

Métodos Estatísticos Aplicados à Engenharia de Software Experimental

Marco Antônio P. Araújo
COPPE/UFRJ
maraujo@cos.ufrj.br

Márcio de O. Barros
PPGI / UNIRIO
marcio.barros@uniriotec.br

Leonardo G. P. Murta
COPPE/UFRJ
murta@cos.ufrj.br

ESE
Engenharia de
Software
Experimental

Guilherme H. Travassos
COPPE/UFRJ
ght@cos.ufrj.br



Objetivo

- Apresentar as principais técnicas estatísticas utilizadas no planejamento e análise de estudos experimentais em Engenharia de Software
- Utilizar uma abordagem prática, apresentando as técnicas estatísticas no contexto de exemplos reais
- Utilizar informações de estudos experimentais já realizados pelos autores e publicados na literatura para apoiar as discussões e apresentação de aplicação das técnicas estatísticas
- Apresentar uma breve introdução aos conceitos de estudos experimentais e estatística, passando para indicações concretas da aplicabilidade destas técnicas



SBES 2006 - Métodos Estatísticos Aplicados à Engenharia de Software Experimental
Copyright ARAÚJO, BARROS, MURTA, TRAVASSOS



Agenda

- Experimentação e Engenharia de Software
- O processo de experimentação
- Escalas numéricas e operações aplicáveis
- Tabulação, preparação e análise visual dos dados
- Medidas de tendência, dispersão e dependência
- Análise de *outliers* e quartis
- Testes estatísticos aplicáveis aos tipos de estudo
- Testes paramétricos e não paramétricos
- Análise de regressão
- Testes ANOVA e Mann-Whitney
- Exemplos

Experimentação e ES

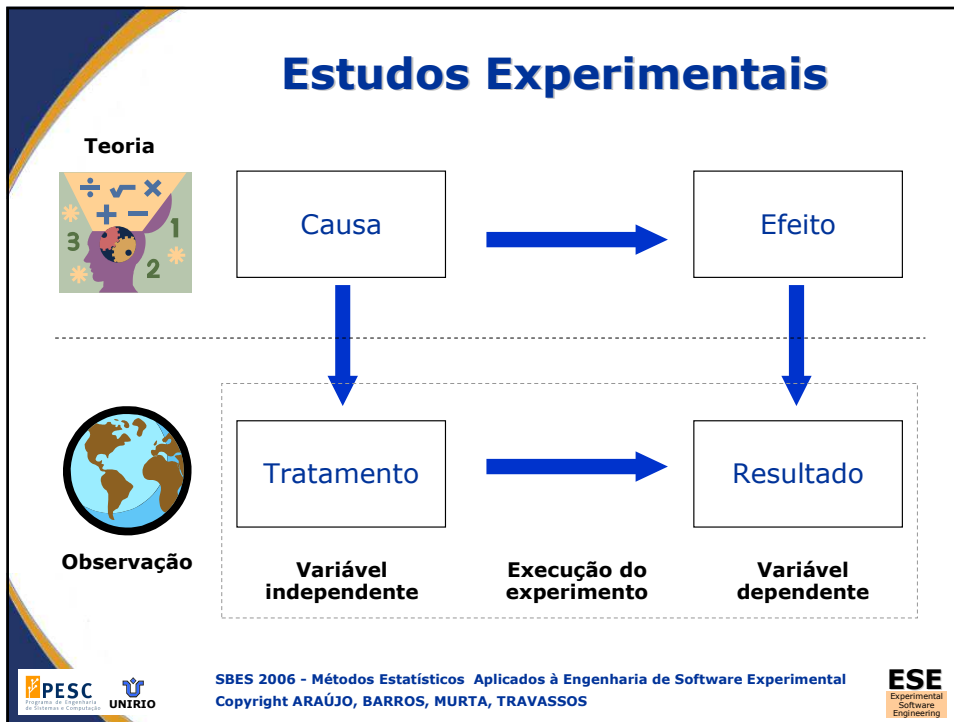
- As pesquisas em Engenharia de Software devem seguir as orientações de pesquisas realizadas em outros tipos de engenharia
- As engenharias se baseiam no uso de pesquisas científicas para construir produtos e serviços economicamente viáveis
- Assim, além das propostas de inovações técnicas, as pesquisas devem incluir uma avaliação dos resultados atingidos pela aplicação destas inovações

Experimentação e ES

- Existem diferentes métodos de avaliação para pesquisas:
 - **Método científico:** se baseia na observação do mundo e na construção de um modelo baseado nestas observações
 - **Método de engenharia:** as técnicas atuais são analisadas, suas fraquezas são identificadas, inovações são propostas e comparadas com as técnicas que as precederam
 - **Método experimental:** um modelo para o mundo real é proposto e avaliado através de um conjunto de estudos experimentais
 - **Método analítico:** uma teoria formal é proposta, resultados são derivados e comparados com observações do mundo real
- Os métodos de engenharia e experimental são considerados derivações do método científico

Experimentação e ES

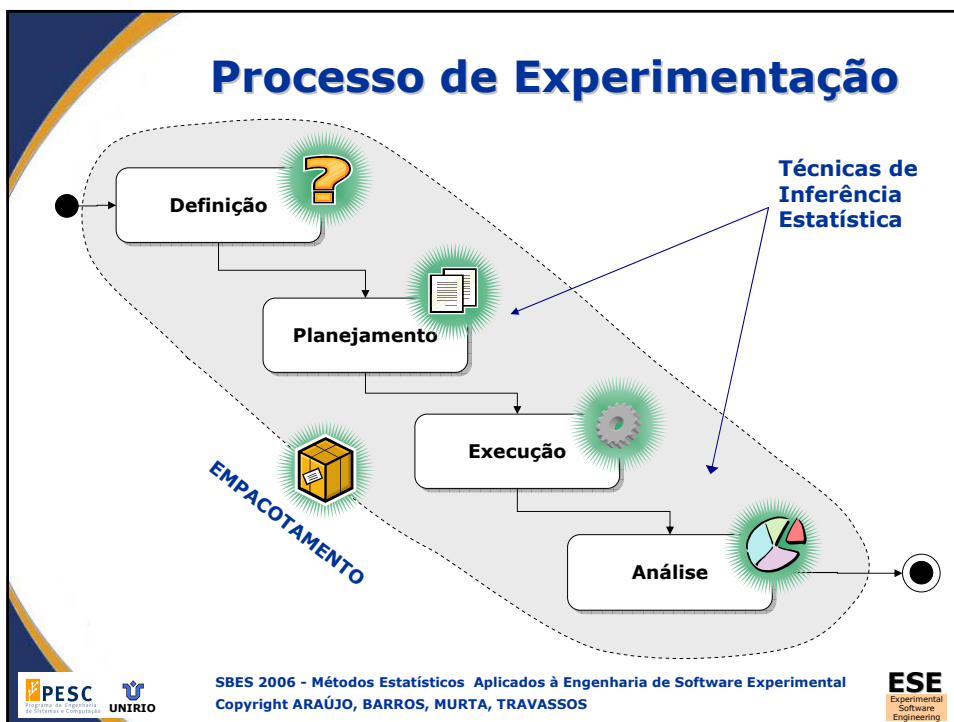
- O método científico (e suas derivações) é tradicionalmente aplicado com sucesso em outras ciências, destacando-se as sociais, onde raramente é possível estabelecer leis da natureza, como na física ou matemática
- Como o fator humano é muito importante na construção e manutenção de software, a Engenharia de Software se aproxima destas ciências sociais
- Assim, o método científico é comumente aplicado para avaliar os benefícios providos por uma nova técnica, teoria ou método relacionado com software






SBES 2006 - Métodos Estatísticos Aplicados à Engenharia de Software Experimental
Copyright ARAÚJO, BARROS, MURTA, TRAVASSOS





Processo de Experimentação

- Definição
 - Identificação dos objetivos do estudo
 - Identificação dos objetos e grupos de estudo

- 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*)

Processo de Experimentação

- Execução do estudo experimental
 - Treinamento
 - Execução do estudo pelos participantes

- Análise de dados
 - Análise gráfica dos dados
 - Estatísticas descritivas
 - Eliminação de *outliers*
 - Análise de distribuições
 - Aplicação da análise estatística

- Empacotamento
 - Apresentação de resultados
 - Preparação do pacote para repetição do estudo

Hipóteses, Variáveis e Escalas



- Planejamento e hipóteses
- Hipóteses
- Escolha de variáveis
- Escalas
- Nível de informação na escala
- Escalas e operações básicas

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 levam 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 % de requisitos encontrados que são corretos

Variáveis e seus Valores

- As variáveis de um estudo podem ser:
 - Qualitativas: os tratamentos representam tipos, formas e procedimentos
 - Quantitativas: os tratamentos representam doses ou níveis de aplicação da variável
- Os valores das variáveis são coletados em escalas:
 - Existem diversas escalas para coleta e representação destes 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)
 - Diferentes graus de coesão (funcional, procedural, temporal, seqüencial, ...)
- 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 é possível porque toda escala intervalar possui um zero arbitrário (no caso das datas, o ano zero)

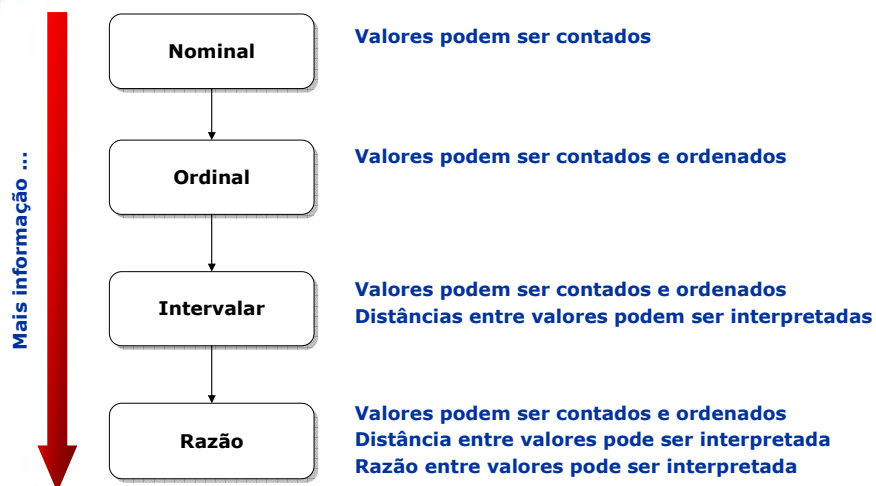
Escalas: Intervalar

- As escalas de Likert são um exemplo de escala intervalar muito utilizado em estudos relacionados à 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
 - Com isso, as operações aplicáveis sobre uma escala intervalar podem ser utilizadas nesta avaliação

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 é duas vezes menor que um software de 2X linhas de código

Informação nas Escalas



Escalas e Operações

- De acordo com a escala das variáveis, podemos executar diferentes operações sobre seus valores

Escala	Nominal	Ordinal	Intervalar	Razão
Contagem de valores	X	X	X	X
Ordenação de valores		X	X	X
Soma de valores			X	X
Subtração de valores			X	X
Divisão de valores				X

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 % de requisitos encontrados que são corretos → Escala razão

Análise Tabular e Gráfica



- Variáveis e execução
- Tabulação
- Análise gráfica
- Histogramas
- Gráficos de torta
- Gráficos de dispersão

Variáveis e Execução

- A execução de um estudo experimental 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	% Encontrados Corretos
1	Y	10	83%	90%
2	Y	13	73%	92%
3	Y	12	87%	99%
4	Y	13	78%	98%
5	Y	10	74%	99%
6	Y	14	74%	96%
7	Y	14	87%	99%
8	Y	13	75%	95%
9	Y	14	86%	100%
10	Y	14	82%	95%
11	Y	13	77%	98%
12	X	13	90%	87%
13	X	9	89%	83%
14	X	11	88%	81%
15	X	14	87%	87%
16	X	9	97%	87%
17	X	12	81%	82%
18	X	9	82%	83%
19	X	12	86%	80%
20	X	11	92%	88%
21	X	14	96%	84%
22	X	13	98%	83%

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 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 frequência
- A frequência 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

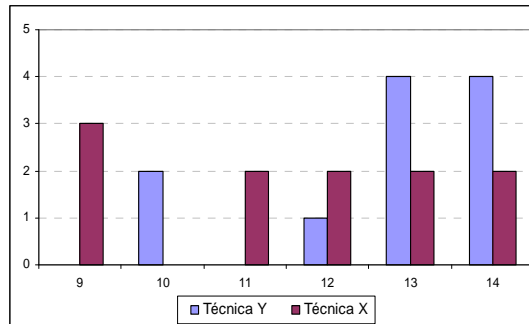
- É um método comum de apresentação de dados numéricos e em qualquer escala, pois envolve apenas contagem
- Os histogramas também permitem relacionar os dados observados com distribuições de frequência conhecidas
 - Estas distribuições possuem propriedades matemáticas das quais foram derivados os testes 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 freqüência 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

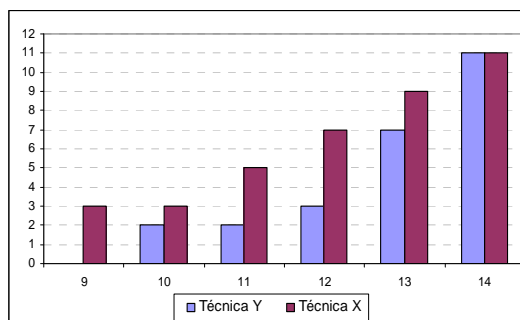


Gráfico de Torta

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

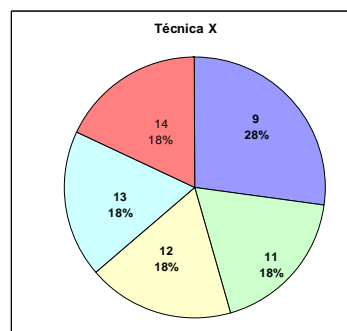
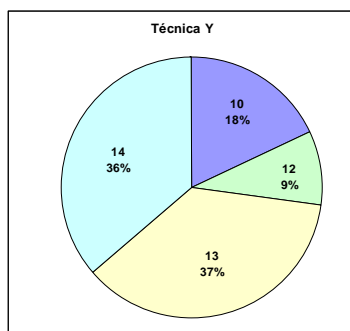
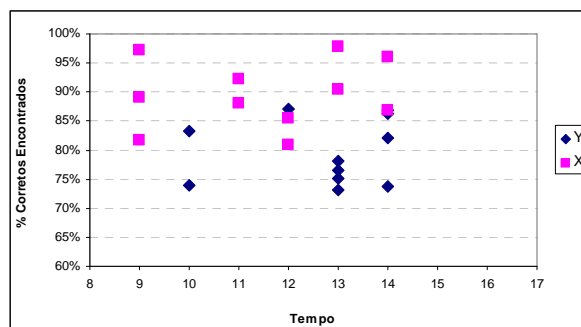


Diagrama de Dispersão

- 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



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

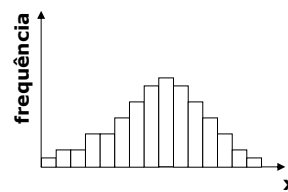
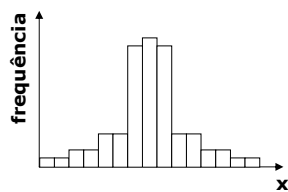
- Indicam o meio do conjunto de valores observados
 - **Média** (aritmética): a média pode ser considerada como o centro de gravidade dos dados coletados. É calculada pelo somatório dos valores coletados, dividido por sua quantidade
 - **Mediana**: valor do meio de um conjunto de dados, ou seja, o número de valores coletados que está abaixo da mediana deve ser o mesmo que está acima. É calculada colocando os valores em ordem crescente ou decrescente e selecionando o elemento central. Em caso de número par de valores, é calculada pela média dos valores centrais
 - **Moda**: representa o valor mais comum dentre o conjunto de valores coletados. É calculada pela contagem do número de ocorrências (frequência) de cada valor, selecionando o mais comum. Se dois ou mais valores ocorrem com a maior frequência, os valores coletados possuem diversas modas

Medidas de Tendência Central

- Outras medidas relevantes
 - **Valor mínimo**: representa o menor valor entre os dados que foram coletados
 - **Valor máximo**: representa o maior valor entre os dados que foram coletados
 - **Percentil**: é o caso geral da mediana, que é conhecida como percentil 50%. 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
 - **Quartil**: são os valores que representam o percentil 25% (ou primeiro quartil), a mediana (segundo quartil) e o percentil 75% (terceiro quartil)

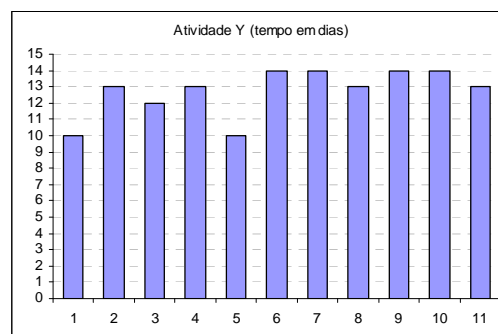
Medidas de Dispersão

- Medem o quanto os valores coletados estão dispersos ou concentrados em torno de seu valor central
 - **Faixa:** é a diferença entre o maior e o menor valor dentre os 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
 - **Desvio Padrão:** é a raiz da variância, sendo a medida de dispersão mais comumente utilizada

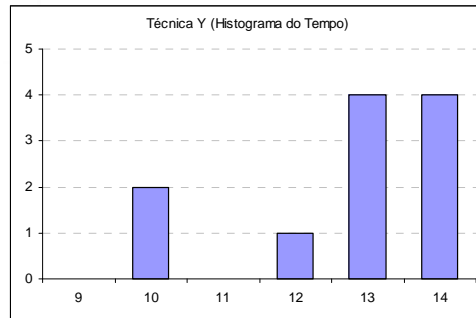


Estatística Descritiva

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



Estatística Descritiva



Medidas de Tendência

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

Estatística Descritiva

- Existem outras medidas (como curtose, assimetria, ...), mas elas estão fora do escopo deste tutorial

$$\text{Média: } \mu = \frac{\sum x_i}{n}$$

$$\text{Variância: } \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

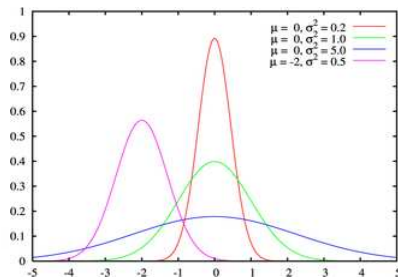
$$\text{Desvio Padrão: } \sigma = \sqrt{\sigma^2}$$

Distribuições de Frequência

- 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

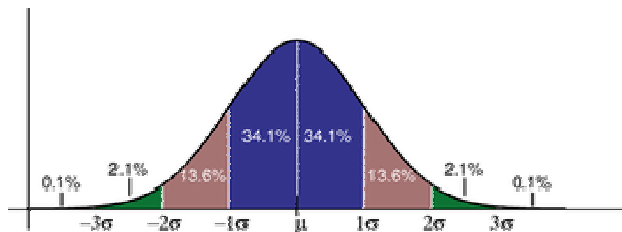
- A distribuição normal possui o formato de um sino, com as pontas se estendendo a direita e esquerda do centro
 - A curva é simétrica em relação a sua média e a largura do sino é proporcional ao seu desvio padrão
 - Assim, a curva pode ser descrita matematicamente apenas com base em sua média e desvio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



$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Distribuição Normal

- Se um conjunto de dados numéricos seguem 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



Medidas de Dependência

- Quando duas ou mais variáveis estão relacionadas em um estudo, é útil calcular o grau de dependência entre elas
- As medidas de dependência determinam a força e direção do relacionamento entre duas ou mais variáveis avaliadas quantitativamente
 - 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 simétrica

Medidas de Dependência

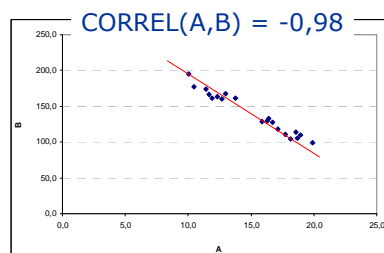
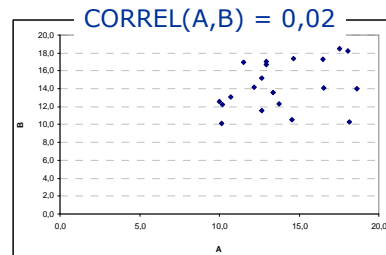
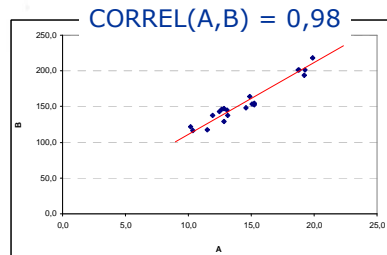
- A correlação entre duas variáveis varia entre -1 e 1
 - A correlação -1 indica que um valor alto em uma variável normalmente ocorre em conjunto com um valor baixo da segunda variável
 - A correlação 1 indica que um valor alto em uma variável normalmente ocorre em conjunto com um valor alto da segunda 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

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

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

Correlação de Spearman

- Considere os valores A_i e B_i de duas variáveis A e B
 - Calcule $R(A_i)$ a posição relativa de cada A_i em relação ao seu conjunto de valores ordenados de forma crescente (ranking)
 - Calcule $R(B_i)$ a posição relativa de cada B_i em relação ao seu conjunto de valores ordenados de forma crescente (ranking)
 - O coeficiente de correlação de Spearman é calculado segundo a fórmula abaixo

$$\rho = 1 - \frac{6 \cdot \sum_i R(A_i) - R(B_i)}{N(N^2 - 1)}$$

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*



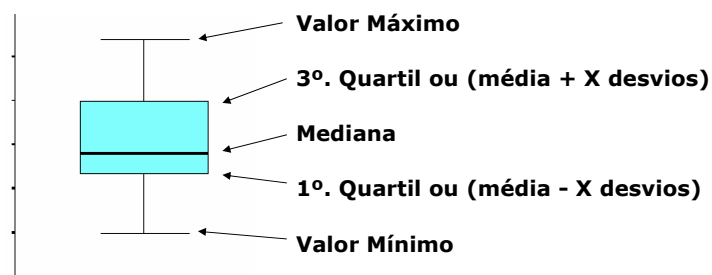
- Conceito
- Condições de ocorrência
- Identificação visual
- Identificação numérica

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 a 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

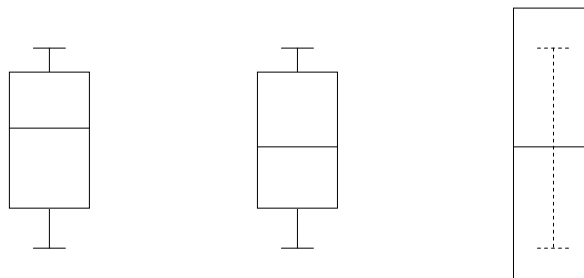
Identificação Visual

- 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

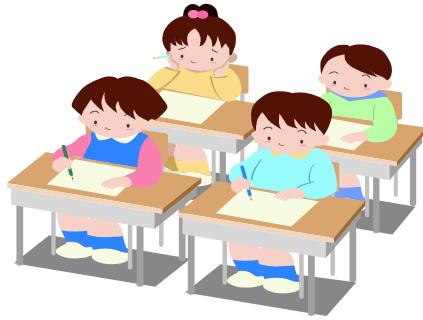
Identificação Numérica

- 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%

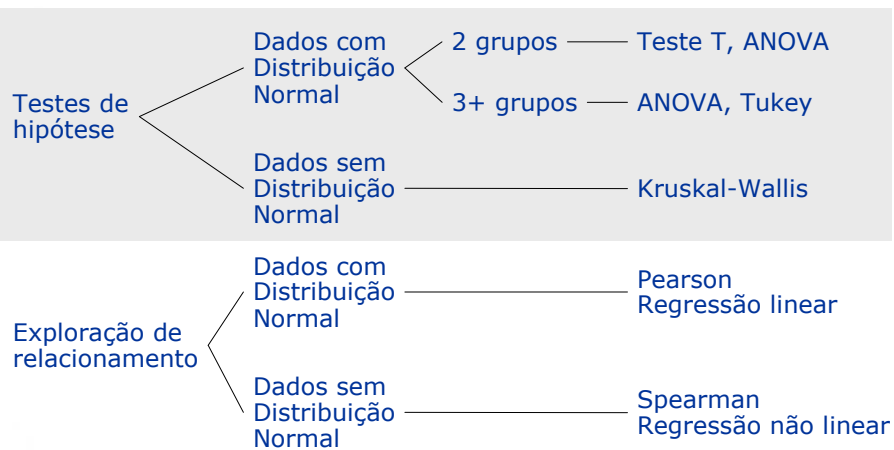
Participante	Tempo (dias)	% Corretos Encontrados	% Encontrados Corretos
1	10	83%	90%
2	13	73%	92%
3	12	87%	99%
4	13	78%	98%
5	10	74%	99%
6	14	74%	96%
7	14	87%	99%
8	13	75%	95%
9	14	86%	100%
10	14	82%	95%
11	13	77%	98%

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 a hipótese
- Em geral, são definidas duas hipóteses
 - **Hipótese nula (H0)**: indica que as diferenças observadas no estudo são coincidentais, ou seja, é a hipótese que o analista deseja rejeitar com a maior significância possível
 - **Hipótese alternativa (H1)**: 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

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

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

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

$$\text{Potência} = 1 - \beta$$

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

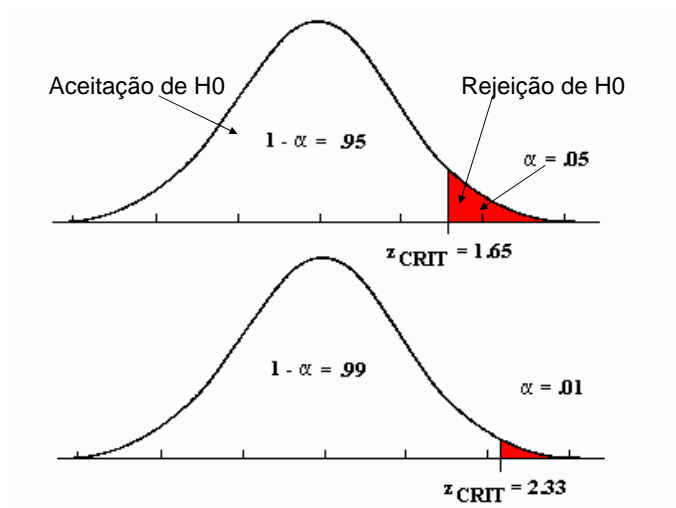
Nível de Significância

- 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 freqüente 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

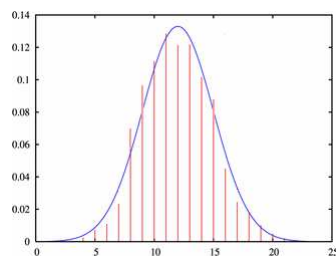
- Os teste de hipótese se dividem em testes paramétricos e testes não paramétricos
- Testes Paramétricos
 - Utilizam fórmulas fechadas, derivadas de propriedades de distribuições de frequência conhecidas (tais como equação da curva, da curva acumulada, simetria, ...)
 - Por conta disso, exigem algumas 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 frequência 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

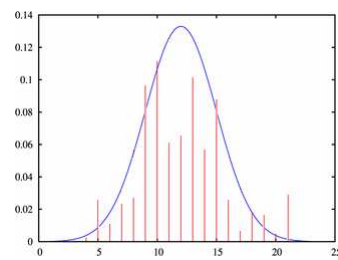
- Testes Não-Paramétricos
 - Devem ser utilizados quando os dados coletados não atendem aos pressupostos esperados pelos testes paramétricos
 - São menos poderosos que os testes paramétricos, mas não presumem distribuições de probabilidade nos dados
 - Utilizam rankings dos valores observados ao invés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

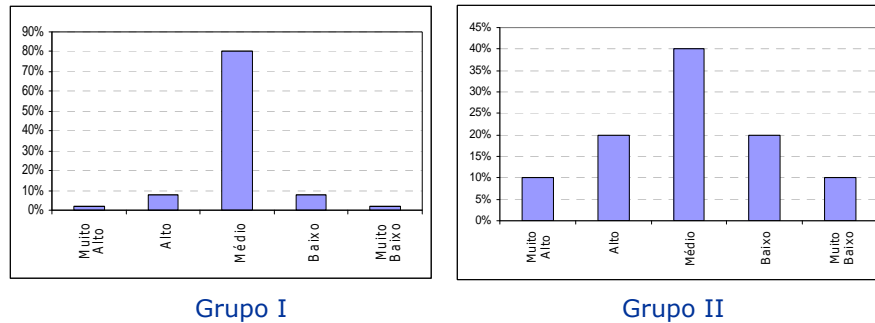
- Teste de Shapiro-Wilk
 - Calcula o valor W , que avalia se uma amostra x_i segue a distribuição normal
 - Valores pequenos calculados para W indicam que a amostra não segue a distribuição normal
 - Os valores de a_i são publicados em tabelas e estão presentes em pacotes estatísticos
 - Teste usado 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 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



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 abaixo for menor do que o valor da distribuição F com k-1 e N-1 graus de liberdade para um nível de significância α
 - Os valores da distribuição F estão disponíveis em tabelas e softwares estatísticos

$$W = \frac{(N-k) \sum_{i=1}^k N_i (\mu_{z_i} - \mu_z)^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - \mu_{z_i})^2} \quad \therefore Z_{ij} = |Y_{ij} - \mu_{y_i}|$$

Tipos de Teste de Hipótese

- Testes de hipótese paramétricos
 - Binomial
 - Chi-2
 - Teste T
 - Teste F
 - ANOVA e MANOVA

- Testes de hipótese não-paramétricos
 - Teste de Wilcoxon
 - Teste de Kruskal-Wallis
 - Teste de Mann-Whitney
 - Teste de Kolmogorov-Smirnov

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
 - 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 T Independente

- Aplicação do teste T
 - Seja uma amostra X com n valores e variância S_X
 - Seja uma amostra Y com m valores e variância S_Y
 - Calcula-se o valor t_0 conforme a equação abaixo
 - Compara-se t_0 com valores da distribuição T (publicados em tabelas) com $n+m-2$ graus de liberdade

$$t_0 = \frac{\mu_X - \mu_Y}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \therefore S_p = \sqrt{\frac{(n-1) \cdot S_X + (m-1) \cdot S_Y}{n+m-2}}$$

$$|t_0| > t_{\alpha/2, n+m-2} \Rightarrow H_1 : \mu_X \neq \mu_Y$$

$$|t_0| > t_{\alpha, n+m-2} \Rightarrow H_1 : \mu_X > \mu_Y$$

Teste T Pareado

➤ Aplicação do teste T

- Sejam duas amostras X e Y, pareadas e com n valores
- Calcula-se o valor t_0 conforme a equação abaixo
- Compara-se t_0 com valores da distribuição T (publicados em tabelas) com n-1 graus de liberdade

$$t_0 = \frac{\mu_D \cdot \sqrt{n}}{\sigma_D} \quad \therefore D_i = X_i - Y_i$$

$$|t_0| > t_{\alpha/2, n-1} \Rightarrow H_1 : \mu_X \neq \mu_Y$$

$$|t_0| > t_{\alpha, n-1} \Rightarrow H_1 : \mu_X > \mu_Y$$

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

$$U = N_A \cdot N_B + \frac{N_A(N_A + 1)}{2} - T \quad N_A = \min(n, m)$$

$$U' = N_A \cdot N_B - U \quad N_B = \max(n, m)$$

$$Z = \min(U, U')$$

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
 - No ANOVA, cada observação é modelada como:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

— Média global das observações

— Variação decorrente do tratamento

— Variação decorrente de erro aleatório

ANOVA – Análise de Variância

- 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

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

- Com isto, o método calcula alguns termos:
 - SSL: soma dos quadrados dentro de um tratamento
 - SSE: soma dos quadrados entre tratamentos
 - SST: soma dos quadrados totais
- A partir destes termos e seus graus de liberdade, calcula-se um termo F0 que é comparado com a estatística F

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
 - A hipótese nula é rejeitada com nível de significância α se K for maior que o valor da distribuição Chi-2 com $N-g$ graus de liberdade para α

$$K = \frac{12}{N(N+1)} \sum_{i=1}^g n_i (\mu_{ri} - \mu_r)^2$$

N é o número total de observações

n_g é o número de elementos no grupo G

R_{ij} é o ranking do i -ésimo valor do grupo j

μ_{ri} é a média dos rankings do grupo i

μ_r é média dos rankings de todos os grupos

Estudos de Caso



- Defeitos em requisitos
- Análise de tempo
- Análise de defeitos
- Testes paramétricos
- Fatores de risco
- Análise com múltiplos fatores
- Testes não-paramétricos

Estudo de Caso 1

- O estudo de caso 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

Estudo de Caso 1

- 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

Estudo de Caso 1

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

Estudo de Caso 1

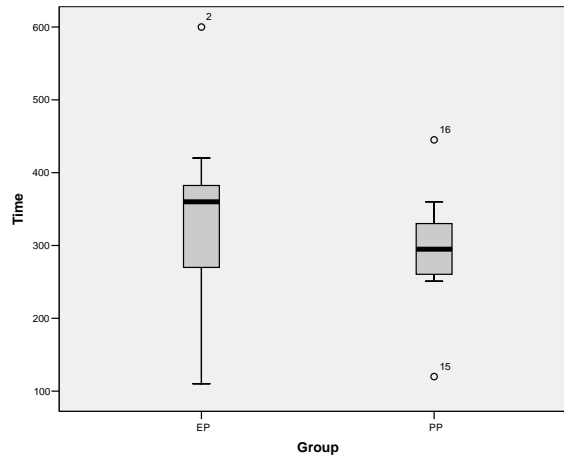
- Utilização do Teste T para duas amostras independentes:
 - uma vez que as variáveis são quantitativas
 - não são conhecidas as variâncias populacionais dos grupos
 - os dados estão distribuídos em duas amostras independentes, pois nenhum participante participou simultaneamente dos dois grupos
- As análises foram feitas utilizando o pacote estatístico SPSS

Estudo de Caso 1 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

Estudo de Caso 1 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



Estudo de Caso 1 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

Estudo de Caso 1 Análise para a Variável Tempo

Tests of Normality

Group	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Time EP	,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

Estudo de Caso 1 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 ($\sigma^2_{\text{GrupoEP}} \neq \sigma^2_{\text{GrupoPP}}$)

Estudo de Caso 1 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		
								Lower	Upper	
Time	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.

Estudo de Caso 1 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:
H0: As médias são iguais ($\mu_{\text{GrupoEP}} = \mu_{\text{GrupoPP}}$)
H1: As médias são diferentes ($\mu_{\text{GrupoEP}} \neq \mu_{\text{GrupoPP}}$)

Estudo de Caso 1 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		
								Lower	Upper	
Time	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.

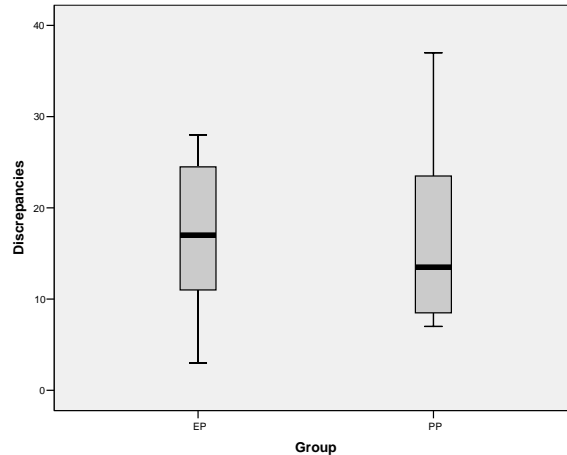
Estudo de Caso 1 Análise para a Variável Tempo

- 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.

Estudo de Caso 1 Análise para a Variável Discrepâncias

➤ Análise Visual Inicial:

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



Estudo de Caso 1 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

Estudo de Caso 1

Análise para a Variável Discrepâncias

- Teste de Levene para igualdade de variâncias, a um nível de significância de 5%, sendo:
 H_0 : As variâncias são iguais ($\sigma^2_{GrupoEP} = \sigma^2_{GrupoPP}$)
 H_1 : As variâncias são diferentes ($\sigma^2_{GrupoEP} \neq \sigma^2_{GrupoPP}$)

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.

Estudo de Caso 1

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

Independent Samples Test

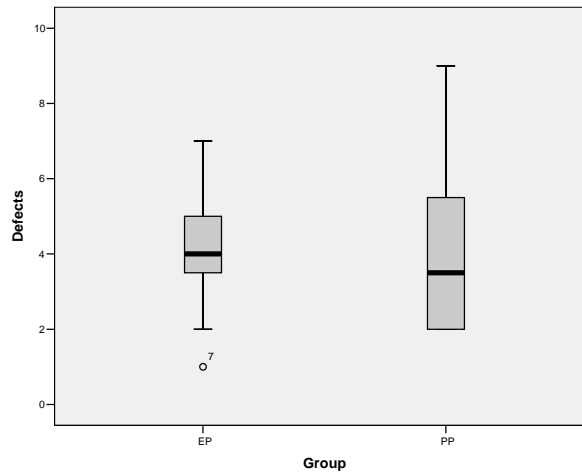
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Discrepancies	Equal variances assumed	,569	,461	,008	17	,994	,034	4,279	-8,994	9,062
	Equal variances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

Estudo de Caso 1 Análise para a Variável Defeitos

➤ Análise Visual Inicial:

- Outlier moderado
- Aparente grande variabilidade entre as duas amostras



Estudo de Caso 1 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

Estudo de Caso 1 Análise para a Variável Defeitos

- Teste de Levene para igualdade de variâncias, a um nível de significância de 5%, sendo:
 H_0 : As variâncias são iguais ($\sigma^2_{GrupoEP} = \sigma^2_{GrupoPP}$)
 H_1 : As variâncias são diferentes ($\sigma^2_{GrupoEP} \neq \sigma^2_{GrupoPP}$)

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	
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

- 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

Estudo de Caso 1 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 ($\mu_{GrupoEP} = \mu_{GrupoPP}$)
 H_1 : As médias são diferentes ($\mu_{GrupoEP} \neq \mu_{GrupoPP}$)

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	
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

Estudo de Caso 1 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

Estudo de Caso 2

- Baseado em um questionário de levantamento de riscos de projetos, foi criada uma técnica para a quantificação dos riscos relacionados com projetos de software (COSTA, 2005)
- A técnica proposta requer o conhecimento da importância relativa entre os fatores de risco que afetam os projetos de software
- A fim de determinar esta informação, um estudo experimental foi realizado e seus resultados permitiram determinar a importância dos fatores de risco de acordo com três tamanhos distintos de projetos de software:
 - projetos pequenos: até 100 Homens/Mês
 - projetos médios: até 300 Homens/Mês
 - projetos grandes: acima de 300 Homens/Mês
- O questionário possuiu 211 questões que foram classificadas em dez grupos, denominados fatores de risco, aplicado a 50 participantes, ponderados em função de sua caracterização

Estudo de Caso 2

- Fatores de Risco considerados:
 - Análise: problemas relacionados com o levantamento dos requisitos, sua estabilidade, nível de dificuldade de implementação, validação e complexidade do sistema
 - Projeto: problemas relacionados à correta concepção da arquitetura, interfaces, algoritmos e mecanismos que facilitem a implementação do sistema
 - Codificação: problemas relacionados à complexidade de implementação dos algoritmos, inadequação de linguagem ou hardware e reutilização de código
 - Teste: problemas relacionados ao planejamento e execução, condições de realização, tipos e abrangência dos testes do sistema.
 - Planejamento: problemas relacionados à experiência dos gerentes, capacidade de elaboração de planejamento e estimativas do projeto, bem como aspectos de definição, utilização e adequação do processo de desenvolvimento de sistemas

Estudo de Caso 2

- Fatores de Risco considerados (cont.):
 - Controle: problemas relacionados à condução do projeto, aprovação de artefatos, resolução de conflitos e apoio à equipe de desenvolvimento, bem como as atividades de acompanhamento e replanejamento ao longo do projeto e a avaliação do processo de desenvolvimento
 - Equipe: problemas relacionados à capacidade, estabilidade, treinamento, maturidade e forma de trabalhar da equipe, bem como o ambiente de desenvolvimento, e o grau com que a equipe segue os planejamentos e os processos
 - Política e Estrutura: problemas relacionados à Política e à Estrutura Organizacional, apoio da Alta Gerência ao projeto, metas e conflitos de interesses
 - Contratos: problemas relacionados aos contratos, subcontratos, fornecedores e dependências externas do projeto.
 - Clientes: estão citados problemas relacionados ao envolvimento do cliente no projeto, número de usuários e nível de mudanças que serão provocadas pelo sistema

Estudo de Caso 2

- Em função do estudo experimental ter sido planejado para capturar informações de 50 indivíduos de forma independente, pode-se afirmar que os dados são independentes e, desta forma, o pressuposto da independência é aceito
- Análise de Dados Univariada: olha-se para cada variável independentemente das demais. Neste caso, não está sendo observada a influência das demais variáveis, podendo não ser o método ideal quando se deseja analisar o impacto de uma variável nas demais
- Método utilizado: ANOVA (Análise de Variância), para testar igualdades de três ou mais médias
 - Pressupostos: normalidade e igualdade de variâncias
 - Delineamento: 1 fator (variável sob análise) e 3 tratamentos (tamanho do projeto)

Estudo de Caso 2

- Normalidade:
 - H0 : Dados tem distribuição normal
 - H1 : Dados não tem distribuição normal

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ANALISE	.160	50	.003	.908	50	.010**
DESIGN	.141	50	.015	.937	50	.018
IMP	.127	50	.043	.907	50	.010**
TESTE	.113	50	.151	.936	50	.017
PLAN	.121	50	.067	.956	50	.126
CONTROLE	.106	50	.200*	.947	50	.051
EQUIPE	.174	50	.001	.926	50	.010**
CONTRATO	.247	50	.000	.746	50	.010**
POL_EST	.088	50	.200*	.959	50	.186
CLIENTE	.094	50	.200*	.945	50	.044

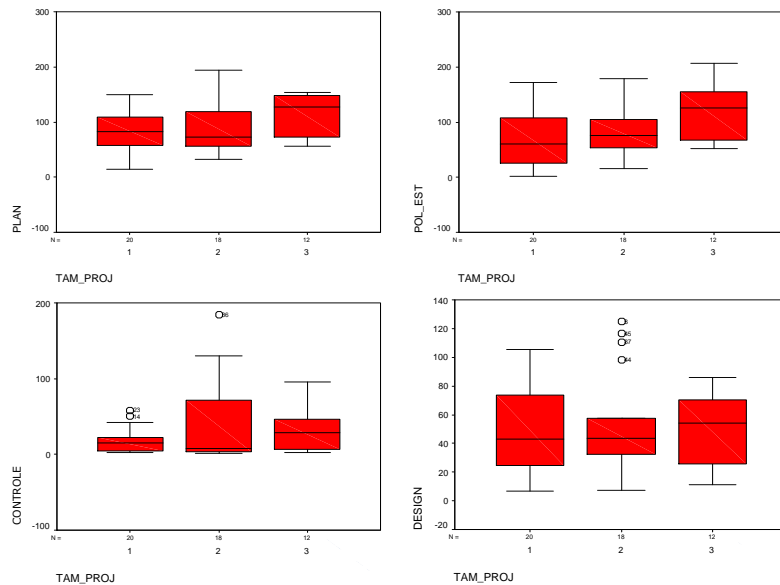
** . This is an upper bound of the true significance.

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

a. Lilliefors Significance Correction

- Para amostras com até 50 elementos (como é o caso), o teste de Shapiro-Wilk é o mais indicado. Assim foram selecionadas as seguintes variáveis para esta análise, que permitirão a utilização de diferentes análises estatísticas:
 - Plan e pol_est possuem distribuição normal (Sig > 0.05)
 - Controle e design não possuem distribuição normal (Sig < 0.05)

Estudo de Caso 2



Estudo de Caso 2

- Para homocedasticidade (variância constante)

H0: dados são homecedásticos

H1: dados não são homocedásticos

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
ANALISE	1.477	2	47	.239
DESIGN	.196	2	47	.823
IMP	.497	2	47	.612
TESTE	3.351	2	47	.044
PLAN	.433	2	47	.651
CONTROLE	.996	2	47	.377
EQUIPE	3.153	2	47	.052
CONTRATO	13.834	2	47	.000
POL_EST	1.567	2	47	.219
CLIENTE	1.107	2	47	.339

- Para a variável Pol_est , como Sig > 0.05, não há indícios para rejeitar H0, assim a distribuição é homocedástica, devendo-se utilizar o método ANOVA para sua análise (é independente, normal e homocedástica). A mesma análise pode ser feita para a variável Plan.
- Já as variáveis Controle e Design são homocedásticas mas não são normais. Neste caso, deve-se utilizar um método não paramétrico uma vez que um dos pressupostos foi violado

Estudo de Caso 2

Análise para a Variável Planejamento

- Teste Paramétrico ANOVA: para estes testes, será verificado se o tamanho do sistema influencia na variável Planejamento
 H_0 : não há diferença de médias: $\mu_{PlanTam1} = \mu_{PlanTam2} = \mu_{PlanTam3}$
 H_1 : pelo menos duas médias são diferentes
- Como Sig = 0.149 (> 0.05), não há indícios para rejeitar H_0 logo, para a variável Plan, o tamanho do projeto não influencia nesta variável

ANOVA

PLAN

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5760.092	2	2880.046	1.982	.149
Within Groups	68307.547	47	1453.352		
Total	74067.639	49			

Estudo de Caso 2

Análise para a Variável Política/Estrutura

- Teste Paramétrico ANOVA: para estes testes, será verificado se o tamanho do sistema influencia na variável Política/Estrutura
 H_0 : não há diferença de médias: $\mu_{Pol_EstTam1} = \mu_{Pol_EstTam2} = \mu_{Pol_EstTam3}$
 H_1 : pelo menos duas médias são diferentes
- Como Sig = 0.018 (< 0.05), deve-se rejeitar H_0 : há diferença de médias

ANOVA

POL_EST

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20221.457	2	10110.729	4.370	.018
Within Groups	108739.3	47	2313.602		
Total	128960.8	49			

Estudo de Caso 2

Análise para a Variável Política/Estrutura

- Desta forma, deve-se proceder com análise para verificar diferenças de médias em função do tamanho do projeto
- Neste caso, pode-se utilizar tanto o método de Tukey quanto o método de Bonferroni. Será utilizado o método de Tukey nesta análise

Multiple Comparisons

Dependent Variable: POL_EST

	(I) TAM_PROJ	(J) TAM_PROJ	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	-14.0641	15.6273	.643	-51.8841	23.7560
		3	-51.5150*	17.5636	.014	-94.0211	-9.0089
	2	1	14.0641	15.6273	.643	-23.7560	51.8841
		3	-37.4509	17.9258	.103	-80.8335	5.9317
	3	1	51.5150*	17.5636	.014	9.0089	94.0211
		2	37.4509	17.9258	.103	-5.9317	80.8335
Bonferroni	1	2	-14.0641	15.6273	1.000	-52.8620	24.7338
		3	-51.5150*	17.5636	.016	-95.1200	-7.9099
	2	1	14.0641	15.6273	1.000	-24.7338	52.8620
		3	-37.4509	17.9258	.126	-81.9552	7.0533
	3	1	51.5150*	17.5636	.016	7.9099	95.1200
		2	37.4509	17.9258	.126	-7.0533	81.9552

*. The mean difference is significant at the .05 level.

Estudo de Caso 2

Análise para a Variável Política/Estrutura

- Comparando Tam1 e Tam2:
 H0: não há diferença de médias: $\mu_{Pol_EstTam1} = \mu_{Pol_EstTam2}$
 H1: há diferença de médias
 - Como Sig = 0.643 (> 0.05), não há indícios para rejeitar H0, portanto, as médias são iguais para esta variável para projetos pequenos (Tam1) e projetos médios (Tam2)
- Comparando Tam2 e Tam3:
 H0: não há diferença de médias: $\mu_{Pol_EstTam2} = \mu_{Pol_EstTam3}$
 H1: há diferença de médias
 - O mesmo pode ser observado para Tam2 e Tam3, uma vez que Sig = 0.103 (> 0.05), onde as médias para projetos médios (Tam2) e projetos grandes (Tam3) são iguais
- Comparando Tam1 e Tam3:
 H0: não há diferença de médias $\mu_{Pol_EstTam1} = \mu_{Pol_EstTam3}$
 H1: há diferença de médias
 - Para Sig = 0.014 (< 0.05) as médias para projetos pequenos (Tam1) são diferentes para projetos grandes (Tam3)

Estudo de Caso 2 Análise para a Variável Projeto

- Teste Não Paramétrico: Kruskal-Wallis (alternativa não paramétrica ao teste ANOVA para amostras independentes)

H0: não há diferença de médias: $\mu_{\text{DesignTam1}} = \mu_{\text{DesignTam2}} = \mu_{\text{DesignTam3}}$
H1: pelo menos duas médias são diferentes

- Como Asymp. Sig. = 0.984 (> 0.05), não há indícios para rejeitar H0, logo as médias são iguais, ou seja, o risco associado à variável Projeto independe do tamanho do projeto

	TAM_PROJ	N	Mean Rank
DESIGN	1	20	25.10
	2	18	25.94
	3	12	25.50
Total		50	

	DESIGN
Chi-Square	.032
df	2
Asymp. Sig.	.984

a. Kruskal Wallis Test

b. Grouping Variable: TAM_PROJ

Estudo de Caso 2 Análise para a Variável Controle

- Teste Não Paramétrico: Kruskal-Wallis

H0: não há diferença de médias: $\mu_{\text{ControleTam1}} = \mu_{\text{ControleTam2}} = \mu_{\text{ControleTam3}}$
H1: pelo menos duas médias são diferentes

- Como Asymp. Sig. = 0.059 é muito próximo de 0.05, deve-se rejeitar H0, logo, existem diferenças de médias
- Assim, deve-se fazer comparações dois a dois a partir do tamanho do projeto para esta variável. Para isso, será utilizado o Teste de Mann-Whitney, que é uma alternativa não paramétrica ao Teste T de amostras independentes.

	TAM_PROJ	N	Mean Rank
CONTROLE	1	20	20.40
	2	18	26.17
	3	12	33.00
Total		50	

	CONTROLE
Chi-Square	5.662
df	2
Asymp. Sig.	.059

a. Kruskal Wallis Test

b. Grouping Variable: TAM_PROJ

Estudo de Caso 2 Análise para a Variável Controle

- Variável Controle para Tamanhos de Projetos Pequenos (Tam1) e Médios (Tam2)
H0: não há diferença de médias: $\mu_{\text{ControleTam1}} = \mu_{\text{ControleTam2}}$
H1: há diferença de médias
- Como Asymp. Sig. (2-tailed) = 0.152 (> 0.05), não há indícios para rejeitar H0, logo as médias iguais, portanto, para a variável Controle, não há diferença de média entre projetos pequenos e médios

	TAM_PROJ	N	Mean Rank	Sum of Ranks
CONTROLE	1	20	17.05	341.00
	2	18	22.22	400.00
	Total	38		

	CONTROLE
Mann-Whitney U	131.000
Wilcoxon W	341.000
Z	-1.433
Asymp. Sig. (2-tailed)	.152
Exact Sig. [2*(1-tailed Sig.)]	.158 ^a

a. Not corrected for ties.

b. Grouping Variable: TAM_PROJ

Estudo de Caso 2 Análise para a Variável Controle

- Variável Controle para Tamanhos de Projetos Médios (Tam2) e Grandes (Tam3)
H0: não há diferença de médias: $\mu_{\text{ControleTam2}} = \mu_{\text{ControleTam3}}$
H1: há diferença de médias
- Como Asymp. Sig. (2-tailed) = 0.117 (> 0.05), não há indícios para rejeitar H0, logo as médias iguais, portanto, para a variável Controle, não há diferença de média entre projetos médios e grandes

	TAM_PROJ	N	Mean Rank	Sum of Ranks
CONTROLE	2	18	13.44	242.00
	3	12	18.58	223.00
	Total	30		

	CONTROLE
Mann-Whitney U	71.000
Wilcoxon W	242.000
Z	-1.566
Asymp. Sig. (2-tailed)	.117
Exact Sig. [2*(1-tailed Sig.)]	.124 ^a

a. Not corrected for ties.

b. Grouping Variable: TAM_PROJ

Estudo de Caso 2 Análise para a Variável Controle

- Variável Controle para Tamanhos de Projetos Pequenos (Tam1) e Grandes (Tam3)
 - H0: não há diferença de médias: $\mu_{\text{ControleTam1}} = \mu_{\text{ControleTam3}}$
 - H1: há diferença de médias
- Como Asymp. Sig. (2-tailed) = 0.039 (< 0.05), deve-se rejeitar H0 e, portanto, as médias são diferentes, portanto, para a variável Controle, há diferença de média entre projetos pequenos e grandes

	TAM_PROJ	N	Mean Rank	Sum of Ranks
CONTROLE	1	20	13.85	277.00
	3	12	20.92	251.00
	Total	32		

	CONTROLE
Mann-Whitney U	67.000
Wilcoxon W	277.000
Z	-2.063
Asymp. Sig. (2-tailed)	.039
Exact Sig. [2*(1-tailed Sig.)]	.040 ^a

a. Not corrected for ties.

b. Grouping Variable: TAM_PROJ

Estudo de Caso 2 Análise Multivariada

- Neste estudo de caso poderiam ter sido utilizadas técnicas de análise multivariada, comparando a relações entre as variáveis de forma conjunta, como por exemplo:
 - Regressão Linear Simples
 - MANOVA (Análise Multivariada de Variância)
 - Análise de Componentes Principais (PCA)
 - Análise de Agrupamento
 - Regressão Logística
- Estas técnicas são mais complexas e não serão aqui abordadas

Estudo de Caso 2 Conclusões

- Para as variáveis consideradas, pode-se concluir, a respeito do impacto dos fatores de risco avaliados:
 - Tamanho do Projeto não influencia no risco associado ao Planejamento
 - Tamanho do Projeto não influencia no risco associado à Política/Estrutura entre projetos pequenos e médios e entre projetos médios e grandes. Existe diferença para esta variável entre projetos pequenos e grandes
 - Tamanho do projeto não influencia no risco associado ao Projeto do sistema
 - Tamanho do projeto não influencia no risco associado ao Controle entre projetos pequenos e médios e entre projetos médios e grandes. Existe diferença para esta variável entre projetos pequenos e grandes

Referências Bibliográficas

- 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 (to be published), 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.
- Montgomery, D. C., "Estatística Aplicada e Probabilidade para Engenheiros", Ed. LTC, 2003.
- Montgomery, D. C., "Design and Analysis of Experiments", Ed. IE-Wiley, 2000.

Referências Bibliográficas

- 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.