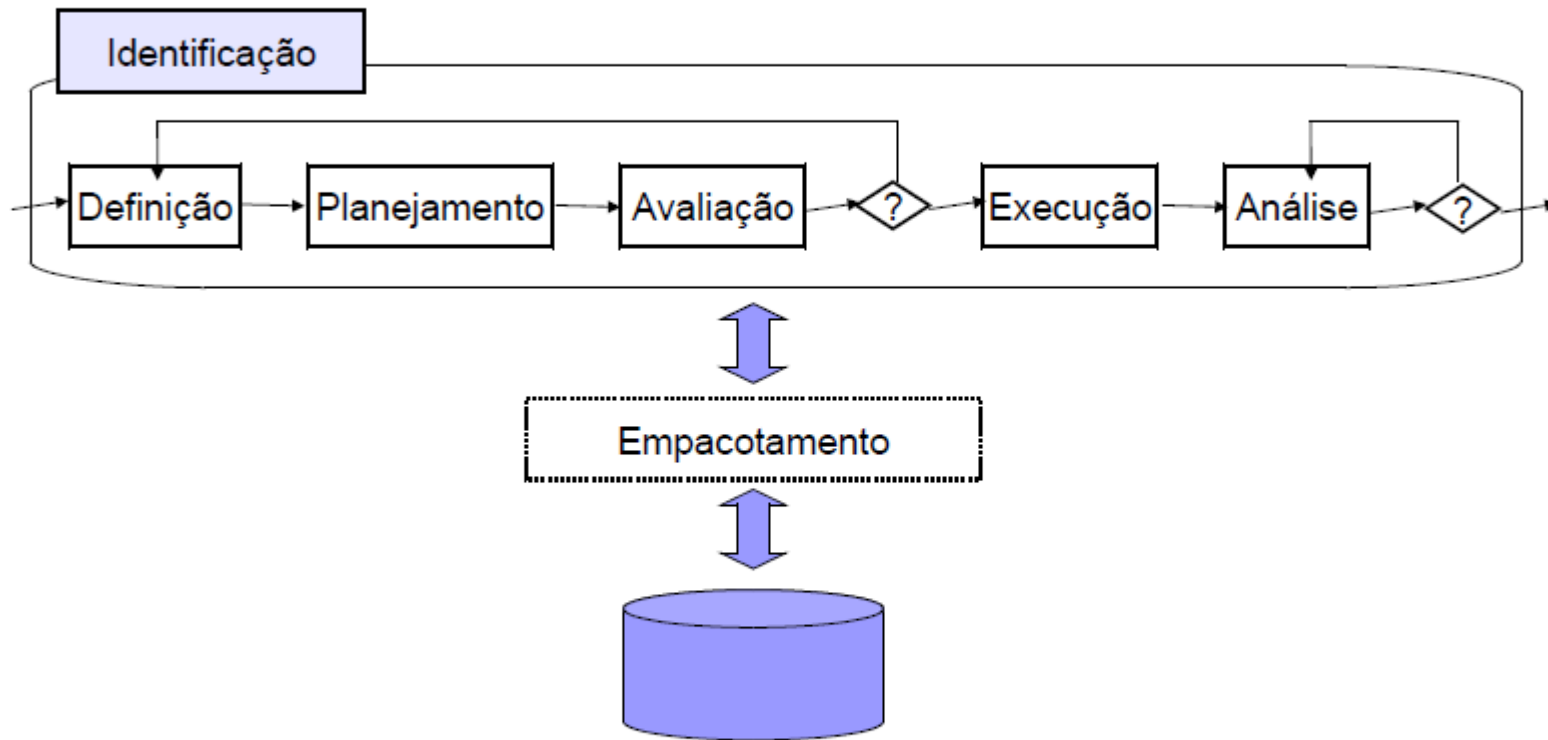


Análise e Interpretação dos Dados Coletados

Prof. Paulo C. Masiero
Cap. 8

Processo de Experimentação



Análise dos Resultados

- Se possível, **entreviste os participantes** para obter feedback:
 - Sobre os artefatos
 - Sobre o processo experimental
 - Para capturar sua impressão sobre os resultados
- **Revise os dados coletados** para verificar se eles são úteis e válidos
- Organize os dados em conjuntos para análise de validade, exploração e teste das hipóteses
- **Analise os dados** com base em princípios estatísticos válidos
- Verifique se as **hipóteses** são aceitas ou rejeitadas
- O processo de análise pode ser **iterativo**.

Análise e interpretação dos resultados

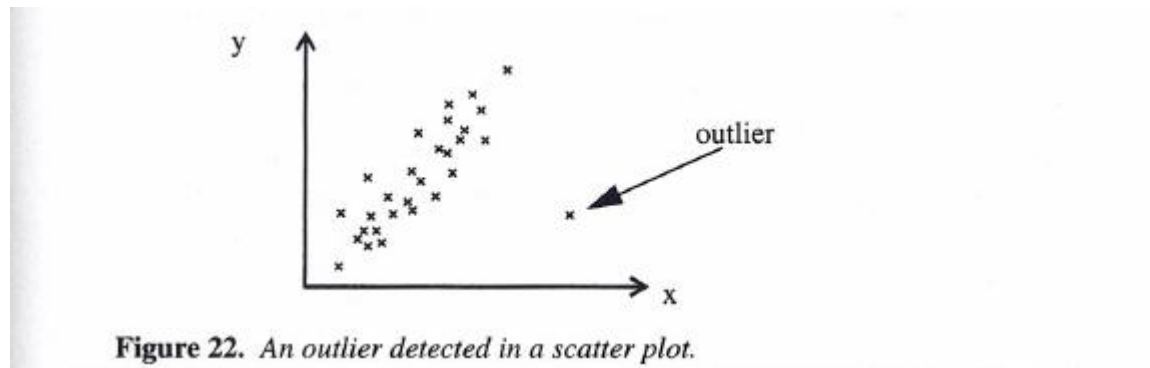
- Redução do conjunto de dados
- Análise descritiva dos dados
- Testes de hipóteses

Redução do conjunto de dados

- Todos os métodos estatísticos dependem da qualidade dos dados usados.
- Se os dados não representam o que nós pensamos que eles representam, então a conclusão que obtemos dos resultados dos métodos não são corretas
- Os erros podem ocorrer de forma sistemática ou como outliers.
- Redução do conjunto de dados é relacionado com validação dos dados.

Outliers

- Diagramas de espalhamento são efetivos para identificar outliers, assim como box plots.



Outliers (Cont.)

- Os outliers devem ser identificados com base na execução do experimento na forma dos dados coletados, do conjunto dos dados e da análise descritiva.
- Quando outliers são identificados, o importante é decidir o que fazer com eles, analisando também porque eles ocorreram.

Outliers: diretrizes

- Um evento estranho ou raro que não deve ocorrer novamente : exclua o dado da amostra
 - Ex. Dado não entendido, ou errado
- Um evento raro que pode ocorrer novamente, é mais sensato não excluir, pois o outlier tem uma informação.
 - Ex. Resultado de sujeito inexperiente.
 - Se essa variável não foi considerada antes (ex. experiência, pode-se também dividir a amostra com base nela e fazer duas (ou mais) análises. Isso deve ser feito caso a caso

Outliers: diretrizes

- Não são apenas dados inválidos que podem ser retirados da amostra.
- Muitas vezes não é efetivo analisar dados redundantes se forem muitos.
 - Técnicas para identificar redundância são a análise de componentes e identificação de fatores ortogonais (não tratadas no livro)
 -

Análise Descritiva

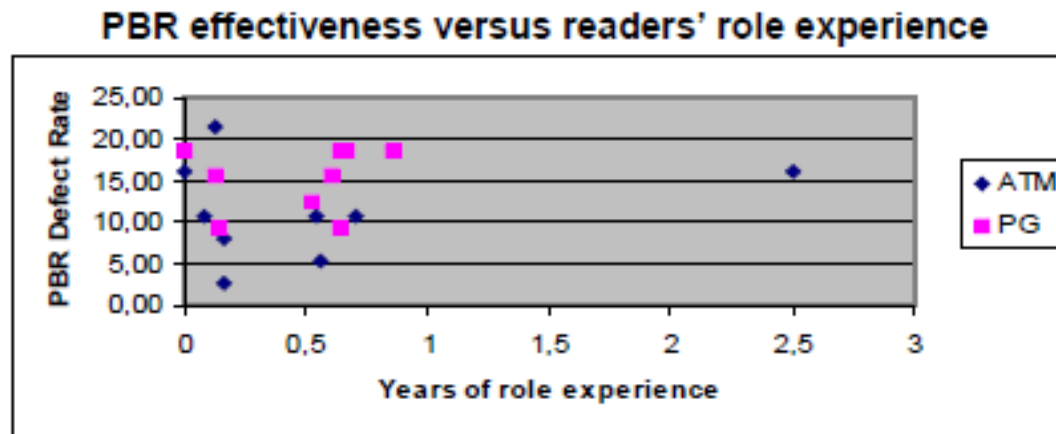
- Estatística Descritiva
 - Medidas de Tendência Central (média, mediana, moda)
 - Medidas de dispersão (desvio padrão, variância)
 - Correlações (Pearson, Spearman)
- Análise Gráfica
 - Diagramas de dispersões
 - Histogramas e Gráficos de Pizza
 - Box Plots

Metas da Análise Descritiva

- Identificar tendências centrais de variáveis e seus tratamentos
- Identificar o grau de dispersão
- Identificar pontos fora da curva (outliers)
- Identificar Correlações

Exemplo de Análise Descritiva

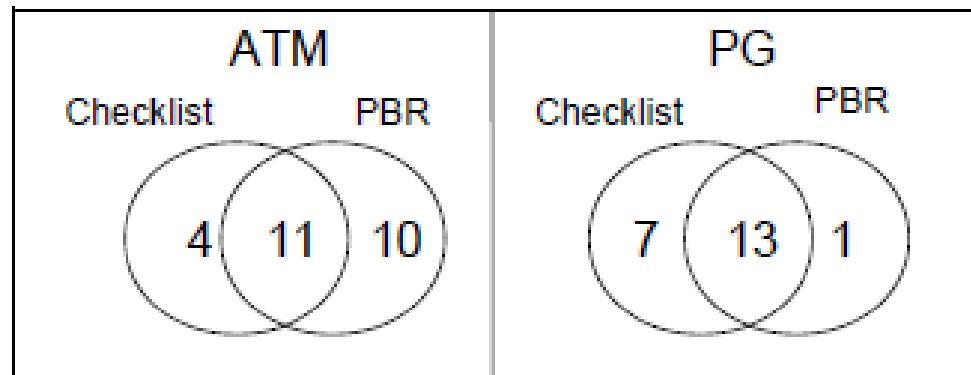
- O3') Does the reviewer's experience affect his or her effectiveness?



Analysis

- We used a questionnaire to measure the subject's experience in their assigned perspective. The relationship between experience and effectiveness is weak
- Reviewers with more experience do not perform better than reviewers with less experience
- This conclusion is supported by the results of the Spearman's and Pearson's correlation tests that showed numbers smaller than 14%, far from indicating a high degree of correlation

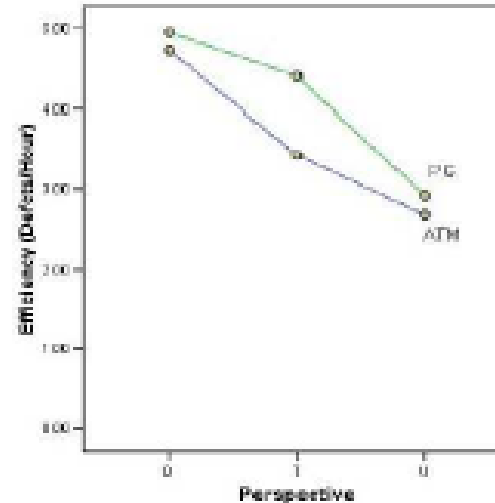
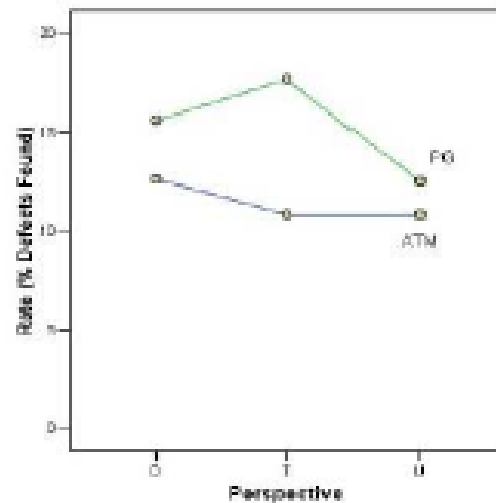
- **R1) Do individual reviewers using PBR and Checklist find different defects?**



Analysis

- **ATM:** the two techniques appear to be complementary in that users of each technique found defects that were not found by the other technique
- **PG:** the techniques do not appear to be complementary, because the PBR users only found 1 defect not found by the checklist users

➤ **R2) Do the PBR perspectives have the same effectiveness and efficiency?**

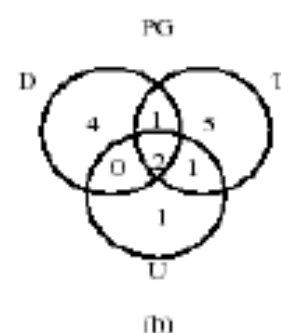
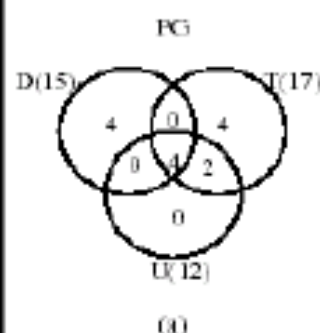
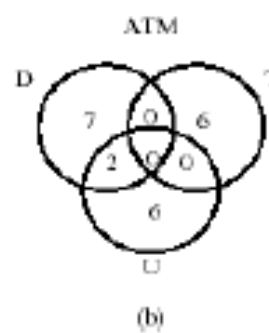
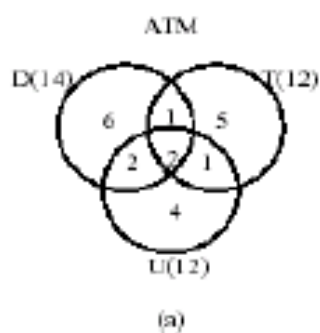


Rate and Efficiency by Perspective

Analysis

- Each point represents the mean of the 3 reviewers composing the group.
- ATM: Designer were the most effective and efficient
- PG: Tester were the most effective and Designer were the most efficient
- The perspectives had no significant effect on either effectiveness ($p=.654$) or efficiency ($p=.182$)

➤ **R3) Do the PBR perspectives find different defects?**



Analysis

- **ATM:**
 - each perspective identified unique defects with little overlap
 - the three perspectives were more likely to find different defects
 - The perspectives identified a similar number of occurrences overall
- **PG:**
 - the Designer and Tester perspectives appear to be complementary, but the User perspective does not provide much added benefit
 - The perspectives identified a similar number of occurrences

Testes de Hipóteses

- Hipóteses avaliadas por testes estatísticos definidos pelos pesquisadores da estatística inferencial
- Normalmente são definidas duas hipóteses
 - **Hipótese nula (H_0):** 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 (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

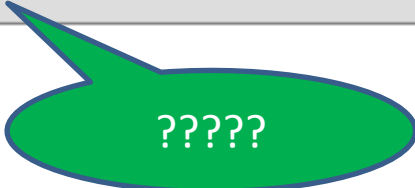
Testes de Hipótese

- Os teste comparam médias entre grupos de participantes realizando tratamentos diferentes

“Utilizando a técnica XYZ, os desenvolvedores concluem a atividade de projeto em menos tempo do que utilizando a técnica ABC”

Hipótese Nula: $\mu (\text{Tempo}_{XYZ}) = \mu (\text{Tempo}_{ABC})$

Hipótese Alternativa: $\mu (\text{Tempo}_{XYZ}) \neq \mu (\text{Tempo}_{ABC})$



?????

Teste Estatístico

- Calculados fundamentalmente a partir de uma função de teste que considera três valores:
 - Diferença entre os valores “médios” das estatísticas para os tratamentos
 - “Dispersão” dos valores da estatística
 - Número de amostras
- A função de teste, $F(m, \sigma, N)$, depende do:
 - tipo de distribuição dos dados, e.x., normalidade e homocedasticidade.
 - Número de fatores e tratamentos



Homogeneidade da
variância

Exemplo

- Quer-se testar a Hipótese “homens são mais altos que mulheres”
 - Determina-se uma amostra da população utilizando um fator e dois tratamentos
 - A certeza depende do:
 - Número de pessoas amostradas
 - Da diferença entre a altura média nos tratamentos
 - Da dispersão da altura nos tratamentos

Tipos de Erros

- A verificação das hipóteses sempre lida com o risco de um erro de análise acontecer
 - O erro do tipo I (a) acontece quando o teste estatístico indica um relacionamento entre causa e efeito e o relacionamento real não existe
 - O erro do tipo II (b) acontece quando o teste estatístico não indica o relacionamento entre causa e efeito, mas existe este relacionamento

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

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

Nível de Significância

- Indica a probabilidade de se cometer um erro tipo-I
 - Os níveis de significância (α) mais comumente utilizados são 10%, 5%, 1% e 0.1%
 - Chama-se 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

Procedimento para o Teste de Hipótese

- Fixar o nível de significância do teste
- Obter uma estatística (estimador do parâmetro que se está testando) que tenha distribuição conhecida sob H_0
- A estatística de teste e o nível de significância constroem a região crítica pela qual o teste passa
- 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

Teste de Hipótese na Prática

- Na prática:
 - escolhe-se a estatística de teste
 - escolhe-se o valor P (significância)
 - Usa-se uma ferramenta estatística para aplicar o teste e verificar o valor de P
- A escolha do teste depende da determinação do tipo de distribuição dos dados e de quantos fatores e tratamentos vão ser analisados no teste
 - Testes paramétricos: assumem uma distribuição e são mais poderosos
 - Testes não paramétricos: não assumem uma distribuição . Têm uma aplicação mais abrangente

Alguns Tipos de Teste

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

Exemplo

Teste de Hipóteses

➤ **O1') Do PBR teams detect a more defects than Checklist teams?**

- **H0:** There is no difference in the defect detection rates of teams applying PBR compared to teams applying the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer has only random effects on team scores.
- **Ha:** The defect detection rates of teams applying PBR are higher compared to teams using the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer decreases the effectiveness of the team.

Analysis:

- Doing a permutation test as done in the original experiment, there were 48620 distinct ways to assign the reviewers into groups of 9.
- The group with no dilution (all PBR reviewers) had the 24769th highest test statistic, corresponding to a p-value of 0.51.
- Therefore, unlike the original study, we cannot reject the hypothesis H0.

➤ **O2') Do individual PBR or Checklist reviewers find more defects?**

- Group effect (RT X DOC interaction)
- H0: There is no difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency.
- Ha: There is a difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency

- Main effect RT
- H0: There is no difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.
- Ha: There is a difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.

- Main effect DOC
- H0: There is no difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.
- Ha: There is a difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.

Because the experimental groups had the same number of subjects, the ANOVA for balanced design was used. This analysis involved two different factors, or treatments: the reading technique (RT), the requirement document (DOC).

➤ O2') Do individual PBR or Checklist reviewers find more defects?

ANOVA summary table with respect
to the individual effectiveness

Independent Variables	Effectiveness (average percentage MINITAB)	P
RT X DOC	-	0.275
RT	Checklist= 11.417; PBR= 13.346	0.404
DOC	ATM= 9.310; PG= 15.453	0.005✓

ANOVA summary table with relation
to the individual efficiency

Independent Variables	Efficiency (average)	P
RT X DOC	-	0.417
RT	Checklist= 2.775; PBR= 3.856	0.101
DOC	ATM= 2.817; PG= 3.814	0.131

➤ O2') Do individual PBR or Checklist reviewers find more defects?

Document	ATM		PG	
Technique	Checklist	PBR	Checklist	PBR
Defects Found/Total defects	15/37 (40.5%)	21/37 (56.8%)	20/32 (60.5%)	14/32 (43.75%)
Occurrences of Defects/Total occurrences	24/333	38/333	45/288	44/288
Effectiveness	7.21	11.41	15.63	15.28
Efficiency	2.00	3.62	3.53	4.10

Analysis

- ATM: PBR found a higher percentage of the defects than checklist.
p-value = 0.143 (not statistically significant at the .05 level)
- PG: Checklist found a higher percentage of the defects than PBR.
p-value = 0.911 (not statistically significant at the .05 level)
- Efficiency (errors/hour): PBR were more efficient for both documents.
ATM p-value=.107, PG pvalue=.51 (not statistically significant at the .05 level)