação simples das métricas poderia explicar custo, qualidade, etc.

➤ Não possui uma base de conhecimento observado



Engenharia de Software Experimental

➤ Tipos de estudo primário que tem sido executados

- 2) O estudo pode ser realizado com novatos, especialistas ou ambos
 - novatos: estudantes ou indivíduos sem experiência no domínio
 - especialistas: profissionais da prática ou pessoas com experiência no domínio
 - o Fatores humanos: investiga as diferenças entre novatos e especialistas
 - o Baseado em Projetos: mais estudos com especialistas, especialmente estudos descritivos sobre organizações e projetos.
- 3) O estudo pode ser executado *in vivo*, *in vitro*, *in virtuo* ou *in silico*
 - *in vivo*: no campo sob condições normais
 - *in vitro*: no laboratório sob condições controladas
 - *in virtuo*: no laboratório/campo usando modelos computacionais do objeto de estudo e do ambiente
 - *in silico*: no laboratório com modelos computacionais para participantes, objeto de estudo e ambiente
 - o Fatores humanos: mais *in vitro*
 - o Baseado em projetos: mais *in vivo*
 - o *Baseados em simulação*: *in virtuo/in silico*
- 4) O estudo primário pode ser experimental ou de observação
 - Experimento: pelo menos um tratamento ou variável controlada
 - Estudo de observação: sem tratamento ou variável controlada



Engenharia de Software Experimental

➤ Tipos de estudo experimental que tem sido executados

- Experimentos podem ser
 - Experimentos controlados
 - *quasi*-experimentos ou projetos pré-experimentais
- Experimentos controlados típicos:
 - Pequeno objeto de estudo
 - *In vitro*
 - Uma combinação de tratamentos de novatos (maioria) e especialistas
- Algumas vezes, indivíduos novatos são usados para “depurar” o projeto do experimento
- ***quasi-experimentos*** ou Projetos Pré-experimentais típicos:
 - Grandes projetos
 - *In vivo*
 - Com especialistas
 - **Sem escolha aleatória da população**
- Estes experimentos tendem a envolver um componente de análise qualitativa, incluindo pelo menos alguma forma de entrevista



Engenharia de Software Experimental

- Tipos de estudo experimental que tem sido executados

Classes de Estudos Experimentais

# de equipes por projeto	# de projetos	
	Um	Mais que um
Uma	Projeto único	Variação de múltiplos projetos
Mais que uma	Projeto Repetido	Grupamento de Assunto/Projeto



Engenharia de Software Experimental

- Tipos de Estudo de Observação que tem sido executados
 - Estudos de Observação
 - Estudos qualitativos/quantitativos
 - Estudos qualitativos puros
 - Estudos Qualitativo/Quantitativos:
 - observador identifica *a priori* um conjunto de variáveis para observação
 - Existe um grande número de estudos de caso e alguns estudos em campo
 - *In vivo*
 - Descritivo
 - Especialistas
 - Estudos qualitativos puros:
 - sem variáveis isoladas/identificadas *a priori*, observação aberta
 - Deduções realizadas a partir de lógica formal não matemática
 - Exemplo, proposições verbais



Engenharia de Software Experimental

- Tipos de estudo de observação que tem sido executados

Estudos de Observação

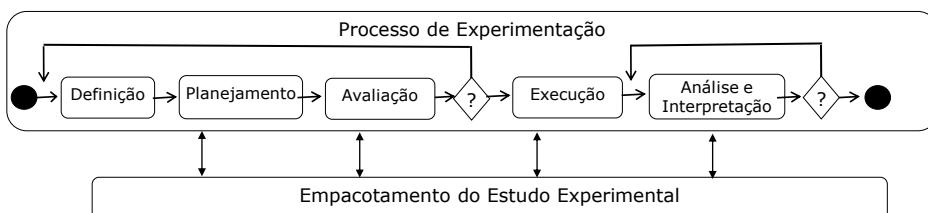
Escopo das Variáveis

# de locais	Definidas a priori		Não definidas a priori
	Um	Estudo de Caso	Estudo de Caso Qualitativo
Mais que um	Estudo de Campo	Estudo de Campo Qualitativo	

ESE
ese.cos.ufrj.br

PESC
Programa de Engenharia de Sistemas e Computação

Processo de Experimentação (estudos primários)



TRAVASSOS, G. H. ; SANTOS, P. S. M. ; MIAN, P. ; DIAS NETO, A. C. ; BIOLCHINI, J. (2008). An Environment to Support Large Scale Experimentation in Software Engineering. In: Proc. of XIII IEEE International Conference on Engineering of Complex Computer Systems, Belfast.

a seguir discutiremos o planejamento de estudos experimentais...

ESE
ese.cos.ufrj.br

PESC
Programa de Engenharia de Sistemas e Computação

Definição

- Determina a base para o estudo experimental
- Objetivo:
 - Definir as metas do estudo experimental de acordo com o *framework* estabelecido
 - Assegurar que a intenção do estudo possa ser realizada pela execução do estudo experimental
 - Caracterizar o tipo de estudo pelo número de participantes e objetos envolvidos :
 - *Estudo com Objeto Único* – 1 objeto e 1 participante
 - *Estudo com variação multi-objeto* – Mais de 1 objeto e 1 participante por objeto
 - *Estudo com multi-teste num mesmo objeto de estudo* – 1 objeto e mais de 1 participante por objeto
 - *Estudo em bloco participante-objeto de estudo* – Mais de 1 objeto e mais de participante por objeto



Definição

- Propósito:
 - Garantir que os aspectos importantes do estudo experimental sejam definidos antes do planejamento e da execução. Indica o que será efetivamente o estudo.

Analisar <produtos, processos, modelos, métricas, teorias ...>

com o propósito de <caracterizar, entender, avaliar, predizer, melhorar, controlar >

com respeito a <foco de qualidade: eficácia, custo, confiabilidade, manutenibilidade, portabilidade, ... >

do ponto de vista do < pontos de vista pelos quais os resultados do experimento são interpretados. Exemplo: desenvolvedor, gerente de projeto, gerente de corporação, cliente, usuário, pesquisador >

no contexto de <ambiente onde o estudo será realizado. Define os participantes e os artefatos (objetos) que serão usados no experimento>



Planejamento

- Descreve **como** o estudo experimental deve ser conduzido
 - Seleção de contexto
 - Formulação de hipóteses
 - Seleção de variáveis
 - Seleção de participantes
 - Projeto do experimento
 - Instrumentação
 - Validação (*validity evaluation*)



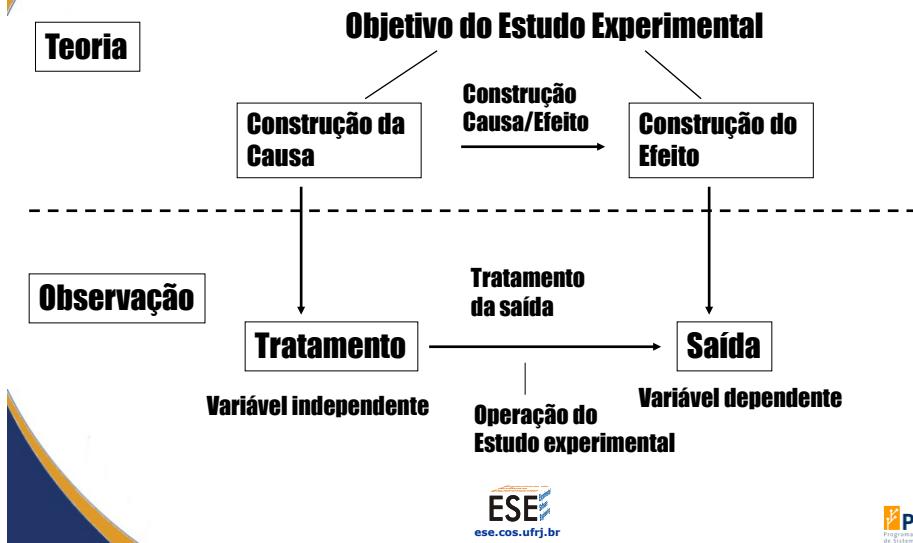
Seleção de Contexto

- Envolve o equilíbrio entre fazer estudos válidos para um contexto específico ou válidos para o domínio de engenharia de software
- Pode ser caracterizado de acordo com quatro dimensões:
 - *Off-line* x *on-line*
 - Estudantes x profissionais
 - Problemas fictícios (*toy*) x problemas reais
 - Específico x geral



Seleção de Contexto

- O Contexto de um Estudo Experimental



Formulação de Hipóteses

- As hipóteses devem ser formalmente descritas
- Os dados coletados durante a execução do estudo devem ser utilizados para colocar as hipóteses à prova
- Declaração de hipóteses:
 - **Hipótese nula:** deve ser rejeitada (refutada), caso contrário
 - o afirma que não há tendência ou padrão real nos dados coletados pelo estudo experimental
 - o as razões para diferenças na observação são apenas coincidências
 - **Hipótese alternativa:**
 - o é a hipótese que pode ser colocada à prova quando a hipótese nula é rejeitada

Riscos no Teste de Hipóteses

➤ Riscos ao testar hipóteses:

- Risco do teste rejeitar uma hipótese verdadeira
 - **Type-I-error (α)**: ocorre quando um teste estatístico indica um padrão ou uma relação, que na verdade não é um padrão real. ($p\text{-value...}$)
- Risco do teste não rejeitar uma hipótese falsa
 - **Type-II-error (β)**: ocorre quando um teste estatístico não indica um padrão ou uma relação mesmo que haja um padrão real.

➤ Potência (power)

- Capacidade do teste estatístico revelar um padrão verdadeiro ao tratar os dados

Dyba, T.; Kampenes, V.; Sjoberg, D. (2005). A Systematic Review of Statistical Power in Software Engineering Experiments. *Information and Software Technology*. Elsevier.



Seleção de Variáveis

➤ Escolha das variáveis independentes e dependentes

- Variável independente:
 - Influencia os resultados do estudo.
 - Quando seu valor pode ser modificado (tratamento), é chamada de Fator.
 - Deve ter algum efeito sobre as variáveis dependentes
 - Sua escolha inclui selecionar as escalas de medidas, a faixa para as variáveis e os níveis específicos nos quais os testes de hipótese serão realizados
- Variável dependente:
 - O efeito do tratamento é medido na variável dependente
 - Deve existir pelo menos uma variável dependente
 - A variável dependente deve ser derivada diretamente da hipótese.



Planejamento e Hipótese

➤ Planejamento

- Formulação de hipótese
- Identificação das variáveis dependentes (resposta)
- Identificação das variáveis independentes (fatores)
- Seleção dos participantes
- Projeto do estudo
- Seleção dos métodos de análise
- Definição dos instrumentos
- Análise de ameaças (*validity threats*)



Hipóteses

- Uma hipótese é uma teoria ou suposição que pode explicar um determinado comportamento de interesse da pesquisa

“Utilizando a técnica Y os desenvolvedores concluem a atividade de análise de requisitos em menos tempo e com um conjunto de requisitos mais completo do que utilizando a técnica X”

- Um estudo experimental tem como objetivo colher dados, em um ambiente controlado, para confirmar ou negar a hipótese



Hipóteses e Variáveis

- Hipóteses devem levar à definição de variáveis
- **Variáveis independentes (ou fatores, quando controladas)**
 - Referem-se à entrada do processo de experimentação, podendo ser controladas durante este processo
 - Representam a causa que afeta o resultado do processo de experimentação. Quando é possível seu controle, os valores são chamados de "tratamentos"
- **Variáveis dependentes**
 - Referem-se à saída do processo de experimentação, sendo afetadas durante o processo de experimentação
 - Representam o efeito da combinação dos valores das variáveis independentes (incluindo os fatores). Seus possíveis valores são chamados de "resultados"



Hipóteses e Variáveis

"Utilizando a técnica Y os desenvolvedores concluem a atividade de análise de requisitos em menos tempo e com um conjunto de requisitos mais completo do que utilizando a técnica X"

Variáveis Independentes	Técnica utilizada (tratamentos: Y e X) Caracterização do desenvolvedor Caracterização da aplicação
Variáveis Dependentes	Tempo de execução da atividade % de requisitos corretos encontrados



Variáveis e seus Valores

➤ As variáveis de um estudo podem ser:

- Qualitativas: os tratamentos representam tipos
- Quantitativas: os tratamentos representam níveis de aplicação da variável

➤ Os valores das variáveis são coletados em escalas:

- Existem diversas escalas para coleta e representação de valores: nominal, ordinal, intervalar e razão
- As escalas determinam as operações que podem ser aplicadas sobre os valores das variáveis



Escalas: Nominal

➤ Os valores de uma escala nominal representam diferentes tipos de um elemento, sem interpretação numérica e de ordenação entre eles

➤ Exemplos em software incluem:

- Diferentes medidas de tamanho de software (linhas de código, pontos por função, pontos por caso de uso, ...)
- Diferentes linguagens de programação (Java, C++, C#, Pascal, ...)

➤ A escala não nos permite dizer, por exemplo, que linhas de código é maior do que pontos por função ou que Java é menor que C#



Escalas: Ordinal

- Os valores de uma escala ordinal representam diferentes tipos de um elemento que podem ser ordenados, ainda que sem qualquer interpretação numérica
- Exemplos em software incluem:
 - Diferentes níveis no CMMI (Nível 1, ..., Nível 5) ou MPS.BR (Nível G, ..., Nível A)
- A escala permite dizer que, no CMMI, "Nível 2" é menor do que "Nível 3", mas não permite dizer que a diferença de qualidade entre empresas do "Nível 2" e empresas do "Nível 3" é a mesma entre empresas do "Nível 3" e "Nível 4"



Escalas: Intervalar

- Os valores de uma escala intervalar podem ser ordenados e distâncias entre valores consecutivos possuem a mesma interpretação, porém a razão entre estes valores não tem significado
- Por exemplo: embora possamos dizer que 2006 é um ano após 2005 e um ano antes de 2007, não faz sentido calcular a razão entre 2006 e 2007.
- Isto ocorre porque toda escala intervalar possui um zero arbitrário (no caso das datas, o ano zero)



Escalas: Intervalar

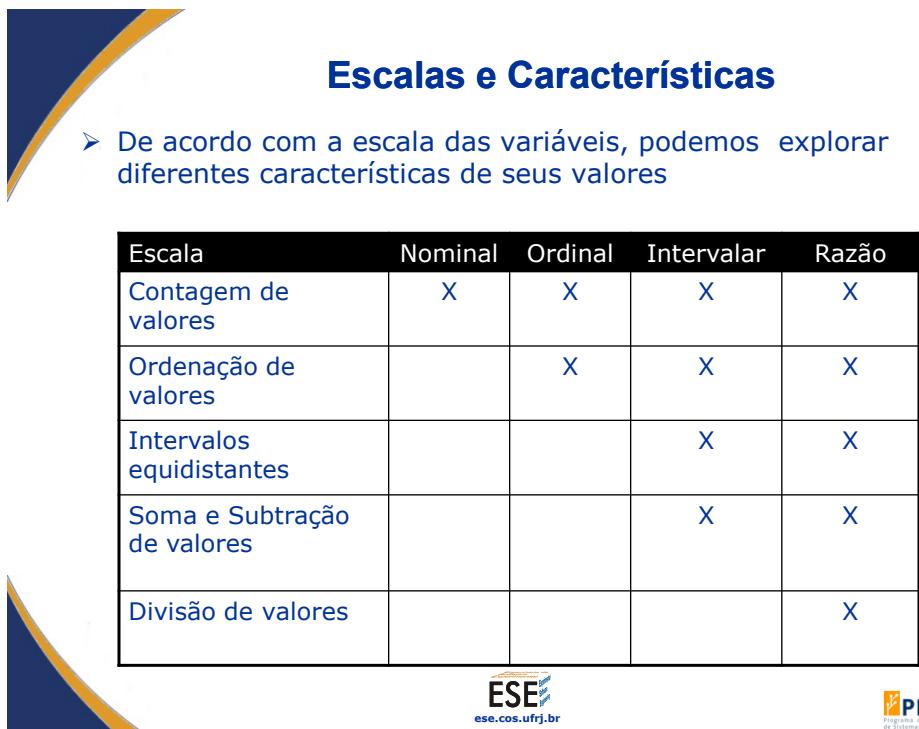
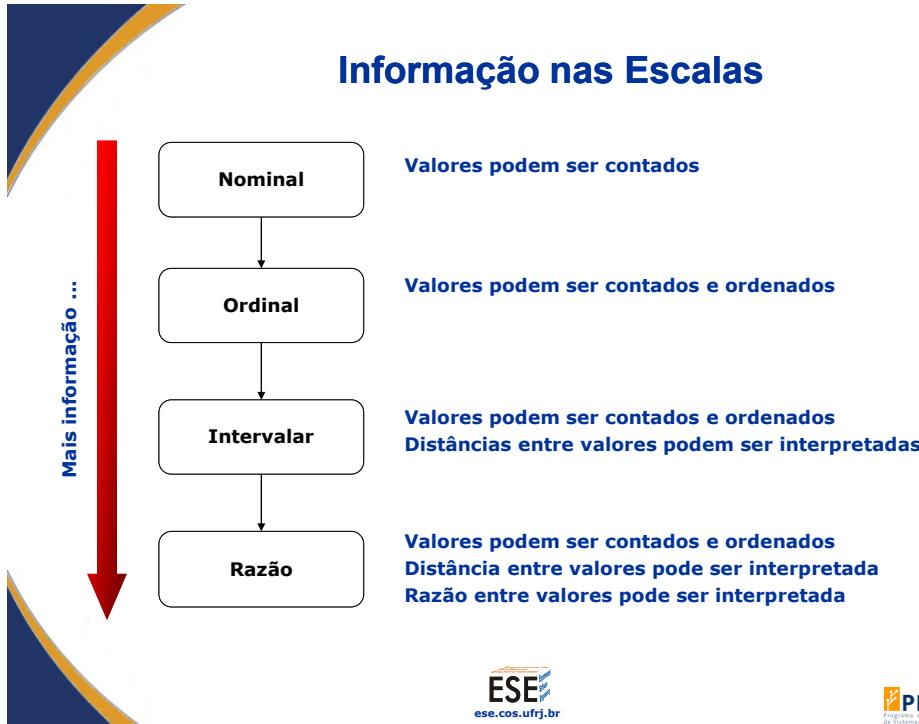
- As escalas de *Likert* são um exemplo de escala intervalar muito utilizadas em estudos relacionados ao software
 - Em uma escala de *Likert* definimos diversos nomes que representam, em geral, a intensidade de uma propriedade que não pode ser diretamente medida
 - Por exemplo, podemos construir uma escala de *Likert* para avaliar o impacto de um risco usando os seguintes valores: muito alto, alto, médio, baixo e muito baixo
 - Embora seja impossível averiguar a distância entre estes valores no mundo real, assume-se que esta distância é muito próxima entre os diferentes valores



Escalas: Razão

- Os valores de uma escala razão podem ser ordenados, distâncias entre valores consecutivos possuem o mesmo significado e a razão entre valores pode ser interpretada
- Exemplos em software incluem o tamanho de um sistema, o esforço necessário para a sua construção e o tempo de realização do projeto que resultou no sistema
- A escala permite dizer, por exemplo, que um software com X linhas de código numa determinada linguagem é duas vezes menor que um software de 2X linhas de código na mesma linguagem





Exemplo

"Utilizando a técnica Y os desenvolvedores concluem a atividade de análise de requisitos em menos tempo e com um conjunto de requisitos mais completo do que utilizando a técnica X"

Variáveis Independentes	Técnica utilizada (tratamentos: Y e X) → Escala nominal com dois tratamentos Caracterização dos participantes e aplicação → Escala nominal ou ordinal
Variáveis Dependentes	Tempo de execução da atividade → Escala razão % de requisitos corretos encontrados → Escala razão



Seleção de Participantes

- Está diretamente ligada a possibilidade de generalização dos resultados do estudo experimental
 - Amostra deve ser representativa da população desejada para a qual se quer generalizar os resultados
- Tipos de amostragem:
 - Probabilística
 - Amostragem aleatória simples: participantes são selecionados aleatoriamente a partir de uma lista da população
 - Amostragem sistemática: o primeiro participante é selecionado aleatoriamente a partir de uma lista da população e, a partir daí, todo enésimo participante é selecionado
 - Amostragem aleatória estratificada : a população é dividida em um número de grupos com distribuição conhecida entre os mesmos
 - Não-probabilística
 - Amostragem por conveniência: os participantes mais próximos e mais convenientes são selecionados da população
 - Amostragem por cota: utilizada para conseguir participantes de diferentes elementos de uma população



Projeto do Estudo Experimental

- Decide quais análises estatísticas podem ser aplicadas e quais escalas de mensuração podem ser utilizadas
 - O projeto de um estudo experimental descreve como os testes são organizados e executados
 - A escolha do projeto afeta a análise do estudo experimental e vice-versa
- Para projetar o estudo experimental é necessário contemplar as hipóteses
 - Para descobrir quais análises estatísticas deverão ser realizadas para rejeitar a hipótese nula



Princípios Gerais para o Projeto Experimental

- Aleatoriedade (*Randomization*)
 - Todos os métodos estatísticos utilizados para analisar os dados requerem que as observações sejam de variáveis independentes aleatórias
- Agrupamento (*Blocking*)
 - Para casos onde existe um Fator sobre o qual não há interesse mas que pode influenciar a resposta
 - Normalmente utilizado para eliminar efeitos indesejados em um estudo
- Balanceamento
 - Atribuir para cada tratamento um número igual de participantes



Princípios Gerais para o Projeto Experimental

- Arranjos experimentais:
 - Um fator e dois tratamentos
 - Um fator e mais de dois tratamentos
 - Dois fatores e dois tratamentos
 - Mais de dois fatores e tratamentos
- Para estudos experimentais com 1 fator e dois tratamentos:
 - Arranjo completamente aleatório
 - *Arranjo de comparação pareado (paired comparison design)*
- Para estudos experimentais com 1 fator e mais de dois tratamentos
 - Arranjo completamente aleatório
 - *Arranjo completamente aleatório em bloco*



Princípios Gerais para o Projeto Experimental

- Para estudos experimentais com 2 fatores e 2 tratamentos
 - arranjo em fatorial 2×2
 - arranjo aninhado em 2 estágios
- Para estudos experimentais com mais de 2 fatores e tratamentos
 - arranjo em fatorial 2^k
 - arranjo em fatorial fracionado 2^k
 - arranjo de meio fatorial fracionado em arranjo de fatorial fracionado de 2^k
 - arranjo em $\frac{1}{4}$ de fatorial fracionado em arranjo de fatorial fracionado de 2^k





Instrumentação

- No planejamento, os instrumentos para o estudo experimental são identificados e preparados
 - Antes da execução, os instrumentos são desenvolvidos para o estudo específico
- Objetivo geral dos instrumentos:
 - Prover *meios* para realizar e monitorar o estudo experimental, sem afetar o controle do mesmo
- Tipos de instrumento para um estudo experimental:
 - Objetos do estudo
 - Formulário de Consentimento
 - Diretrizes (*guidelines*)
 - Instrumentos de medição




Instrumentação

- Instrumentos para um estudo experimental:
 - Objetos do estudo
 - o É importante escolher os objetos apropriados
 - Exs.: documentos de especificação ou código
 - Formulário de Consentimento
 - o Considera as questões éticas e formaliza a participação no estudo (usado quando os participantes são indivíduos)
 - Diretrizes (*guidelines*)
 - o São importantes para guiar os participantes durante o estudo experimental
 - Exs.: descrições de processos e *checklists* de execução
 - Instrumentos de medição
 - o As medições em um estudo são conduzidas via coleta de dados
 - É necessário preparar os formulários e questões para entrevista e avaliá-los





Validade do Estudo

- Questão fundamental para um estudo:
 - Quão válidos são os resultados do estudo?
- A questão da validade deve ser considerada na fase de planejamento
 - Para que seja planejada a validação adequada para os resultados do estudo experimental
- Validade adequada não implica em validade geral!
 - Se são desejadas conclusões mais gerais, a validade deve cobrir um escopo mais geral também




Validade do Estudo

- Tipos de ameaças à validade aos resultados experimentais:
 - Validação de Conclusão (*conclusion validity*)
 - o Diz respeito à relação entre o tratamento e ao resultado
 - Validação Interna (*internal validity*)
 - o Diz respeito se o tratamento realmente **causa** o resultado (relação de causa e efeito)
 - Validação do Constructo (*construct validity*)
 - o Diz respeito à relação entre a teoria e a observação
 - Se o tratamento reflete corretamente a construção da causa
 - Se o resultado reflete corretamente a construção do efeito
 - Validação Externa (*external validity*)
 - o Diz respeito a possibilidade de generalização dos resultados



Validade do Estudo

- Prioridade entre os tipos de ameaças à validade:
 - Há um conflito entre alguns tipos de ameaças à validade
 - Quando se aumenta um tipo, outro tipo pode diminuir
 - Trata-se de um problema de otimização
 - Prioridade comum em testes de teoria:
 1. Validade Interna
 2. Validade do Constructo
 3. Validade de Conclusão
 4. Validade Externa
 - Prioridade em pesquisa aplicada (maioria dos estudos em engenharia de software):
 1. Validade Interna
 2. Validade Externa
 3. Validade do Constructo
 4. Validade de Conclusão



Fase de Execução

- Identifica o momento no qual o estudo será executado
- Os tratamentos são aplicados aos participantes
- Três passos:
 - Preparação – quando os participantes são escolhidos e os formulários (e outros objetos) preparados
 - Execução – quando os participantes realizam suas tarefas de acordo com os diferentes tratamentos e os dados são coletados
 - Validação dos dados – quando os dados coletados são validados



Preparação

- Existem dois aspectos importantes na preparação:

1. Selecionar e informar os participantes

- o Pessoas que estão dispostas a atuar como participantes devem ser encontradas
- o É essencial que as pessoas estejam motivadas e dispostas a participar durante todo o experimento
- o Caso as pessoas escolhidas não componham um conjunto representativo das pessoas alvo dos resultados do experimento, isto pode ser uma ameaça a validade externa do experimento

2. Preparar material (formulários e ferramentas)

- o Antes da execução em si, todos os instrumentos do estudo experimental devem estar prontos
 - Isto inclui: objetos do estudo, diretrizes para o estudo, formulários de consentimento, medição e ferramentas
- o Os instrumentos necessários são determinados pelo projeto do estudo experimental e pelo método utilizado para coleta de dados
 - Em alguns casos é apropriado preparar um conjunto pessoal de instrumentos para cada participante



Execução

- Estudos experimentais podem ser executados de diversas maneiras:

- Alguns estudos podem ser realizados em uma única rodada (*trial*), na qual todos os participantes estão reunidos
 - o Uma vantagem é que o pesquisador pode estar presente durante a execução e, caso surjam questões, elas podem ser tratadas imediatamente
- Outros estudos são executados durante um longo período de tempo
 - o Nesses casos, o pesquisador não pode participar de todos os detalhes da execução do estudo e da coleta de dados

- Coleta de dados:

- Dados podem ser coletados manualmente ou automaticamente durante a execução do estudo experimental

- Ambiente de experimentação

- Se um estudo experimental é realizado em um projeto de desenvolvimento regular, o estudo não deve afetar o projeto mais que o necessário
- No entanto, existem casos, nos quais é apropriado haver algumas interações entre o estudo experimental e o projeto



Validação dos Dados

- Quando os dados são coletados, o pesquisador deve verificar se os dados são razoáveis e se foram coletados corretamente
 - Isto se relaciona com o fato dos participantes terem entendido e preenchido corretamente os formulários
 - Necessário avaliar a qualidade dos dados e garantir que foram capturados de acordo com o planejado



Variáveis e Execução

- A execução de um estudo experimental usualmente consiste de uma série de rodadas (*trials*)
 - Em cada rodada, um participante utiliza um tratamento do conjunto de variáveis independentes e gera resultados para cada variável dependente
 - Estes resultados são colhidos em tuplas do tipo $A_i = \{T_i, R_i\}$, onde T_i é o conjunto ordenado de cada tratamento aplicado a cada variável independente pelo participante i e R_i é o conjunto ordenado de cada resultado observado pelo mesmo participante para cada variável dependente
 - Estes resultados serão motivo da análise de dados do estudo experimental



Variáveis e Execução

- Alguns dados tabulados após a execução de um estudo hipotético. Estes dados serão usados nos próximos exemplos.

Participante	Técnica	Tempo (dias)	% Corretos Encontrados
1	Y	10	83%
2	Y	13	73%
3	Y	12	87%
4	Y	13	78%
5	Y	10	74%
6	Y	14	74%
7	Y	14	87%
8	Y	13	75%
9	Y	14	86%
10	Y	14	82%
11	Y	13	77%
12	X	13	90%
13	X	9	89%
14	X	11	88%
15	X	14	87%
16	X	9	97%
17	X	12	81%
18	X	9	82%
19	X	12	86%
20	X	11	92%
21	X	14	96%
22	X	13	98%

ese.cos.ufrj.br



Programa de Engenharia

de Sistemas e Computação

Variáveis e Execução

- Após a tabulação dos dados, medidas de tendência central, dispersão e dependência podem ser utilizadas em conjunto com a análise gráfica para que o analista tenha um melhor "entendimento" sobre os dados
- Este entendimento será útil na seleção e aplicação das técnicas de inferência estatística, que por sua vez avaliarão a aceitação ou rejeição das hipóteses



Visualização Gráfica

- Um gráfico representa visualmente a informação tabulada
 - Gráficos são normalmente mais fáceis de entender do que grandes quantidades de dados tabulados
 - A apresentação espacial dos dados ajuda na identificação de grupos e visualização de relacionamentos entre eles
 - Os gráficos geralmente podem ser lidos mais rapidamente que a informação tabulada
- Métodos de representação gráfica
 - Histogramas
 - Gráficos de torta (ou pizza)
 - Diagramas de dispersão



Visualização Gráfica

- Os métodos de visualização gráfica podem depender da classificação de suas variáveis como contínuas ou discretas
- Variáveis discretas podem assumir qualquer valor dentro de um conjunto finito de valores
 - Elas são comuns nas escalas nominal e ordinal, mas também podem ocorrer nas escalas intervalar e razão
- Variáveis contínuas podem assumir qualquer valor dentro de um número infinito de valores em um intervalo
 - Elas são comuns nas escalas intervalar e razão



Histograma

➤ Apresenta os valores observados para uma variável de interesse no domínio da freqüência

- A freqüência indica o número ou percentual de ocorrências de cada valor no conjunto de valores coletados
- Se os dados são discretos, cada informação é representada em uma barra, cuja altura representa o número de vezes que o valor ocorre nos valores coletados
- Se os dados são contínuos, eles devem ser discretizados, ou seja, separa-se os dados em regiões equidistantes e conta-se quantas vezes valores de cada região são encontrados dentre os valores coletados. Em seguida, uma barra é traçada como no caso de dados discretos



Histograma

➤ É um método comum de apresentação de dados numéricos e em qualquer escala, pois envolve apenas contagem

➤ Os histogramas também permitem relacionar os dados observados com distribuições de freqüência conhecidas

- Estas distribuições possuem propriedades matemáticas das quais foram derivados os testes de inferência estatística
- Se os dados observados não seguem estas propriedades, não podemos confiar nos resultados dos testes
- Nestes casos, temos que utilizar outros tipos de testes, mais adequados à distribuição observada nos dados

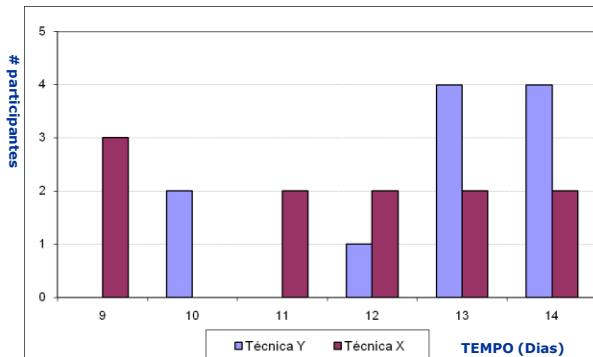


Histograma

- Histograma do tempo consumido pelos participantes na atividade de análise, de acordo com a técnica utilizada

Tempo (dias)	Técnica Y	Técnica X
9	0	3
10	2	0
11	0	2
12	1	2
13	4	2
14	4	2

* Tabela com a distribuição dos dados



Histograma Cumulativo

- Um histograma cumulativo apresenta a frequência de ocorrência de valores menores ou iguais a um dado valor

- Cada barra no gráfico representa o somatório das barras anteriores em um histograma convencional
- Em diversas situações, já é possível ter alguma sugestão sobre a aceitação ou rejeição da hipótese observando o histograma cumulativo dos dados (entretanto apenas os testes estatísticos poderão confirmar ou não a hipótese)
- Como é necessário ordenar os valores, os histogramas cumulativos somente podem ser aplicados em variáveis em escala ordinal, intervalar ou razão



Histograma Cumulativo

- Histograma cumulativo do tempo usado pelos participantes na atividade de análise com as técnicas X e Y

Tempo (dias)	Técnica Y	Técnica X
9	0	3
10	2	3
11	2	5
12	3	7
13	7	9
14	11	11

* Tabela com a distribuição acumulada dos dados

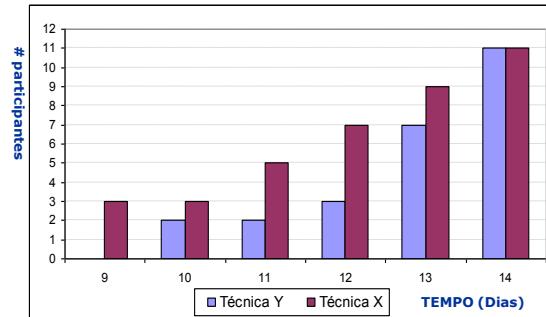


Gráfico de Torta

- Um gráfico de torta (ou pizza) apresenta a freqüência relativa (ou percentual) de ocorrência dos dados, dividindo os dados em um conjunto de classes distintas e apresentando-os como fatias de um círculo

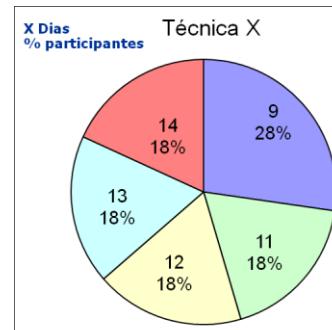
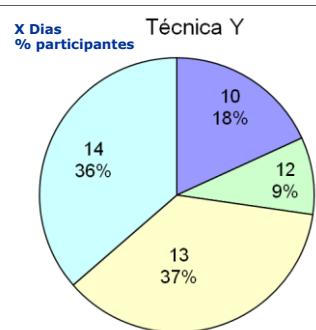


Diagrama de Dispersão

- É a representação dos valores observados para duas ou mais variáveis através de gráficos cartesianos
 - Cada eixo do gráfico representa uma das variáveis, formando tuplas (pares ou mais dimensões) entre elas
 - Essa forma de representar dados ajuda a identificar padrões que possam sugerir a natureza da relação entre as variáveis
 - Os gráficos de dispersão também ajudam a identificar valores que estejam muito distantes do comportamento normal dos dados do conjunto observado (*outliers*)
 - Estes *outliers* podem distorcer a análise estatística e usualmente são eliminados antes dos testes de inferência



Diagrama de Dispersão

- Dispersão entre o percentual de requisitos corretos que foram encontrados e o tempo de execução da atividade de análise, para as técnicas X e Y

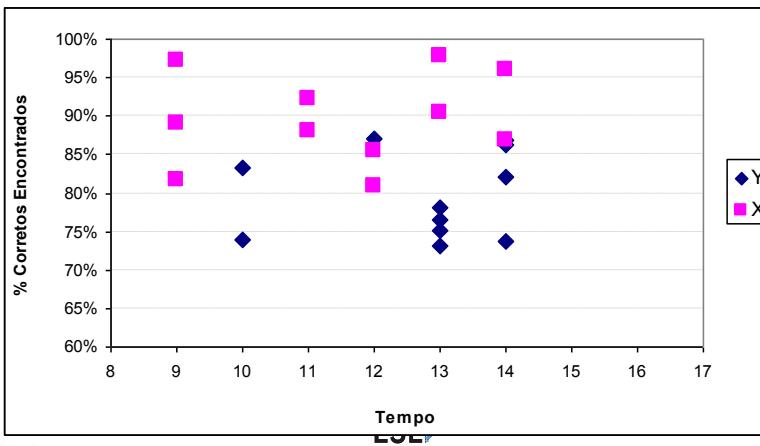


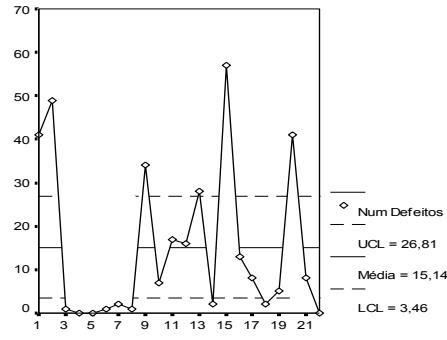
Gráfico de Controle

- Ferramenta estatística que permite observar o comportamento dos dados quantitativos que representam as características sob investigação
- Um Gráfico de Controle consiste de três linhas paralelas:
 - uma linha média, que reflete o comportamento médio apresentado pelos dados
 - uma linha externa denominada Limite Superior de Controle (UCL – *Upper Control Limit*)
 - outra linha externa chamada Limite Inferior de Controle (LCL – *Lower Control Limit*)



Gráfico de Controle

- Se o comportamento da característica está sob controle, os valores irão oscilar em torno da linha central (por exemplo, o número médio de defeitos por versão), dentro da faixa determinada por LCL e UCL
- Uma vez que o comportamento da característica esteja sob controle, é pequena a probabilidade de obter-se um valor fora destes limites



Estatística Descritiva



- Objetivos
- Medidas de tendência central
- Medidas de dispersão
- Distribuição de frequência
- Exemplo
- Medidas de dependência



Objetivos

- Após a coleta dos dados de um estudo experimental, a estatística descritiva é utilizada para descrever algumas características relevantes dos dados coletados
- Junto com a análise gráfica, a estatística descritiva apóia a análise inicial dos dados, medindo as dependências e relacionamentos entre eles
- A estatística descritiva tem como meta passar uma visão geral de como o conjunto de dados está distribuído



Medidas de Tendência Central

- Indicam o meio do conjunto de valores observados
 - **Média** (aritmética): a média pode ser considerada como o centro de gravidade dos dados coletados.
 - o É calculada pelo somatório dos valores coletados, dividido pela quantidade de valores

$$\mu = \frac{\sum x_i}{n}$$
 - **Mediana**: valor que divide ao meio um conjunto de dados, ou seja, o número de valores coletados que está abaixo da mediana deve ser o mesmo que está acima.
 - o É calculada colocando os valores em ordem crescente ou decrescente e selecionando o elemento central. Em caso de número par de valores é calculada a média dos valores centrais
 - **Moda**: representa o valor mais comum dentre o conjunto de valores coletados.
 - o É calculada pela contagem do número de ocorrências (frequência) de cada valor, selecionando o mais comum. Se dois ou mais valores ocorrem com maior frequência, os valores coletados possuem diversas modas



Medidas de Tendência Central

- Outras medidas relevantes
 - **Valor mínimo**: representa o menor valor entre os dados coletados
 - **Valor máximo**: representa o maior valor entre os dados coletados
 - **Percentil**: Em uma amostra de 100 elementos, o percentil X% é o valor que divide a amostra em X valores menores que ele e (100-X) valores maiores que ele. A mediana é conhecida como percentil 50%.
 - **Quartil**: são os valores que representam o percentil 25% (ou primeiro quartil), a mediana (segundo quartil) e o percentil 75% (terceiro quartil)

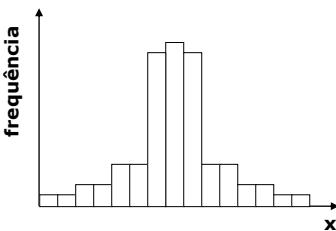


Medidas de Dispersão

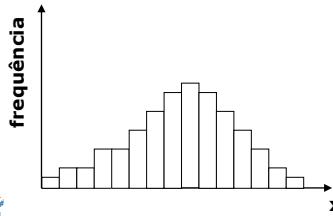
- Medem o quanto os valores coletados estão dispersos ou concentrados em torno de seu valor central
 - **Faixa:** é a diferença entre o **valor máximo** e o **valor mínimo** dos valores coletados
 - **Variância:** é a soma do quadrado da diferença entre cada valor e a média dos valores coletados, dividida pelo número de valores coletados menos 1

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$
 - **Desvio Padrão:** é a raiz quadrada da variância, sendo a medida de dispersão mais comumente utilizada

$$\sigma = \sqrt{\sigma^2}$$



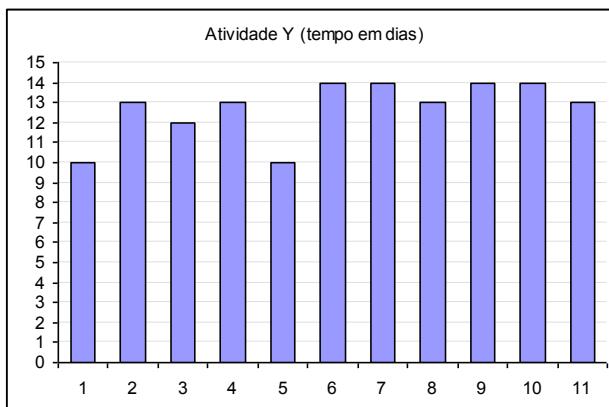
ESE
ese.cos.ufrj.br



PESC
Programa de Engenharia de Sistemas e Computação

Estatística Descritiva

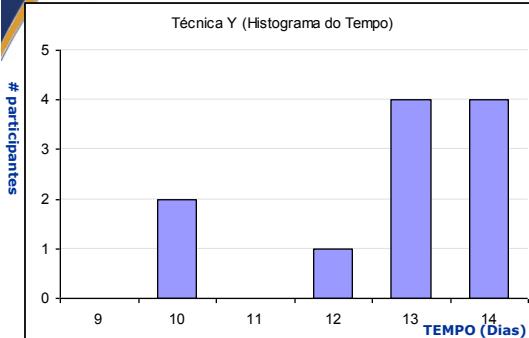
- Gráfico de barras com o tempo que foi consumido por cada participante que aplicou a técnica Y na atividade de análise



ESE
ese.cos.ufrj.br

PESC
Programa de Engenharia de Sistemas e Computação

Estatística Descritiva



Medidas de Tendência Central	
Média	12,73
Mediana	13
Modas	13 e 14
Faixa	4
Mínimo	10
Máximo	14
1º Quartil	12,5
3º Quartil	14
Variância	2,22
Desvio Padrão	1,49



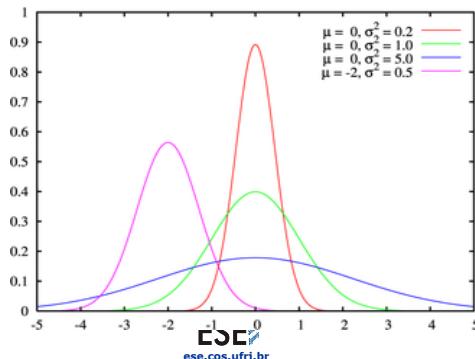
Distribuições de Freqüência

- Como vimos na visualização gráfica, um conjunto de dados pode ser mapeado no domínio da frequência e apresentado na forma de histogramas
- Os histogramas permitem verificar se a distribuição dos dados segue uma distribuição clássica, como normal, uniforme, beta, entre outras
- A distribuição normal, em particular, é importante para alguns testes estatísticos, que exigem que os dados que serão analisados sigam uma distribuição normal



Distribuições de Frequência

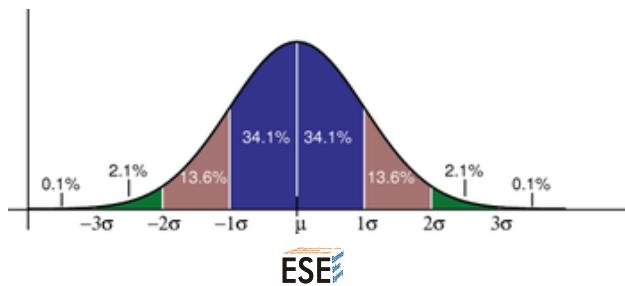
- A distribuição normal possui o formato de um sino, com as pontas se estendendo a direita e esquerda do centro
 - A curva é simétrica em relação a sua média e a largura do sinal é proporcional ao seu desvio padrão
 - Assim, a curva pode ser descrita matematicamente apenas com base em sua média e desvio padrão



PESC
Programa de Engenharia
de Sistemas e Computação

Distribuição Normal

- Se um conjunto de dados numéricos segue a distribuição normal, é possível afirmar que
 - 68% de todas as observações estão entre um desvio padrão a mais ou a menos da média
 - 95,5% de todas as observações estão entre 2 desvios padrão a mais ou a menos da média
 - 99,7% de todas as observações estão entre 3 desvios padrão a mais ou a menos da média



PESC
Programa de Engenharia
de Sistemas e Computação

Medidas de Dependência

- Quando duas ou mais variáveis estão relacionadas em um estudo pode ser útil calcular o grau de dependência entre elas
- As medidas de dependência determinam a força e direção do relacionamento entre duas ou mais variáveis quantitativamente avaliadas
 - A medida de dependência mais comumente utilizada é o coeficiente de correlação
 - Se o estudo relaciona duas variáveis, a correlação entre elas é representada como um número simples
 - Se o estudo relaciona mais de duas variáveis, a correlação é representada como uma matriz



Medidas de Dependência

- A correlação entre duas variáveis varia entre -1 e 1
- A correlação -1 indica que um valor alto em uma variável normalmente ocorre associado a um valor baixo na segunda variável (↑↓)
- A correlação 1 indica que um valor alto em uma variável normalmente ocorre associado a um valor alto da segunda variável (↑↑)
- A correlação próxima de zero indica que não podemos inferir nenhum relacionamento entre as variáveis



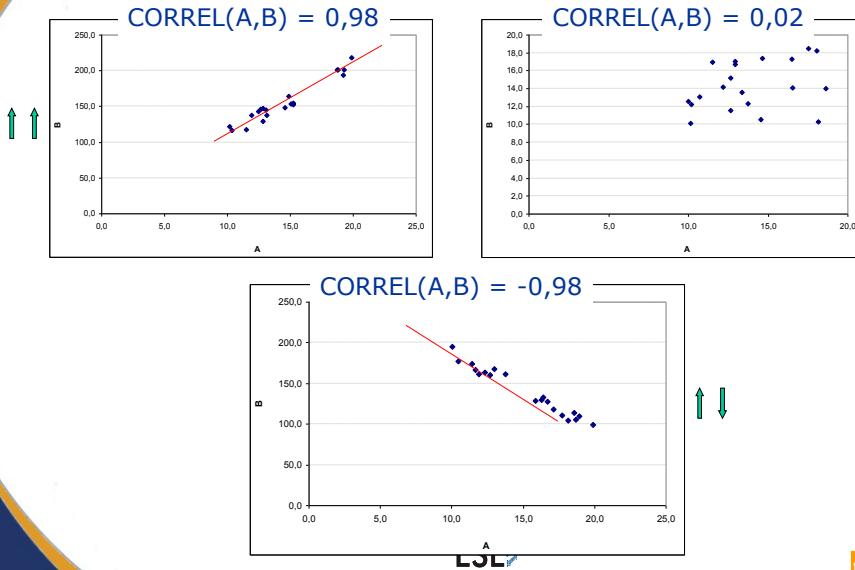
Correlação de Pearson

➤ Coeficiente de correlação mais comum

- Quantifica a força de associação linear entre duas variáveis e descreve o quanto uma linha reta se ajustaria através da representação cartesiana de seus valores
- O coeficiente assume que os valores das variáveis seguem aproximadamente distribuições normais
- Devido a forma da distribuição normal, esta condição é indicada pela formação de uma nuvem em forma de elipse em um gráfico de dispersão que apresente estes valores



Medidas de Dependência



Correlação de Spearman

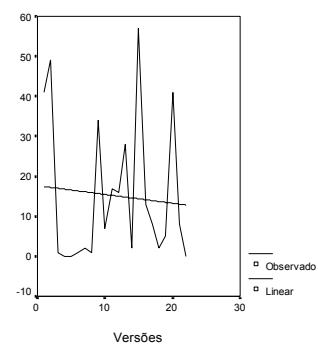
- Outro exemplo de correlação é o coeficiente de *Spearman*
 - O método se baseia no *ranking* dos valores coletados em seu conjunto, não nos valores propriamente ditos
 - Com isto, este método pode ser aplicado sobre valores em uma escala ordinal (não apenas intervalar e razão)

- A correlação de *Spearman* também é aplicável quando os valores não parecem seguir uma distribuição normal
 - Por exemplo, exibir uma relação crescente ou decrescente num formato de curva (ou seja, não linear)
 - No caso específico de uma curva exponencial, a correlação pode ser aplicada sobre os logaritmos dos valores



Análise de Regressão

- Enquanto um coeficiente de correlação mede apenas a intensidade e direção da relação entre duas variáveis, a análise de regressão fornece uma equação que descreve a natureza do relacionamento.
- Em Análise de Regressão Linear Simples, o interesse é predizer o valor de variáveis dependentes baseado no valor de apenas uma variável independente.
 - O método dos mínimos quadrados produz uma reta calculada de forma a minimizar a soma dos quadrados dos desvios, objetivando encontrar uma reta que mais se aproxime do conjunto de dados em questão



Estatística Descritiva

- De acordo com a escala das variáveis, podemos calcular as seguintes medidas da estatística descritiva

Escala	Nominal	Ordinal	Intervalar	Razão
Média			X	X
Mediana		X	X	X
Moda	X	X	X	X
Faixa		X	X	X
Variância			X	X
Desvio Padrão			X	X
Corr Pearson			X	X
Corr Spearman		X	X	X

Análise de *Outliers*

- Conceito
- Condições de ocorrência
- Identificação visual
- Identificação numérica

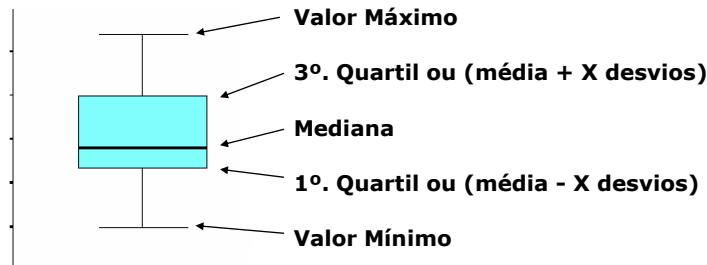


Remoção de *Outliers*

- Valores extremos (ou *outliers*) são valores observados que estão muito distantes dos demais valores observados
 - Estes dados podem representar erros no conjunto de valores observados e usualmente são removidos deste conjunto antes de se aplicar as técnicas de inferência estatística
 - Os *outliers* podem ocorrer por problemas de aplicação da sistemática prevista no projeto do estudo, por erros de digitação, problemas de interpretação ou motivação dos participantes
 - É importante verificar as origens de cada *outlier*, pois eles podem ser efetivamente observações válidas e que deveriam ser consideradas no universo de estudo

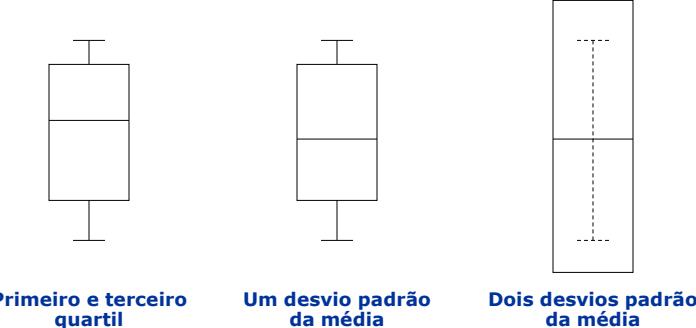
Identificação Visual

- Os *outliers* podem ser identificados visualmente, através de gráficos de dispersão e *box-plots*
 - Os diagramas de *box-plot* foram idealizados para apresentar a distribuição de dados quantitativos
 - Eles utilizam medidas de tendência central e dispersão para caracterizar esta distribuição



Identificação Visual

- Box-plots do percentual de requisitos corretos encontrados por participantes que aplicaram a técnica Y



Identificação Numérica

- Métodos de eliminação de *outliers* geralmente removem valores que estão acima de uma determinada distância da média ou da mediana
 - Valores muito próximos destes limites nem sempre precisam ser removidos do conjunto de dados (subjetividade)
 - A distância normalmente é determinada por um quartil, um percentil ou um número de desvios padrão

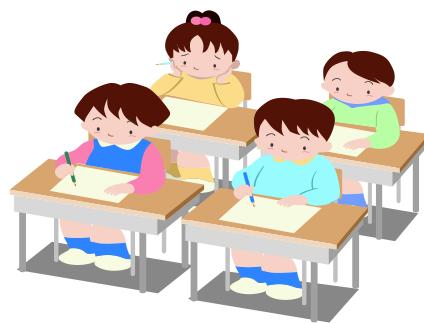
Identificação Numérica

- Remoção de *outliers* do percentual de requisitos corretos encontrados por participantes que aplicaram a técnica Y usando um desvio padrão

Medida	Valor
Mínimo	73%
Média - 1dp	74%
Média	80%
Média + 1dp	86%
Máximo	87%

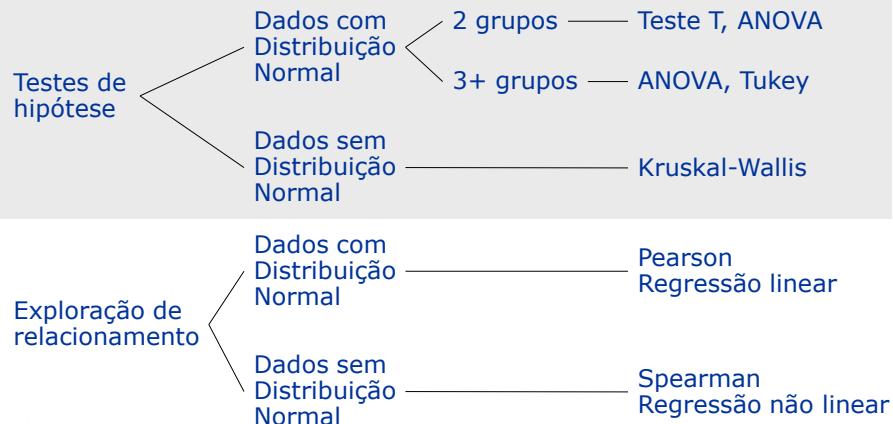
Participante	Tempo (dias)	% Corretos Encontrados
1	10	83%
2	13	73%
3	12	87%
4	13	78%
5	10	74%
6	14	74%
7	14	87%
8	13	75%
9	14	86%
10	14	82%
11	13	77%

Testes de Hipótese



- Tipos de estudo experimental
- Testes de hipótese
- Erro, potência e p-value
- Tipos de teste de hipótese
- T-test
- Mann-Whitney
- ANOVA
- Kruskal-Wallis

Tipos de Estudo Experimental



Testes de Hipótese

- Conforme já vimos, um estudo experimental tem como objetivo colher dados para confirmar ou negar uma hipótese
- Em geral, são definidas duas hipóteses
 - **Hipótese nula (H_0)**: indica que as diferenças observadas no estudo são coincidentes, ou seja, é a hipótese que o analista deseja rejeitar com a maior significância possível
 - **Hipótese alternativa (H_1)**: é a hipótese inversa à hipótese nula, que será aceita caso a hipótese nula seja rejeitada
- Os testes estatísticos verificam se é possível rejeitar a hipótese nula, de acordo com um conjunto de dados observados e suas propriedades estatísticas



Testes de Hipótese

- Em geral, os testes realizados em Engenharia de Software comparam médias entre grupos de participantes realizando tratamentos diferentes

"Utilizando a técnica Y os desenvolvedores concluem a atividade de análise de requisitos em menos tempo e com um conjunto de requisitos mais completo do que utilizando a técnica X"

Hipótese Nula: $\mu (\text{Tempo}_Y) = \mu (\text{Tempo}_X)$

Hipótese Alternativa: $\mu (\text{Tempo}_Y) \neq \mu (\text{Tempo}_X)$



Testes de Hipótese

➤ Procedimentos

- Fixar o nível de significância do teste
- Obter uma estatística (estimador do parâmetro que se está testando) que tenha distribuição conhecida sob H_0
- Através da estatística de teste e do nível de significância, construir a região crítica
- Usando as informações amostrais, obter o valor da estatística (estimativa do parâmetro)
- Se valor da estatística pertencer à região crítica, rejeita-se a hipótese nula, aceitando-se a hipótese alternativa
- Caso contrário, não se rejeita a hipótese nula e nada se pode dizer a respeito da hipótese alternativa



Tipos de Erro

➤ A verificação das hipóteses sempre lida com algum tipo de risco, que implica que um erro de análise pode acontecer

- O erro do tipo I (α) acontece quando o teste estatístico indica um relacionamento entre causa e efeito e o relacionamento real não existe
- O erro do tipo II (β) acontece quando o teste estatístico não indica o relacionamento entre causa e efeito, mas existe este relacionamento

$$\alpha = P(\text{erro-tipo-I}) = P(H_{\text{NULA}} \text{ é rejeitada} \mid H_{\text{NULA}} \text{ é verdadeira})$$

$$\beta = P(\text{erro-tipo-II}) = P(H_{\text{NULA}} \text{ não é rejeitada} \mid H_{\text{NULA}} \text{ é falsa})$$



Tipos de Erro

➤ A Hipótese Nula é normalmente construída de forma a minimizar erros do tipo I

➤ Considere:

- H_0 : medicamento A = medicamento B
- H_1 : medicamento A é melhor que medicamento B

➤ Erros:

- Tipo I: medicamento A é melhor que B, quando isso não é verdade (são iguais)
- Tipo II: medicamento A é igual a B, quando isso não é verdade (A seria melhor)



Potência do Teste

- Indica a probabilidade de rejeitar a hipótese nula quando esta é falsa, ou seja, a probabilidade de decisão correta baseada na hipótese alternativa
 - O tamanho do erro durante a verificação das hipóteses depende da potência do teste estatístico
 - A potência do teste implica a probabilidade de que o teste vai encontrar o relacionamento quando a hipótese nula for falsa
 - Um teste estatístico com a maior potência possível deve ser escolhido para avaliar uma hipótese

$$\text{Potência} = 1 - \beta$$

$$\text{Potência} = P(H_{\text{NULA}} \text{ rejeitada} \mid H_{\text{NULA}} \text{ é falsa})$$



Nível de Significância

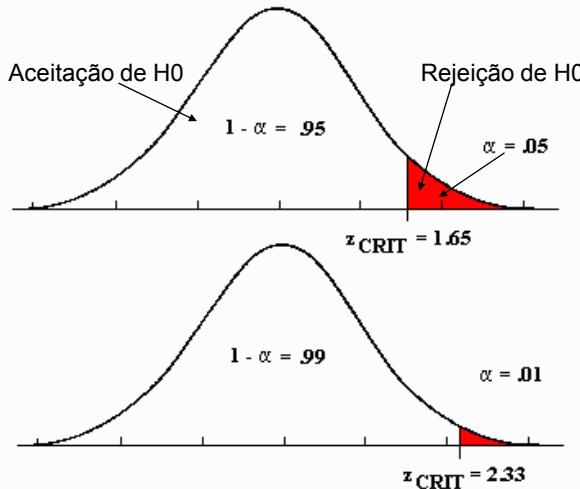
- Indica a probabilidade de cometer um erro tipo-I
 - Os níveis de significância (α) mais comumente utilizados são 10%, 5%, 1% e 0.1%
 - Chamamos de *p-value* o menor nível de significância com que se pode rejeitar a hipótese nula
 - Dizemos que há significância estatística quando o *p-value* é menor que o nível de significância adotado
 - Por exemplo, quando $p=0.0001$ pode-se dizer que o resultado é bastante significativo, pois este valor é muito inferior aos níveis de significância usuais
 - Porém, se $p=0.048$ pode haver dúvida pois, embora o valor seja inferior, ele está muito próximo ao nível usual de 5%



Região Crítica

- A designação hipótese nula advém do uso frequente do teste de hipótese na comparação de dois tratamentos, em que H_0 é a hipótese de igualdade dos tratamentos, ou seja, nulidade da superioridade do tratamento alternativo
 - A hipótese nula deve ser escrita de forma que o erro considerado mais sério seja do tipo I, ou seja, quando se rejeita H_0 sendo ela verdadeira
 - A probabilidade de se cometer um erro tipo I depende dos valores da população e é designada por α
 - O maior valor de α , para H_0 verdadeira, é chamado de nível de significância de um teste
 - Assim, o nível de significância de um teste é a probabilidade máxima com que se deseja correr o risco de um erro do tipo I

Nível de Significância



Tipos de Teste de Hipótese

- Os testes de hipótese se dividem em testes paramétricos e testes não paramétricos
- Testes Paramétricos
 - Utilizam fórmulas fechadas, derivadas de propriedades de distribuições de freqüência conhecidas (tais como equação da curva, da curva acumulada, simetria, ...)
 - Por conta disso, exigem algumas premissas sobre os dados que serão testados:
 - o **Normalidade:** os valores se concentram simetricamente em torno de uma média e quanto maior a distância desta média, menor a freqüência das observações
 - o **Homocedasticidade:** implica em variância constante entre os conjuntos de dados que serão testados, ou seja, a variância de um subgrupo não é maior que a de outro



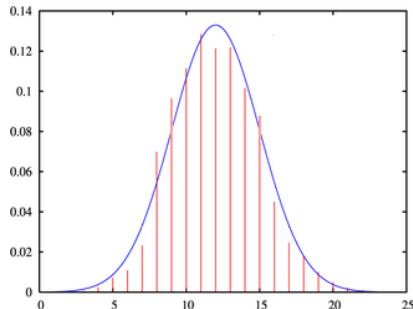
Tipos de Teste de Hipótese

- Testes Não-Paramétricos
 - Devem ser utilizados quando os dados coletados não atendem aos pressupostos esperados pelos testes paramétricos
 - São menos poderosos que os testes paramétricos, mas não presumem distribuições de probabilidade nos dados
 - Utilizam *rankings* dos valores observados ao invés dos valores propriamente ditos

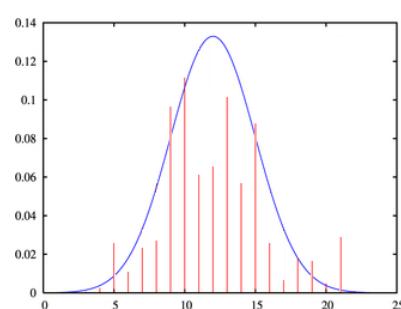


Normalidade

- Gráficos de distribuição de frequência da curva normal (em azul) e de dados hipotéticos (linhas verticais vermelhas)



Dados com distribuição
próxima à normal



Dados com distribuição
não normal

Testes de Normalidade

- Teste de Kolmogorov-Smirnov (K-S)

- Avalia se duas amostras têm distribuições semelhantes ou se uma amostra tem distribuição semelhante a uma distribuição clássica (como a distribuição normal, por exemplo)
- Frequentemente utilizado para identificar normalidade em variáveis com pelo menos 30 valores
- Detecta diferenças em relação à tendência central, dispersão e simetria, mas é muito sensível a caudas longas

Testes de Normalidade

➤ Teste de Shapiro-Wilk

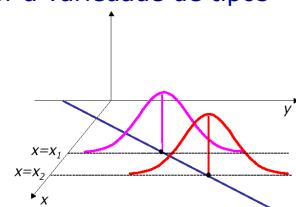
- Calcula o valor W , que avalia se uma amostra x_i segue a distribuição normal
- Frequentemente utilizado para identificar normalidade em variáveis com menos de 50 valores
- Valores pequenos calculados para W indicam que a amostra não segue a distribuição normal
- Os valores de a_i são publicados em tabelas e estão presentes em pacotes estatísticos
- Teste usado para pequenos conjuntos de dados, onde valores extremos podem dificultar o uso de K-S



Homocedasticidade

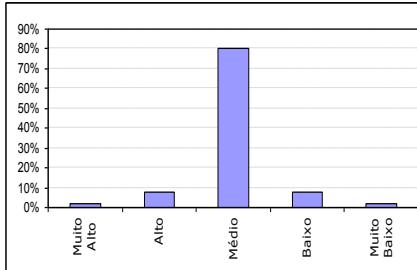
➤ Um conjunto de variáveis é homocedástico se as variáveis possuem variâncias similares

- Um exemplo clássico da falta de homocedasticidade é a relação entre o tipo de alimento consumido e o salário:
 - A medida que o salário de uma pessoa aumenta, a variedade de tipos de alimento que ela pode consumir também aumenta
 - Uma pessoa pobre geralmente gasta um valor constante em alimentação, consumindo os produtos similares
 - Uma pessoa mais rica pode eventualmente consumir produtos mais simples, mas também pode consumir produtos sofisticados
 - Assim, quanto mais rica a pessoa, maior a variedade de tipos de alimento que ela pode consumir

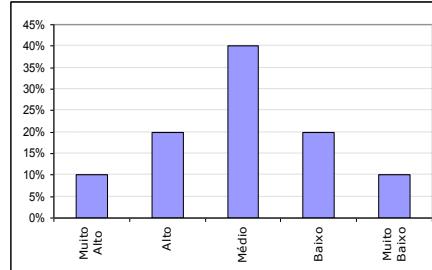


Homocedasticidade

- Valores observados em um estudo hipotético, demonstrando heterocedasticidade entre dois grupos



Grupo I



Grupo II

Testes de Homocedasticidade

- Teste de Levene

- Considere uma variável Y , com N valores divididos em k grupos, onde N_i é o número de valores no grupo i
- O teste de Levene aceita a hipótese de que as variâncias são homogêneas se o valor W for menor do que o valor da distribuição F
- Os valores da distribuição F estão disponíveis em tabelas e softwares estatísticos

Tipos de Teste de Hipótese

➤ Testes paramétricos

- Teste T
- Teste F
- ANOVA e MANOVA
- Teste de Tukey

➤ Testes não-paramétricos

- Binomial
- Chi-2
- Teste de Mann-Whitney
- Teste de Kruskal-Wallis



Tipos de Teste de Hipótese

Projeto	Teste paramétrico	Teste não-paramétrico
Um fator, um tratamento	--	Binomial Chi-2
Um fator, dois tratamentos aleatórios	Teste T Teste F	Mann-Whitney Chi-2
Um fator, dois tratamentos pareados	Teste T pareado	Wilcoxon
Um fator, mais de dois tratamentos	ANOVA	Kruskal-Wallis Chi-2



Teste T ou Student-T

- Teste paramétrico utilizado para comparar médias de duas amostras independentes
 - Trata-se de uma categoria de testes, onde diferentes testes são aplicados de acordo com diferenças nas variâncias das amostras (homocedásticas ou não)
 - Diferentes testes também são aplicados se as amostras são independentes ou pareadas
 - Dizemos que duas amostras são pareadas quando existe uma relação única entre um valor em uma amostra e um valor na segunda amostras
 - o Exemplo: uma amostra antes de um treinamento e uma amostra após o treinamento
 - Todos os tipos de teste T assumem normalidade nos valores que estão sendo testados



Teste de Mann-Whitney

- Alternativa não paramétrica para o teste T
 - Requer que as amostras sejam independentes, com dados contínuos e nas escalas ordinal, intervalar ou razão
 - Para a realização do teste as observações das amostras são reunidas em um único grupo, que é ordenado
 - As amostras são transformadas em *rankings* dentro do grupo e calcula-se o somatório dos *rankings* da menor amostra (T)
 - Finalmente, calcula-se o valor Z que é comparado com uma tabela de valores



ANOVA – Análise de Variância

- Técnica estatística cujo objetivo é testar a igualdade entre as médias de dois ou mais grupos
 - Permite comparar as médias de diversos tratamentos, sendo usada como uma extensão dos testes T
 - Avalia se a variabilidade dentro dos grupos é maior do que a existente entre os grupos
 - A técnica supõe independência, normalidade e igualdade entre as variâncias dos grupos
- Como o objetivo do método é avaliar se as médias são iguais, independente do fator, a hipótese nula do ANOVA estabelece que as variações dependentes de fator devem ser iguais a zero

$$H_0 : \tau_1 = 0; \tau_2 = 0; \dots; \tau_a = 0$$



Teste de Tukey

- Teste para comparação de médias
 - Comumente utilizado junto com o teste ANOVA quando este último acusa diferença entre as médias de múltiplas amostras
 - O teste de Tukey auxilia na identificação das amostras cujas médias diferem



Teste de Kruskal-Wallis

➤ Alternativa não-paramétrica para a análise de variância

- Como grande parte dos testes paramétricos, este teste se baseia na substituição dos valores por seus *rankings* no conjunto de todos os valores



Exemplo de Aplicação

- Defeitos em requisitos
- Análise de tempo
- Análise de defeitos
- Testes paramétricos



Exemplo de Aplicação

- O exemplo baseia-se na análise da utilização de uma técnica para detecção de defeitos em requisitos de software
- Esta técnica foi desenvolvida originalmente na língua inglesa e tinha-se dúvida se deveria ser traduzida para o português para ser utilizada no contexto dos cursos de pós-graduação em engenharia de software da COPPE/UFRJ
- Desta forma, o objetivo principal deste estudo experimental é verificar se existem diferenças significativas na utilização da técnica nas versões em inglês e português, em relação ao tempo, número de discrepâncias e defeitos encontrados



Exemplo de Aplicação

- Neste sentido foi conduzido um estudo experimental envolvendo 19 participantes, que foram divididos em dois grupos com o objetivo de aplicar a técnica de detecção de defeitos num mesmo documento de requisitos de software escrito em português
- O primeiro grupo (EP) foi composto por 11 participantes e utilizou a técnica na sua versão em inglês
- O segundo grupo (PP) foi composto por 8 participantes e utilizou uma versão traduzida da técnica para o português



Exemplo de Aplicação

Grupo	Participante	Tempo	Discrepâncias	Defeitos
EP	EP01	360	13	4
EP	EP02	600	17	5
EP	EP03	240	25	5
EP	EP04	420	25	5
EP	EP05	160	19	7
EP	EP06	110	15	4
EP	EP07	300	8	1
EP	EP08	405	3	2
EP	EP09	300	24	4
EP	EP10	360	9	3
EP	EP11	360	28	7
PP	PP02	360	20	2
PP	PP03	251	37	4
PP	PP04	300	13	5
PP	PP05	120	7	2
PP	PP07	445	27	9
PP	PP08	300	9	2
PP	PP10	270	14	6
PP	PP11	290	8	3



Exemplo de Aplicação

- Utilização do Teste T para duas amostras independentes:
 - uma vez que as variáveis são quantitativas
 - não são conhecidas as variâncias populacionais dos grupos
 - os dados estão distribuídos em duas amostras independentes, pois nenhum participante participou simultaneamente dos dois grupos

- As análises foram feitas utilizando o pacote estatístico SPSS

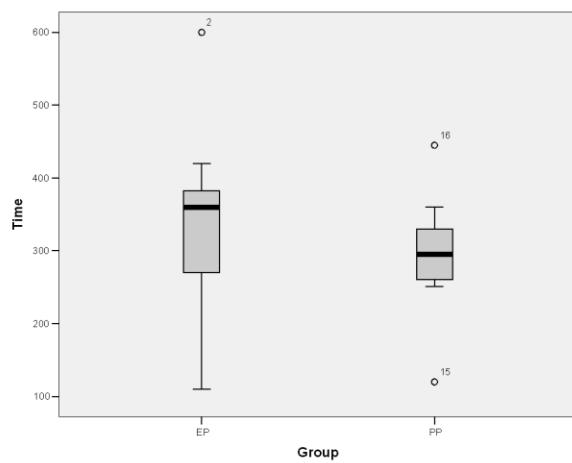


Exemplo de Aplicação Análise para a Variável Tempo

- A primeira consideração a ser feita ao conjunto de dados é relativa à normalidade e homocedasticidade (variância constante) das amostras utilizadas
- Uma análise visual inicial da distribuição é eficiente para o conhecimento do comportamento das amostras

Exemplo de Aplicação Análise para a Variável Tempo

- Outliers moderados: a princípio não apresentam problemas
- Aparente grande variabilidade entre as duas amostras
- Para analisar corretamente esta questão, deve-se executar um teste estatístico apropriado





Exemplo de Aplicação Análise para a Variável Tempo

- O Teste T para duas amostras independentes exige que as amostras sigam uma distribuição normal
- Desta forma, tem-se um primeiro teste de hipóteses a ser feito, considerando um nível de significância de 5%, sendo:
 - H₀: A distribuição é normal
 - H₁: A distribuição não é normal




Exemplo de Aplicação Análise para a Variável Tempo

Tests of Normality

Group	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Time	.155	11	,200*	,956	11	,715
PP	,216	8	,200*	,937	8	,582

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- Teste de Kolmogorov-Smirnov: teste de normalidade em amostras com mais de 30 elementos
- Teste de Shapiro-Wilk: teste de normalidade em amostras com menos de 50 elementos
- Assim, através da análise do Teste de Shapiro-Wilk, observa-se que ambas as amostras possuem o valor de significância (Sig.) superior a 0.05 e, portanto, não há indícios para rejeitar a hipótese nula a um nível de significância de 5%
- Desta forma, a distribuição das amostras para a variável Tempo é normal, logo poderá ser utilizado o teste paramétrico T para duas amostras independentes





Exemplo de Aplicação Análise para a Variável Tempo

- O Teste T pode ter duas expressões diferentes em função das variâncias poderem ou não ser assumidas como iguais, conclusão que se retira diretamente do nível de significância do Teste de Levene
- Desta forma, tem-se um outro teste de hipóteses a ser considerado, a um nível de significância de 5%, sendo:
 - H0: As variâncias são iguais ($\sigma^2_{\text{GrupoEP}} = \sigma^2_{\text{GrupoPP}}$)
 - H1: As variâncias são diferentes ($\sigma^2_{\text{GrupoEP}} \neq \sigma^2_{\text{GrupoPP}}$)




Exemplo de Aplicação Análise para a Variável Tempo

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Time	Equal variances assumed	1,066	,316	,669	17	,513	36,636	54,796	-78,974	152,247
	Equal variances not assumed			,708	16,990	,488	36,636	51,715	-72,478	145,751

- Pelas colunas do Teste de Levene para Igualdade de Variâncias, verifica-se pela primeira linha dos resultados, que as amostras possuem variâncias iguais, uma vez que a significância (Sig.= 0.316) é maior que 0.05, não havendo indícios para rejeitar a hipótese nula. Logo, as variâncias são iguais.





Exemplo de Aplicação Análise para a Variável Tempo

- Por fim, satisfeito o pressuposto de normalidade e uma vez que as variâncias são iguais, pode-se proceder com a análise de comparação das médias das duas amostras, gerando um novo teste de hipóteses, a um nível de significância de 5%, sendo:

H_0 : As médias são iguais ($\mu_{\text{GrupoEP}} = \mu_{\text{GrupoPP}}$)

H_1 : As médias são diferentes ($\mu_{\text{GrupoEP}} \neq \mu_{\text{GrupoPP}}$)




Exemplo de Aplicação Análise para a Variável Tempo

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Time	Equal variances assumed	1,066	,316	,669	17	,513	36,636	54,796	-78,974	152,247
	Equal variances not assumed			,708	16,990	,488	36,636	51,715	-72,478	145,751

- Percebe-se na primeira linha, que corresponde à confirmação de que as variâncias das amostras são iguais, que a significância do Teste T (Sig. (2-tailed) = 0.513) também é superior a 0.05 e, desta forma, não existem indícios para rejeitar a hipótese nula, concluindo-se que as médias são iguais a um nível de significância de 5%
- Uma outra maneira de verificar esta situação é a constatação de que o valor zero está entre os limites inferior e superior do intervalo de confiança, também não sendo possível rejeitar a hipótese nula.



Exemplo de Aplicação Análise para a Variável Tempo

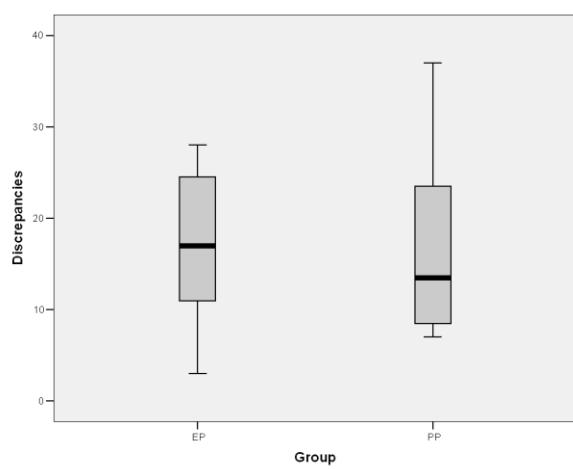
- Por estas análises efetuadas, pode-se concluir que estatisticamente não existe diferença significativa em relação à variável Tempo na utilização da técnica de detecção de defeitos na suas versões em inglês e em português.



Exemplo de Aplicação Análise para a Variável Discrepâncias

➤ Análise Visual Inicial:

- Não apresenta Outliers
- Aparente grande variabilidade entre as duas amostras



Exemplo de Aplicação Análise para a Variável Discrepâncias

- Teste T para duas amostras independentes exige que as amostras sigam uma distribuição normal. Desta forma, tem-se um primeiro teste de hipóteses a ser feito, considerando um nível de significância de 5%:

H0: A distribuição é normal

H1: A distribuição não é normal

Tests of Normality

Group	Kolmogorov-Smirnov ^a			Shapiro-Wilk			
	Statistic	df	Sig.	Statistic	df	Sig.	
Discrepancies	EP	,172	11	,200*	,951	11	,656
	PP	,232	8	,200*	,878	8	,179

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- Através da análise do Teste de Shapiro-Wilk, observa-se que ambas as amostras possuem o valor de significância (Sig.) superior a 0,05 e, portanto, não há indícios para rejeitar a hipótese nula a um nível de significância de 5%
- Desta forma, a distribuição das amostras para a variável Discrepâncias é normal, logo poderá ser utilizado o teste paramétrico T para duas amostras independentes



Exemplo de Aplicação Análise para a Variável Discrepâncias

- Teste de Levene para igualdade de variâncias, a um nível de significância de 5%, sendo:

H0: As variâncias são iguais ($\sigma^2_{\text{GrupoEP}} = \sigma^2_{\text{GrupoPP}}$)

H1: As variâncias são diferentes ($\sigma^2_{\text{GrupoEP}} \neq \sigma^2_{\text{GrupoPP}}$)

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means							
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference		
								Lower	Upper	
Discrepancies	Equal variances assumed	,569	,461	,008	,17	,994	,034	4,279	-8,994 9,062	
	Equal variances not assumed			,008	12,710	,994	,034	4,465	-9,634 9,702	

- Verifica-se que as amostras possuem variâncias iguais, uma vez que a significância (Sig. = 0,461) é maior que 0,05, não havendo indícios para rejeitar a hipótese nula.



Exemplo de Aplicação Análise para a Variável Discrepâncias

- Satisfeitos os pressupostos, aplica-se um Teste T de comparação de médias, a um nível de significância de 5%, sendo:
 H_0 : As médias são iguais ($\mu_{\text{GrupoEP}} = \mu_{\text{GrupoPP}}$)
 H_1 : As médias são diferentes ($\mu_{\text{GrupoEP}} \neq \mu_{\text{GrupoPP}}$)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Discrepancies	Equal variances assumed	,569	,461	,008	17	,994	,034	4,279	-8,994	9,062
	Equal variances not assumed			,008	12,710	,994	,034	4,465	-9,634	9,702

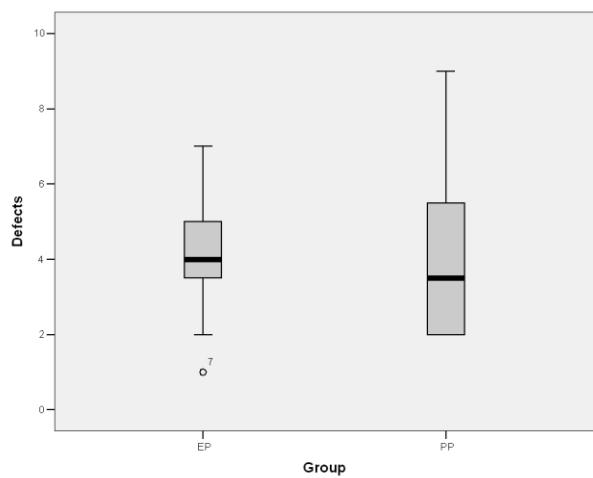
- A significância do Teste T (Sig. (2-tailed) = 0.994) é superior a 0.05 e, desta forma, não existem indícios para rejeitar a hipótese nula, concluindo-se que as médias são iguais a um nível de significância de 5%
- Por estas análises efetuadas, pode-se concluir que estatisticamente não existe diferença significativa em relação à variável Discrepâncias na utilização da técnica de detecção de defeitos na suas versões em inglês e em português



Exemplo de Aplicação Análise para a Variável Defeitos

➤ Análise Visual Inicial:

- Outlier moderado
- Aparente grande variabilidade entre as duas amostras



Exemplo de Aplicação Análise para a Variável Defeitos

- Teste T para duas amostras independentes exige que as amostras sigam uma distribuição normal. Desta forma, tem-se um primeiro teste de hipóteses a ser feito, considerando um nível de significância de 5%:

H0: A distribuição é normal

H1: A distribuição não é normal

Tests of Normality

Group	Kolmogorov-Smirnov ^a			Shapiro-Wilk			
	Statistic	df	Sig.	Statistic	df	Sig.	
Defects	EP	,169	11	,200*	,945	11	,578
	PP	,195	8	,200*	,863	8	,128

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- Através da análise do Teste de Shapiro-Wilk, observa-se que ambas as amostras possuem o valor de significância (Sig.) superior a 0.05 e, portanto, não há indícios para rejeitar a hipótese nula a um nível de significância de 5%
- Desta forma, a distribuição das amostras para a variável Defeitos é normal, logo poderá ser utilizado o teste paramétrico T para duas amostras independentes



Exemplo de Aplicação Análise para a Variável Defeitos

- Teste de Levene para igualdade de variâncias, a um nível de significância de 5%, sendo:

H0: As variâncias são iguais ($\sigma^2_{\text{GrupoEP}} = \sigma^2_{\text{GrupoPP}}$)

H1: As variâncias são diferentes ($\sigma^2_{\text{GrupoEP}} \neq \sigma^2_{\text{GrupoPP}}$)

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference			
							Lower	Upper		
Defects	Equal variances assumed	,788	,387	,149	,17	,883	,148	,989	-1,939	2,235
				,142	12,405	,889	,148	1,037	-2,105	2,400

- Verifica-se que as amostras possuem variâncias iguais, uma vez que a significância (Sig. = 0.387) é maior que 0.05, não havendo indícios para rejeitar a hipótese nula



Exemplo de Aplicação Análise para a Variável Defeitos

- Satisfeitos os pressupostos, aplica-se um Teste T de comparação de médias, a um nível de significância de 5%, sendo:
 H_0 : As médias são iguais ($\mu_{\text{GrupoEP}} = \mu_{\text{GrupoPP}}$)
 H_1 : As médias são diferentes ($\mu_{\text{GrupoEP}} \neq \mu_{\text{GrupoPP}}$)

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper	
Defects	Equal variances assumed	,788	,387	,149	,17	,883	,148	,989	-1,939	2,235
	Equal variances not assumed			,142	12,405	,889	,148	1,037	-2,105	2,400

- A significância do Teste T (Sig. (2-tailed) = 0.883) é superior a 0.05 e, desta forma, não existem indícios para rejeitar a hipótese nula, concluindo-se que as médias são iguais a um nível de significância de 5%
- Por estas análises efetuadas, pode-se concluir que estatisticamente não existe diferença significativa em relação à variável Defeitos na utilização da técnica de detecção de defeitos na suas versões em inglês e em português



Exemplo de Aplicação Conclusões

- Pelos resultados das análises efetuadas, constatou-se que estatisticamente não existe diferença significativa na utilização das versões em inglês e em português desta técnica, em nenhuma das variáveis consideradas: Tempo, Discrepâncias e Defeitos
- Assim, não existem indícios para que as técnicas sejam traduzidas do inglês para o português para serem utilizadas no contexto dos cursos de pós-graduação em engenharia de software da COPPE/UFRJ



Ampliando a classificação de Estudos em ESE

Estudos Secundários

- Dependem dos resultados dos estudos primários para serem realizados
- Usualmente utilizados para revelar evidências e construir conhecimento
 - o Exemplos de estudos secundários
 - Meta-análise
 - Revisão Sistemática
 - *quasi*-revisão sistemática (mapeamento sistemático de estudos)

Estudos Terciários

- Também chamados de Revisão Terciária
- Uma revisão de estudos secundários relacionados a uma mesma questão de pesquisa.



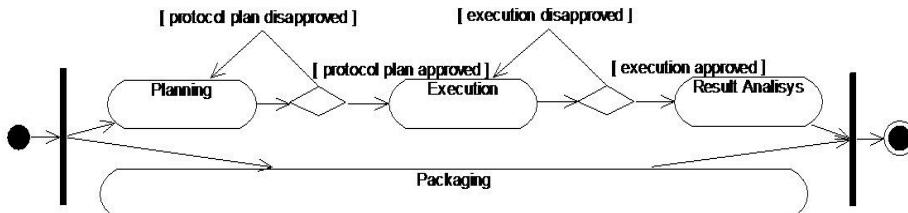
Diferenças entre Revisões Tradicionais e Sistemáticas

Característica	Revisão Tradicional	Revisão Sistemática
Questão	Escopo abrangente, formulação genérica	Foco definido, formulação específica
Identificação da Pesquisa	Usualmente não explicitamente especificada, podendo ser tendenciosa	Fontes abrangentes, estratégia de busca definida e explícita
Seleção	Usualmente não explicitamente especificada, podendo ser tendenciosa	Baseada em critérios explícitos e uniformemente aplicados
Análise	Variável	Rigorosa nos objetivos, no método, e crítica
Síntese	Geralmente um sumário qualitativo	Síntese qualitativa e/ou quantitativa
Inferências	Algumas vezes baseada em evidência	Usualmente baseada em evidência

Dybå, T. et al. (2007) Applying Systematic Reviews to Diverse Study Types: An Experience Report. ACM/IEEE First International Symposium on Empirical Software Engineering and Measurement. DOI 10.1109/ESEM.2007.59.



Processo de Experimentação Estudos Secundários



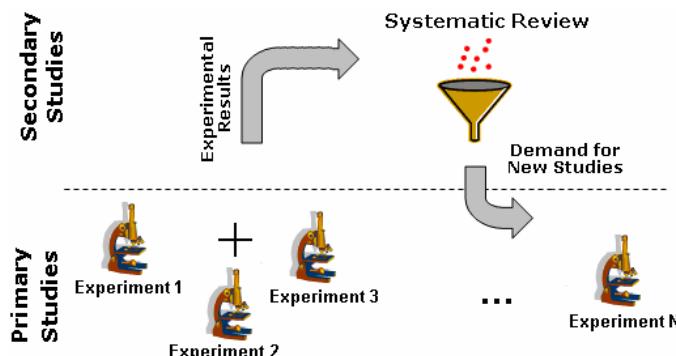
Biolchini, J., Mian, P.G., Natali, A.C. and Travassos, G.H. (2005) "Systematic Review in Software Engineering: Relevance and Utility", Technical Report ES67905, PESC - COPPE/UFRJ. Disponível em <http://cronos.cos.ufrj.br/publicacoes/reltec/es67905.pdf>



Realimentação do Conhecimento dos Estudos em ES

O processo de experimentação apresenta natureza recursiva

Conhecimento adquirido de estudos *primários* alimentam os estudos secundários, que por sua vez permitem identificar a necessidade de novos estudos primários...



TRAVASSOS, G. H. ; SANTOS, P. S. M. ; MIAN, P. ; DIAS NETO, A. C. ; BIOLCHINI, J. (2008). An Environment to Support Large Scale Experimentation in Software Engineering. In: Proc. of XIII IEEE International Conference on Engineering of Complex Computer Systems, Belfast.



Engenharia de Software Experimental

➤ Como está a maturidade da experimentação?

- Sinal de maturidade de uma área:
 - o Nível de sofisticação dos objetivos de um estudo experimental
 - o Interesse no entendimento das “coisas” sobre a disciplina
 - o Para Engenharia de Software isto deveria significar:
 - Podemos construir **modelos** que permitem ser utilizados **para medir e diferenciar processos e produtos?**
 - Podemos **medir o efeito de** uma modificação numa **variável** particular do processo na variável do produto?
 - Podemos **predizer as características de um produto** (valores da variável do produto) baseado no modelo do processo (valores das variáveis do processo), dentro de um contexto em particular?
 - Podemos **controlar os efeitos no produto**, baseado em objetivos, dado um conjunto particular de variáveis do contexto?



Engenharia de Software Experimental

➤ Como está a maturidade da experimentação?

- Sinal de maturidade de uma área:
 - o um **padrão de conhecimento** construído a partir de uma **série de estudos experimentais**

A disciplina foi construída a partir de conhecimento, modelos ou experimentos anteriores?

O estudo representa um evento isolado?

Ele levou a outros estudos experimentais que utilizaram suas informações ?
Estudos experimentais vem sendo repetidos sob condições similares ou diferentes?

A construção de conhecimento existe em um grupo de pesquisa ou ambiente, ou tem sido disseminada para outros – pesquisadores construindo e compartilhando seus trabalhos?

Por exemplo, inspeções, em geral, estão bem estudadas experimentalmente

Entretanto, existe muito pouca combinação de resultados, repetição e análise de diferenciação das variáveis



Engenharia de Software Experimental

- Como está a maturidade da experimentação?
 - Existe alguma evidência que pesquisadores estão
 - o Utilizando **questões** mais **sophisticadas**
 - o **Estudando relacionamentos** entre características de processo/produto
 - Realizando mais estudos **no campo (in vivo)** que em laboratório
 - **Combinando** diferentes **classes de estudos experimentais** para construir conhecimento
 - Experimentação pode nos fornecer
 - o Melhor **base científica e de engenharia** para a engenharia do software
 - o **Melhores modelos** de
 - Processos de software e produtos
 - Fatores diversos do ambiente, por exemplo, as pessoas, a organização
 - o Melhor **compreensão da interação** entre estes modelos



Engenharia de Software Experimental

- Com que o nosso futuro se parece:
 - Engenheiros de Software terão à sua disposição
 - o A habilidade de controlar e manipular soluções de projeto
 - Baseado no ambiente e conjunto de objetivos para o projeto
 - o O conhecimento baseado em evidência experimental
 - Do que funciona e não funciona e sob que condições
 - Pesquisadores estarão providos de laboratórios para experimentação
 - o Isto necessitará de um plano de pesquisa posto em prática durante vários anos
 - Coordenando estudos experimentais
 - Evoluindo com novos conhecimentos



Comentários

- Decisões sobre tecnologias devem ser baseadas em conhecimento científico e evidência experimental.
- Investigações em Engenharia de Software compartilham alguns dos mesmos problemas de Ciências Sociais:
 - Difícil coletar dados
 - Não-repetível
 - Difícil de controlar
- Quanto mais cuidado tivermos no projeto e controle dos estudos
 - Mais confidentes poderemos estar com seus resultados
 - Melhor será nosso entendimento da área
 - Mais efetivas serão nossas ações



ESE
ese.cos.ufrj.br

PESC
Programa de Engenharia de Sistemas e Computação

Comentários Para desenvolvermos ciência (em larga escala), precisamos de recursos humanos!

OPINIÃO • 7

O GLOBO

Terça-feira, 6 de junho de 2006

Ciência no ensino fundamental

JORGE WERTHEIM

Muito se fala sobre a necessidade de melhorar a qualidade da educação básica no Brasil, e são várias as estratégias defendidas. Um fim é dar mais atenção à formação dos docentes, aumentar o tempo de permanência na escola, melhorar a infraestrutura e equipamentos, entre outras. Todas são válidas e, certamente, se colocadas em prática, colaborarão para melhorar a educação. Contudo, existe uma alternativa de grande impacto que é pouco lembrada: a incorporação do ensino de ciências ao currículo desde os primeiros anos.

No Brasil, o ensino de ciências tem pouca ênfase dentro da educação básica, apesar de forte presença no cotidiano das pessoas e do lugar central que a inovação tecnológica desempenha entre as empresas.

Entretanto, a falta de atenção dispensada à formação na área de ciências neste país é o enorme déficit do docente de física, química e biologia, que se avolumou em 200 mil segundo o próprio Ministério da Educação.

Enquanto isso, em diversas partes do mundo, inclusive da América Latina, tem-se experimentado o impac-

to positivo do ensino de ciências sobre a qualidade da educação. Argentina, Uruguai, Chile, Costa Rica, Cuba detêm os melhores indicadores educacionais da região, e são exemplos da importância de perceber que o ensino das ciências pode ser muito importante e produtivo.

O impacto do ensino de ciências sobre a educação é óbvio, mas deve

ao fato de que ele envolve um exerci-

cio extremamente importante de raciocínio,

que desperta na criança seu interesse, melhorando a aprendizagem e a capacidade de resolu-

ção de situações. Por isso,

a criança que se familiariza com as ciências tem mais chances de se desenvol-

ver neste campo e em outros. Somente assim

uma sociedade justa terá

maior atenção ao

ensino de ciências por parte dos for-

mulaadores de políticas públicas na

escola e no governo, mas existem ou-

tras. Uma segunda razão é que apena-

com um bom ensino de ciências para

todas as crianças é que elas terão

maior chance de se tornarem profis-

sionais para as carreiras científicas.

A terceira, e última lembrada no

âmbito deste artigo, é o fato de que o

conhecimento científico e as novas tecnologias são fundamentais para que a população possa se posicionar frente a propostas de desenvolvimento que precisam ter uma opinião a fim de legitimá-las. É o caso do uso de alimentos geneticamente modificados, da exploração biológica e o cadáver da energia nuclear.

Nesse sentido, o domínio do co-

científico da cidadania no contexto da democra-

cia é fundamental.

Costuma-se dizer

que, no mundo contem-

porâneo, o capital mais

importante é o capital

e o conhecimento. O

conhecimento, contudo,

depende da forma-

ção de pessoas capazes

de produzi-lo e nuan-

çá-lo. Só assim, o pa-

ísco do Brasil, com um

número elevado de

crianças e adolescentes em idade esco-

lar, não é exagerado dizer que este

o bem mais valioso que se tem à dis-

posição, o qual pode ser

valiosíssimo para o futuro.

Assim, se esse poten-

cial for aproveitado por meio de

uma educação de qualidade.

Mas justamente por se almejar

uma educação de qualidade é preci-

so atentar para um aspecto fundamen-

tal: o ensino para as ciências não

está no currículo. Vida o que aconte-

ce no ensino médio, em que a educa-

ção para as ciências é mínima e ma-

isca. E evidentemente defi-

ciente devido, entre outras coisas, à

falta de capacitação dos docentes.

Assim, a melhoria das ciências des-

envolvidas fundamental deve ser as-

sociada, necessariamente, a uma po-

lítica de formação de docentes de-

niados que possam ensinar e apren-

der a proporcionar aos alunos apren-

dizimentos significativos.

Não existe fantasma algum no en-

tre. Para se apenas preciso

transmitir conhecimentos que são

até elementares e que gerem interes-

se e engajamento. O

conhecimento, contudo,

depende da forma-

ção de pessoas capazes

de produzi-lo e nuan-

çá-lo. Só assim, o pa-

ísco do Brasil, com um

número elevado de

crianças e adolescentes em idade esco-

lar, não é exagerado dizer que este

o bem mais valioso que se tem à dis-

posição, o qual pode ser

valiosíssimo para o futuro.

JORGEBERTHHEIM é assessor do diretor-

geral da Organização dos Estados Ibero-

Américas.

Desafios para ESE

- Experimentação pode apoiar o desenvolvimento e evolução de tecnologias e processos de software
 - Poderosa ferramenta para a construção de um núcleo de conhecimento baseado em evidência
- Demandas para os pesquisadores:
 - Definir uma agenda de pesquisa em comum
 - Identificar as teorias e construir modelos para elas
 - Discutir uma terminologia/taxonomia em comum
 - Aumentar a disseminação de conceitos de experimentação, entre alunos, pesquisadores e engenheiros de software
 - Melhor disseminar os benefícios que a experimentação (e não o empirismo!) podem trazer para a área de Engenharia de Software e tecnologias correlatas
 - Aumentar a compreensão sobre estudos quantitativos, qualitativos e semi-quantitativos aplicados a ES
 - Planejar, executar e tornar disponíveis estudos secundários evidenciando comportamentos em ES.
 - Ser mais **experimental** e menos **empírico...**



Referências Bibliográficas

- Araújo, M.A.P., Barros, M.O., Murta, L.G.P., Travassos, G.H. **Métodos Estatísticos Aplicados à Engenharia de Software Experimental**. SBES, 2006.
- Araújo, M.A.P., Travassos, G.H. **A Utilização de Métodos Estatísticos no Planejamento e Análise de Estudos Experimentais em Engenharia de Software**. ESELAW, 2009.
- Cochran, W. G., Cox, G. M., "Experimental Designs". John Wiley & Sons, 1957.
- Costa, H.R., Barros, M.O., Travassos, G.H., "Evaluating Software Project Portfolio Risks", Journal of Systems and Software, 2006
- Dyba, T.; Kampenes, V.; Sjoberg, D., "A Systematic Review of Statistical Power in Software Engineering Experiments", Information and Software Technology, Elsevier, 2005
- Juristo, N.; Moreno, A. M.; "Basics of Software Engineering Experimentation". Kluwer Academic Publishers, 2001.
- Kitchenham, B.A. et al, **Preliminary guidelines for empirical research in software engineering** - IEEE Transactions on Software Engineering, Volume: 28 No.: 8 , Page(s): 721 -734, Aug. 2002.
- Miller, J., Dali, J., Wood, M., Roper, M., Brooks, A., **Statistical power and its Subcomponents – Missing and Misunderstood Concepts in Empirical Software Engineering Research**, Information and Software Technology, Vol. 39, No. 4, pp. 285-295, 1997.



Referências Bibliográficas

- Montgomery, D. C., "Estatística Aplicada e Probabilidade para Engenheiros", Ed. LTC, 2003.
- Montgomery, D. C., "Design and Analysis of Experiments", Ed. IE-Wiley, 2000.
- National Institute of Standards and Technology, acessado em outubro/2006 na URL <http://www.nist.gov>
- Pfleeger, Shari L., **Albert Einstein and Empirical Software Engineering**. IEEE Computer: 32-37, 1999.
- Sobral, A. P. B., **Curso de Técnicas Estatísticas Experimentais**. Março, 1996.
- Tichy, W. F., **Should Computer Scientists Experiment More?**, IEEE Computer: 32-40, May, 1998.
- Vieira, S., "Estatística Experimental", 2a. Edição, Ed. Atlas, 1999.
- Wohlin, C. et al. "Experimentation in Software Engineering – An Introduction". Kluwer Academic Publishers, USA, 2000.
- Maxwell, K. D., "Applied Statistics for Software Managers". Prentice Hall PTR, 2002.

