Engenharia de Software Experimental – Testes de Inferência

Prof. Márcio Barros
PPGI / UNIRIO

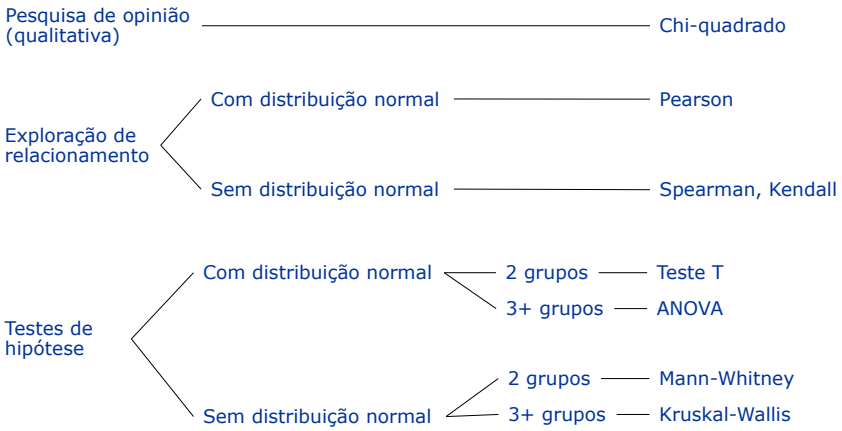
Testes de Hipótese

- Um estudo experimental tem como objetivo colher dados para confirmar ou rejeitar uma hipótese
  - » Os testes estatísticos verificam se é possível rejeitar a hipótese nula, de acordo com um conjunto de dados observados e suas propriedades estatísticas
  - » Os estudos de Engenharia de Software se dividem em qualitativos e quantitativos
  - » Estudos qualitativos se baseiam na contagem de respostas (ex. surveys) ou relacionamentos entre estas respostas
  - » Estudos quantitativos comparam médias entre os grupos de participantes que realizam tratamentos distintos

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

Hipótese nula: μ (Tempo$_Y$) = μ (Tempo$_X$)
Hipótese alternativa: μ (Tempo$_Y$) $\neq$ μ (Tempo$_X$)

Testes de Hipótese: Escolha do Tipo de Teste



Testes de Hipótese: Procedimento

- Fixar o nível de significância do teste, α, geralmente como 0.05 ou 0.01
- Obter uma estatística (estimador do parâmetro que está sendo testado) que tenha distribuição conhecida
- Através da estatística de teste e do nível de significância, construir a região crítica para o teste
- Utilizando as informações amostrais (dados coletados), obter o valor da estatística (estimativa do parâmetro)
- Se o valor da estatística pertencer à região crítica, rejeita-se a hipótese nula e aceita-se a hipótese alternativa
- Caso contrário, não se rejeita a hipótese nula e nada se pode dizer a respeito da hipótese alternativa
- Calcular uma medida de tamanho de efeito para entender a diferença prática dos resultados observados

Tipos de Erro

- O teste de hipótese sempre lida com algum tipo de risco, ou seja, as chances com 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

α = P (erro-tipo-I) = P (H_{NULA} é rejeitada | H_{NULA} é verdadeira)

β = P (erro-tipo-II) = P (H_{NULA} não é rejeitada | H_{NULA} é falsa)

Potência do Teste

- Indica a probabilidade de rejeitar a hipótese nula se esta for 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

Potência = 1 - β

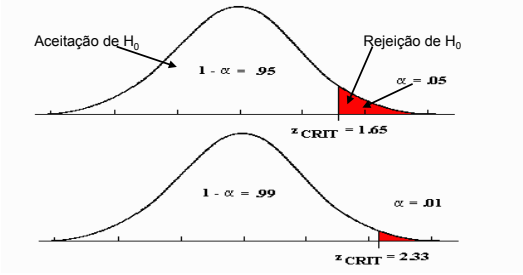
Potência = P (H_{NULA} rejeitada | H_{NULA} é falsa)

Nível de Significância

- Indica a probabilidade de cometer um erro tipo-I
  - » Os níveis de significância (α) mais comumente utilizados são 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, se o *p-value* for 0.0001 pode-se dizer que o resultado é significativo, pois este valor é muito inferior aos níveis de significância usuais
  - » Porém, se o *p-value* for igual a 0.048 pode haver dúvida pois, embora o valor seja inferior, ele está muito próximo ao nível usual de 5%

Nível de Significância

- O nome “hipótese nula” vem do uso frequente dos testes de hipótese na comparação de dois tratamentos
  - » Nestes casos, H_0 é a hipótese de igualdade dos tratamentos, ou seja, de falta de superioridade do tratamento alternativo
  - » 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

- Cuidados com o *p-value*!
  - » O *p-value* foi definido em torno de 1920 como uma "*maneira de julgar se as evidências disponíveis são significativas e merecem atenção do pesquisador*"
  - » O *p-value* não é um resultado definitivo e inquestionável da existência de diferença estatística, pois ele depende do tamanho da população
  - » Com um grande volume de dados, é possível obter resultados com diferença estatisticamente significativa, mesmo que a diferença seja tão pequena que não tenha significado prático
  - » Se fizermos um número suficientemente grande de comparações, em algum momento iremos encontrar alguma diferença!
  - » O *p-value* também não mostra se a diferença observada é relevante na prática, devendo ser complementado com uma medida de *tamanho de efeito*

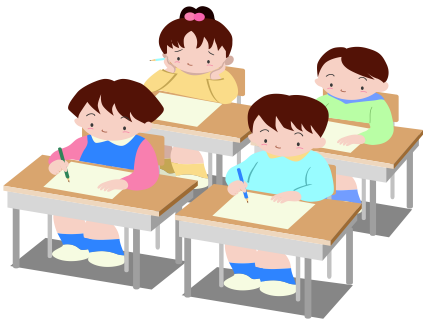
Tamanho de Efeito (*Effect-Size*)

- Tamanho de efeito é uma medida estatística descritiva da força do relacionamento entre duas variáveis em uma população ou amostra
  - » O tamanho de efeito expressa a magnitude do relacionamento sem indicar se esta o relacionamento entre os dados é casual ou significativo
  - » Como um exemplo de tamanho de efeito *absoluto*, podemos dizer que a técnica de teste A identifica 30 erros a mais que a técnica de teste B
  - » Como um exemplo de tamanho de efeito *padronizado*, podemos dizer que resultados identificados pelo algoritmo A são mais eficientes do que os resultados obtidos pelo algoritmo B em 80% das execuções
  - » Medidas de tamanho de efeito padronizados são preferenciais à medidas absolutas por serem independentes do contexto (no exemplo acima, 30 erros a mais é muito ou pouco? depende do número total de erros ...)
  - » Cada tipo de teste estatístico tem um ou mais tipos de tamanho de efeito

Tamanho de Efeito (*Effect-Size*)

- Algumas medidas de tamanho de efeito
  - » Cramer's V
  - » Odds ration
  - » Correlação
  - » Cohen's d
  - » Vargha and Delaney's $\hat{A}_{12}$

Mais detalhes adiante ...



TESTES QUALITATIVOS

(Testes baseados em contagem)

Testes Qualitativos

- São testes baseados exclusivamente na contagem de valores
  - » Diferentes grupos de participantes respondem a um questionário
  - » Cada pergunta possui um número limitado de respostas
  - » Para cada pergunta, queremos saber se há diferença nas respostas entre os diferentes grupos
- Exemplo de estudo
  - » Diversas empresas de software dos segmentos de jogos e sistemas de saúde foram convidadas a responder um questionário sobre o uso de métodos ágeis
  - » Entre as perguntas, foi questionado se as empresas usam programação em pares: cada empresa respondeu sim ou não
  - » Em outra pergunta, foi questionado que sistema operacional as empresas usam para o desenvolvimento de software: Windows, Mac ou Linux

Testes Qualitativos

- A primeira pergunta é um estudo de 1 fator e 2 tratamentos
  - » Os dados podem ser tabulados conforme a tabela abaixo

Número de empresas	Usa programação em pares	Não usa programação em pares
Jogos	10	30
Saúde	20	20

- Um teste de chi-quadrado (χ^2) pode ser utilizado para saber se os dados dos dois grupos são estatisticamente diferentes
  - » Os dados devem ter sido colhidos de forma independente
  - » Os dados não podem ser pareados (ex.: duas coletas com a mesma pessoa)

Testes Qualitativos

```
> data <- matrix(c(10, 30, 20, 20), ncol=2, byrow=TRUE);
> chisq.test(data)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 4.32, df = 1, p-value = 0.03767
```

Conclusão: p-value < 0.05 indica que existem diferenças na distribuição de dados com 95% de confiança

Testes Qualitativos

- O tamanho de efeito padronizado do teste χ^2 pode ser calculado pela estatística Cramer's V

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

onde *n* é o número total de opiniões

k é o número total de colunas

r é o número de linhas

	Pequeno	Médio	Grande
Cramer's V	> 0.10	> 0.30	> 0.50

Fonte: Cohen, J. (1992), "A Power Primer", Psychological Bulletin, 112, pp. 155-159

Testes Qualitativos

```
> cramerV <- function(data)
> {
>    tempchi <- chisq.test(data);
>    chi2 <- unname(tempchi$statistic["X-squared"]);
>    pvalue <- unname(tempchi$p.value);
>    cv <- sqrt(chi2 / sum(data) / (min(length(data), nrow(data))-1));
>    c(effsize = cv, p.value = pvalue, chi2 = chi2);
> }
> cramerV(data)
   effsize    p.value       chi2
0.23237900 0.03766692 4.32000000
```

Tamanho de efeito pequeno

Testes Qualitativos

- Como apresentar o resultado na Dissertação ou nos artigos?

O teste de chi-quadrado indica que o percentual de uso de programação em pares é significativamente diferente entre os segmentos das empresas com tamanho de efeito pequeno (χ^2 = 4.32, p-value = 0.03767, cramer-V = 0.23).

Testes Qualitativos

- Uma alternativa ao teste de chi-quadrado é o teste de Fisher
  - » O teste de Fisher é útil para pequenos volumes de dados
  - » É difícil saber o que é um "volume de dados pequeno"!
  - » Se recomenda usar o teste de Fisher quando há menos de 10 observações em qualquer categoria de dados observados

```
> data <- matrix(c(10, 30, 20, 20), ncol=2, byrow=TRUE);
> fisher.test(data)

        Fisher's Exact Test for Count Data

data:  data
p-value = 0.03683
```

Conclusão: p-value < 0.05 indica que existem diferenças na distribuição de dados com 95% de confiança

Testes Qualitativos

- Se tivermos apenas duas variáveis binárias, podemos usar o *odds ratio* como medida de tamanho de efeito não padronizada
  - » Por exemplo, na primeira pergunta temos 10 empresas de jogos que usam programação em pares para cada 30 que não usam, gerando uma relação de 1:3, ou seja, 0.33
  - » Temos ainda 20 empresas de saúde que usam programação em pares para cada 20 que não usam, gerando uma relação de 1:1 (ou seja, 1.0)
  - » O *odds ratio* é calculado como a divisão da relação entre os grupos: 0.33/1.0 ou 0.33, indicando que é 3 vezes mais raro a programação em pares em empresas de desenvolvimento de jogos

Testes Qualitativos

```
> data <- matrix(c(10, 30, 20, 20), ncol=2, byrow=TRUE);
> fisher.test(data)

        Fisher's Exact Test for Count Data

data:  data
p-value = 0.03683
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1148838 0.9433281
sample estimates:
odds ratio
 0.3381386
```


Testes Qualitativos

- A segunda pergunta representa um estudo com 1 fator e 3 tratamentos

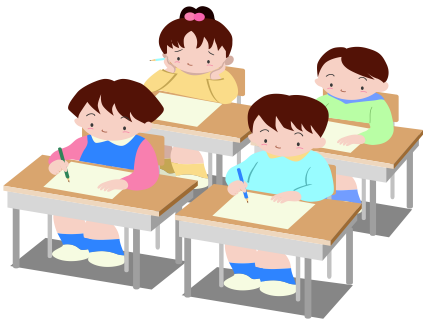
# Empresas	Windows	Mac	Linux
Jogos	16	11	3
Saúde	21	8	1

```
> data2 <- matrix(c(16, 11, 3, 21, 8, 1), ncol=3, byrow=TRUE);
> cramerV(data)
  effsize   p.value     chi2
0.1892683 0.3414070 2.1493599
```

Conclusão: p-value > 0.05 ➔ não é possível afirmar que existe diferença entre os valores dos fatores

Exercício 06

- Foi realizada uma pesquisa com diversas empresas da região Sul do país visando identificar o perfil de uso de tecnologia Java por estas empresas
- Perguntamos para cada empresa se usa *Hibernate* e qual framework de desenvolvimento web suas equipes utilizam (*Struts*, *JSF* ou *Spring*)
- Os dados colhidos foram disponibilizados para sua análise, junto com o estado em que se localiza cada empresa
- Existem diferenças entre o Rio Grande do Sul, Santa Catarina e Paraná?



TESTES DE CORRELAÇÃO

(Relações entre variáveis)

Medidas de Dependência

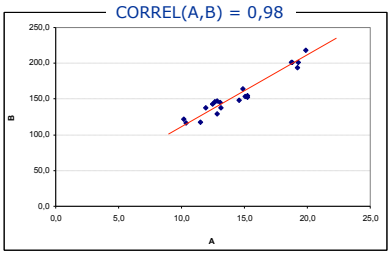
- Quando duas ou mais variáveis estão relacionadas em um estudo, é útil calcular seu grau de dependência
  - » As medidas de dependência determinam a força e direção do relacionamento entre duas ou mais variáveis
  - » A medida de dependência mais comumente utilizada é o coeficiente de correlação
  - » Se o estudo relaciona duas variáveis, a correlação entre elas é representada como um número entre -1 e +1
  - » Se o estudo relaciona mais de duas variáveis, a correlação é representada como uma matriz simétrica onde cada célula assume um valor 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 um relacionamento entre as variáveis

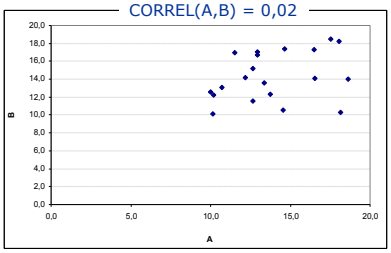
Medidas de Dependência: Correlação de Pearson

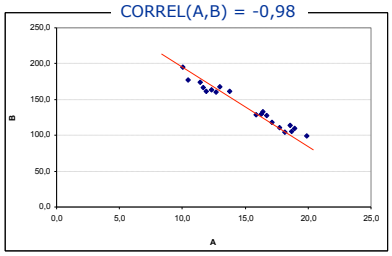
- Coeficiente de correlação mais conhecido
  - » Quantifica a força de associação linear entre duas variáveis e descreve o quanto uma linha reta se ajustaria na representação cartesiana de seus valores
  - » O coeficiente de Pearson assume que os valores assumidos pelas variáveis sob análise seguem aproximadamente distribuições normais
  - » Devido à distribuição normal, esta condição é indicada pela formação de uma nuvem elíptica 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 Pearson







Medidas de Dependência: Correlações em R

- A hipótese nula do teste de correlação é que não existe correlação entre os dados (r = 0)

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> cor.test(xcorretos, ycorretos, method="pearson")

        Pearson's product-moment correlation

data:  xcorretos and ycorretos
t = -0.2526, df = 9, p-value = 0.8063
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6510033  0.5433282
sample estimates:
       cor
-0.0838885
```

Conclusão: p-value > 0.05 indica que é possível que não exista correlação entre os dados observados

Medidas de Dependência: Correlações em R

- O teste também informa o intervalo de confiança para a correlação com 95% de certeza (note o intervalo grande no exemplo abaixo)

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> cor.test(xcorretos, ycorretos, method="pearson")

        Pearson's product-moment correlation

data:  xcorretos and ycorretos
t = -0.2526, df = 9, p-value = 0.8063
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6510033  0.5433282
sample estimates:
       cor
-0.0838885
```


Medidas de Dependência

- O tamanho de efeito de um teste de correlação é a própria correlação

	Pequeno	Médio	Grande
r	> 0.10	> 0.30	> 0.50

Fonte: Cohen, J. (1992), "A Power Primer", Psychological Bulletin, 112, pp. 155-159

- Como apresentar o resultado?

> Observamos que os dados não são fortemente correlacionados e não podemos rejeitar a hipótese de que não existe correlação entre eles (Pearson r = -0.08, p = 0.80).

Medidas de Dependência: Correlação de Spearman

- Outra medida para calcular correlação é o coeficiente de Spearman ou *rank-order correlation*
  - » 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) ou dados sem 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

Escala	Nominal	Ordinal	Intervalar	Razão
Correlação de Pearson			X	X
Correlação de Spearman		X	X	X

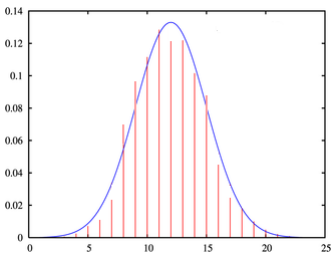
Medidas de Dependência: Correlação de Spearman

- 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

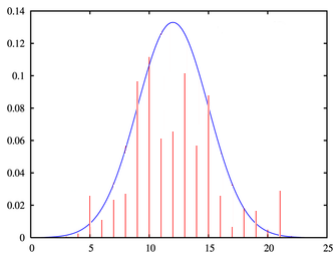
$$\rho = 1 - \frac{6 \cdot \sum_i R(A_i) - R(B_i)}{N(N^2 - 1)}$$

Mas como saber se os dados têm distribuição normal?

- 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 K-S

- O teste de *Kolmogorov-Smirnov* avalia se duas amostras têm distribuições semelhantes ou se uma amostra tem distribuição semelhante a uma distribuição contínua clássica (como a 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 K-S

```
> y <- runif(1000, 2, 4)
> ks.test(y, "pnorm", mean=mean(y), sd=sd(y))

        One-sample Kolmogorov-Smirnov test

data:  y
D = 0.0754, p-value = 2.289e-05
alternative hypothesis: two-sided

> x <- rnorm(1000, 5, 3)
> ks.test(x, "pnorm", mean=mean(x), sd=sd(x))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.016, p-value = 0.9609
alternative hypothesis: two-sided
```

Conclusão: como p-value < 0.05, rejeitamos a hipótese de normalidade

Conclusão: como p-value > 0.05, aceitamos a hipótese de normalidade

PPGI - UNIRIO

Testes de Normalidade: Teste K-S

```
> v1 ← c(23.4, 30.9, 18.8, 23.0, 21.4, 1, 24.6, 23.8, 24.1, 18.7, 16.3, 20.3, 14.9, 35.4, 21.6, 21.2,
         21.0, 15.0, 15.6, 24.0, 34.6, 40.9, 30.7, 24.5, 16.6, 1, 21.7, 1, 23.6, 1, 25.7, 19.3, 46.9,
         23.3, 21.8, 33.3, 24.9, 24.4, 1, 19.8, 17.2, 21.5, 25.5, 23.3, 18.6, 22.0, 29.8, 33.3, 1,
         21.3, 18.6, 26.8, 19.4, 21.1, 21.2, 20.5, 19.8, 26.3, 39.3, 21.4, 22.6, 1, 35.3, 7.0, 19.3,
         21.3, 10.1, 20.2, 1, 36.2, 16.7, 21.1, 39.1, 19.9, 32.1, 23.1, 21.8, 30.4, 19.62, 15.5);
> ks.test (v1, "pnorm", mean = mean(v1), sd = sd(v1))

        One-sample Kolmogorov-Smirnov test

data:  v1
D = 0.1604, p-value = 0.03266
alternative hypothesis: two-sided
```

Conclusão: com 95% temos α=5%. Como p-value < 0.05, rejeitamos a hipótese de normalidade

PPGI - UNIRIO

Testes de Normalidade: Teste de Shapiro-Wilk

- Usado para pequenos conjuntos de dados, onde valores extremos podem dificultar o uso de K-S

```
> v1 ← c(23.4, 30.9, 18.8, 23.0, 21.4, 1, 24.6, 23.8, 24.1, 18.7, 16.3, 20.3, 14.9, 35.4, 21.6, 21.2,
         21.0, 15.0, 15.6, 24.0, 34.6, 40.9, 30.7, 24.5, 16.6, 1, 21.7, 1, 23.6, 1, 25.7, 19.3, 46.9,
         23.3, 21.8, 33.3, 24.9, 24.4, 1, 19.8, 17.2, 21.5, 25.5, 23.3, 18.6, 22.0, 29.8, 33.3, 1,
         21.3, 18.6, 26.8, 19.4, 21.1, 21.2, 20.5, 19.8, 26.3, 39.3, 21.4, 22.6, 1, 35.3, 7.0, 19.3,
         21.3, 10.1, 20.2, 1, 36.2, 16.7, 21.1, 39.1, 19.9, 32.1, 23.1, 21.8, 30.4, 19.62, 15.5);
> shapiro.test (v1)

        Shapiro-Wilk normality test

data:  v1
W = 0.923, p-value = 0.0001315
```

Conclusão: como p-value < 0.05, rejeitamos a hipótese de normalidade

Testes de Normalidade: Teste de Shapiro-Wilk

```
> x <- rnorm(1000, 5, 3)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9983, p-value = 0.412

> y <- runif(1000, 2, 4)
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.9495, p-value < 2.2e-16
```

p-value > α → NORMAL

p-value < α → NÃO NORMAL

Medidas de Dependência: Correlações em R

- Se houver muitos empates entre os rankings, devemos usar a correlação de *Kendall* (uma variante do índice de *Spearman*)

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> cor(xcorretos, ycorretos, method="spearman")
[1] -0.06392761

> cor.test(xcorretos, ycorretos, method="kendall")

        Kendall's rank correlation tau

data:  xcorretos and ycorretos
z = -0.235, p-value = 0.8142
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.05556508

Mensagens de aviso perdidas:
In cor.test.default(xcorretos, ycorretos, method = "kendall") :
  Impossível calcular o valor exato de p com empates
```


Medidas de Dependência: Correlações em R

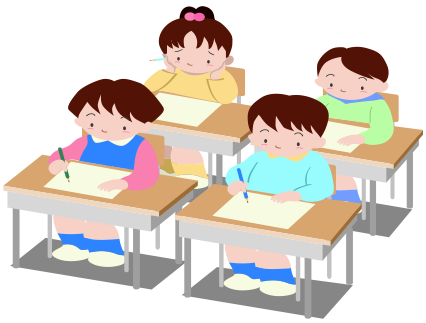
- Identificando o número de dados duplicados em uma sequência e calculando rankings ...

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> rank(ycorretos)
[1]  8.0  1.0 10.5  6.0  2.5  2.5 10.5  4.0  9.0  7.0  5.0

> dup <- duplicated(ycorretos)
> length(dup[dup == TRUE])
 [1] 2
```

Exercício 07

- Podemos dizer que os dados da tabela de precipitação de Nova York para o mês de janeiro têm distribuição normal?
- Quantos dados duplicados existem na tabela de precipitação de Nova York no mês de janeiro?
- Calcule a correlação entre as chuvas e a temperatura em Nova York no mês de janeiro.
- Calcule a matriz de correlação de precipitação para todos os meses do ano.
- Estes resultados são diferentes se considerarmos apenas do ano de 1950 até 2012?



TESTES DE HIPÓTESE COM DOIS GRUPOS

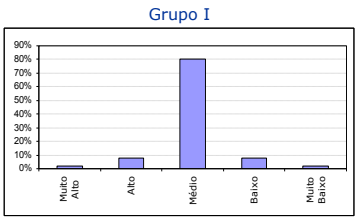
(Testes baseados em comparação)

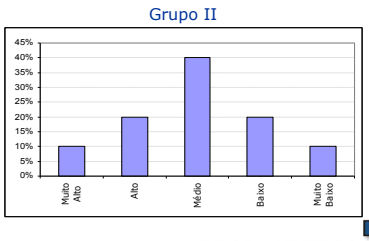
Tipos de Testes de Hipótese

- Os testes de hipótese se dividem em paramétricos e não paramétricos
- Testes paramétricos utilizam fórmulas fechadas, derivadas de propriedades de distribuições de frequência conhecidas (ex.: equação da curva normal)
  - » Estes testes têm maior potência que as alternativas não-paramétricas, mas exigem premissas sobre os dados que serão testados
  - » Normalidade: os valores se concentram simetricamente em torno de uma média e quanto maior a distância desta média, menor a freqüência das observações
  - » Homocedasticidade: implica em variância constante entre os conjuntos de dados testados, ou seja, a variância de um subgrupo não é maior que a de outro
- Testes não-paramétricos devem ser usados quando os dados coletados não atendem aos pressupostos esperados pelos testes paramétricos

Homocedasticidade

- Duas variáveis são homocedásticas se 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 e consome produtos similares. Uma pessoa rica pode consumir produtos mais simples, mas também pode consumir produtos sofisticados





Testes de Homocedasticidade

- O teste de *Levene* (pacote *lawstat* do R) recebe dois grupos de dados com o mesmo número de informações e verifica se eles têm a mesma variância

```
> library(lawstat)
> v1 ← c(23.4, 30.9, 18.8, 23.0, 21.4, 1, 24.6, 23.8, 24.1, 18.7, 16.3, 20.3, 14.9, 35.4, 21.6, 21.2,
         21.0, 15.0, 15.6, 24.0, 34.6, 40.9, 30.7, 24.5, 16.6, 1, 21.7, 1, 23.6, 1, 25.7, 19.3, 46.9,
         23.3, 21.8, 33.3, 24.9, 24.4, 1, 19.8, 17.2, 21.5, 25.5, 23.3, 18.6, 22.0, 29.8, 33.3, 1,
         21.3, 18.6, 26.8, 19.4, 21.1, 21.2, 20.5, 19.8, 26.3, 39.3, 21.4, 22.6, 1, 35.3, 7.0, 19.3,
         21.3, 10.1, 20.2, 1, 36.2, 16.7, 21.1, 39.1, 19.9, 32.1, 23.1, 21.8, 30.4, 19.62, 15.5);
> v2 ← c(16.5, 1, 22.6, 25.3, 23.7, 1, 23.3, 23.9, 16.2, 23.0, 21.6, 10.8, 12.2, 23.6, 10.1, 24.4,
         16.4, 11.7, 17.7, 34.3, 24.3, 18.7, 27.5, 25.8, 22.5, 14.2, 21.7, 1, 31.2, 13.8, 29.7, 23.1,
         26.1, 25.1, 23.4, 21.7, 24.4, 13.2, 22.1, 26.7, 22.7, 1, 18.2, 28.7, 29.1, 27.4, 22.3, 13.2,
         22.5, 25.0, 1, 6.6, 23.7, 23.5, 17.3, 24.6, 27.8, 29.7, 25.3, 19.9, 18.2, 26.2, 20.4, 23.3,
         26.7, 26.0, 1, 25.1, 33.1, 35.0, 25.3, 23.6, 23.2, 20.2, 24.7, 22.6, 39.1, 26.5, 22.7, 10.0);
> levene.test (v1, v2)

 modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median

data:  v1
Test Statistic = 1.1205, p-value = 0.4032
```

p-value > α, aceitamos a hipótese de homocedasticidade

Teste T ou Student-T

- Teste paramétrico para comparar médias de duas amostras independentes
- Restrições nos dados submetidos ao teste
  - » As séries comparadas devem ter distribuição próxima a normal
  - » As séries comparadas devem ter variância similar (homocedasticidade)
  - » As séries devem ter sido coletadas de forma independente
- Existe uma versão do teste T para dados pareados (não independentes)
  - » Dizemos que duas amostras são pareadas quando existe uma relação única entre um valor em uma amostra e um valor na segunda amostra
  - » Por exemplo, uma amostra antes e outra amostra após um treinamento

Teste T ou Student-T

```
> ytempo <- c(10, 13, 12, 13, 10, 14, 14, 13, 14, 14, 13);
> xtempo <- c(13, 9, 11, 14, 9, 12, 9, 12, 11, 14, 13);
> t.test(ytempo, xtempo, var.equal=TRUE);

        Two Sample t-test

data:  ytempo and xtempo
t = 1.6149, df = 20, p-value = 0.122
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3446983  2.7083347
sample estimates:
mean of x mean of y
 12.72727  11.54545
```

Conclusão: não é possível afirmar com 95% de certeza que existe diferença no tempo de realização da atividade de acordo com os tratamentos X e Y (*p-value* > 0.05)

Teste T ou Student-T

- O teste de Welch é uma variante do teste T que introduz alguma liberdade em relação às variâncias

```
> ytempo <- c(10, 13, 12, 13, 10, 14, 14, 13, 14, 14, 13);
> xtempo <- c(13, 9, 11, 14, 9, 12, 9, 12, 11, 14, 13);
> t.test(ytempo, xtempo);

        Welch Two Sample t-test

data:  ytempo and xtempo
t = 1.6149, df = 18.857, p-value = 0.1229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3506866  2.7143230
sample estimates:
mean of x mean of y
 12.72727  11.54545
```

Conclusão: não é possível afirmar com 95% de certeza que existe diferença no tempo de realização da atividade de acordo com os tratamentos X e Y (*p-value* > 0.05)

Teste T ou Student-T

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> t.test(ycorretos, xcorretos);

        Welch Two Sample t-test

data:  ycorretos and xcorretos
t = -4.1749, df = 19.976, p-value = 0.0004684
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.14996779 -0.05003221
sample estimates:
mean of x mean of y
0.7963636 0.8963636
```

Conclusão: é possível afirmar, com pelo menos 95% de certeza, que existe diferença no número de requisitos corretos encontrados pelos participantes (p-value < 0.05)

Teste T ou Student-T

- Teste T comparando uma série a um valor constante

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> t.test(ycorretos, um=0.70);

        One Sample t-test

data:  ycorretos
t = 47.861, df = 10, p-value = 3.825e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.7592894 0.8334379
sample estimates:
mean of x
0.7963636
```

Teste T ou Student-T

- Medida de tamanho de efeito (Cohen's d)
  - » O cálculo do s não é claramente definido, podendo ser calculado pelas três equações abaixo
  - » Na prática, isto não faz muita diferença porque os desvios padrão devem ser aproximadamente iguais (homocedasticidade)

$$d = \frac{|\mu_1 - \mu_2|}{s}$$

$$\begin{cases} s = \sqrt{\dfrac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}} \\ s = \sigma_1 \\ s = \sigma_2 \end{cases}$$

Teste T ou Student-T

- Um problema com o Cohen's d é a falta de interpretação física ...

```
> cohenD <- function(data1, data2)
+ {
+         n1 <- length(data1);
+         n2 <- length(data2);
+         mean1 <- mean(data1);
+         mean2 <- mean(data2);
+         sd1 <- sd(data1);
+         sd2 <- sd(data2);
+         s <- sqrt(((n1 - 1) * sd1 * sd1 + (n2 - 1) * sd2 * sd2) / (n1 + n2 - 2));
+         abs(mean1 - mean2) / s;
+ }
>
> mean(ycorretos);
[1] 0.7963636

> mean(xcorretos);
[1] 0.8963636

> cohenD(ycorretos, xcorretos);
[1] 1.780201
```

Teste T ou Student-T

- Interpretação do tamanho de efeito

	Pequeno	Médio	Grande
d	> 0.20	> 0.50	> 0.80

Fonte: Cohen, J. (1992), "A Power Primer", Psychological Bulletin, 112, pp. 155-159

- Como apresentar o resultado?

> Aplicando um teste de Welch observamos diferenças significativas no número de requisitos encontrados por participantes aplicando as duas técnicas com tamanho de efeito grande (p-value = 0.0004, Cohen's d = 1.78) favorecendo a técnica X.

Teste T ou Student-T: Cálculo da Amostra

- O tamanho de efeito está relacionado com a potência do teste
  - » A potência é a probabilidade do teste rejeitar a hipótese nula se esta for falsa, ou seja, a probabilidade de decisão correta baseada na hipótese alternativa
  - » A potência do teste T depende de três fatores: a diferença entre as médias, a variância dos resíduos e o número de observações
  - » Ela também pode ser calculada dados o nível de significância (5% na tabela abaixo), o número de observações e o tamanho de efeito (Cohen's d)
  - » O número de observações é dado por grupo, ou seja, o número total de observações deve ser o dobro
  - » Se um pesquisador fizer um estudo com 20 pessoas em cada grupo e esperar tamanho de efeito médio, as chances de rejeitar a hipótese nula são de 33%
  - » Ou seja, para cada 3 vezes que o experimento for feito, a hipótese nula será rejeitada 1 vez!!!

	Effect Size (Cohen's *d*)		
n	.2	.5	.8
10	7	18	39
20	9	33	69
40	14	60	94
80	24	88	99
100	29	94	99
200	51	99	99

Fonte: Cohen, J. (1977)

Teste T ou Student-T: Cálculo da Amostra

- A tabela abaixo pode ser usada para calcular o número de observações desejadas no estudo
  - » A tabela assume α = 0.05 e indica o número de observações por grupo para um determinado tamanho de efeito
  - » O pesquisador estima (a.k.a. chuta) o tamanho de efeito que espera observar, indica a potência do teste e calcula o número de pessoas que deve envolver
  - » Por exemplo, um estudo com tamanho de efeito moderado deveria contar com 64 pessoas. Abaixo disso, a potência do teste será muito baixa ...

	Effect Size (Cohen's *d*)		
Power	.2	.5	.8
.25	84	14	6
.50	193	32	13
.60	246	40	16
.70	310	50	20
.80	393	64	26
.90	526	85	34
.95	651	105	42
.99	920	148	58

Fonte: Cohen, J. (1977)

Teste de Mann-Whitney (Wilcoxon)

- Alternativa não paramétrica para o teste T
  - » Requer que as amostras sejam independentes, com dados 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 publicados em livros de Estatística

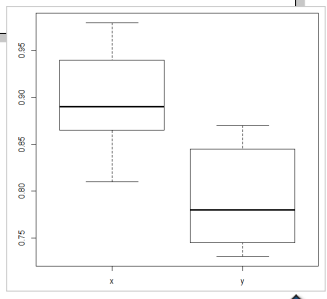
Teste de Mann-Whitney (Wilcoxon)

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> wilcox.test(ycorretos, xcorretos);

 Wilcoxon rank sum test with continuity correction

data:  xcorretos and ycorretos
W = 108, p-value = 0.001986
alternative hypothesis: true location shift is not equal to 0
```

Conclusão: é possível afirmar, com pelo menos 95% de certeza, que existe diferença no percentual de requisitos corretos encontrados pelos participantes (p-value < 0.05)



Teste de Mann-Whitney (Wilcoxon)

- Comparando uma série com um valor constante ...

```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> wilcox.test(ycorretos, um=0.75);

        Wilcoxon signed rank test with continuity correction

data:  ycorretos
V = 66, p-value = 0.003822
alternative hypothesis: true location is not equal to 0

Warning message:
In wilcox.test.default(ycorretos, um = 0.75) :
  cannot compute exact p-value with ties
```

Teste de Mann-Whitney (Wilcoxon)

- Medida de tamanho de efeito (Vargha & Delaney's $\hat{A}_{12}$)
  - » O tamanho de efeito é calculado a partir dos rankings dos valores observados
  - » O valor varia de 0% a 100%
  - » Sua interpretação física é muito simples: representa o número de vezes em que a série A foi maior do que a série B
  - » $\hat{A}_{12}$ = 50% indica que os dois tratamentos têm chances iguais de gerar o melhor resultado
  - » $\hat{A}_{12}$ = 80% indica que o primeiro tratamento será melhor que o segundo em 80% das vezes que for aplicado

$$\hat{A}_{12} = \frac{\frac{R_1}{n_1} - \frac{n_1 + 1}{2}}{n_2}$$

onde R1 é o somatório dos rankings da primeira série sob análise quando se considera a ordem dos dados disponíveis nas duas séries

Teste de Mann-Whitney (Wilcoxon)

- Interpretação do tamanho de efeito

	Pequeno	Médio	Grande
d	< 0.60	< 0.75	> 0.75

- Como apresentar o resultado?

> Aplicando um teste de Wilcoxon observamos diferenças significativas no número de requisitos encontrados por participantes aplicando as duas técnicas com tamanho de efeito grande (p-value = 0.0019, $\hat{A}_{12}$ = 0.89) favorecendo a técnica X.

Teste de Mann-Whitney (Wilcoxon)

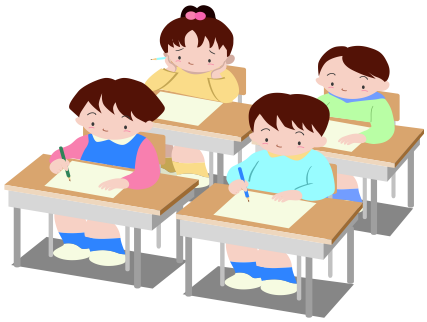
```
> ycorretos <- c(0.83, 0.73, 0.87, 0.78, 0.74, 0.74, 0.87, 0.75, 0.86, 0.82, 0.77);
> xcorretos <- c(0.90, 0.89, 0.88, 0.87, 0.97, 0.81, 0.82, 0.86, 0.92, 0.96, 0.98);
> ax <- vargha.delaney(xcorretos, ycorretos);
> ay <- vargha.delaney(ycorretos, xcorretos);
> ax; ay;

[1] 0.892562
[1] 0.1074380
```

Conclusão: o método X gerará melhores resultados em 89% das aplicações, enquanto o método Y gerará melhores resultados em 11% das aplicações

Exercício 08

- Considerando os dados sobre clusterização de projetos de software ...
- Verifique se os dados de *hypervolume* usando os tipos *mq* e *none* possuem distribuição normal para a instância *javacc*
- Verifique se os dados de *hypervolume* usando os tipos *mq* e *none* são homocedásticos para a instância *javacc*
- Compare as médias entre as duas série acima usando um teste T ou teste de Mann-Whitney, conforme as distribuições dos dados
- Calcule o tamanho de efeito do teste descrito no item anterior



TESTES DE HIPÓTESE COM MAIS DE DOIS GRUPOS

(Testes baseados em comparação)

Teste ANOVA

- Teste de hipótese que avalia a existência de diferença significativa entre as médias observadas entre dois ou mais grupos
  - » Permite comparar as médias de diversos tratamentos, sendo usada como uma extensão dos testes T quando existem mais de dois tratamentos
  - » Avalia se a variabilidade dentro dos grupos é maior do que a existente entre os grupos
  - » A técnica supõe independência das observações, normalidade e igualdade entre as variâncias dos grupos
  - » A técnica somente deve ser usada se os grupos possuírem aproximadamente o mesmo número de participantes

Teste ANOVA

- Porque não utilizar um conjunto de Testes T?
  - » Considere que você têm três técnicas de levantamento de requisitos (A, B e C) e quer saber se existe diferença de desempenho (ex.: número de requisitos corretos encontrados) entre elas
  - » A hipótese nula do teste ANOVA é que **não existe diferença significativa entre os tratamentos**
  - » Para aplicar testes T, teríamos que fazer isto em pares, ou seja, aplicar um teste entre A e B, outro entre B e C e um terceiro entre A e C
  - » Se cada teste tiver $\alpha = 0.05$, teremos um α final igual a 0.142 ($1.0 - 0.95^3$), que é baixo demais para representar um teste de qualidade
  - » Neste caso, o teste ANOVA manterá $\alpha = 0.05$

Teste ANOVA

- Mas existe alternativa?
  - » Sempre que uma série de dados for comparada com mais de uma série de dados, devemos usar um teste apropriado (ANOVA) ou aplicar testes de pares com correções
  - » Um exemplo é a correção de Bonferroni, que divide o α do teste pelo número de vezes em que a série é avaliada (número de testes)
  - » No entanto, a correção de Bonferroni é muito severa quando fazemos mais de 3 comparações com a mesma série
  - » Existem correções mais suaves (ex.: Holm), mas na prática o ideal é partir para um teste que considere mais do que dois grupos
  - » De qualquer forma, as correções vão ser aplicadas depois que o resultado do teste (múltiplos grupos) for conhecido (*post-hoc analysis*)

Teste One-Way ANOVA

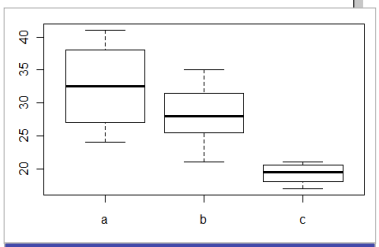
- Teste realizado quando temos uma variável independente com diversos tratamentos. O teste determina se os tratamentos são importantes para predizer o valor da variável dependente usada na comparação.

```
> data = read.table("http://personality-project.org/r/datasets/R.appendix1.data", header=T);
> data

   Dosage Alertness
1       a        30
...
6       a        24
7       b        32
...
14      b        25
15      c        17
...
18      c        19

> boxplot(Alertness~Dosage,data=data)
> result <- aov(Alertness~Dosage,data=data)
> summary(result)

            Df  Sum Sq  Mean Sq  F value    Pr(>F)
Dosage       2  426.25   213.12   8.7887  0.002977 **
Residuals   15  363.75    24.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Conclusão: é possível afirmar,que existe diferença na média de atenção com as diferentes dosagens (p-value < 0.05)

Teste One-Way ANOVA: Tamanho de Efeito

- O tamanho de efeito de um teste *one-way* ANOVA é o eta-quadrado

$$\eta^2 = \frac{SumSq_{indep}}{SumSq_{indep} + SumSq_{residuals}}$$

	Pequeno	Médio	Grande
η^2	> 0.01	> 0.06	> 0.14

Aplicando um teste one-way ANOVA observamos diferenças significativas na atenção dos pacientes de acordo com a dosagem aplicada, com tamanho de efeito grande (p-value = 0.00298, η^2 = 0.54).

Teste *Two-Way* ANOVA: Introdução

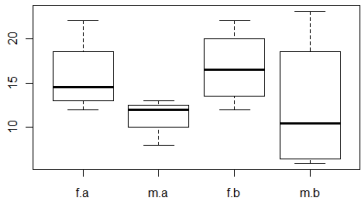
- Grosso modo, usado quando há duas variáveis independentes

```
> data = read.table("http://personality-project.org/r/datasets/R.appendix2.data", header=T);
> data

   Observation Gender Dosage Alertness
1            1      m      a         8
2            2      m      a        12
3            3      m      a        13
4            4      m      a        12
5            5      m      b         6
6            6      m      b         7
7            7      m      b        23
8            8      m      b        14
9            9      f      a        15
10          10      f      a        12
11          11      f      a        22
12          12      f      a        14
13          13      f      b        15
14          14      f      b        12
15          15      f      b        18
16          16      f      b        22

> boxplot(Alertness~Gender*Dosage, data=data)
> result <- aov(Alertness~Gender*Dosage,data=data)
> summary(result)

             Df  Sum Sq Mean Sq F value Pr(>F)
Gender        1  76.562  76.562  2.9518 0.1115
Dosage        1   5.062   5.062  0.1952 0.6665
Gender:Dosage 1   0.063   0.063  0.0024 0.9617
Residuals    12 311.250  25.938
```



Conclusão: não é possível afirmar que existe diferença na média de atenção com as diferentes dosagens, gêneros e sua combinação (p-value > 0.05)

Teste *Two-Way* ANOVA: Introdução

- O cálculo do tamanho de efeito é similar ao usado no one-way ANOVA
- Porque uma introdução?
  - » Porque existe muito mais complicação por trás dos testes ANOVA ...
  - » E porque eles raramente são aplicados em dados da Engenharia de Software
- Existe um three-way ANOVA?
  - » Matematicamente sim, mas os experimentos começam a ficar muito complexos e o número de participantes começa a ficar impraticável
  - » É melhor tentar isolar os efeitos que estão sendo analisados e se limitar a *one-way* ou *two-way* ANOVA

Teste de Kruskal-Wallis

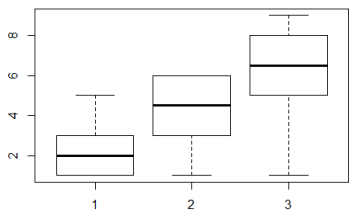
- É o equivalente não-paramétrico do teste ANOVA
  - » Não presume normalidade e homocedasticidade
  - » Permite comparar dados em escala ordinal (além de razão e intervalar)

```
> value <- c(1, 2, 5, 3, 2, 1, 1, 3, 2, 1, 4, 3, 6, 5, 2, 6, 1, 6, 5, 4, 9, 6, 7, 7, 5, 1, 8, 9, 6, 5);
> group <- factor(c(rep(1, 10), rep(2, 10), rep(3, 10));
> data <- data.frame(group, value);
> kruskal.test(value ~group, data)

        Kruskal-Wallis rank sum test

data:  value by group
Kruskal-Wallis chi-squared = 13.6754, df = 2, p-value = 0.001073
```

p-value < 0.05 ➔ observamos um efeito significativo do grupo no valor



Teste de Kruskal-Wallis

- Agora que observamos um efeito, temos que identificar onde ele está
  - » Para identificar este efeito, fazemos um teste chamado *post-hoc*
  - » O teste consiste na avaliação de pares de comparações usando o teste de Mann-Whitney com correção (Bonferroni ou Holm)

```
> pairwise.wilcox.test(value, group, p.adj="bonferroni", exact=FALSE);

        Pairwise comparisons using Wilcoxon rank sum test

data:  value and group

  1      2
2 0.0418 -
3 0.0058 0.0791

P value adjustment method: bonferroni
```

p-value < 0.05 ➔ observamos um efeito significativo entre os tratamentos 1 e 2 e 1 e 3

Teste de Kruskal-Wallis

```
> g1 <- subset(data, data$group == 1)
> g2 <- subset(data, data$group == 2)
> g3 <- subset(data, data$group == 3)
>
> vargha.delaney(g1$value, g2$value)
[1] 0.175

> vargha.delaney(g1$value, g3$value)
[1] 0.09

> vargha.delaney(g2$value, g3$value)
[1] 0.205
```

Um teste de Kruskal-Wallis identificou diferenças significativas do grupo sobre o valor (p-value = 0.001073). Um teste post-hoc usando testes de Mann-Whitney com correção de Bonferroni mostrou que as diferenças significativas ocorriam entre os grupos 1 e 2 (p-value = 0.0418, $\hat{A}_{12}$=17.5%) favorável ao grupo 2 e entre os grupos 1 e 3 (p-value = 0.0058, $\hat{A}_{12}$=9%) favorável ao grupo 3.

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, 2006
- Dyba, T.; Kampenes, V.; Sjoberg, D., “A Systematic Review of Statistical Power in Software Engineering Experiments”, Information and Software Technology, 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., “Design and Analysis of Experiments”, Ed. IE-Wiley, 2000.
- National Institute of Standards and Technology - http://www.nist.gov
- Statistical Methods for HCI Research - http://yatani.jp/teaching/doku.php?id=hcistats:start
- Pfleeger, Shari .L., Albert Einstein and Empirical Software Engineering. IEEE Computer: 32-37, 1999.
- Tichy, W. F., Should Computer Scientists Experiment More?, IEEE Computer: 32-40, May, 1998.
- 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.