Inferência em Dados Categoricos

André M Ribeiro-dos-Santos

13/03/2017

Objetivos

- Reconhecer variáveis categoricas.
- Identificar quando aplicar teste binomial.
- Identificar quando aplicar um teste de qui-quadrado e/ou Fisher.
- Compreender qual a pergunta estatística resolvida por cada teste.
- Realizar os testes.
- Visualizar e Ilustrar os resultados.

Testes de Proporção

Imagine...

Você estuda Fibrose Cística na população de Belém. Você começou a reunir amostras de casos da doença no HUJBB. Na primeira coleta, você obteve 10 amostras das quais 7 (70%) apresentam um quadro de hipoproteinemia. No entanto, espera-se que apenas 45% dos casos aprensetem hipoproteinemia. *A elevada incidência pode indicar um importante fator genético atuando na população.*

A frequência de hipoproteinemia diferencia-se da esperada?

$$H_o : \hat{p} = p$$

$$H_a = \hat{p} \neq p$$

Avaliando o problema

Quais as características do experimento?

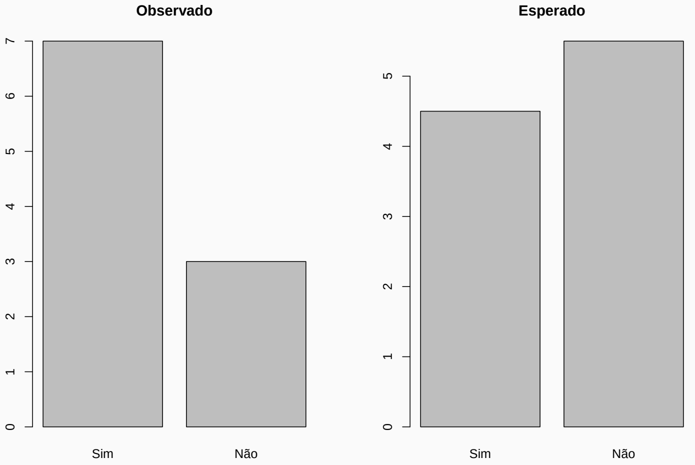
- Qual Tamanho da amostra?
- Qual a variável em questão?
- A variável é numérica ou categorica?
- Quantos casos foram observados e quantos seriam esperados?

Avaliando o problema

Quais as características do experimento?

- Qual Tamanho da amostra? **10**
- Qual a variável em questão? **presença de hipoproteinemia**
- A variável é numérica ou categorica? **categorica, com duas respostas**
- Quantos casos foram observados e quantos seriam esperados? **Foram observados 7, mas esperaria-se 4.5**

```
> par(mfrow = c(1,2))
> barplot(c("Sim" = 7,   "Não" = 3),   main = "Observado")
> barplot(c("Sim" = 4.5, "Não" = 5.5), main = "Esperado")
```



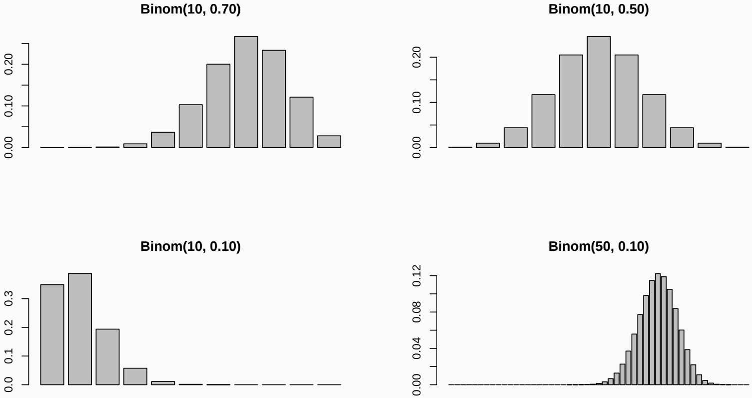
Distribuição Binomial $B(n, p)$

A distribuição de **variáveis categoricas com duas altertivas (ou grupos)** são descritas por uma **distribuição binomial** (ou de bernoulli), a qual é a probabilidade *discreta* de observa-se x "sucessos" em n independente experimentos com probabilidade p.

$$P(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

```
> dbinom(x, size, prob, log = FALSE)
> pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
> qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
> rbinom(n, size, prob)
```

```r
> par(mfrow = c(2,2))
> barplot(dbinom(0:10, size = 10, p = 0.7), main = "Binom(10, 0.70)")
> barplot(dbinom(0:10, size = 10, p = 0.5), main = "Binom(10, 0.50)")
> barplot(dbinom(0:10, size = 10, p = 0.1), main = "Binom(10, 0.10)")
> barplot(dbinom(0:50, size = 50, p = 0.7), main = "Binom(50, 0.10)")
```



Teste Binomial

Uma vez que a variável em questão segue a *distribuição binomial*, podemos avaliar se frequencia observada diferencia-se do esperado usando um **teste binomial**.

```
> ? binom.test
> ## Exact Binomial Test
> ##
> ## Description:
> ##    Performs an exact test of a simple null
> ##    hypothesis about the probability of
> ##    success in a Bernoulli experiment.
> ## Usage:
> ##    binom.test(x, n, p = 0.5,
> ##      alternative = c("two.sided", "less", "greater"),
> ##      conf.level = 0.95)
> ## ...
```

Teste Binomial

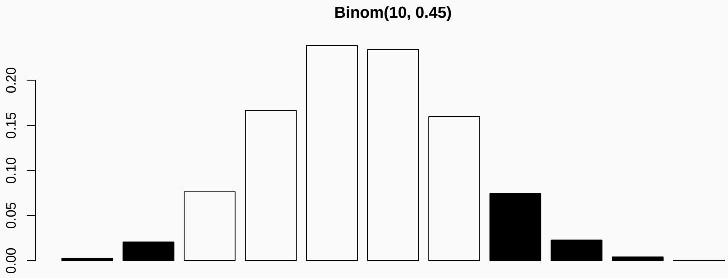
A frequência de hipoproteinemia diferencia-se da esperada?

$$H_o : \hat{p} = p; \quad H_a = \hat{p} \neq p$$

```
> binom.test(7, n = 10, p = 0.45)
```

```
##
##  Exact binomial test
##
## data:  7 and 10
## number of successes = 7, number of trials = 10, p-value = 0.1253
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.3475471 0.9332605
## sample estimates:
## probability of success
##                    0.7
```

```r
> marked <- (0:10 < floor(7 - 4.5)) | (0:10 >= 7)
> probs  <- dbinom(0:10, size = 10, p = 0.45)
> barplot(probs, col = marked, main = "Binom(10, 0.45)")
```




```r
> sum(probs[marked])
```

```
## [1] 0.125252
```

Exercícios - Teste de Proporção

Questão (1)

Após a segunda coleta de amostras, os pesquisadores observaram mais 11 casos de hipoproteinemia entre as 15 amostras coletadas. Após reunir as amostras, qual a frequêcia de hipoproteinemia observada? Os pesquisadores constataram diferenças significativas?

```r
> x    <-  7 + 11
> n    <- 10 + 15
> x / n
```

```
## [1] 0.72
```

```r
> binom.test(x, n, p = 0.45)
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 18, number of trials = 25, p-value =
## 0.008143
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.5061232 0.8792833
## sample estimates:
## probability of success
##                   0.72
```

Questão (2)

Qual foi o *Odds-Ratio* ou risco relativo de hipoproteinemia em relação ao esperado?

$$
\begin{aligned}
O(p) &= p/(1-p) \\
OR &= \frac{O(p_o)}{O(p_e)} \\
&= \frac{p_o/(1-p_o)}{p_e/(1-p_e)}
\end{aligned}
$$

```r
> p_o <- x / n
> p_e <- 0.45
>
> (p_o / (1 - p_o))
```

```
## [1] 2.571429
```

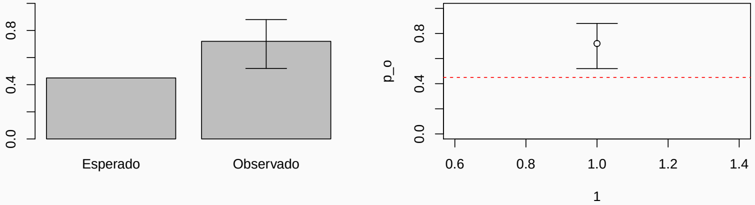
Questão (3)

Qual o intervalo de confiança da proporção observada? Represente o valor em gráficos.

```
> (cnfint <- qbinom(c(0.025, 0.975), n, p_o) / 25)
```

```
## [1] 0.52 0.88
```

```
> par(mfrow = c(1,2))
>
> barplot(c("Esperado" = p_e, "Observado" = p_o), ylim=c(0, 1))
> arrows(1.9, cnfint[1], 1.9, cnfint[2], code = 3, angle = 90)
>
> plot(1, p_o, ylim=c(0,1))
> abline(h = p_e, lty = 'dashed', col = 'red')
> arrows(1,  p_o + c(-0.02, 0.02), 1, cnfint, angle=90)
```



Questão (4)

A hipoproteinemia ocorre com frequência associada a anemia, os pesquisadores decidiram avaliar se anemia também estava alterada na amostra. Dado a tabela em anexo, avalie se houve mudança em relação a proporção de anemicos esperados (35%). Qual a conclusão?

```
> cftr <- read.table('cftr-ex.tsv', header = T)
```

```
> table(cftr$anemia)
```

```
##
## Não Sim
##  27  23
```

```
> binom.test(table(cftr$anemia), p = 1 - 0.35)
```

```
##
##  Exact binomial test
##
## data:  table(cftr$anemia)
## number of successes = 27, number of trials = 50, p-value = 0.1052
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.3932420 0.6818508
## sample estimates:
## probability of success
##                   0.54
```

Sua vez...

5. Dada uma variável x corresponde ao número de sucessos em uma série de 50 experimentos de Bernoulli com probabilidade de sucesso 30%. Calcule:
   - $P(x = 15)$, $P(x >= 15)$, e \$P(x < 15).
   - $P(15 < x <= 35)$, ilustre num gráfico da distribuição àrea selecionada.
   - Quantiles $x_{0.025}$ e $x_{0.975}$.
6. Qual a frequência do sexo e responsta ao tratamento (*response*) na amostra de **cftr**?
7. A amostra apresenta uma proporção similar de homens e mulheres?

$P(x = 15)$, $P(x >= 15)$, e $P(x < 15)$?

```
> dbinom(15, 50, 0.3)
```

```
## [1] 0.1223469
```

```
> ## sum(dbinom(15:50, 50, 0.3))
> pbinom(14, 50, 0.3)
```

```
## [1] 0.4468316
```

```
> ## sum(dbinom(0:14,  50, 0.3))
> pbinom(14, 50, 0.3, lower.tail = F)
```

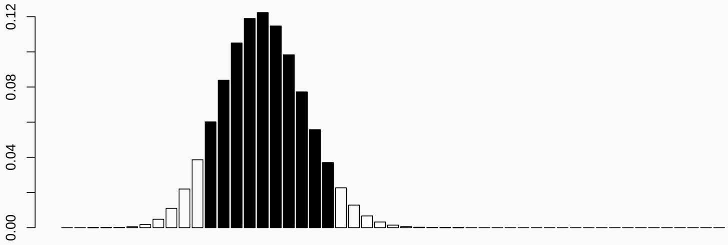
```
## [1] 0.5531684
```

$P(15 < x <= 35)$, ilustre num gráfico da distribuição a àrea selecionada.

```
> ## sum(dbinom(9:25, size = 50, p = 0.3))
> pbinom(25, 50, 0.3) - pbinom(10, 50, 0.3)
```

```
## [1] 0.9202162
```

```
> barplot(dbinom(0:50, 50, 0.3), col = (0:50 > 10) & (0:50 <= 20))
```



Quantiles $x_{0.025}$ e $x_{0.975}$.

```
> qbinom(c(0.025, 0.975), 50, 0.3)
```

```
## [1]  9 22
```

Qual a frequência do sexo e responsta ao tratamento (*response*) na amostra de **cftr**?

```
> prop.table(table(cftr$sexo))
```

```
##
##    F    M
## 0.58 0.42
```

```
> prop.table(table(cftr$response))
```

```
##
##    I   II  III
## 0.22 0.52 0.26
```

A amostra apresenta uma proporção similar de homens e mulheres?

```
> binom.test(table(cftr$sexo))
```

```
##
##  Exact binomial test
##
## data:  table(cftr$sexo)
## number of successes = 29, number of trials = 50, p-value = 0.3222
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.4320604 0.7181178
## sample estimates:
## probability of success
##                   0.58
```

Testes de Independência

Uma questão de independência

Ao pesquisar os sintomas de *Fibrose Cistíca*, os pesquisadores constataram que *anemia* geralmente acompanha os pacientes com *hipoproteinemia*. Sendo a ocorrência de anemia muito mais comun entre pacientes com *hipoproteinemia*. Na amostra estudada, eles observaram a seguinte tabela de confunsão:

	Anemia	
Hipoproteinemia	Não	Sim
Não	2	6
Sim	11	7

A anemia é distribuida independentemente da hipoproteinemia?

Avaliando o problema

Quais as características do experimento?

- Qual Tamanho da amostra?
- Quais as variáveis em questão?
- As variáveis são numérica ou categorica?
- Descreva o problema estatístico em questão.

Avaliando o problema

Quais as características do experimento?

- Qual Tamanho da amostra? **25**
- Quais as variáveis em questão? **presença de hipoproteinemia e anemia**
- As variáveis são numérica ou categorica? **ambas categoricas, com duas respostas**2
- Descreva o problema estatístico em questão. **deseja-se avaliar se o conhecimento da presença de anemia afeta a chance do individuo apresentar hipoproteinemia**

Teste de Fisher

Uma vez que desejamos avaliar a independência entre duas variáveis catergoricas com duas possíveis respostas, podemos utilizar o **teste de Fisher**.

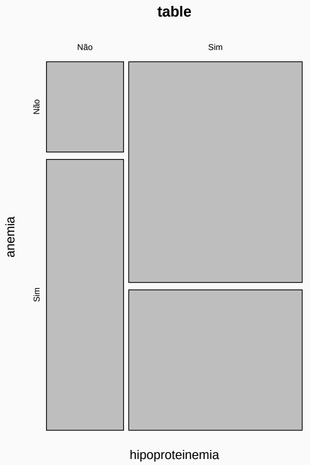
$$Ho : P(B|A) = P(B); \quad Ha : P(B|A) \neq P(B)$$

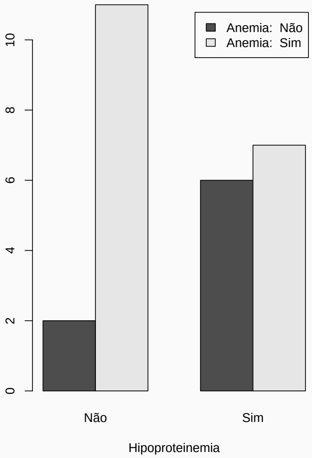
```
> ? fisher.test
> ## Fisher's Exact Test for Count Data
> ## Description:
> ##      Performs Fisher's exact test for testing the null of
> ##      independence of rows and columns in a contingency
> ##      table with fixed marginals.
> ## Usage:
> ##      fisher.test(x, y = NULL, workspace = 200000,
> ##                  hybrid = FALSE, control = list(),
> ##                  or = 1, alternative = "two.sided",
> ##                  conf.int = TRUE, conf.level = 0.95,
> ##                  simulate.p.value = FALSE, B = 2000)
> ##...
```

```r
> table <- matrix(c(2, 11, 6, 7), nrow=2,
+                 dimnames = list("hipoproteinemia" = c("Não", "Sim"),
+                                 "anemia"          = c("Não", "Sim")))
> addmargins(table)
```

```
##                anemia
## hipoproteinemia Não Sim Sum
##             Não   2   6   8
##             Sim  11   7  18
##             Sum  13  13  26
```

```
> par(mfrow=c(1, 2))
> mosaicplot(table)
> barplot(table, beside=TRUE, xlab = "Hipoproteinemia",
+         legend=paste("Anemia: ", rownames(table)))
```





```r
> fisher.test(table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table
## p-value = 0.2016
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0175156 1.7390386
## sample estimates:
## odds ratio
##  0.2257303
```

Uma nova dúvida

Os pesquisadores também acreditam que a presença de *hipoproteinemia* está associado ao *risco* de resposta adversa ao tratamento. O risco é dividido em três clases (I, II, e III). Ao avaliar os pacientes estudados, obtiveram a seguinte tabla de confusão:

	Hipoproteinemia	
Resposta	Não	Sim
I	0	6
II	3	10
III	3	3

O risco é independente da hipoproteinemia?

Teste de Qui-quadrado

Neste caso, uma da variavéis em questão possui três possíveis respostas, não sendo mais adequado utilizar o **teste de Fisher**, no lugar podemos utilizar o **teste de Qui-quadrada** para avaliar a independência.

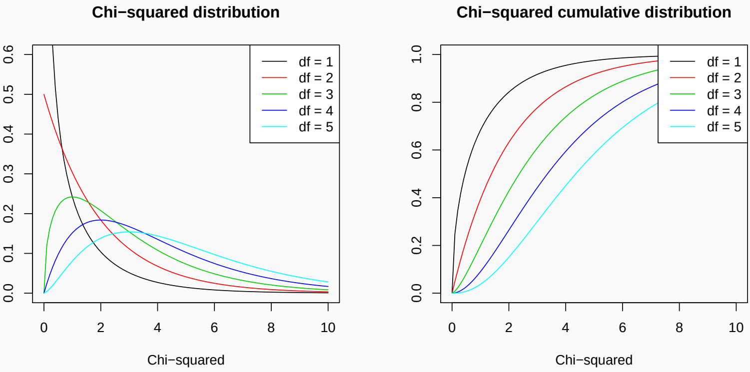
$$Ho : P(B|A) = P(B); \quad Ha : P(B|A) \neq P(B)$$

O teste usa a dsitribuição qui-quadrada ou χ^2 para avaliar o teste.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

A distribuição Qui-quadrada ou $\chi^2(k)$ com k graus de liberdade representa a distribuição da soma dos quadrados de k independentes curvas normais.

$$df = (|linhas| - 1) * (|colunas| - 1)$$



```
> ? chisq.test
> ## Pearson's Chi-squared Test for Count Data
> ## Description:x
> ##      ‘chisq.test’ performs chi-squared contingency table
> ##      tests and goodness-of-fit tests.
> ## Usage:
> ##      chisq.test(x, y = NULL, correct = TRUE,
> ##                 p = rep(1/length(x), length(x)),
> ##                 rescale.p = FALSE,
> ##                 simulate.p.value = FALSE, B = 2000)
```

```r
> obs  <- matrix(c(0, 3, 3, 6, 10, 3), nrow = 3,
+               dimnames = list("Resposta" = c("I", "II", "III"),
+                               "Hipoproteinemia" = c("Não", "Sim")))
> exp  <- rowSums(obs) %*% t(colSums(obs)) / sum(obs)
> (chi <- sum((obs-exp)^2/exp))
```

```
## [1] 4.124494
```

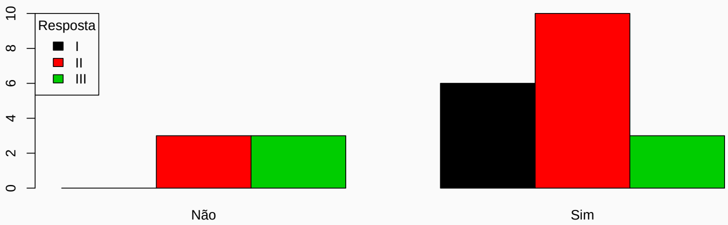
```r
> df   <- (nrow(obs) - 1) * (ncol(obs) - 1)
> 1 - pchisq(chi, df = df)
```

```
## [1] 0.1271679
```

```r
> chisq.test(obs)
```

```
##
##  Pearson's Chi-squared test
##
## data:  obs
## X-squared = 4.1245, df = 2, p-value = 0.1272
```

```r
> barplot(obs, beside = T, col = 1:3)
> legend("topleft", rownames(obs), fill = 1:3, title = "Resposta")
```



Quando usar Fisher ou Qui-quadrado?

- Mais indicado **Teste de Fisher**:
  - Tabela de contigência 2x2.
  - Preferido para pequenas amostragens.
- Mais indicado **Qui-quadrado**:
  - alguma das variáveis possui 3 ou mais categorias.
  - o tamanho amostral maior do que 1000.

Exercícios

Resolução

Ao Final. . .

O que vimos?

Até a próxima