Matéria de Análise de Dados – Ciências Biomédicas

Aula 14: Analise de Variância – ANOVA

James Hunter

7 de abril de 2017

Carregar os Pacotes Necessários

```
suppressMessages(library(tidyverse))
suppressPackageStartupMessages(library(DescTools))
suppressPackageStartupMessages(library(knitr))
suppressPackageStartupMessages(library(car))
suppressPackageStartupMessages(library(broom))
suppressPackageStartupMessages(library(coefplot))
suppressMessages(library(mosaic))
options(scipen = 5)
pvalaov <- function(model) { # função para extrair o valor p
  x <- summary(model)
  return(unlist(x[[1]][,5][1]))
}
R2 <- function(model) { # função para extrair o R quadrado
  x <- summary(model)
  SST <- sum(x[[1]][,2])
  SSR <- x[[1]][,2][1]
  return(SSR/SST)
}
```

Nesta aula, vamos fazer uma introdução a análise de variância (ANOVA). Aqui, nós vamos tratar os conceitos básicos dos modelos de ANOVA e como construir os modelos em R. Também, nós vamos homenagear o início de temporada de beisebol no meu país e usar um exemplo daquele esporte.

Proposito de ANOVA

Nós usamos ANOVA para analisar comparações entre três ou mais grupos de uma variável. Para dois grupos, usamos um teste-t ou o equivalente não-paramétrico. Teoricamente, podemos fazer comparações entre a média de cada par de grupos quando temos mais que dois. Por exemplo, podemos fazer um teste de grupo 1 contra grupo 2, grupo 2 contra grupo 3 e grupo 1 contra grupo 3, no caso que temos três grupos. Esse faz um total de três comparações. Esse pode criar um grande dificuldade porque é provável que acharemos uma comparação das médias que é significativa por acaso, mesmo se não há uma diferença verdadeira na população.

ANOVA evita este problema porque usa um teste de hipótese único para ver se as médias entre todos os grupos na amostra são iguais.

Teste de Hipótese de ANOVA

A ANOVA teste a hipótese nula que o resultado média de todos os grupos é o mesmo contra a alternativa que pelo menos uma das médias é diferente. Uma maneira alternativa de falar da hipótese nula é que qualquer diferença entre as médias dos grupos

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \ \ (onde\ \mu_i\ é\ a\ média\ das\ observações\ grupo\ i)$$

$$H_1 : ao\ menos\ 1\ \mu\ é\ diferente$$

A presença de grandes diferenças entre as médias dos grupos é evidência em favor da rejeição da hipótese nula.

Porque esta técnica é chamada análise de variância quando estamos testando diferenças entre médias e não os desvios padrões? A resposta é que o modelo avalia a variação entre as médias dos grupos relativo a variação entre observações individuais dentro dos grupos para determinar o grau de diferença entre médias.

Premissas de ANOVA

Como com os outros testes paramétricos, ANOVA tem os seguintes premissas:

1. As observações devem ser independentes dentro e entre os grupos
2. Os dados dentro de cada grupo devem ser quase normais
3. A variância dos grupos deve ser quase igual

Dados para ANOVA – Homenagem a Nova Temporada de Beisebol

- Início de nova temporada no último domingo

Figure 1: Big Swing

Figure 2: LA Dodger Stadium

Figure 3: Minha Equipe

Rebatadores – Carregar Dados

```
load("bat2015.RData")
kable(head(bat, 8))
```

name	R	H	HR	RBI	POS	avg	OBP
Rico Noel	5	1	0	0	DH	0.5000000	0.5000000
Miguel Cabrera	64	145	18	76	IF	0.3379953	0.4562738
Slade Heathcott	6	10	2	8	OF	0.4000000	0.4285714
Mike Trout	104	172	41	90	OF	0.2991304	0.4137931
Max Stassi	4	6	1	2	DH	0.4000000	0.4117647
Shawn O'Malley	10	11	1	7	OF	0.2619048	0.4035088
Mike Napoli	9	23	5	10	IF	0.2948718	0.4021739
Ryan Raburn	22	52	8	29	OF	0.3005780	0.4019608

O Que Nos Interesse?

Nossa questão hoje é se tem diferenças entre a OBP para os jogadores nos posições de campo diferentes no American League em 2015. Estou usando dados da base de dados de beisebol de Lahman que são disponíveis no pacote do mesmo nome. Simplifiquei as posições para reduzir a complexidade do exercício. Todos os três tipos de 'outfielders' (right, left e center) são combinados em `OF` (outfielder). Também, como os jogadores que cuidam do "infield", combinei os cinco posições em `IF`. Receptor, "Catcher", a posição que recebe a bola do arremesador, deixei assim. E porque estamos trabalhando com a American League, existe nesta liga um jogador que só rebate (não faz defesa), um "Designated Hitter" (DH).

OBP

OBP é a abreviação para "on base percentage", ou seja, a porcentagem das vezes que o rebatedor aparece que ele consegue ganha um base. Muitas analistas acham que este é uma medida mais precisa sobre a habilidade de um rebatedor que a "batting average", que mede só a porcentagem das rebatidas válidas.

Os Dados

```
Desc(OBP ~ POS, data = bat, plotit = FALSE)
```

```
## ------------------------------------------------------------------------------
## OBP ~ POS
##
## Summary:
## n pairs: 295, valid: 295 (100.0%), missings: 0 (0.0%), groups: 4
##
##
##                C       DH       IF       OF
## mean       0.279    0.305    0.306    0.312
## median     0.296    0.312    0.308    0.313
## sd         0.058    0.096    0.049    0.047
## IQR        0.076    0.106    0.062    0.051
## n             42       20      132      101
## np       14.237%   6.780%  44.746%  34.237%
## NAs            0        0        0        0
```

```
## 0s            0        0        0        0
##
## Kruskal-Wallis rank sum test:
##   Kruskal-Wallis chi-squared = 9.0551, df = 3, p-value = 0.02857
```

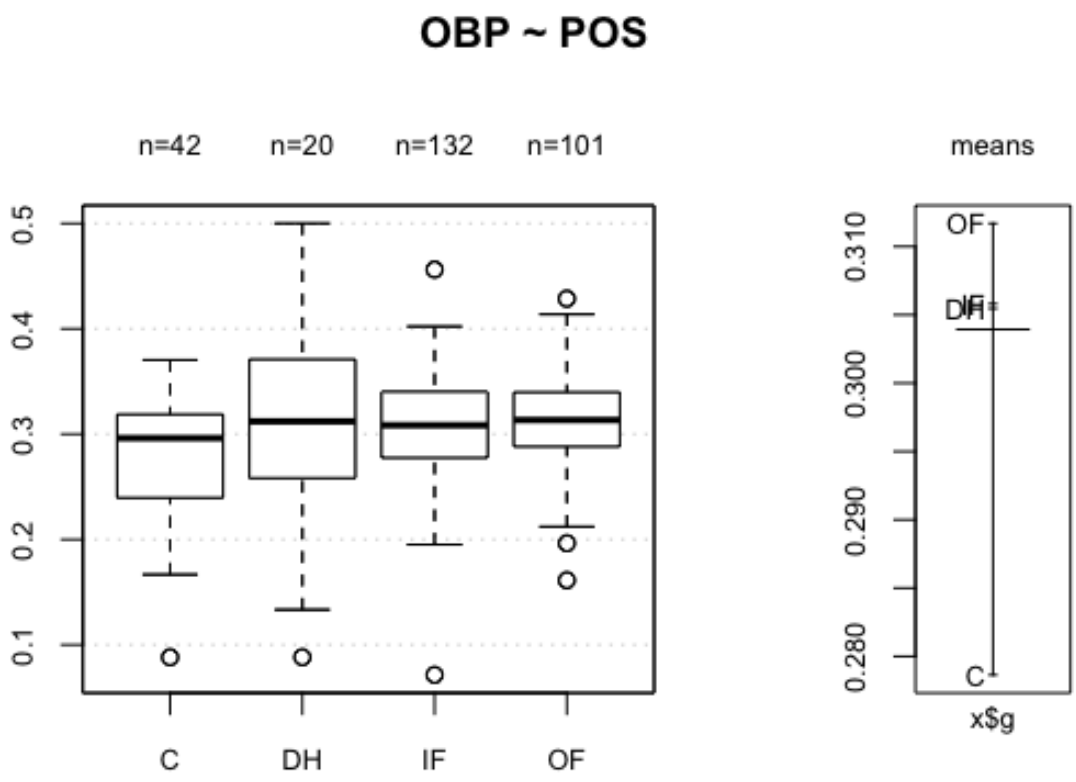


Figure 4: Boxplot dos dados

Esses estatísticas mostram que a variação entre os grupos é muito parecido e podemos sentir confortáveis que a premissa #3 está sendo respeitada. O boxplot revela que há um outlier longe da caixa para os "infielders", mas com uma amostra dentro deste grupo de 132, o outlier não causa preocupação.

Teoria de ANOVA

Sem entrar em muitos detalhes, ANOVA responde a uma pergunta única:

> É a variação nas médias das amostras tão grande que parece improvável que surge de acaso sozinho.

(Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics**, 3ª Ed, p. 250.)

Em ANOVA, testamos todas as diferenças entre grupos simultaneamente. ANOVA funciona pela divisão da variação em componentes diferentes utilizando a soma dos quadrados que vemos primeiro em regressão. O algoritmo calcula primeiro um soma dos quadrados total que é quadrado das diferenças de todos os valores, não importa o grupo, da média de todos os valores (*grand mean*). O primeiro componente disso é a soma

dos quadrados das diferenças entre a média dos grupos e a grand mean. É conhecido como a variação entre os grupos. ("SSG") O que sobra da variação é por causa dos residuais, ou seja, a soma dos quadrados das diferenças entre todos os valores dentro de um grupo e a média desse grupo. Este representa a variação dentro dos grupos. ("SSE")

Cada uma dessas somas de quadrados tem um grau de liberdade associada. Para SSG, é o número dos grupos (k) menos 1. O 1 representa a grand mean, que nós não podemos variar.

$$df_G = k - 1$$

O grau de liberdade associado com a SSE é o tamanho de amostra (n) menos o número dos grupos:

$$df_E = n - k$$

Estatística F

A estatística que teste a hipótese mede a relação entre os dois componentes divididos pelos graus de liberdade. Esta divisão cria duas médias de soma dos quadrados ("MSG" e "MSE"). A estatística é conhecido pelo nome de distribuição que ela segue, "F". A formula para calcular F é o seguinte:

$$F_{df_1, df_2} = \frac{MSG}{MSE}$$

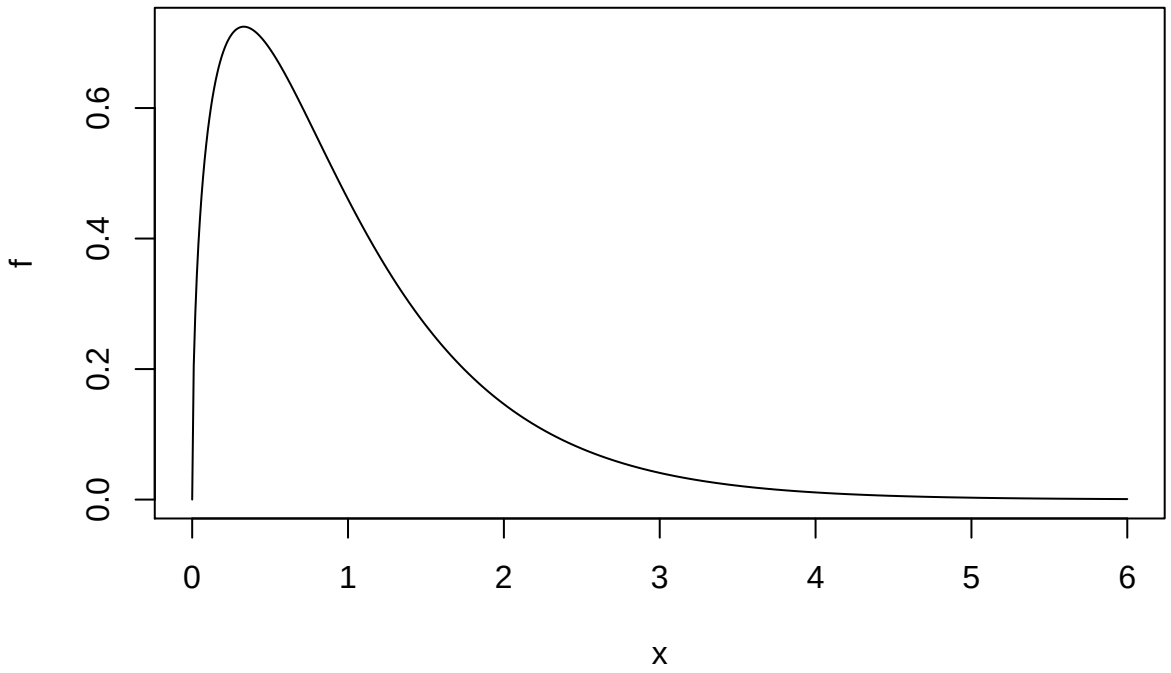
O teste F precisa utilizar os graus de liberdade de grupos e de erro. Então, em nossa exemplo de beisebol, temos 4 grupos e um n de 295. Assim, os graus de liberdade são:

```
grupos <- length(unique(bat$POS))
df1 <- grupos - 1
df2 <- nrow(bat) - grupos
paste("grupos:", grupos, " df1:", df1, " df2:", df2)
```

```
## [1] "grupos: 4  df1: 3  df2: 291"
```

e a distribuição F tem a forma:

```
x <- seq(0, 6, .01)
f <- df(x, df1, df2)
plot(x, f, type = "l")
```



Como em regressão, se a variação entre os grupos (MSG) fica maior relativa a variação dentro dos grupos, maior vai ser o valor de F e a evidência seria mais forte para rejeitar a hipótese nula.

R^2 para ANOVA

Lembrando que o R^2, o coeficiente de determinação, mede a proporção da variação total que existe por causa do modelo (neste caso, nossos grupos) invés do acaso, aqui podemos usar a SSG e a SST para fazer este calculo. Porque a função `summary()` para modelos de aov não relata o valor de R^2, criei uma função `R2(modelo)` que faz isso para modelos de ANOVA "one-way", ou seja com uma variável independente. Para usar esta função, chame ela e dê o nome que você usou para gravar o modelo.

$$R^2 = \frac{SSG}{SST}$$

ANOVA em R

Para fazer uma análise ANOVA em R, podemos usar ou a função `aov()` ou `lm()`. A última é a mesma função que usamos para fazer regressão linear. A diferença é como as duas funções apresentam os resultados. `aov()` foca no modelo em si, as somas de quadrados e o teste F. A `lm()` fornece mais informação sobre os parâmetros das variáveis independentes. A sintaxe é o mesmo que usamos para especificar um modelo de regressão.

A função `summary()` mostra os resultados. Porque a ANOVA usa um modelo linear para fazer seus calculo, você pode acessar os parâmetros com a função `summary.lm()`.

```
modela <- aov(OBP ~ POS, data = bat)
summary(modela)
```

```
##              Df Sum Sq  Mean Sq F value Pr(>F)
## POS           3 0.0335 0.011152    3.82 0.0104 *
## Residuals   291 0.8494 0.002919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Este resultado mostra que existe uma diferença entre as posições em OBP. O valor p do teste F (0.0104) é abaixo do valor padrão, 0.05. Podemos rejeitar a hipótese nula e dizer que as diferenças entre as médias são significativas. Antes de determinar quais das posições têm médias de OBP maiores ou menores que a média para todos os jogadores no estudo, precisamos ver se o modelo cumpriu as necessidades das premissas.

Resumo `lm` de um Modelo de ANOVA

Apesar que é possível de mostrar um resumo no formato de um modelo linear (regressão) de um modelo ANOVA, muito da informação não é útil para analise. Este resumo está disponível com a função `summary.lm()'. A ilustração abaixo mostra os componentes da apresentaçãolm‘ de um modelo[1] e depois podemos ver nosso modelo neste formato.

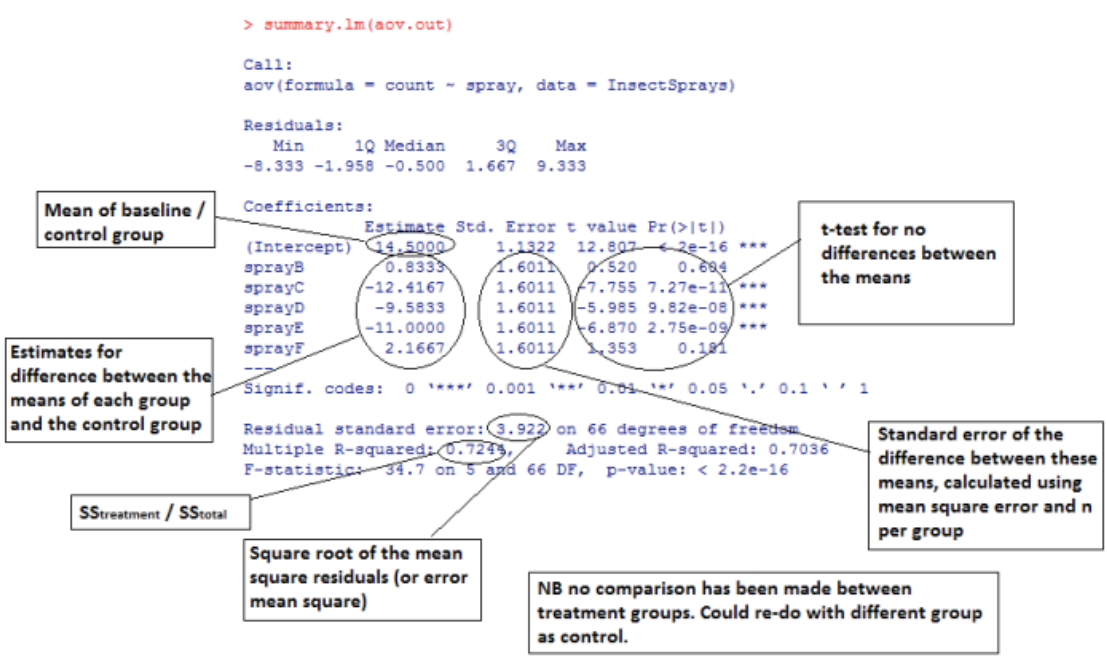


Figure 5:

```
summary.lm(modela)
```

```
##
## Call:
## aov(formula = OBP ~ POS, data = bat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.234383 -0.028034  0.005212  0.035166  0.194540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.278641   0.008337  33.424  < 2e-16 ***
## POSDH       0.026819   0.014678   1.827 0.068703 .
```

[1] Webb, B.; Pajak, M., “ANOVA in R”, course handout from School of Informatics, University of Edinburgh, n.d.

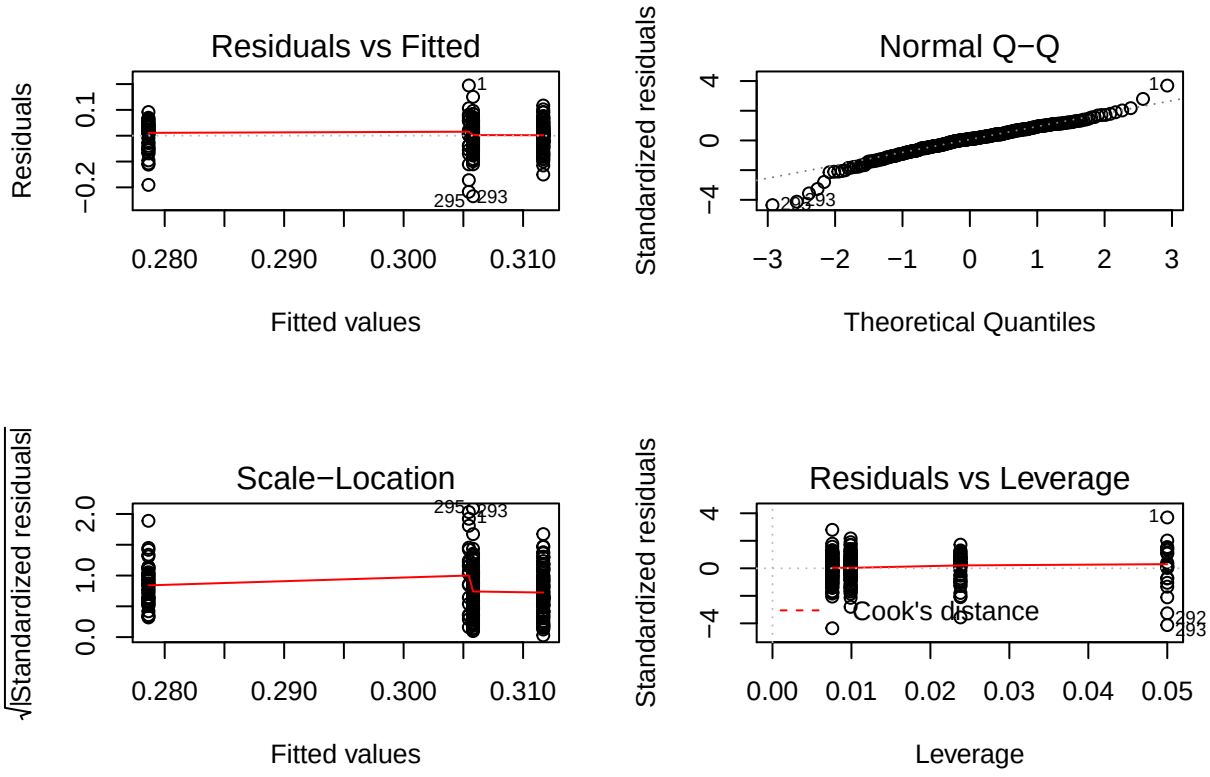
```
## POSIF       0.027171   0.009571   2.839 0.004848 **
## POSOF       0.033048   0.009920   3.332 0.000975 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05403 on 291 degrees of freedom
## Multiple R-squared:  0.03789,    Adjusted R-squared:  0.02798
## F-statistic:  3.82 on 3 and 291 DF,  p-value: 0.0104
```

No caso do modelo de beisebol, o "baseline/control group" é a categoria de catcher. Os outros três linhas descrevem os resultados para as outras três posições. Mas as estatísticas com Estimate e Std. Error não têm muito utilidade para nós porque as categorias não são preditivas para OBP, a variável dependente. Vamos ver, abaixo, uma maneira de comparar as categorias que tem mais relação com testes de comparação entre médias que com regressão.

Validade do Modelo

Fazemos isso com gráficos como fizemos com regressão linear. A função `plot()` produz os mesmos quatro gráficos para um modelo de ANOVA que modelos de regressão produziram.

```
par(mfrow=c(2,2))
plot(modela)
```




```
par(mfrow=c(1,1))
```

Os gráficos mostram que as premissas de independência e igualdade de variância estão compridas. Não mostram qualquer padrão ou tendência dos residuais. A normalidade de todos os grupos podemos presumir porque os grupos de interesse principal, outfielders and infielders, tem suficiente casos para ter confiança na teorema de limite central. Também, a plotagem "Normal Q-Q" mostra uma linha reta exceto nas caudas, que mostra que o conjunto dos dados inteiro tem uma distribuição normal.

R^2 para Modelo de beisebol

O R^2 de nosso modelo avisa que o poder explicatória de nossa variável independente é muito baixo ($R^2 = 0.038$). Este resultado é muito comum em modelos de ANOVA, especialmente com um número pequeno de variáveis que tem múltiplas categorias. O propósito de ANOVA é de julgar se diferenças existem. Mostramos aqui que eles existem entre posições no campo, sim. Mas se nós quisemos focar em quais são as causas dessas diferenças, seria melhor de construir um modelo de regressão com um mix de variáveis categóricas e numéricas que tem a ver com habilidade de rebater a bola.

Como um exemplo, um fator separando os melhores rebatedores e os outros é a habilidade de rebater um arremesso chamado "curve ball". O ex-Governador do Estado de Nova Iorque, falecido em 2015, Mario Cuomo jogou beisebol nas ligas de treinamento de beisebol profissional americano para times afiliados com o Pittsburgh Pirates. Ele nunca subiu além do segundo grau dos "minor leagues" porque, como ele disse, "Eu não podia rebater um curve ball." Invés, ele serviu como governador de um estado, um advogado bem respeitado e quase um candidato para a presidência dos EUA.

Comparações das Categorias – Comparações Múltiplas

Sabemos que alguma diferença entre os grupos existe. Como nós podemos descobrir quais posições são a fonte deste diferença? Ao início, queremos comparar as médias de todos os pares de grupos. Temos quatro grupos (C, DH, IN, OF). Podemos fazer 6 comparações (C vs. DH, C vs. IF, C vs. OF, DH vs. IF, DH vs. OF, IF vs. OF) utilizando um teste-t de duas amostras, mas temos de considerar uma modificação ao nível de α e uma estimativa combinada do desvio padrão incluindo todos os grupos.

Se nós não levamos em consideração o problema de desvio padrão combinado entre os grupos, podemos super-estimar o número de comparações que são significativas. Este quer dizer, em termos de erros, que estaremos fazendo mais erros de Tipo I, identificando um valor como significativo quando não é (um falso positivo). Quando temos comparações múltiplas, precisa aplicar uma correção que vai controlar o que é chamada a *taxa de erro familiar* ("family-wise error rate", FWER).

Correção Bonferroni

O mais tradicional correção para as comparações múltiplas é a correção Bonferroni.[2] Invés de corrigir a probabilidade do valor-p, a Bonferroni muda o α. O novo α é o resultado da divisão da α original por o número de comparações ("C"). Equivalentemente, a correção pode ser calculado em termos de valores-p como o produto do C vezes o valor-p da comparação.

$$\alpha_{Bf} = \frac{\alpha}{C}$$

Para estimar a correção em R, precisa fazer um teste-t para todas as comparações, que podemos fazer com a função `pairwise.t.test()` e usar o argumento `p.adjust.method = "Bonferroni"`.

```
grpmeans <- tapply(bat$OBP, bat$POS, mean)
grpmeans
```

```
##         C        DH        IF        OF
## 0.2786409 0.3054598 0.3058120 0.3116890
```

```
pairwise.t.test(bat$OBP, bat$POS, p.adjust.method = "bonferroni")
```

[2] Dunn, Olive Jean. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56 (March 1961): 52–64. doi:10.1080/01621459.1961.10482090.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  bat$OBP and bat$POS
##
##    C      DH     IF
## DH 0.4122 -      -
## IF 0.0291 1.0000 -
## OF 0.0059 1.0000 1.0000
##
## P value adjustment method: bonferroni
```

Com a correção, podemos ver que catchers são diferentes que infielders e outfielders. Olhando nos OBP's para estas posições, podemos ver que as outras posições tipicamente consegue ganhar um base mais frequentemente que os catchers (valores p de 0.029 e 0.006, os dois abaixo de $\alpha = 0.05$). Mas os outfielders e infielders não mostram alguma diferença com o outro o com os DH's (que normalmente também jogam no outfield ou infield).

Alternativas a Bonferroni

A correção Bonferroni está considerada muito conservadora e pode eliminar muitas comparações significativas incorretamente. Existem duas alternativas que merecem atenção aqui: *a taxa de descoberta falso* ("false discovery rate" FDR), também conhecido pelos nomes d os estatísticos que elaboraram ele, a correção Benjamini-Hochberg.[3] A segunda correção é a *Tukey Diferenças Significativas Honestas* ("Tukey Honest Significant Differences" HSD).[4] A correção de Tukey segue o padrão de um FWER, reduzindo a probabilidade de erros de Tipo I. Mas a FDR tenta de controlar a proporção das descobertas que são falso (rejeições da hipóteses nulas incorretas). Apesar as diferenças entre as correções são teoricamente um pouco abstratas, em operação, o HSD e a FDR são menos rígidos que a Bonferroni e vão aceitar mais comparações como significativas.

Benjamini-Hochberg FDR

Em R, podemos fazer a correção de FDR para ANOVA na mesma maneira que fizemos a Bonferroni. Invés de `p.adjust.method = "Bonferroni"`, agora usamos `p.adjust.method = "BH"`.

```
grpmeans <- tapply(bat$OBP, bat$POS, mean)
grpmeans
```

```
##         C        DH        IF        OF
## 0.2786409 0.3054598 0.3058120 0.3116890
```

```
pairwise.t.test(bat$OBP, bat$POS, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  bat$OBP and bat$POS
##
##    C      DH     IF
## DH 0.1374 -      -
## IF 0.0145 0.9783 -
```

[3] Benjamini, Yoav; Hochberg, Yosef "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *J. Roy. Statist. Soc. Ser. B* 57 (1995): 289–300.

[4] Tukey, John W. "Comparing Individual Means in the Analysis of Variance." *Biometrics* 5 (June 1949): 99. doi:10.2307/3001913.

```
## OF 0.0059 0.7655 0.6169
##
## P value adjustment method: BH
```

Apesar que os resultados grossos ficam o mesmo, agora a diferença entre infielders e outfielders tem um valor-p muito menor que com a Bonferroni.

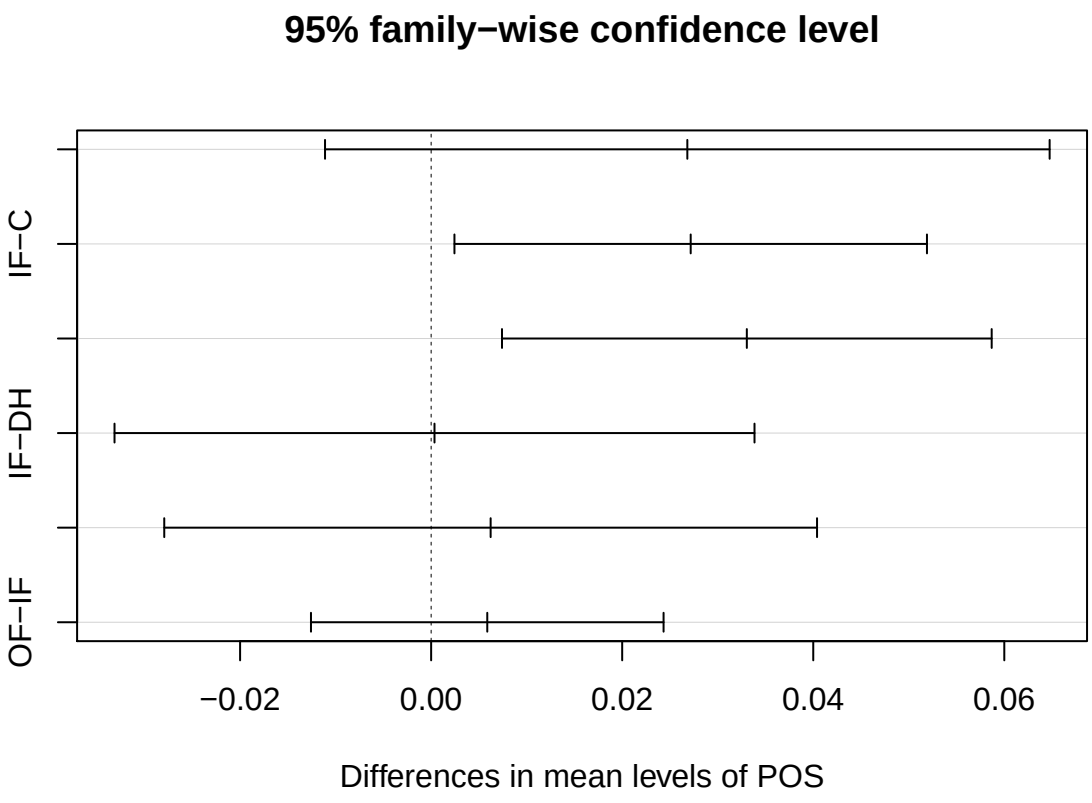
Tukey HSD

A correção de Tukey tem uma função especial que pode trabalhar diretamente com o modelo de ANOVA, como nosso `modela`. A função fornece uma tabela das diferenças entre categorias, um intervalo de confiança e um valor-p ajustado. Além disso, tem um método para `plot` para fazer um gráfico da tabela.

```
TukeyHSD(modela)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = OBP ~ POS, data = bat)
##
## $POS
##               diff          lwr        upr     p adj
## DH-C  0.0268188799 -0.011108364 0.06474612 0.2626619
## IF-C  0.0271711026  0.002439167 0.05190304 0.0248781
## OF-C  0.0330480680  0.007416347 0.05867979 0.0053615
## IF-DH 0.0003522227 -0.033145472 0.03384992 0.9999928
## OF-DH 0.0062291882 -0.027938224 0.04039660 0.9653620
## OF-IF 0.0058769654 -0.012578517 0.02433245 0.8436396
```

```
plot(TukeyHSD(modela))
```



De novo, somente os catchers são abaixo dos infielders e outfielders na categoria de OBP. Você pode ver que os ajustes de valor-p são muito mais próximos aos da Bonferroni que aos de FDR.

Minha preferência é para uso da correção Benjamini-Hochberg. Acho a ideia de simplesmente dividir o α pelo número de comparações não suficiente sofisticada para justificar um castigo tão severo que a Bonferroni impõe.

Outros Tipos de Modelos de ANOVA

Aqui, só tratamos de um tipo de ANOVA, one-way, o modelo mais simples. Existem muitos outros modelos com múltiplas variáveis categóricas independentes. Esses modelos todos precisam de pesos para controlar as diferenças entre tamanhos de categorias das variáveis independentes e ficam para uma aula mais avançada.