MAD-CB

Analise de Variância – ANOVA

Proposito de ANOVA

- Analisar comparações entre três ou mais grupos de um variável
  - Para 2 grupos, usamos testes-t para comparação das médias
- Variável dependente - NUMÉRICA
- Variável(eis) independente(s) - CATEGÓRICA

Quando Têm Mais de 2 Grupos a Comparar

- Testes-t criam dificuldade
  - Provável que achará uma comparação significativa por acaso
  - Mesmo se não existe
  - Aumenta risco de erro Tipo I (falso positivo)
- ANOVA evita esse risco
  - Teste de hipótese é que todos as médias dos grupos são iguais
  - Rejeição da hipótese nula significa que pelo menos uma diferença existe

Teste de Hipótese de ANOVA

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

$$H_1 : \textit{ao menos } 1 \ \mu \textit{ é diferente}$$

- Presença de grandes diferenças entre as médias dos grupos é evidência em favor da rejeição da hipótese nula

Porque Análise de **Variancia**?

- Porque esta técnica é chamada análise de variância quando estamos testando diferenças entre médias e não os desvios padrões?
- Resposta:

  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

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 Beisbol

- Início de nova temporada no último domingo

Fotos para Motivação

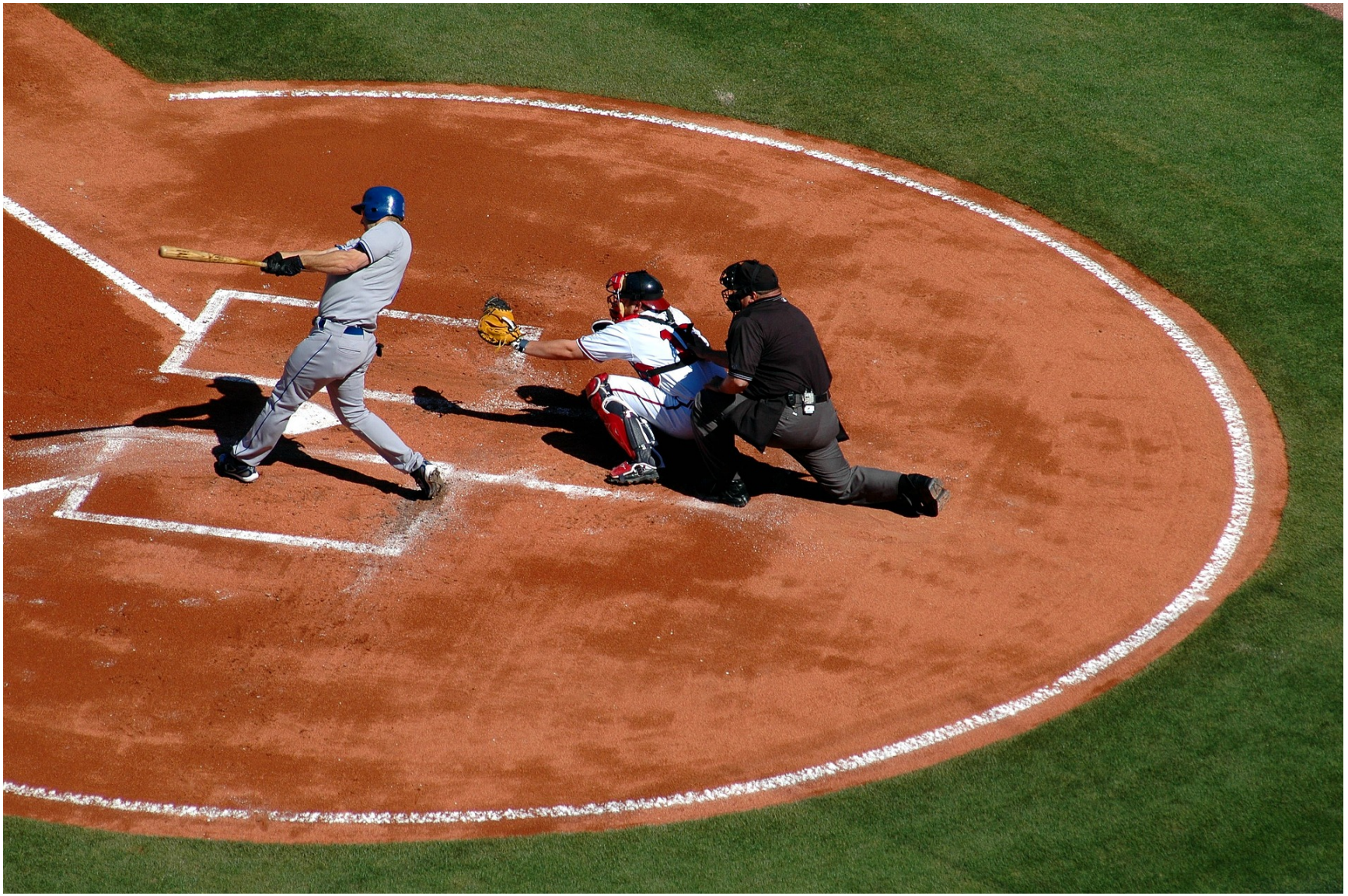


Photo 2 – Dodger Stadium – Minha Equipe

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

Questão que Tentaremos Responder

- Existe diferenças entre a OBP para jogadores nos posições de campo diferentes
- Limitado a American League em 2015
- Dados vêm de base de dados de beisebol "Lahman" (versão em R)
- Simplifiquei as posições para os seguintes:
  - OF – outfielder (left, right ou center)
  - IF – infielder (1B, 2B, 3B ou SS)
  - C – catcher ("receptor", quem recebe os lances do lançador)
  - DH – designated hitter (rebatedor que não joga defesa – só na AL)

OBP?

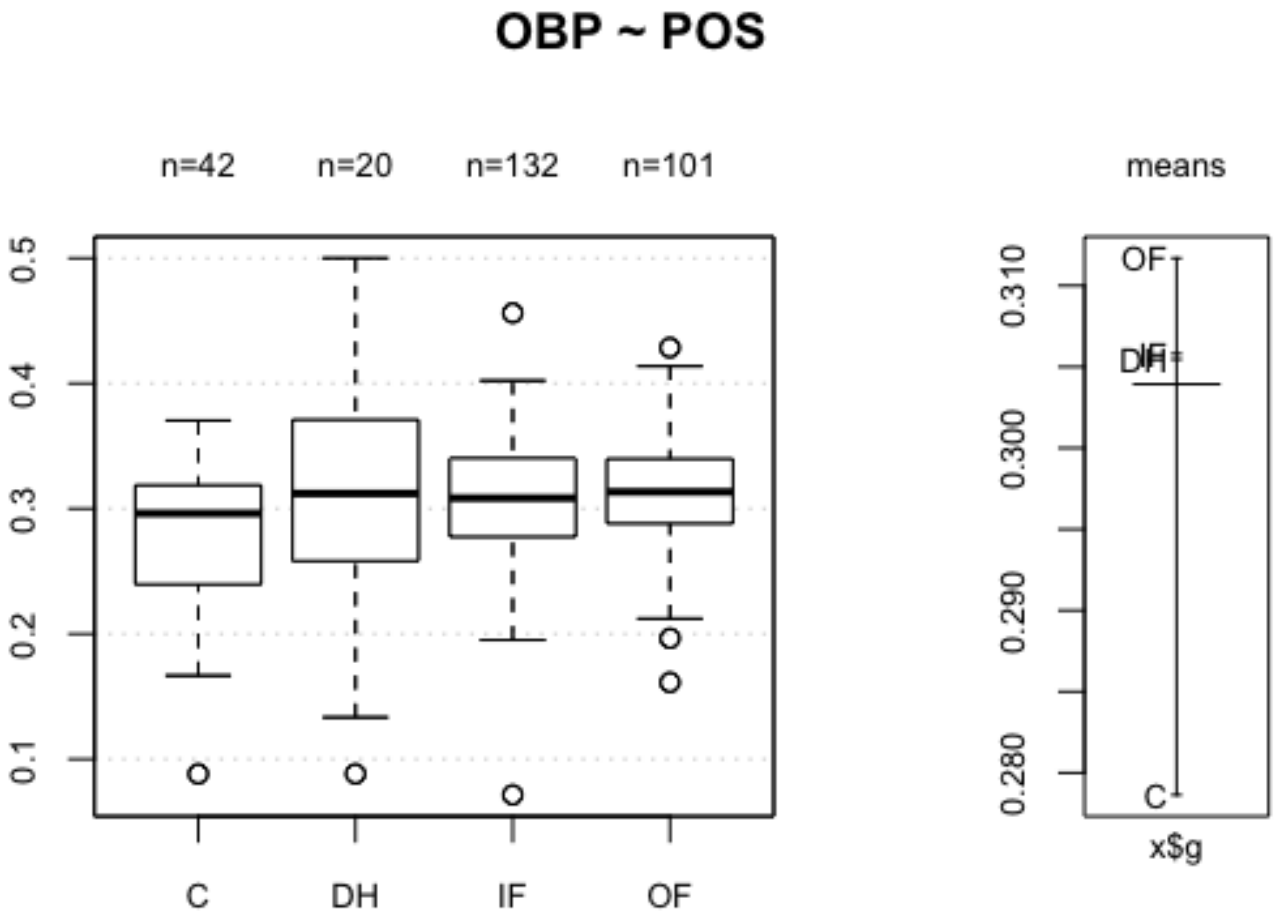
- Porcentagem das vezes que aparece como rebatedor que ganhe pelo menos um base
- Considerado um melhor indicador da habilidade de um rebatedor
- BA – número de rebatidas válidas por at-bats
  - Só uma maneira de ganhar um base
  - at-bats – definição artificial de quantas vezes jogador aparece como rebatedor

Estatísticas Descritivas de OBP por Posição

```
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
```

Boxplot


OBP ~ POS
n=42
n=20
n=132
n=101
0.1
0.2
0.3
0.4
0.5
C
DH
IF
OF
means
0.280
0.290
0.300
0.310
OF
IF
DH
C
x$g

Observações

1. Variação entre os grupos é muito parecido e podemos sentir confortáveis que a premissa #3 está sendo respeitada
2. Boxplot revela que há um outlier longe da caixa para os "infielders"
   - Com uma amostra dentro deste grupo de 132, o outlier não causa preocupação

Teoria de ANOVA

Questão Principal da Técnica

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.)

Como Funciona ANOVA

- Testamos todas as diferenças entre grupos simultaneamente
- Divisão da variação em componentes diferentes
- Usa soma dos quadrados
  - Que vemos primeiro em regressão
- Calcula primeiro soma dos quadrados total (SST)
  - Quadrado das diferenças de todos os valores, não importa o grupo, da média de todos os valores (*grand mean*)

Componentes de Soma de Quadrados

- SSG
  - Soma dos quadrados das diferenças entre a média dos grupos e a grand mean
  - Variação entre os grupos
- SSE
  - O que sobra da variação é por causa dos residuais
  - Soma dos quadrados das diferenças entre todos os valores dentro de um grupo e a média desse grupo
  - Variação dentro dos grupos

Graus de Liberadade (*df*)

- Cada uma das somas de quadrados tem um grau de liberdade associada
- SSG – número de grupos menos 1
  - O 1 representa o *grand mean* que não pode ser variada
- SSE – tamanho de amostra (n) menos o número dos grupos

Formulas para *dfs*

$$df_G = k - 1$$

$$df_E = n - k$$

Estatística F

- A estatística que teste a hipótese nula mede a relação entre os dois componentes divididos pelos graus de liberdade
  - MSG e MSE
- Estatística tem a distribuição “F”
- Formula para calcular F:

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

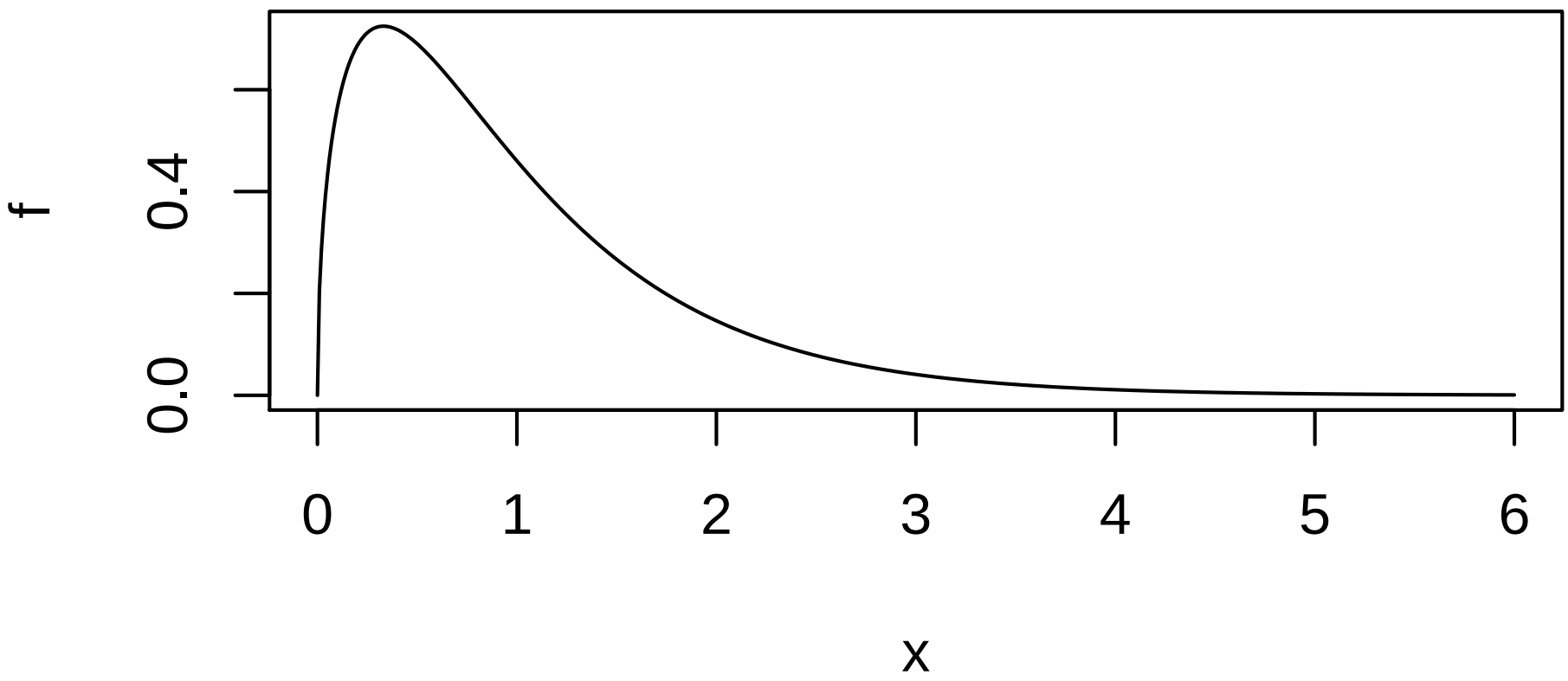
Calculo de *dfs* para OBP e POS

```
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"
```

Forma da Distribuição F

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



ANOVA em R

- Pode usar 2 funções
  - `aov()`**
  - `lm()`
- Diferença entre as 2
  - Na apresentação dos resultados
  - `aov()` foca no modelo e o teste F
  - `lm()` foca mais sobre parâmetros das variáveis independentes
- Especificação do modelo
  - Mesmo que usamos em regressão

ANOVA de OBP e POS

```
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
```

2 Funções para Ajudar Interpretação dos Resultados

```r
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)
}
```

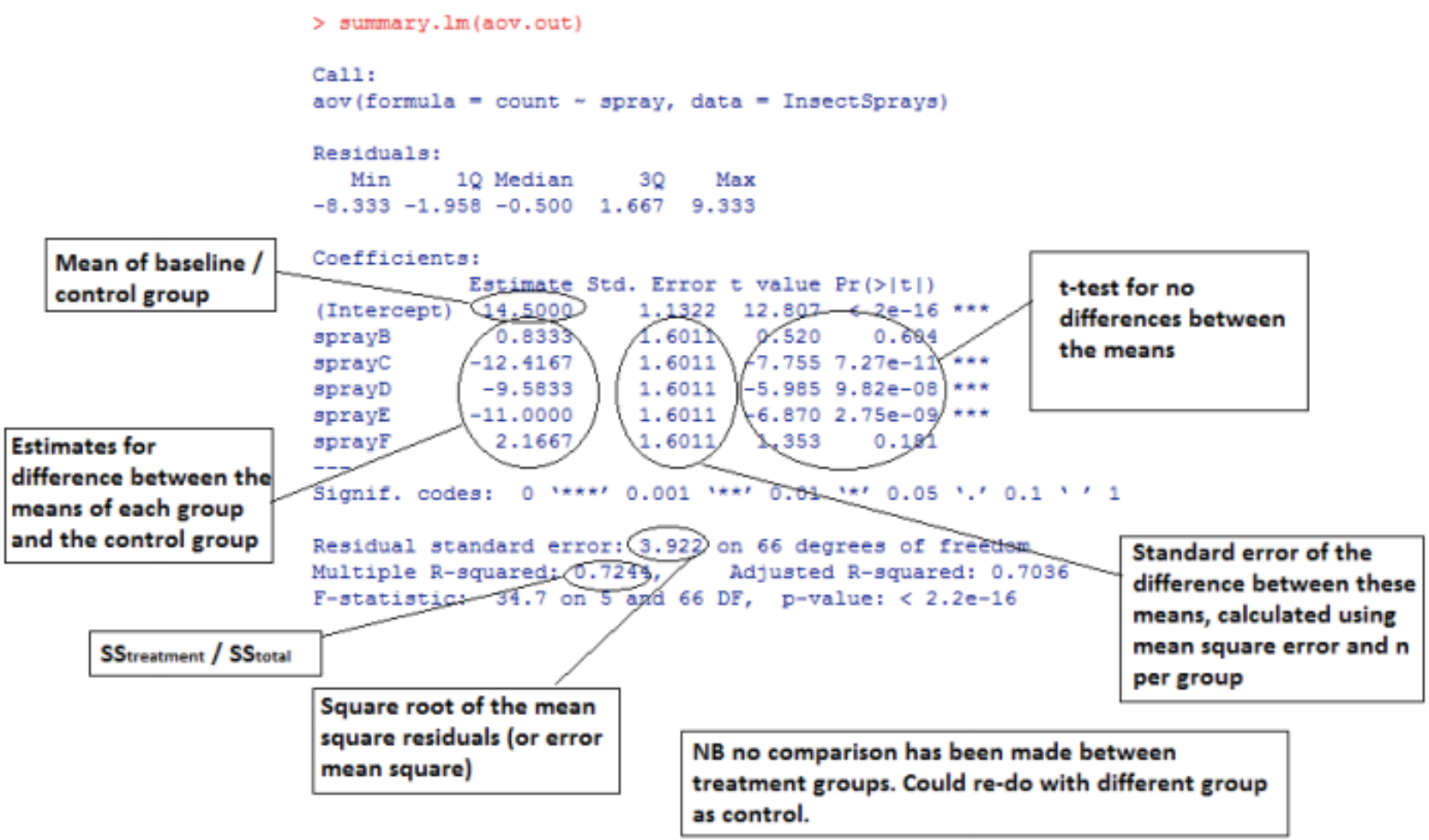
Interpretação dos Resultados

- Existe uma diferença significativa entre as posições em OBP.
- Valor p do teste-F (0.0104)
  - Abaixo do valor de α (0.05)
- Pode rejeitar a hipótese nula (H_0)
  - As diferenças entre as médias são significativas
- Antes de determinar quais diferenças são significativas
  - Precisa ver se o modelo cumpriu as premissas

Resumo `lm` de um Modelo de ANOVA

- Pode mostrar um resumo no formato de um modelo linear (regressão)
- Porém, muito da informação não é útil para analise
- Resumo está disponível com a função 'summary.lm()'

Elementos de 'summary.lm()'



'summary.lm()' do Modelo OBP~POS

```
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 .
## 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
```

Interpretação do Modelo

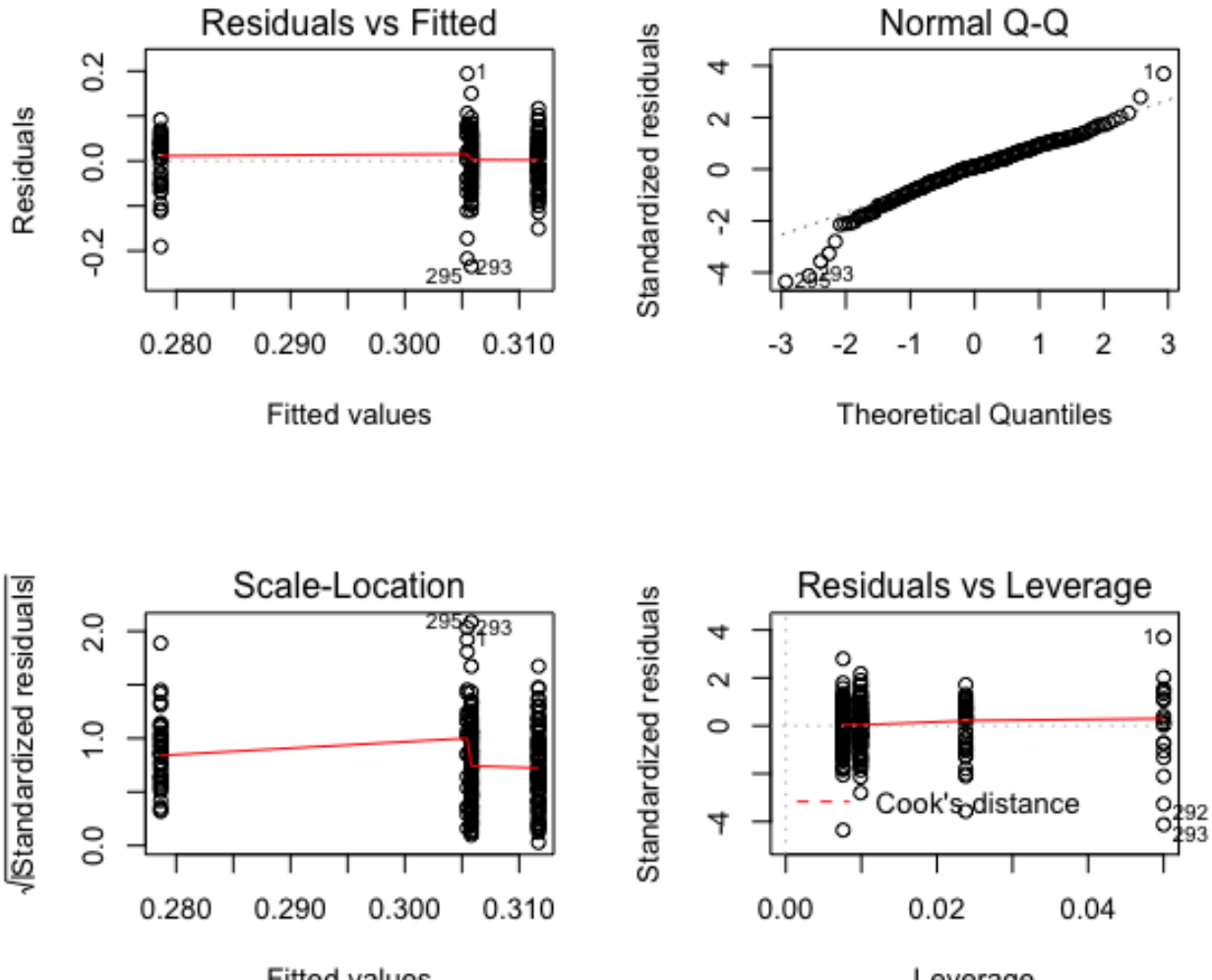
- “baseline/control group” é a categoria de catcher
- Estatísticas com Estimate e Std. Error não têm muito utilidade
  - categorias não são preditivas para OBP
- Temos outra maneira de comparar as categorias

Validade do Modelo

- Fazemos isso com gráficos
  - Como na regressão linear
- Função `plot()` produz os mesmos 4 gráficos

4 Gráficos

```
par(mfrow=c(2,2))
plot(modela)
par(mfrow=c(1,1))
```



Modelo Válido

- Premissas de *independência* e *igualdade de variância* estão compridas
  - Não mostram qualquer padrão ou tendência dos residuais
- Normalidade dos grupos
  - Podemos presumir porque grupos de interesse principal, outfielders and infielders, tem suficiente casos para ter confiança na teorema de limite central
- Plotagem "Normal Q-Q" mostra uma linha reta exceto nas caudas

R^2 para Modelo de Beisbol

- $R^2 = 0.038$
- Resultado é muito comum em modelos de ANOVA
  - Número pequeno de variáveis que tem múltiplas categorias
- Propósito de ANOVA é de julgar se diferenças existem
  - Aqui, SIM
- Se quisermos entender as causas dessas diferencias
  - Construir um modelo de regressão
  - Usando mix de variáveis categóricas e numéricas que tem a ver com a habilidade de rebater a bola
  - "I couldn't hit a curve ball" - Gov. Mario Cuomo (NY)

Comparações das Categorias – Comparações Múltiplas

- Sabemos que alguma diferencia existe
- Quais posições são a fonte desta diferença?
- Temos 6 comparações que queremos fazer
  - C vs. DH
  - C vs. IF
  - C vs. OF
  - DH vs. IF
  - DH vs. OF
  - IF vs. OF

Uso de Teste-t

- Pode usar um teste-t (ou equivalente não-paramétrico) de 2 amostras para testar as 6 comparações
- Precisa ajustar o nível de α ou o valor-p para não super-estimar o número de comparações significativas
- Controlar os erros de Tipo I
- Alias, controla a *taxa de erro familiar* ("family-wise error rate", FWER)

Correção Bonferroni

- Mais tradicional correção para as comparações múltiplas
- Bonferroni muda o α
- Novo α é o resultado da divisão da α original por o número de comparações ("C")
- Correção pode ser calculado em termos de valores-p como o produto do C vezes o valor-p da comparação

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

Bonferroni em R

- Precisa fazer um teste-t para todas as comparações
- Utilizar a função `pairwise.t.test()`
- Com argumento `p.adjust.method = "Bonferroni"`
- Não pode calcular diretamente com o `summary.aov()`

Bonferroni com Modelo

```
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")
```

```
##
##  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 Correção

- Catchers são diferentes que infielders e outfielders
- Outras posições tipicamente consegue ganhar um base mais frequentemente que os catchers
  - Valores-p de 0.029 e 0.006, os dois abaixo de $\alpha = 0.05$
  - OF e IF não mostram alguma diferença com o outro o com os DH's

Alternativas a Bonferroni

- Bonferroni considerada muito conservadora
  - Elimina muitas comparações significativas incorretamente
- 2 alternativas
- *taxa de descoberta falso* ("false discovery rate" FDR)
- Também conhecido como correção Benjamini-Hochberg
- *Tukey Diferenças Significativas Honestas* ("Tukey Honest Significant Differences" HSD)
- Tukey segue o padrão de FWER – reduzir erros de Tipo I
- FDR tenta de controlar a proporção das descobertas que são falso (rejeições da hipóteses nulas incorretas)

Benjamini-Hochberg FDR

- Utiliza a mesma função para ANOVA que usamos para ver a Bonferroni.
- `p.adjust.method = "BH"`

FDR para Modelo

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

Resultados de FDR

- Conclusões são as mesmas
- Diferenças mostram valor-p muito menor que com a Bonferroni

Tukey HSD

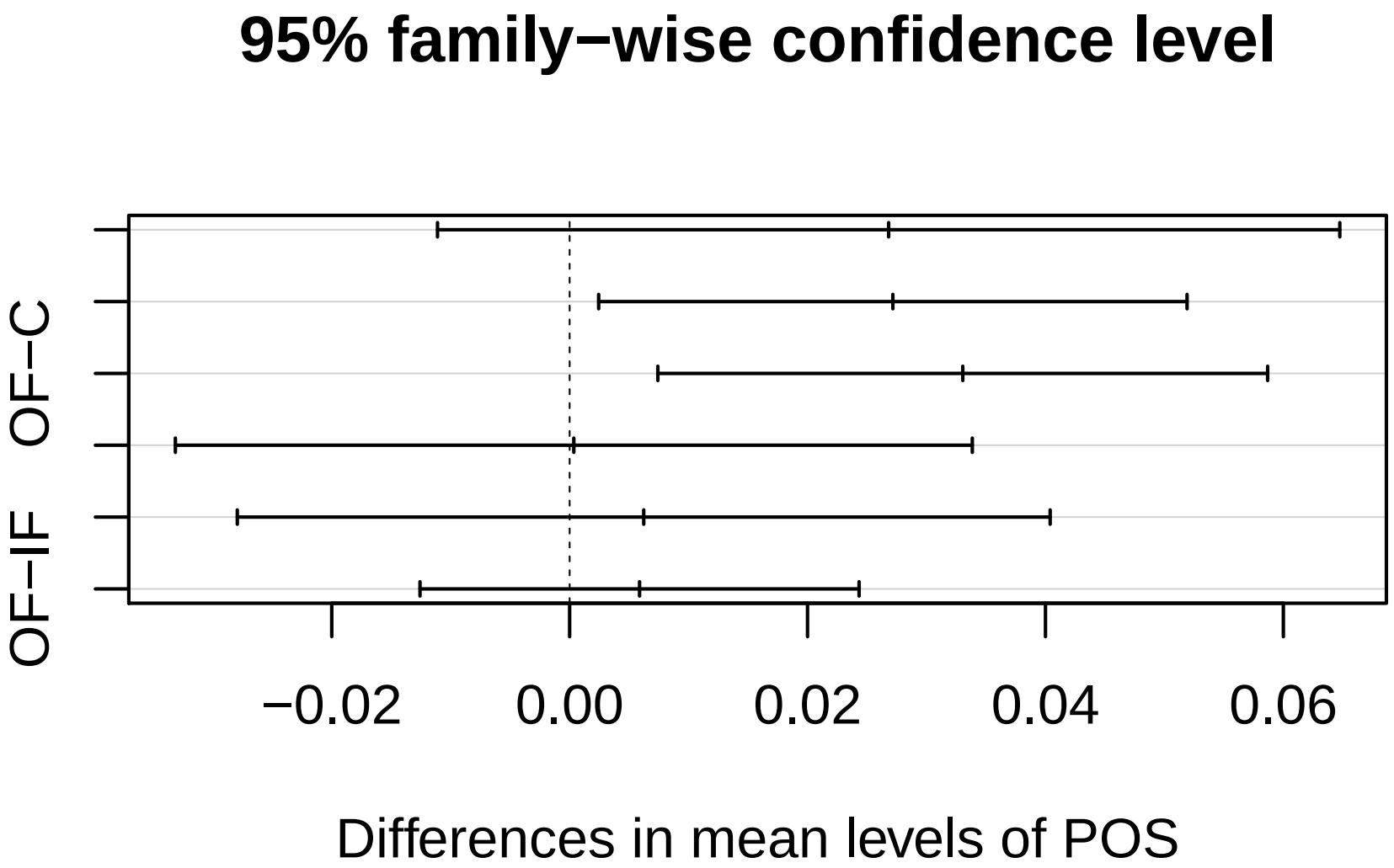
- Tem uma função especial
  - Trabalha diretamente com o modelo de ANOVA
  - `TukeyHSD()`
- Produz para cada comparação
  - Tabela das diferenças entre categorias
  - Intervalo de confiança
  - Valor-p ajustado
- Tem um método para `plot()` que produz gráfico da tabela

Tukey HSD para Modelo

```
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))
```



Resultados de Tukey HSD

- Mesmas conclusões
- Valores-p mais perto a Bonferroni que a FDR
  - Por causa de FWER

Qual Comparação Eu Uso

- Prefiro FDR
- Acho FWER um approach antigo e um castigo em que perdemos informação importante
- FDR muito mais sofisticado como approach

Outros Tipos de Modelos de ANOVA

- Só tratamos um tipo de ANOVA
  - "One way"
- Há muitos outros tipos
- Multiplás variáveis independentes
- Todos esses modelos precisam "pesos"
  - Controle das diferenças entre tamanhos de categorias independentes

Próxima Semana – Machine Learning