

Anova

Bioestatística em R

André M Ribeiro-dos-Santos

12 de Abr, 2017

ANOVA One-Way

- Avaliar a diferença de uma variável quantitativa entre três ou mais grupos.
- Compreender o resultado de uma ANOVA.
- Quando aplica-se uma ANOVA *one-way* ou *two-way*.
- Como investigar a diferença entre os grupos.
- Ilustrar os resultados.

Pesquisadores buscando desenvolver uma nova droga para o tratamento de hipertensão, investigaram o efeito de três compostos sobre a pressão sanguínea de pacientes. Neste estudo eles distribuíram os pacientes entre os três medicamentos totalizando 5 pacientes por composto e 5 pacientes tratados com placebo.

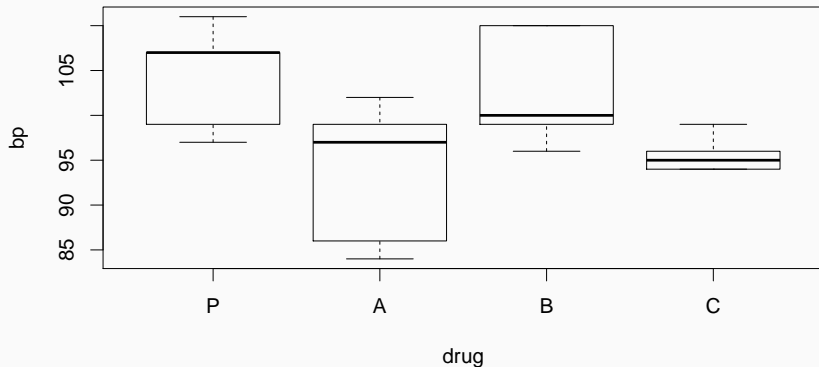
Existe diferença na pressão sanguínea dos pacientes entre os tratamentos?

Table 1: Pressão sanguínea por tratamento

P	A	B	C
107	102	110	96
111	84	99	95
99	86	100	94
107	97	110	99
97	99	96	94

```
> drug <- factor(rep(1:4, each=5), labels = c("P", "A", "B", "C"))  
> bp <- c(107, 111, 99, 107, 97, 102, 84, 86, 97, 99,  
+         110, 99, 100, 110, 96, 96, 95, 94, 99, 94)  
> plot(bp~drug, main = "Pressão sanguínea por tratamento")
```

Pressão sanguínea por tratamento



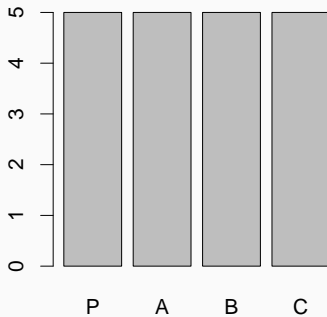
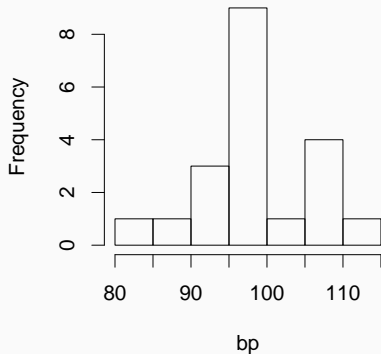
- Quais os tipos de variáveis envolvidas?
- Qual a questão que deseja-se resolver?
- Quantas categorias possui a variável qualitativa?
- Qual a distribuição da variável quantitativa?

- Quais os tipos de variáveis envolvidas? **Uma medida quantitativa e uma categórica.**
- Qual a questão que deseja-se resolver? **Saber se existe diferença entre os tratamentos.**
- Quantas categorias possui a variável qualitativa? **Quatro categorias: três compostos e o placebo**

- Qual a distribuição das variáveis?

```
> par(mfrow = c(1,2))  
> hist(bp)  
> plot(drug)
```

Histogram of bp



Quando deseja-se comparar a distribuição de uma medida quantitativa entre três ou mais grupos usa-se ANOVA (**AN**alysis **Of** **VA**riance).

$$H_0 : \hat{x}_A = \hat{x}_B = \dots = \hat{x}_k$$

H_a : Nem todas as médias são iguais

$$H_a : \exists a, b \in \{A, B, \dots, k\} : \hat{x}_a \neq \hat{x}_b$$

Esta análise compara a variância explicada pela média central com o obtido em cada grupo.

Calcula-se a razão do erro médio quadrático **entre os grupos** e **dentro dos grupo** para obter o valor de F.

Variação	DF	SSQ ¹	MS ²	F
Entre	$k - 1$	$SSQ(treat)$	$SSQ(treat)/(k - 1)$	$MS(treat)/MS(error)$
Dentro	$n - k$	$SSQ(error)$	$SSQ(error)/(n - m)$	
Total	$n - 1$	$SSQ(total)$		

$$SSQ(treat) = \sum (\hat{x} - \hat{x}_i)^2$$

$$SSQ(error) = \sum (x - \hat{x}_i)^2$$

$$SSQ(total) = SSQ(treat) + SSQ(error) = \sum (x - \hat{x})^2$$

¹Soma do erro quadrático

²Erro quadrático médio.

```
> ?aov
> ## Fit an Analysis of Variance Model
> ## Description:
> ##      Fit an analysis of variance model by a call to 'lm' for each
> ##      stratum.
> ## Usage:
> ##      aov(formula, data = NULL, projections = FALSE, qr = TRUE,
> ##          contrasts = NULL, ...)
> ## Arguments:
> ##      formula: A formula specifying the model.
> ##      data: A data frame in which the variables specified in the
> ##            formula will be found. If missing, the variables are
> ##            searched for in the standard way.
```

```
> (model_drug <- aov(bp~drug))
```

```
## Call:
```

```
##      aov(formula = bp ~ drug)
```

```
##
```

```
## Terms:
```

```
##                drug Residuals
```

```
## Sum of Squares  418.6      591.2
```

```
## Deg. of Freedom    3         16
```

```
##
```

```
## Residual standard error: 6.078651
```

```
## Estimated effects may be unbalanced
```

```
> summary(model_drug)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## drug           3  418.6   139.53   3.776 0.0319 *
## Residuals     16  591.2    36.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regressão com Variáveis Discretas

Outra forma de olhar o mesma análise é através de uma regressão linear com variáveis discretas. Neste tipo de regressão substitui-se a variável em questão por várias variáveis binárias indicando cada categoria.

Como numa regressão linear busca-se minimizar o erro, o β obtido em cada caso corresponde ao desvio da média de cada categoria em relação a uma categoria basal (quando incluso α).

```
> (model_lm <- lm(bp ~ drug))
```

```
##
```

```
## Call:
```

```
## lm(formula = bp ~ drug)
```

```
##
```

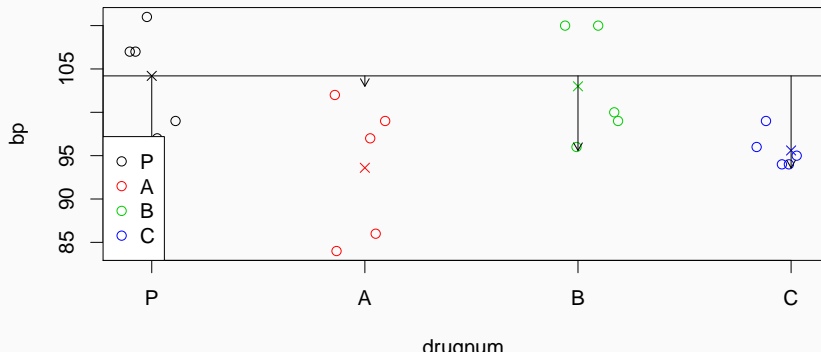
```
## Coefficients:
```

## (Intercept)	drugA	drugB	drugC
## 104.2	-10.6	-1.2	-8.6

```

> cf      <- coef(model_lm)
> drugmn  <- c(cf[1], cf[1] + cf[2:4])
> drugnum <- jitter(as.numeric(drug))
>
> plot(bp~drugnum, col = drug, xaxt="n")
> points(1:4, drugmn, col=1:4, pch=4)
> abline(h = drugmn[1])
> arrows(1:4, drugmn[1], 1:4, drugmn[2:4], length=.1)
> axis(side = 1, at = 1:4, labels = levels(drug))
> legend("bottomleft", levels(drug), col = 1:4, pch = 1, bg = "white")

```



```
> (drugcf <- coef(model_lm))
```

```
## (Intercept)      drugA      drugB      drugC  
##          104.2      -10.6       -1.2      -8.6
```

```
> (drugmean <- c(drugcf[1], drugcf[2:4] + drugcf[1]))
```

```
## (Intercept)      drugA      drugB      drugC  
##          104.2       93.6     103.0     95.6
```

```
> c(mean(bp[drug == "P"]), mean(bp[drug == "A"]),  
+   mean(bp[drug == "B"]), mean(bp[drug == "C"]))
```

```
## [1] 104.2  93.6 103.0  95.6
```

```
> coef(lm(bp ~ 0 + drug))
```

```
## drugP drugA drugB drugC  
## 104.2  93.6 103.0  95.6
```



```
> ssr <- sum((mean(bp) - predict(model_lm))^2)
> sse <- sum((bp - predict(model_drug))^2) # ou sum(resid(model_drug)^2)
> f <- (ssr / (4 - 1)) / (sse / (20 - 4))
> data.frame("SSR" = ssr, "SSE" = sse, "F" = f, "P" = 1 - pf(f, 4-1, 20-4))
```

```
##      SSR   SSE      F      P
## 1 418.6 591.2 3.776274 0.03185873
```

```
> anova(model_drug)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: bp
```

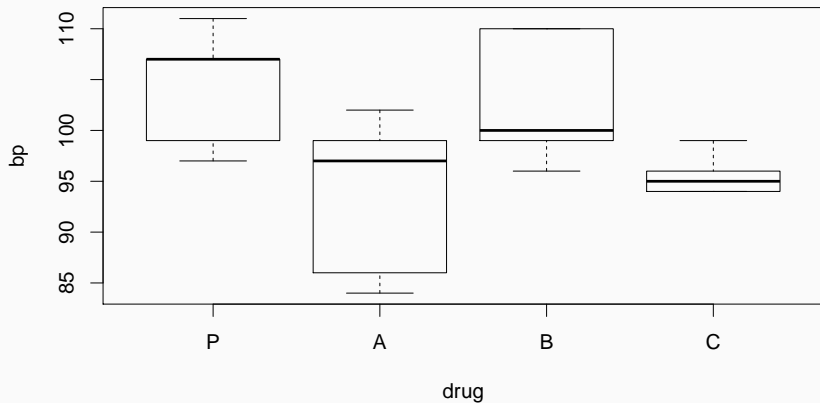
```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## drug        3   418.6   139.53   3.7763 0.03186 *
## Residuals  16   591.2    36.95
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Entre quais grupos da amostragem houve diferença?

Pressão sanguínea por tratamento



Para responder essa questão, poderíamos fazer uma série de testes t entre as amostras. No entanto a medida que aumentamos o número de testes também aumentamos as chances de identificar como significativo uma relação aleatória.

Para cada teste (usando $\alpha = 0.05$) admitimos uma probabilidade de identificar como significativo algo aleatório em 5% dos casos e 95% de identificar corretamente. Portanto, em dois testes temos 90.25% de chance de chamar corretamente e em três testes temos 85.74%.

Para reduzir esse tipo de erro, aplicamos correções do p-value, como **Bonferroni** e **Benjamin-Hockenberg**.

```
> p.adjust(pvalues, method = "bonf")  
> p.adjust(pvalues, method = "fdr")
```

```
> pairwise.t.test(bp, drug, p.adjust.method = "bonf")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data:  bp and drug  
##  
##      P      A      B  
## A 0.084 -      -  
## B 1.000 0.159 -  
## C 0.239 1.000 0.433  
##  
## P value adjustment method: bonferroni
```

Post-hoc (Tukey's HSD)

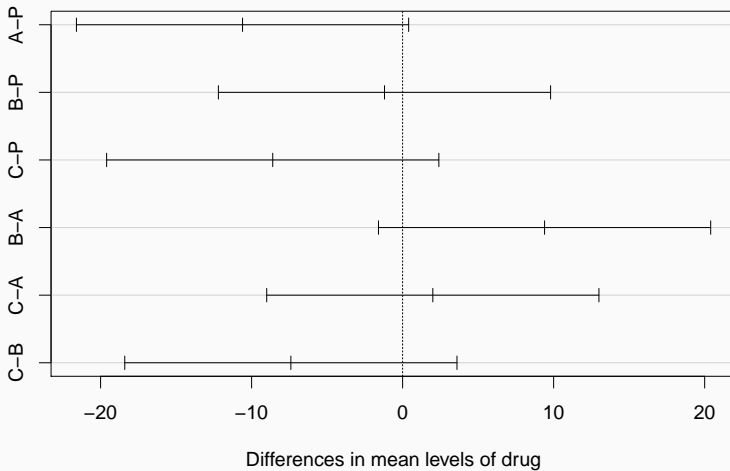
Para **ANOVA** foi desenvolvida um teste mais preciso que utiliza a variação global no lugar de somente os pares (como seria feito).

```
> TukeyHSD(model_drug)
```

```
##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = bp ~ drug)
##
## $drug
##      diff      lwr      upr    p adj
## A-P -10.6 -21.599124  0.3991236 0.0609262
## B-P  -1.2 -12.199124  9.7991236 0.9890559
## C-P  -8.6 -19.599124  2.3991236 0.1554472
## B-A   9.4  -1.599124 20.3991236 0.1082113
## C-A   2.0  -8.999124 12.9991236 0.9529943
## C-B  -7.4 -18.399124  3.5991236 0.2569956
```

```
> plot(TukeyHSD(model_drug))
```

95% family-wise confidence level



ANOVA Two-way

Sabendo que a pressão sanguínea varia entre os sexos (masculino e feminino).

A resposta ao tratamento, varia dependendo do sexo?

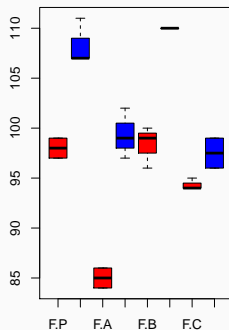
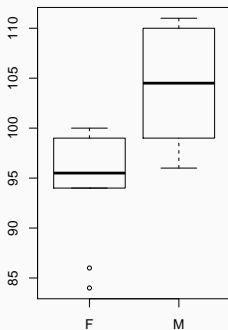
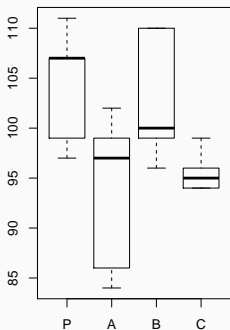
Table 3: Pressão Sanguínea por Tratamento e sexo

Pressão Sanguínea	Tratamento	Sexo	Pressão Sanguínea	Tratamento	Sexo
107	P	M	110	B	M
111	P	M	99	B	F
99	P	F	100	B	F
107	P	M	110	B	M
97	P	F	96	B	F
102	A	M	96	C	M
84	A	F	95	C	F
86	A	F	94	C	F
97	A	M	99	C	M
99	A	M	94	C	F

```

> sex <- factor(c("M", "M", "F", "M", "F", "M", "F", "F", "M", "M",
+                 "M", "F", "F", "F", "M", "F", "M", "F", "F", "M", "F"))
> par(mfrow = c(1,3))
> boxplot(bp~drug)
> boxplot(bp~sex)
> boxplot(bp~sex*drug, col=c('red', 'blue'))

```



ANOVA Two-way

Fonte	DF	SSQ ³	MS ⁴	F
A	$a - 1$	$SSQ(A)$	$SSQ(A)/DF(A)$	$MS(A)/MS(error)$
B	$b - k$	$SSQ(B)$	$SSQ(B)/DF(B)$	$MS(B)/MS(error)$
AB	$(a - 1)(b - 1)$	$SSQ(AB)$	$SSQ(AB)/DF(AB)$	$MS(AB)/MS(error)$
Resid	$n - ab$	$SSQ(err)$	$SSQ(er)/DF(err)$	
Total	$n - 1$	$SSQ(total)$		

$$SSQ(A) = \sum (\hat{y}_i - \hat{y})^2$$

$$SSQ(B) = \sum (\hat{y}_{ij} - \hat{y}_i)^2$$

$$SSQ(AB) = SSQ(total) - SSQ(A) - SSQ(B) - SSQ(err)$$

$$SSQ(err) = \sum (y - \hat{y}_{ijk})^2$$

$$SSQ(total) = \sum (y - \hat{y})^2$$

³Soma do erro quadrático

⁴Erro quadrático médio.

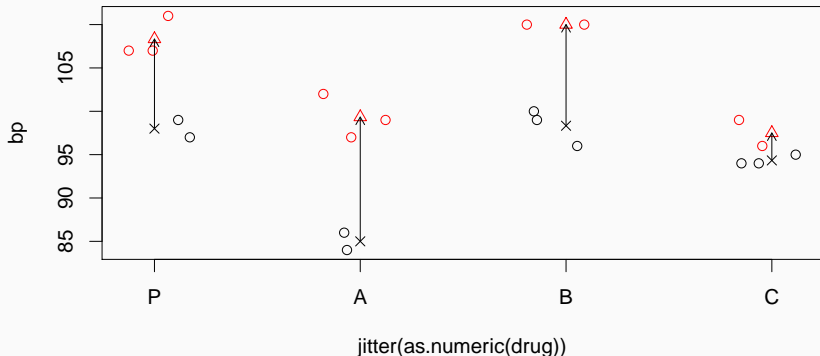
```
> summary(aov(bp ~ sex * drug))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## sex           1  441.8   441.8 128.784 8.99e-08 ***
## drug          3  444.9   148.3  43.227 1.04e-06 ***
## sex:drug       3   82.0    27.3   7.964 0.00346 **
## Residuals     12   41.2     3.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

> cf <- coef(lm(bp~0+drug*sex))
> plot(bp~jitter(as.numeric(drug)), col=sex, xaxt="n")
> axis(side = 1, 1:4, levels(drug))
> points(1:4, cf[1:4], col=1, pch=4)
> points(1:4, cf[5] + cf[1:4] + c(0, cf[6:8]), col=2, pch=2)
> arrows(1:4, cf[1:4],
+       1:4, cf[5] + cf[1:4] + c(0, cf[6:8]), length=.1)

```



```

> sssex <- sum((predict(lm(bp~sex)) - predict(lm(bp~1)))^2)
> ssdrug <- sum((predict(lm(bp~sex+drug)) - predict(lm(bp~sex)))^2)
> ssint <- sum((predict(lm(bp~sex*drug)) - predict(lm(bp~sex+drug)))^2)
> sserr <- sum(resid(lm(bp~sex*drug))^2)
> c("SS SEX"=sssex, "SS DRUG"=ssdrug, "SS SEX:DRUG"=ssint, "SS ERR"=sserr)

```

```

##          SS SEX          SS DRUG SS SEX:DRUG          SS ERR
##    441.80000    444.87500    81.95833    41.16667

```

```

> summary(aov(bp~sex*drug))[[1]]

```

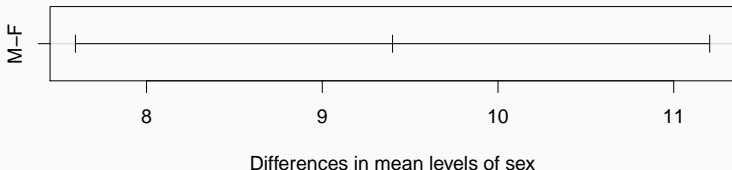
```

##          Df Sum Sq Mean Sq  F value    Pr(>F)
## sex          1  441.80   441.80 128.7838 8.986e-08 ***
## drug          3  444.87   148.29  43.2267 1.043e-06 ***
## sex:drug       3   81.96    27.32   7.9636 0.003459 **
## Residuals     12   41.17     3.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

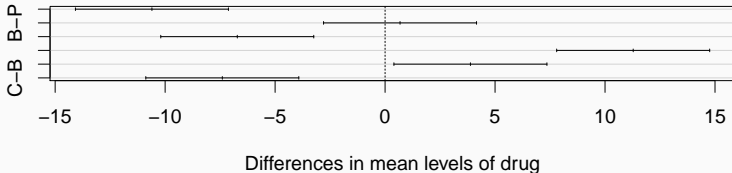
```

```
> par(mfrow = c(2,1))  
> plot(TukeyHSD(aov(bp~sex*drug), "sex"))  
> plot(TukeyHSD(aov(bp~sex*drug), "drug"))
```

95% family-wise confidence level

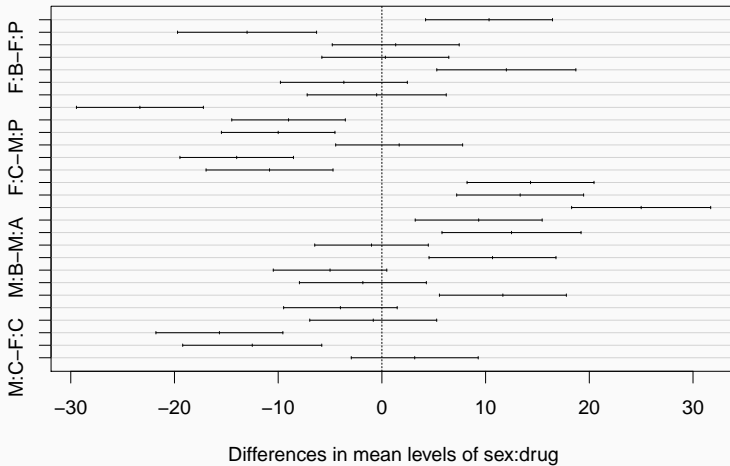


95% family-wise confidence level




```
> plot(TukeyHSD(aov(bp~sex*drug), "sex:drug"))
```

95% family-wise confidence level



Ao final

- Avaliar a diferença de uma variável quantitativa entre três ou mais grupos.
- Compreender o resultado de uma ANOVA.
- Quando aplica-se uma ANOVA *one-way* ou *two-way*.
- Como investigar a diferença entre os grupos.
- Ilustrar os resultados.

Até a próxima
