R Commander ANOVA

Antes, este tutorial ANOVA fora elaborado para Minitab, depois transcrito, adaptado e ampliado para R Commander pelo Prof. Paulo Silveira.

O Minitab é um software comercial para uso em ensino e pesquisa, que fora adotado na disciplina de Métodos Quantitativos em Medicina ao longo de uma década no início deste milênio com sucesso, mas depende de licença paga pela USP, o que restringe seu acesso posterior aos estudantes e pesquisadores. Pensando em provê-los com uma alternativa gratuita, optou-se a partir do ano de 2014 por experimentar com a linguagem de programação R, que é muito utilizada em procedimentos estatísticos por pesquisadores do mundo todo.

Table of Contents

1 Resumo

A análise de variância é um tópico extenso e bastante complexo da estatística. Aqui introduziremos as formas mais simples de comparação de três ou mais médias com somente um fator de comparação (One-way ANOVA e Kruskal-Wallis). Adicionalmene, abordaremos noções sobre a análise de variância de dois fatores e suas diversas possibilidades, por meio de alguns exemplos práticos.

2 Objetivos

Ao final deste tutorial você deve ser capaz de:

- Testar a homogeneidade de variâncias e interpretar os resultados;
- Executar o teste de ANOVA One-way.
- Executar o teste de Kruskal-Wallis.
- Identificar diversos problemas de análise de variância de dois fatores.
- Codificar os dados com dois fatores de comparação.
- Proceder a análise de variância de dois fatores.

3 Análise de Variância: One Way

Para executar a análise de variância pode-se entrar ou estruturar os dados de duas maneiras:

- em colunas em que o conjunto de dados para cada situação experimental ou fator está organizado em colunas, como neste exemplo; e
- em duas colunas de resposta e fator, em que os dados de medida estão numa coluna denominada resposta e a situação experimental que os gerou identificados na coluna fator.

O arquivo que você deve baixar para este tutorial, rcmdr_anova_exemplo.txt, está na primeira forma de estruturação.

Abra o RCmdr e importe os dados em um **dataset** chamado **Exemplo**. Este é um exemplo hipotético, cujo intuito é mostrar o procedimento de cálculo de ANOVA por meio do RCmdr. Dê uma olhada no aspecto dos dados. Encontrará três colunas (três grupos) denominadas **X1**, **X2** e **X3**. Olhando as linhas finais verá que os três grupos têm tamanhos desiguais, aparecendo NA (not available) para completar os espaços vazios.

3.1 Reorganizando os dados

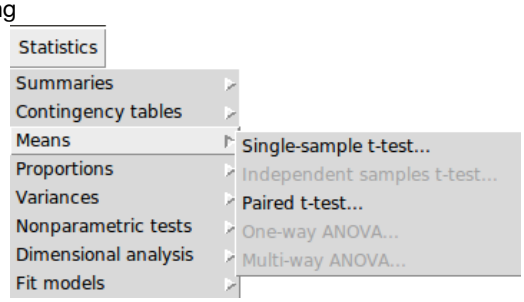
ANOVA One-way é um teste estatístico paramétrico para comparar médias de 3 ou mais grupos simultaneamente. Está, portanto, localizado no RCmdr onde você poderia esperar:

Estatísticas -> Médias -> ANOVA para um fator (one way)...

Statistics -> Means -> One-way ANOVA...

Para sua surpresa, encontrará o teste desativado:

MPT0164 ANOVA grayed.png



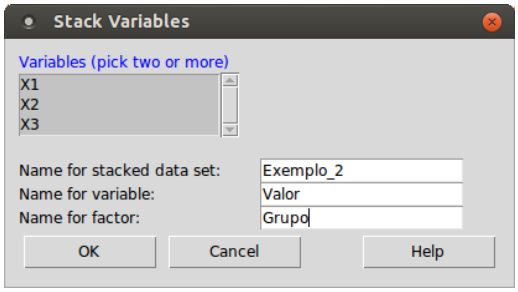
Isto acontece porque o RCmdr sempre assume que cada linha corresponde a um indivíduo, e cada coluna a uma variável diferente. Em nossos dados é o arranjo que não temos. Há três grupos independentes, **X1**, **X2** e **X3**, de forma que os dados de uma mesma linha nada têm a ver uns com os outros; no corpo da planilha estão os diferentes valores de uma mesma variável medidos nos indivíduos de cada um dos grupos.

Nossa primeira providência é rearranjar os dados para que o pacote estatístico possa operar: vamos empilhar (stack) os dados.

Dados -> Conjunto de dados ativos -> Stack variables no conjunto de dados ativos...

Data -> Active data set -> Stack variables in active data set...

Observe o preenchimento do quadro:



- **Exemplo_2** é o nome do conjunto de dados que será criado, empilhando os dados a partir de Exemplo.
- Valor é a variável numérica que, em Exemplo, está espalhada nas colunas X1, X2 e X3.
- Grupo é o fator, que vai herdar um dos três nomes de grupo X1, X2 ou X3.
- Note, por fim, que as três colunas X# foram selecionadas.

Clique OK e observe como ficou o novo conjunto de dados, Exemplo_2. Os valores numéricos foram empilhados na coluna (variável) chamada Valor, e os nomes originais das colunas tornaram-se os identificadores do Grupo de origem dos dados:

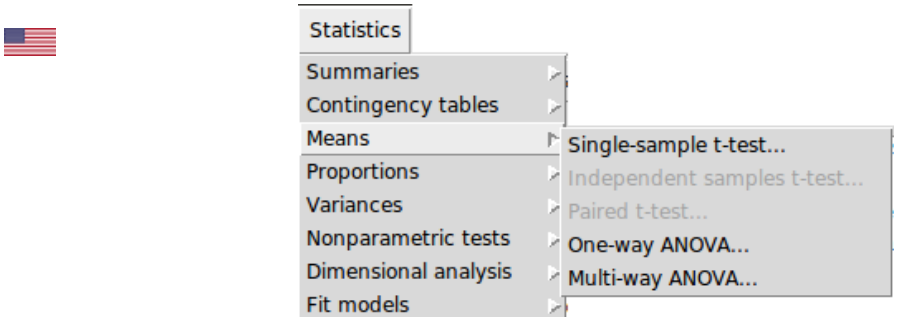
Exemplo_2

	Valor	Grupo
1	20.174215	X1
2	14.528113	X1
3	16.183813	X1
4	13.634431	X1
5	16.227159	X1
6	14.220350	X1
7	20.356473	X1
8	13.605937	X1
9	12.731168	X1
10	12.547583	X1
11	13.806271	X1
12	16.677570	X1
13	11.764149	X1
14	17.443526	X1
15	14.304031	X1
16	10.691904	X1
44	NA	X1
45	NA	X1
46	NA	X1
47	16.214131	X2
48	16.008662	X2
49	15.293592	X2
50	8.826618	X2
51	14.140516	X2
52	14.001721	X2
53	15.642451	X2
54	15.192455	X2
55	13.658362	X2
56	14.168974	X2
90	NA	X2
91	NA	X2
92	NA	X2
93	17.846552	X3
94	15.078566	X3
95	15.705403	X3
96	14.125957	X3
97	15.180095	X3
98	14.204873	X3
99	11.059677	X3
100	12.625985	X3
101	13.340760	X3
102	11.049017	X3
103	11.253983	X3

Agora atente para o que ocorre com:

Estatísticas -> Médias -> ANOVA para um fator (one way...

Statistics -> Means -> One-way ANOVA...



Não execute o teste. Ainda não estamos prontos para ele!

3.2 Um pouco de estatística descritiva

É um bom costume, antes de começar qualquer análise estatística, conhecer os dados: saber os nomes das variáveis, classificá-las (qualitativas nominais ou ordinais, quantitativas discretas ou contínuas), verificar suas medidas de tendência central (média ou mediana) e de dispersão (desvio padrão ou quartis), fazer alguns gráficos exploratórios (como histogramas ou boxplot).

Neste exemplo simplificado temos apenas duas variáveis:

1. **Grupo**: qualitativanominal, usada para denominar os grupos em estudo;
2. **Valor**: quantitativa contínua, uma medida hipotética.

Você deve se perguntar quais os valores que essas variáveis assumem? Como se distribuem? Que valores existem? Há valores espúrios por algum erro em meus dados?

3.2.1 Tendência central e dispersão dos dados

Por exemplo, explore:

Estatísticas -> Resumos -> Resumos numéricos...

Statistics -> Summaries -> Numerical summaries...

Você poderá ver a média, desvio-padrão, mediana e quartis de todos os indivíduos:

```
     mean       sd       0%      25%      50%     75%     100%   n NA
 13.66731 2.442804 6.936922 11.75338 13.63443 15.1801 20.35647 117 21
```

Pode ver as mesmas medidas dos grupos...

```
        mean       sd        0%      25%      50%      75%     100% data:
n
X1 15.12733 2.391789 10.691904 13.62731 14.80856 16.74152 20.35647     4
0
X2 13.60447 1.964106  8.826618 12.32428 13.65836 15.02118 17.11983     3
1
X3 12.44007 2.101316  6.936922 11.10825 11.97699 14.03726 17.84655     4
6
   data:NA
X1       6
X2      15
X3       0
```

... o que, aliás, mostra-lhe que existem três grupos, X1, X2, X3.

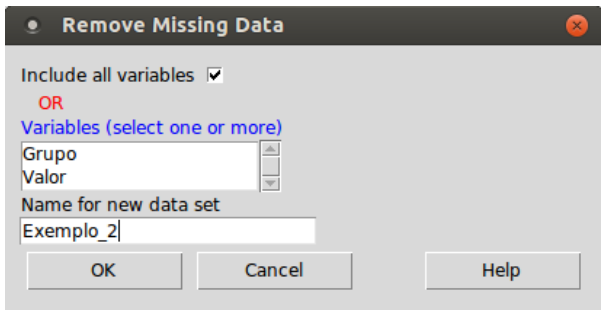
> Neste exemplo, você já sabia que eram três grupos. No entanto, caso recebesse um banco de dados desconhecido, precisaria verificar. Imagine que existisse uma variável com variáveis como sexo (masculino e feminino) ou diabete (sim ou não). Fazer a análise descritiva garante que não há erros de digitação ou informações faltantes. Poderia ser que aparecessem outras variantes de codificação de sexo, como Masculino, M, Masc., etc... é preciso arrumar os dados antes de analisá-los.

Lembre-se, no entanto, que os grupos tinham tamanhos diferentes e campos preenchidos com NA para completar? O stacking, empilhamento, acabou produzindo-os. Podemos nos livrar deles. Experimente:

Dados -> Conjunto de dados ativos -> Remover observações com dados faltantes...

Data -> Active data set -> Remove cases with missing data...

Preencha o quadro selecionando todas as variáveis:



Note que não criaremos um novo conjunto de dados: ao repetir o nome Exemplo_2, substituiremos o data set corrente. Se você prestar atenção, o Exemplo_2 tinha 138 linhas. A estatística descritiva tinha indicado $6 + 15 = 21$ ocorrências de NA. Depois de removermos os "missing data", Exemplo passou a ter 117 linhas. Foram as 21 ocorrências que sumiram. Repita:

Estatísticas -> Resumos -> Resumos numéricos...

Statistics -> Summaries -> Numerical summaries...

verificando todos juntos e separando por grupos. Veja que os NA sumiram...

```
     mean       sd       0%      25%      50%      75%     100%   n
 13.66731 2.442804 6.936922 11.75338 13.63443 15.1801 20.35647 117
       mean       sd       0%      25%      50%      75%     100% data:
n
X1 15.12733 2.391789 10.691904 13.62731 14.80856 16.74152 20.35647     4
0
X2 13.60447 1.964106  8.826618 12.32428 13.65836 15.02118 17.11983     3
1
X3 12.44007 2.101316  6.936922 11.10825 11.97699 14.03726 17.84655     4
6
```

3.2.1.1 Uma olhadinha gráfica

Gerar gráficos é útil para termos uma noção de como são nossos dados.

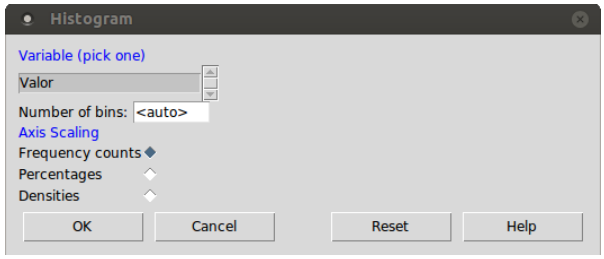
3.2.1.2 Histograma

No caso, como temos uma variável quantitativa contínua, podemos fazer histogramas. Experimente:

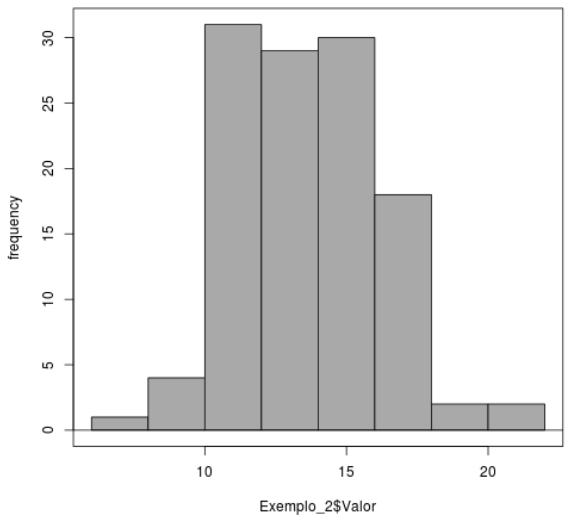
Gráficos -> Histograma...

Graphs -> Histogram...

Na versão do RCmdr 1.8-1 a tela para o histograma tem o seguinte aspecto:



Há opções para explorar, permitindo mostrar as colunas em números absolutos, percentagem e densidade. O aspecto do histograma é algo como:



Este histograma corresponde à média geral de 13.66731 e desvio padrão de 2.442804 que obtivemos quando não separamos os grupos. Pelo menos na versão do RCmdr que usei aqui, ainda não implementaram a possibilidade de separar os subgrupos. É possível, usando sintaxe similar à soma condicional que mostramos quando calculamos ranks nos testes não paramétricos

Repare, na Script Window que o histograma é construído por:

```
Hist(Exemplo_2$Valor, scale="frequency", breaks="Sturges", col="darkgra
y")
```

Modifique a linha para:

```
Hist(Exemplo_2$Valor[Exemplo_2$Grupo=="X1"], scale="frequency", breaks="
Sturges", col="darkgray")
```

para construir o histograma do grupo X1. A expressão [Exemplo_2$Grupo=="X1"] indica é condição, em instrui o R a usar do conjunto de dados Exemplo_1, a variável Grupo, nas linhas em que seu conteúdo é igual (o sinal de igual aparece duas vezes: é um operador de comparação) a texto X1 (entre aspas, porque é uma string).

Similarmente,

Hist(Exemplo_2$Valor[Exemplo_2$Grupo=="X2"], scale="frequency", breaks="Sturges", col="darkgray")
Hist(Exemplo_2$Valor[Exemplo_2$Grupo=="X3"], scale="frequency", breaks="Sturges", col="darkgray") ```
fornece os histogramas dos outros dois grupos.

> Olhe os histogramas: parecem seguir distribuições normais?

Na versão do R Commander disponível há a opção de gerar o histograma por grupos, que gera gráficos diferentes daqueles gerados através da linha de comando do tutorial. Por quê?

Drawing

> Dica: dêem uma olhada em [Aulas_práticas/R_Commander/Estatística_Descritiva_e_Probabilidade#Número_de_categorias](http://sislau.fm.usp.br/wiki/MPT0164/Aulas_pr%C3%A1ticas/R_Commander/Estat%C3%ADstica_Desc

3.2.1.3 Boxplot

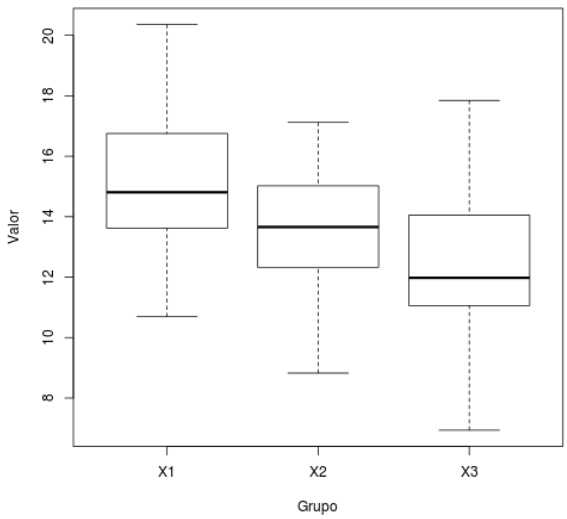
Uma maneira prática e adequada para dados organizados desta forma é usar

Gráficos -> Boxplot...

Graphs -> Boxplot...

A variável **Valor**, que é a única quantitativa, já aparece selecionada. Clique em [??? (Plot by groups)] e escolha a variável Grupo.

Deve obter:



Sabe interpretar um boxplot? Se não, reveja!

Condições para executar One-way ANOVA

Homoscedasticidade

Vamos verificar se existe homogeneidade das variâncias dos dados em análise, pois **ANOVA só pode ser usada se as variâncias das medidas dos grupos forem similares entre si, além de a amostra provir de uma população com distribuição normal para a variável em estudo**.

Como sempre, devemos estabelecer as hipóteses nula e alternativa:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$$

> Não há uma forma boa de escrever a hipótese alternativa. Caso rejeitemos a hipótese nula em favor da alternativa, significa que assumiremos que pelo menos uma das variâncias destoa das demais, e não que todas são diferentes entre si.

Adotaremos nível de significância de 5%.

> Para aqueles que se lembram da Aula de Teste t, é bom ressaltar que existe uma rotina distinta para o teste de homogeneidade das variâncias aplicada a ANOVA, pois neste caso desejamos comparar

mais que duas variâncias simultaneamente. Observe porém que a motivação de testagem de igualdade de variância no teste t não-pareado é distinta. No teste t, o resultado da testagem de variâncias determina a expressão do erro-padrão da média das diferenças, pois se as variâncias forem iguais é necessário o cálculo da variância conjugada (S_0^2, vide aula teórica sobre teste t), mas o teste t continua aplicável. No caso presente, caso rejeitemos a hipótese nula, o teste ANOVA deixa de ser indicado, e precisaremos aplicar seu equivalente não paramétrico, teste de Kruskal-Wallis, que veremos adiante.

Para realizar o teste de homogeneidade de variâncias, execute:

 Estatísticas -> Variância -> Teste de Barlett...

 Statistics -> Variances -> Bartlett's test...

Não há opções, e como temos apenas uma variável qualitativa nominal (Grupo) e uma quantitativa contínua (Valor), o Rcmdr já as mostra. Clique Ok para obter:

```
X1       X2       X3
5.720654 3.857712 4.415528
...
    Bartlett test of homogeneity of variances

data:  Valor by Grupo
Bartlett's K-squared = 1.4215, df = 2, p-value = 0.4913
```

Com $p = 0,4913$, para $\alpha = 0,05$, aceitamos a hipótese nula. Concluímos que as variâncias dos dados dos grupos X1, X2 e X3 são todas similares entre si e, portanto, podemos aplicar o teste de ANOVA para testar a igualdade das médias dos três grupos estudados.

Caso tivéssemos rejeitado H_0 concluiríamos que pelo menos uma das variâncias (σ_1^2, σ_2^2 ou σ_3^2) difere das demais e não poderíamos aplicar o teste de ANOVA. Optaríamos pelo seu equivalente, não paramétrico, [teste de Kruskal-Wallis], descrito adiante.

3.3.2 Distribuição normal dos dados dos grupos amostrais

É menos essencial, especialmente com amostras de maior tamanho e com variáveis quantitativas contínuas, e também com alguma controvérsia sobre a melhor forma de se avaliar, mas espera-se que as distribuições de dados em cada um dos grupos amostrais em estudo sejam bem ajustáveis por uma distribuição normal.

Será que no exemplo corrente os três grupos, X1, X2 e X3, têm seus dados distribuídos de acordo com a normal?

No RCmdr existe um teste para normalidade disponível:

 Estatísticas -> Resumos -> Teste de normalidade de Shapiro-Wilk...

 Statistics -> Summaries -> Shapiro-Wilk test of normality...

Note que já aparece selecionada a única variável quantitativa disponível em Exemplo_2. A hipótese nula é de que a distribuição dos dados não difere de uma distribuição normal e podemos adotar $\alpha = 5\%$; obteremos:

```
    Shapiro-Wilk normality test

data:  Exemplo_2$Valor
W = 0.9898, p-value = 0.5327
```

Aceitaríamos a hipótese nula, mas **CUIDADO !** Que dados foram avaliados? Uma mistura de tudo junto,

X1, X2 e X3. Para testar cada grupo, temos que interferir na linha de comando, novamente usando a sintaxe que aprendemos com a soma condicional discutida no cálculo manual dos ranks nos testes não paramétricos. No caso, repare que a Script Window exibiu

```
shapiro.test(Exemplo_2$Valor)
```

modifique para

```
shapiro.test(Exemplo_2$Valor[Exemplo_2$Grupo=="X1"])
shapiro.test(Exemplo_2$Valor[Exemplo_2$Grupo=="X2"])
shapiro.test(Exemplo_2$Valor[Exemplo_2$Grupo=="X3"])
```

e submeta para obter:

```
        Shapiro-Wilk normality test

data:  Exemplo_2$Valor[Exemplo_2$Grupo == "X1"]
W = 0.9752, p-value = 0.5161

        Shapiro-Wilk normality test

data:  Exemplo_2$Valor[Exemplo_2$Grupo == "X2"]
W = 0.9825, p-value = 0.8789

        Shapiro-Wilk normality test

data:  Exemplo_2$Valor[Exemplo_2$Grupo == "X3"]
W = 0.9836, p-value = 0.7529
```

Concluímos, portanto, que os três subconjuntos de dados podem ser ajustados para uma distribuição normal.

3.4 Cálculo do ANOVA propriamente dito

Devemos antes estabelecer as hipóteses nula e alternativa da ANOVA:

$H_0 : \mu_1 = \mu_2 = \mu_3$
$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$
Adotaremos o nível de significância, $\alpha = 5\%$.

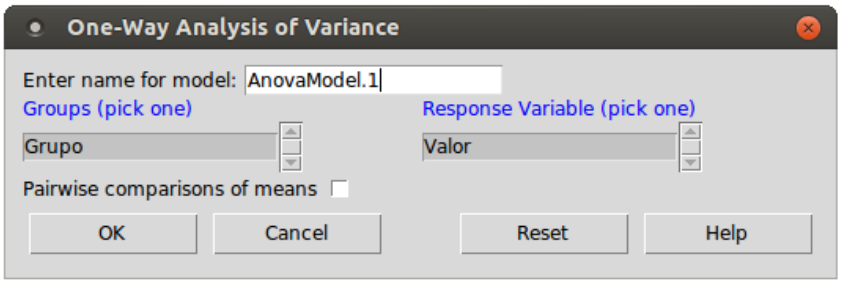
> Note que para rejeitarmos H_0 não significa que todas as médias sejam diferentes entre si. Basta que uma das médias destoe das demais para isso.

Para executar o cálculo da ANOVA escolha:

 Estatísticas -> Médias -> ANOVA para uma amostra (one way) ...

 Statistics -> Means -> One-way ANOVA...

abrindo o diálogo:



Clique OK para obter:

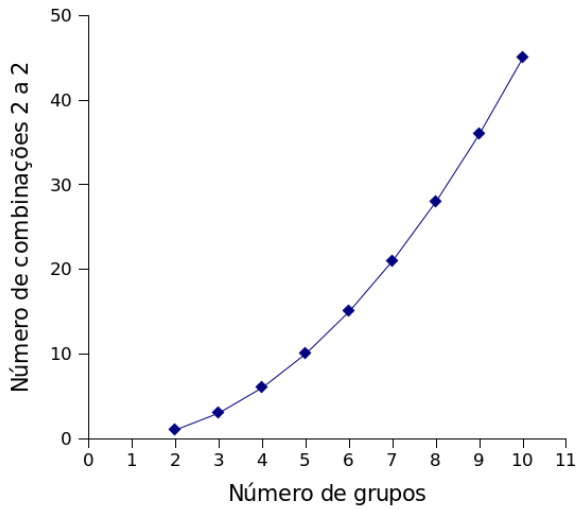
```
             Df Sum Sq Mean Sq F value   Pr(>F)
Grupo         2  154.7   77.34    16.4 5.49e-07 ***
Residuals   114  537.5    4.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       mean       sd data:n
X1 15.12733 2.391789     40
X2 13.60447 1.964106     31
X3 12.44007 2.101316     46
```

Encontre o valor $p = 5.49e-07$. Esta é uma notação usada na área de engenharia e adotada em calculadoras manuais científicas: e não é o [número de Euler (http://en.wikipedia.org/wiki/E_%28mathematical_constant%29)](), mas sim a potência de 10. Leia como $p = 5.49 \cdot 10^{-7}$, $p = \frac{5.49}{10000000}$ ou $p = 0.000000549$. É um valor pequeno, muito menor que α e, portanto, rejeita-se H_0.

> Note, também, as outras colunas: Df e F value, que são, respectivamente, os graus de liberdade (degrees of freedom) e o valor de F (pois o teste de ANOVA é um teste de comparação de variâncias e usa a distribuição F). Sum Sq e Mean Sq, respectivamente soma e média dos quadrados, farão mais sentido quando aplicarmos o teste F em regressão linear.

Logo abaixo aparecem os valores das médias e dos desvios-padrão dos três grupos em análise. A conclusão é que pelo menos uma das três médias -- $\bar{X}1 = 15.12733$, $\bar{X}2 = 13.60447$ ou $\bar{X}3 = 12.44007$ -- difere das demais. Quais delas?

Para descobrir poderíamos fazer um teste t comparando as médias duas a duas ($\bar{X}1$ contra $\bar{X}2$, $\bar{X}1$ contra $\bar{X}3$ e $\bar{X}2$ contra $\bar{X}3$), mas isto exige, como discutimos na aula teórica, correções... além de ser trabalhoso. O número de pares cresce muito rápido quando há mais grupos envolvidos:



O RCmdr cuida disto para você. Basta que se repita o teste, desta vez assinalando o **checkbox** [x]

[x] Comparação de médias 1 a 1

[x] Pairwise comparisons of means

A saída do teste será adicionada dos testes post hoc, comparando as médias dos grupos em pares; localize:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Grupo         2  154.7   77.34    16.4 5.49e-07 ***
Residuals   114  537.5    4.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       mean       sd data:n
X1 15.12733 2.391789     40
X2 13.60447 1.964106     31
X3 12.44007 2.101316     46

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Valor ~ Grupo, data = Exemplo_2)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
X2 - X1 == 0  -1.5229     0.5196  -2.931   0.0113 *
X3 - X1 == 0  -2.6873     0.4695  -5.724   <1e-04 ***
X3 - X2 == 0  -1.1644     0.5046  -2.308   0.0586 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

     Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Valor ~ Grupo, data = Exemplo_2)

Quantile = 2.3738
95% family-wise confidence level

Linear Hypotheses:
             Estimate lwr      upr
X2 - X1 == 0 -1.52286 -2.75627 -0.28946
X3 - X1 == 0 -2.68726 -3.80162 -1.57290
X3 - X2 == 0 -1.16440 -2.36216  0.03337
```

O trecho que mais lhe interessará neste momento é:

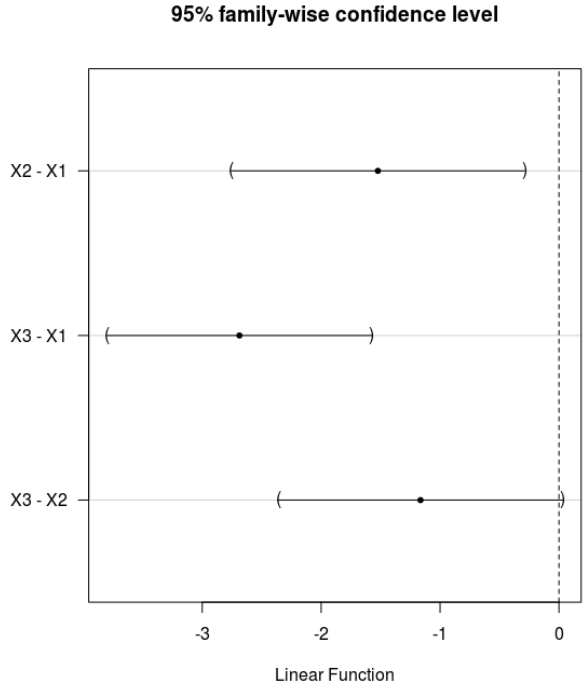
```
             Estimate Std. Error t value Pr(>|t|)
X2 - X1 == 0  -1.5229     0.5196  -2.931   0.0113 *
X3 - X1 == 0  -2.6873     0.4695  -5.724   <1e-04 ***
X3 - X2 == 0  -1.1644     0.5046  -2.308   0.0586 .
```

Na primeira coluna estão as hipóteses nulas (usando o == característico da sintaxe para igualdade entre os operadores de comparação do R, Python, C e outras linguagens de programação), indicando, respectivamente, $H_0 : \mu_1 = \mu_2$ ou, o que é dizer o mesmo, $H_0 : \mu_2 - \mu_1 = 0$ ou (como aparece algo incorretamente) $H_0 : \bar{X}_1 = \bar{X}_2$ ou $H_0 : \bar{X}_2 - \bar{X}_1 = 0$ $H_0 : \mu_1 = \mu_3$ ou, o que é dizer o mesmo,

$H_0 : \mu_3 - \mu_1 = 0$ ou (como aparece algo incorretamente) $H_0 : \bar{X}_1 = \bar{X}_3$ ou $H_0 : \bar{X}_3 - \bar{X}_1 = 0$
$H_0 : \mu_2 = \mu_3$ ou, o que é dizer o mesmo, $H_0 : \mu_3 - \mu_2 = 0$ ou (como aparece algo incorretamente) $H_0 : \bar{X}_2 = \bar{X}_3$ ou $H_0 : \bar{X}_3 - \bar{X}_2 = 0$

levando às conclusões de que, para $\alpha = 0.05$, as médias amostrais dos grupos X1 e X2 diferem entre si ($p = 0.0113$), X1 e X3 diferem entre si ($p < 10^{-4}$), mas não existe diferença estatisticamente significante entre X2 e X3 ($p = 0.0586$). Em resumo, É $\bar{X1} = 15.12733$ que difere das demais, $\bar{X2} = 13.60447$ e $\bar{X3} = 12.44007$, iguais entre si.

Esta opção ainda traz o gráfico de intervalo de confiança de 95% para as diferenças entre as médias, que sumaria o resultado discutido no parágrafo anterior:



que nada mais é do que a expressão gráfica dos intervalos de confiança expressos em outro segmento da saída textual:

```
Linear Hypotheses:
             Estimate lwr      upr
X2 - X1 == 0 -1.52286 -2.75627 -0.28946
X3 - X1 == 0 -2.68726 -3.80162 -1.57290
X3 - X2 == 0 -1.16440 -2.36216  0.03337
```

Na aula sobre Medidas de Risco discutiremos sobre Intervalos de confiança.

4 Teste de Kruskal-Wallis

4.1 Organizando os dados

O teste de Kruskal-Wallis é o equivalente não paramétrico de One-way ANOVA. É utilizado para variáveis qualitativas ordinais ou para quantitativas ordinais que não atendem às premissas exigidas pelos testes paramétricos.

Utilizaremos, para experimentar este teste, os mesmos dados que visitamos no tutorial de Introdução ao R Commander, disponíveis no arquivo **exames1.txt** (arquivo texto delimitado por tabs e com ponto como separador decimal) que você pode baixar para seu computador e importar como um novo conjunto de dados (que, aqui, denominarei de Exames) para o RCmdr.

Uma vez importado com sucesso, deve-se observar que:

```
NOTE: The dataset Exames has 705 rows and 15 columns.
```

e pode-se, também, olhar seu conteúdo. Verá que IDADE é uma das variáveis recebida como Real (i.e. numérica ou quantitativa no "entendimento" do RCmdr). Vamos utilizá-la para criar alguns grupos, desta vez recodificando a variável em uma variável chamada FAIXA_ETARIA:

- 0 a 20,
- 21 a 40,
- 41 a 60,
- 61 a 70 e
- mais de 70 anos

Teremos, com isto, cinco grupos de pacientes.

4.2 Estatística descritiva e gráficos

Vamos analisar como **UREIA** (dosagem sérica de uréia) e **CREAT** (dosagem sérica de creatinina) comportam-se nesses grupos. Antes de prosseguir, você já sabe como iniciar a análise exploratória: faça uma breve estatística descritiva e dê uma olhada nos gráficos. O que lhe parece?

Esta pausa serve para que você faça uma análise exploratória antes de prosseguir neste tutorial.

4.3 Condições para o teste paramétrico

Após esta análise descritiva você deve imaginar que UREIA e CREATININA, embora sejam variáveis quantitativas contínuas, não obedecem a uma distribuição normal e que, mais importante, talvez existam discrepâncias entre as variâncias (i.e., talvez falte homoscedasticidade). Tente confirmar estas hipóteses, utilizando os testes adequados.

Outro momento de relaxamento para que você pense, e então execute, os testes que verificarão as condições para aplicar o teste de ANOVA.

Tendo feito os testes, deve ter concluído que as condições para o teste paramétrico não foram satisfeitas.

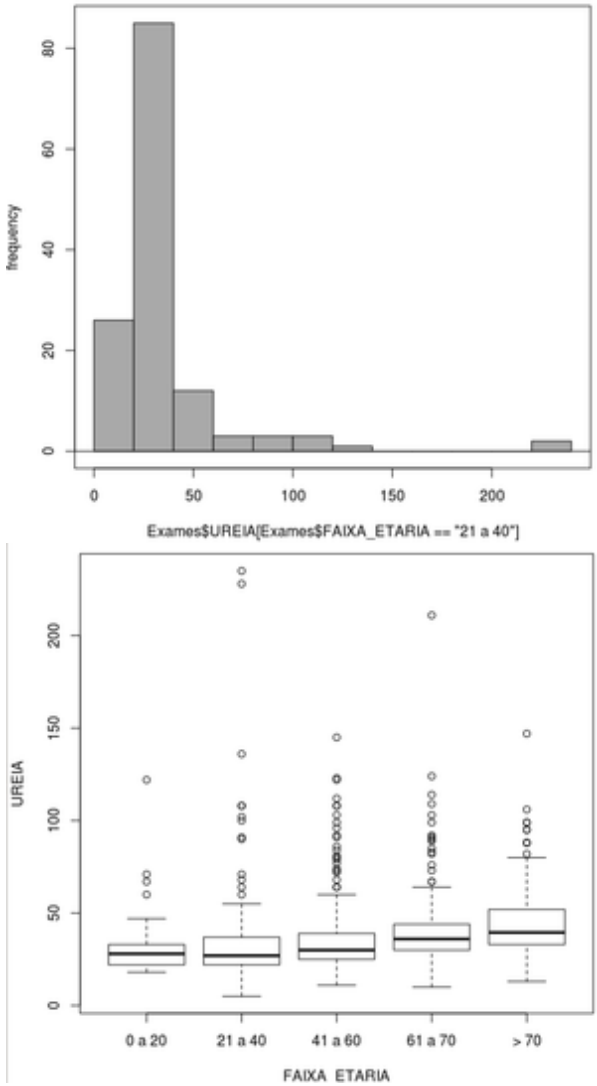
Para sua conferência, obtem-se para uréia sérica:

```
            mean       sd 0% 25%  50%   75% 100% data:n
0 a 20  35.07692 23.05632 18  22 28.0 32.75  122     26
21 a 40 35.83704 31.59497  5  22 27.0 37.00  235    135
41 a 60 36.07874 20.51592 11  25 30.0 39.00  145    254
61 a 70 42.63380 24.74450 10  30 36.0 44.00  211    142
> 70    45.24324 19.71450 13  33 39.5 52.00  147    148

    Bartlett test of homogeneity of variances

data:  UREIA by FAIXA_ETARIA
Bartlett's K-squared = 45.8271, df = 4, p-value = 2.676e-09

    Shapiro-Wilk normality test

data:  Exames$UREIA[Exames$FAIXA_ETARIA == "0 a 20"]
W = 0.6861, p-value = 3.437e-06

data:  Exames$UREIA[Exames$FAIXA_ETARIA == "21 a 40"]
W = 0.547, p-value < 2.2e-16

data:  Exames$UREIA[Exames$FAIXA_ETARIA == "41 a 60"]
W = 0.72, p-value < 2.2e-16

data:  Exames$UREIA[Exames$FAIXA_ETARIA == "61 a 70"]
W = 0.6939, p-value = 7.892e-16

data:  Exames$UREIA[Exames$FAIXA_ETARIA == "> 70"]
W = 0.8528, p-value = 7.297e-11
```

e para creatinina:

```
            mean        sd  0%  25% 50%   75% 100% data:n
0 a 20  1.126923 1.7208272 0.4 0.70 0.8 0.875  9.5     26
21 a 40 1.253333 1.7023601 0.5 0.75 0.9 1.000 13.1    135
41 a 60 1.107087 0.9688575 0.5 0.80 0.9 1.000 10.9    254
61 a 70 1.106338 0.6092915 0.5 0.80 0.9 1.200  5.2    142
> 70    1.218919 0.8392863 0.5 0.90 1.0 1.200  7.6    148

    Bartlett test of homogeneity of variances

data:  CREAT by FAIXA_ETARIA
Bartlett's K-squared = 177.1991, df = 4, p-value < 2.2e-16

    Shapiro-Wilk normality test

data:  Exames$CREAT[Exames$FAIXA_ETARIA == "0 a 20"]
W = 0.2968, p-value = 3.872e-10

data:  Exames$CREAT[Exames$FAIXA_ETARIA == "21 a 40"]
W = 0.327, p-value < 2.2e-16

data:  Exames$CREAT[Exames$FAIXA_ETARIA == "41 a 60"]
W = 0.3989, p-value < 2.2e-16

data:  Exames$CREAT[Exames$FAIXA_ETARIA == "61 a 70"]
W = 0.6163, p-value < 2.2e-16

data:  Exames$CREAT[Exames$FAIXA_ETARIA == "> 70"]
W = 0.5246, p-value < 2.2e-16
```

Além disto, os gráficos tipicamente sugerem distribuições assimétricas. Dois exemplos:



Dica: Ao gerar o boxplot para creatinina dividido por faixas etárias pode ser difícil visualizar as "caixas" por causa da dispersão dos valores. É possível interferir nas escalas dos eixos x e y. Modifique

```
boxplot(CREAT~FAIXA_ETARIA, ylab="CREAT", xlab="FAIXA_ETARIA", data=
Exames)
```

para

```
boxplot(CREAT~FAIXA_ETARIA, ylab="CREAT", xlab="FAIXA_ETARIA", ylim=
c(0, 2), data=Exames)
```

e verifique o que ocorre.

Outra possibilidade é tornar o eixo y logarítmico (mas neste caso convém avisar ao leitor mudando também o rótulo do eixo y):

```
boxplot(CREAT~FAIXA_ETARIA, ylab="log(CREAT)", xlab="FAIXA_ETARIA",
log = "y", data=Exames)
```

Estas alterações de eixos valem para vários tipos de gráfico em R.

4.4 Aplicando o teste de Kruskal-Wallis propriamente dito

Optando pelo teste de Kruskal-Wallis, como sempre, deve-se formular as hipóteses. Atente para o fato de que, sendo este um teste não paramétrico, a análise decritiva são das medianas (respectivamente 28.0, 27.0, 30.0, 36.0 e 39.5) que estão sob análise:

```
             mean       sd 0% 25%  50%   75% 100% data:n
0 a 20   35.07692 23.05632 18  22 28.0 32.75  122     26
21 a 40  35.83704 31.59497  5  22 27.0 37.00  235    135
41 a 60  36.07874 20.51592 11  25 30.0 39.00  145    254
61 a 70  42.63380 24.74450 10  30 36.0 44.00  211    142
> 70     45.24324 19.71450 13  33 39.5 52.00  147    148
```

As hipóteses, portanto, são que:

H_0 : as medianas de uréia (ou de creatinina) são similares em todos os grupos.
H_1 : pelo menos uma das medianas de um grupo difere das medianas dos demais grupos.

e, como sempre, definiremos $\alpha = 0.05$

Experimente, então:

 Estatísticas -> Teste Não-paramétricos -> Teste de Kruskal-Wallis ...

 Statistics -> Nonparametric tests -> Kruskal-Wallis test...

Escolha agrupar pelas faixas etárias e selecione a **variável de uréia sérica** como resposta, obtendo:

```
        Kruskal-Wallis rank sum test

data:  UREIA by FAIXA_ETARIA
Kruskal-Wallis chi-squared = 87.0097, df = 4, p-value < 2.2e-16
```

e, para **creatinina**:

```
        Kruskal-Wallis rank sum test

data:  CREAT by FAIXA_ETARIA
Kruskal-Wallis chi-squared = 38.0178, df = 4, p-value = 1.111e-07
```

5 Exercício

O que concluiria sobre a glicemia? Podem postar suas análises compartilhadamente aqui.

6 Algo para testar (uso avançado)

Voltando à análise de uréia e creatinina com cinco grupos de faixa etária, note que, diferentemente da implementação do ANOVA, no RCmdr não há como automatizar a aplicação do teste de Mann-Whitney aos pares.

Dá para apelar para a linha de comando. Em primeiro lugar, precisamos corrigir o valor de alfa. A mais simples de todas é a correção de Bonferroni (http://en.wikipedia.org/wiki/Bonferroni_correction), que divide o valor de α pelo número de pares (combinatória do número de grupos tomados 2 a 2). A fórmula da combinatória é:

$$ {}^{n}_{x}\mathrm{C} = \frac{n!}{(n-x)! \cdot x!} \tag{1} $$

No exemplo em que temos 5 grupos que, tomados em pares o resultado...

$$ {}_{2}^{5}\mathrm{C} = \frac{5!}{(5-2)! \cdot 2!} = \tag{2} $$

... pode ser preguiça para fazer a conta. Podemos (como já sabem) usar RCmdr como calculadora. Como existe a função `factorial($\dot$)` , experimente

```
factorial(5) / (factorial(5-2) * factorial(2))
```

que lhe retornará

```
[1] 10
```

O valor corrigido, que usaremos para comparar será $\alpha_{\text{Bonferroni}} = \frac{\alpha}{10} = 0.005$.

Também já vimos a sintaxe que o RCmdr aplica para o teste de Mann-Whitney. No conjunto de dados que temos, ao tentarmos:

 Estatísticas -> Teste Não-Paramétrico -> Teste de Wilcoxon (teste 2 amostras)...

 Statistics -> Nonparametric tests -> Two-sample Wilcoxon test...

o único agrupamento disponível é SEXO. Isto seria bem esperado, pois este teste usa, para definir grupos, uma variável qualitativa nominal que tenha somente dois estados (no caso masculino e feminino). Experimente ver se UREIA difere entre os sexos. Obterá:

```
wilcox.test(UREIA ~ SEXO, alternative="two.sided", data=Exames)

    Wilcoxon rank sum test with continuity correction

data:  UREIA by SEXO
W = 47600, p-value = 0.0001275
alternative hypothesis: true location shift is not equal to 0
```

Seguindo novamente a lógica da sintaxe do R que utilizamos com as somas condicionais quando calculamos ranks nos testes não paramétricos, experimentei trocar o sexo por algo que indicasse apenas duas faixas etárias (e parece que funciona):

```
wilcox.test(UREIA[FAIXA_ETARIA=="0 a 20" | FAIXA_ETARIA=="21 a 40"] ~ FA
IXA_ETARIA[FAIXA_ETARIA=="0 a 20" | FAIXA_ETARIA=="21 a 40"], alternativ
e="two.sided",data=Exames)
```

que é equivalente a:

```
wilcox.test(Exames$UREIA[Exames$FAIXA_ETARIA=="0 a 20" | Exames$FAIXA_ET
ARIA=="21 a 40"] ~ Exames$FAIXA_ETARIA[Exames$FAIXA_ETARIA=="0 a 20" | E
xames$FAIXA_ETARIA=="21 a 40"], alternative="two.sided")
```

> Quando endereçando as variáveis explicitamente com Exames$, a cláusula data=Exames do final pode ser removida.

obtém-se:

```
    Wilcoxon rank sum test with continuity correction

data:  UREIA[FAIXA_ETARIA == "0 a 20" | FAIXA_ETARIA == "21 a 40"] by FA
IXA_ETARIA[FAIXA_ETARIA == "0 a 20" | FAIXA_ETARIA == "21 a 40"]
W = 1779.5, p-value = 0.9121
alternative hypothesis: true location shift is not equal to 0
```

e concluímos que as medianas das faixas de idade 0 a 20 e 21 a 40 são iguais.

Para creatinina pode-se pedir:

```
wilcox.test(CREAT[FAIXA_ETARIA=="0 a 20" | FAIXA_ETARIA=="21 a 40"] ~ FA
IXA_ETARIA[FAIXA_ETARIA=="0 a 20" | FAIXA_ETARIA=="21 a 40"], alternativ
e="two.sided",data=Exames)
```

obtendo-se:

```
        Wilcoxon rank sum test with continuity correction

data:  CREAT[FAIXA_ETARIA == "0 a 20" | FAIXA_ETARIA == "21 a 40"] by FA
IXA_ETARIA[FAIXA_ETARIA == "0 a 20" | FAIXA_ETARIA == "21 a 40"]
W = 1318.5, p-value = 0.04262
alternative hypothesis: true location shift is not equal to 0
```

Repare que é com $\frac{\alpha}{10} = 0.005$ que comparamos e, neste caso, as medianas de creatinina destas faixas também são consideradas estatisticamente iguais.

Teríamos que repetir o procedimento para todos os 10 pares possíveis. Fazer todas as combinações é tedioso... é nestas horas que o uso de scripts resolveria. O dois scripts que apresento abaixo não são os mais elegantes, mas funcionam.

> Você não precisará digitar os textos abaixo: ambos estão disponíveis em Arquivo:MPT0164 KW scripts.zip; é só baixar este arquivo e descompactar os dois scripts. Poderá executá-los como mostramos em Salvando_e_lendo_scripts na aula de Introdução ao R Commander.

O primeiro, **UREIA_wilcoxon.R** contém:

```
faixas <- c("0 a 20","21 a 40","41 a 60","61 a 70","> 70")
num_grupos = length(faixas)
combinacoes <- factorial(num_grupos) / (factorial(num_grupos-2)*factoria
l(2))
alfa = 0.05
alfa_bonferroni = alfa/combinacoes

print (paste ("alfa corrigido = ",alfa_bonferroni, " combinacoes = ", co
mbinacoes))
for (par1 in faixas)
{
  for (par2 in faixas)
  {
    if (par2 > par1)
    {
        p <- wilcox.test(Exames$UREIA[Exames$FAIXA_ETARIA==par1 | Exames
$FAIXA_ETARIA==par2] ~
        + Exames$FAIXA_ETARIA[Exames$FAIXA_ETARIA==par1 | Exames$FAIXA_E
TARIA==par2], alternative="two.sided")
        if (p[3] < alfa_bonferroni)
        {
            significante = " (H1) ... medianas diferentes"
        }
        else
        {
            significante = " (H0) ... medianas iguais"
        }
        print (paste (par1," vs. ",par2," p=",p[3],significante))
    }
  }
}
numSummary(Exames[,"UREIA"], groups=Exames$FAIXA_ETARIA,
  statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

... muito similarmente, **CREAT_wilcoxon.R** é:

```
faixas <- c("0 a 20","21 a 40","41 a 60","61 a 70","> 70")
num_grupos = length(faixas)
combinacoes <- factorial(num_grupos) / (factorial(num_grupos-2)*factoria
l(2))
alfa = 0.05
alfa_bonferroni = alfa/combinacoes

print (paste ("alfa corrigido = ",alfa_bonferroni, " combinacoes = ", co
mbinacoes))
for (par1 in faixas)
{
  for (par2 in faixas)
  {
    if (par2 > par1)
    {
        p <- wilcox.test(Exames$CREAT[Exames$FAIXA_ETARIA==par1 | Exames
$FAIXA_ETARIA==par2] ~
        + Exames$FAIXA_ETARIA[Exames$FAIXA_ETARIA==par1 | Exames$FAIXA_E
TARIA==par2], alternative = "two.sided")
        if (p[3] < alfa_bonferroni)
        {
            significante = " (H1) ... medianas diferentes"
        }
        else
        {
            significante = " (H0) ... medianas iguais"
        }
        print (paste (par1," vs. ",par2," p=",p[3],significante))
    }
  }
}
numSummary(Exames[,"CREAT"], groups=Exames$FAIXA_ETARIA,
  statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

Execute os scripts, observe as saídas e as medianas. Consegue ver quais grupos diferem de quais grupos?