Inferência em Dados Categoricos

André M Ribeiro-dos-Santos

13/03/2017

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 amostra diferencia-se da proporção esperada?

Objetivos

- Reconhecer variáveis categoricas.
- Identificar quando aplicar teste binomial.
- Identificar quando aplicar um teste de qui-quadrado e/ou Fisher.
- Compreender qual a questão estatística resolvida por cada teste.
- Reproduzir os testes em R.
- Saber como e com quais ferramentas responder as questões anteriores.

Avaliando o problema

Quais as características do experimento?

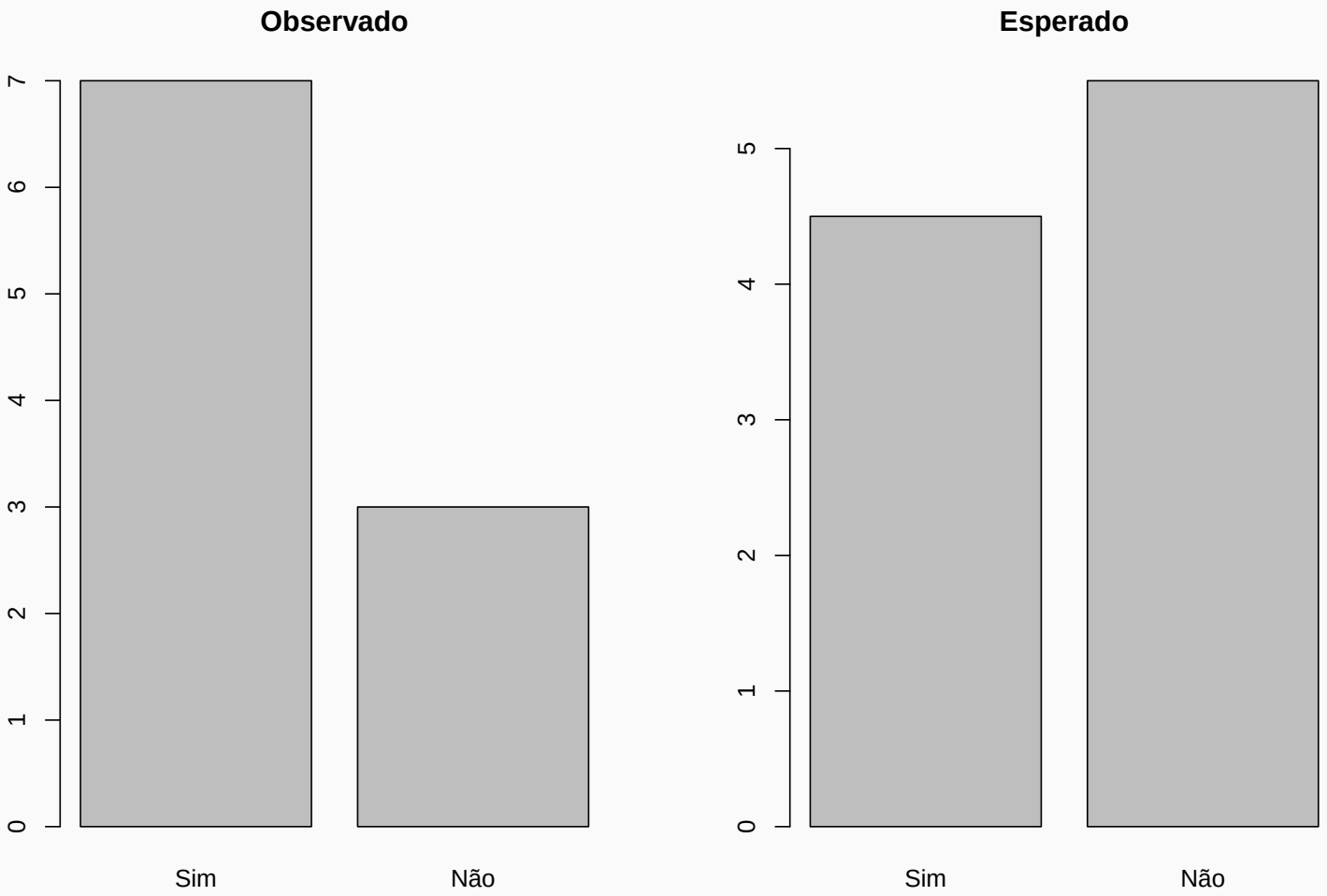
- Qual Tamanho da amostra?
- Qual a variável em questão?
- A variável é numérica ou categorica?

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**

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



Testes de Proporção

Teste Binomial

Como testar se a proporção observada é diferente da esperada?

O teste estatístico mais indicado é o **teste binomial**. Este basei-se nas probabilidades obtidas num experimento de Bernolli (ou *Bernoulli trial*), que consiste em qualquer experimento estatítico com apenas duas respostas (e.g. *sucesso* ou *falha*). O teste avalia, portanto, o número de *sucessos* numa amostra e compara a uma proporção esperada.

$$Ho : \hat{p} = p; \;\; Ha : \hat{p} \neq p \qquad Bilateral$$

$$Ho : \hat{p} = p; \;\; Ha : \hat{p} < p \qquad Menor$$

$$Ho : \hat{p} = p; \;\; Ha : \hat{p} > p \qquad Maior$$

```
> ? 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)
> ## ...
```

A amostra diferencia-se da proporção esperada?

```
> 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
```

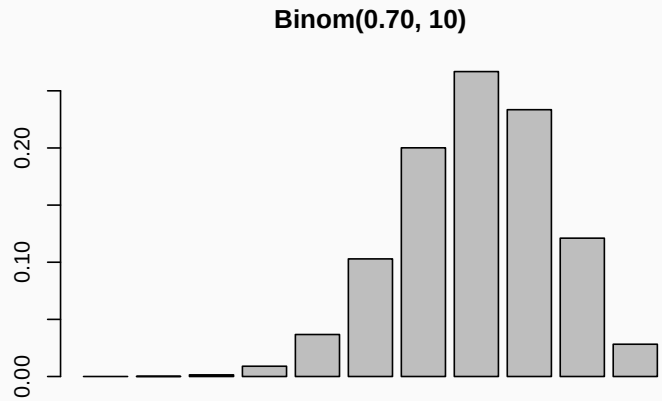
Distribuição Binomial

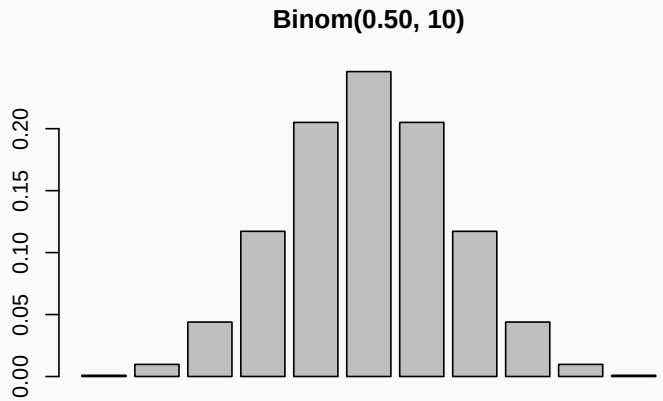
$$Binom(n, p)$$

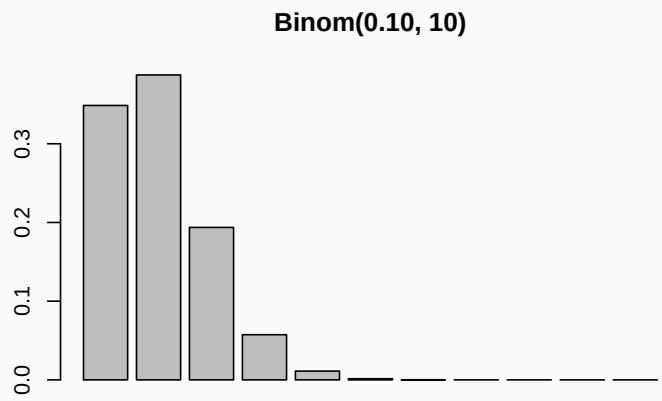
Esta distribuição descreve as probabilidade observadas num experimento de Bernoulli. Corresponde a probabilidade de observar-se x sucessos em n experimentos.

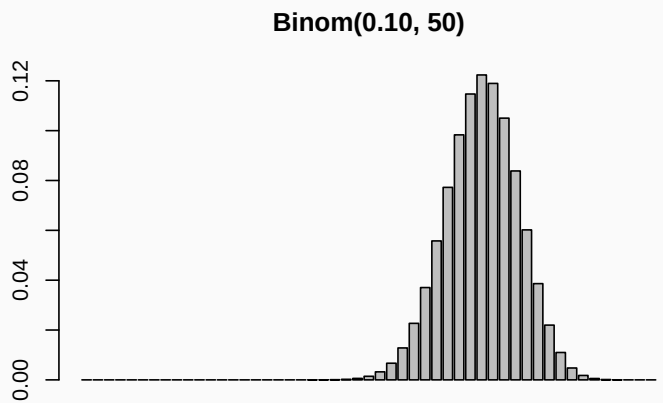
$$P(x; n, p) = \binom{n}{x} p^x q^{n-x}$$

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

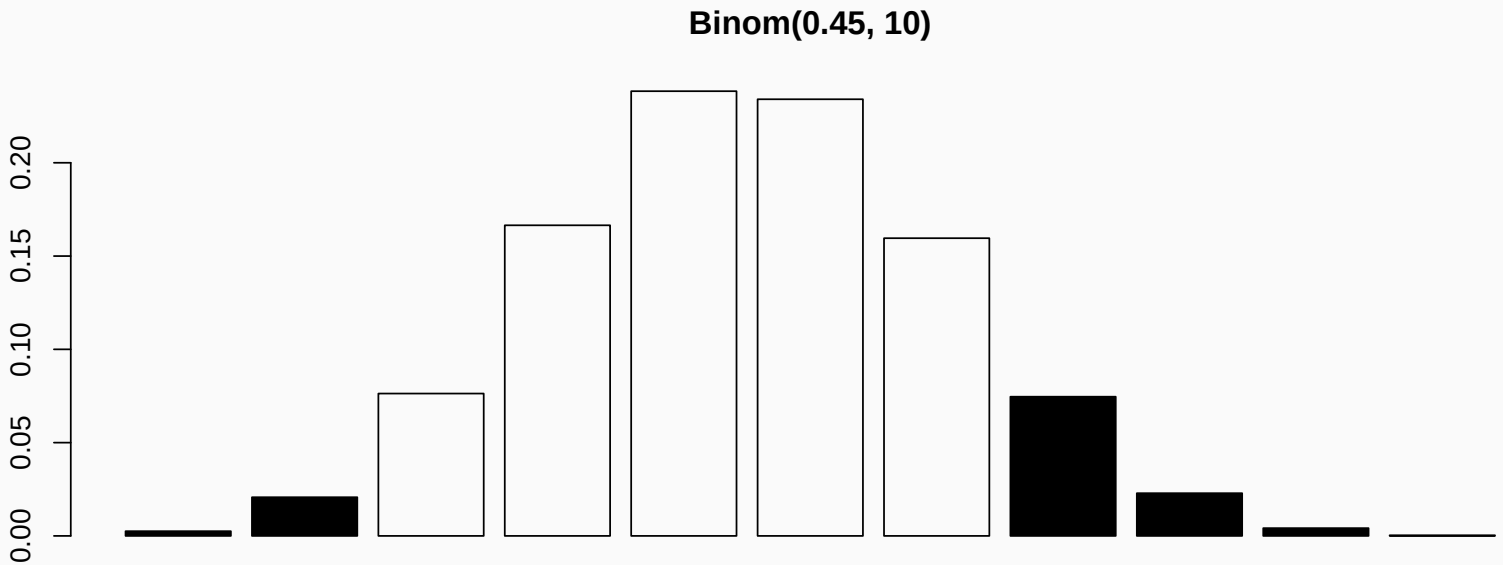








```
> marked <- (0:10 < floor(7 - 4.5)) | (0:10 >= 7)
> probs  <- dbinom(0:10, size = 10, p = 0.45)
> barplot(probs, col = marked, main = "Binom(0.45, 10)")
```




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

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

Exercícios

1. Após a segunda coleta de amostras, os pesquisadores observaram mais 12 casos de hipoproteinemia entre as 15 amostras coletadas. Ao reunir ambas as amostras, os pesquisadores podem constatar diferenças significativas entre os casos de hipoproteinemia observado e esperado?
2. Qual foi o *Odds-Ratio* ou risco relativo da amostra?

$$OR = \frac{p_o/(1-p_o)}{p_e/(1-p_e)}$$

3. Qual o intervalo de confiança da proporção observada? Represente o intervalo com um gráfico.
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?

Resolução (1)

```
> x <- 7 + 12
> n <- 10 + 15
>
> binom.test(x, n, p = 0.45)


##
##  Exact binomial test
##
## data:  x and n
## number of successes = 19, number of trials = 25, p-value =
## 0.002122
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.5487120 0.9064356
## sample estimates:
## probability of success
##                   0.76
```

Resolução (2)

```
> p_obs <- x / n
> p_exp <- 0.45
>
> (p_obs/ (1 - p_obs)) / (p_exp / (1-p_exp))
```

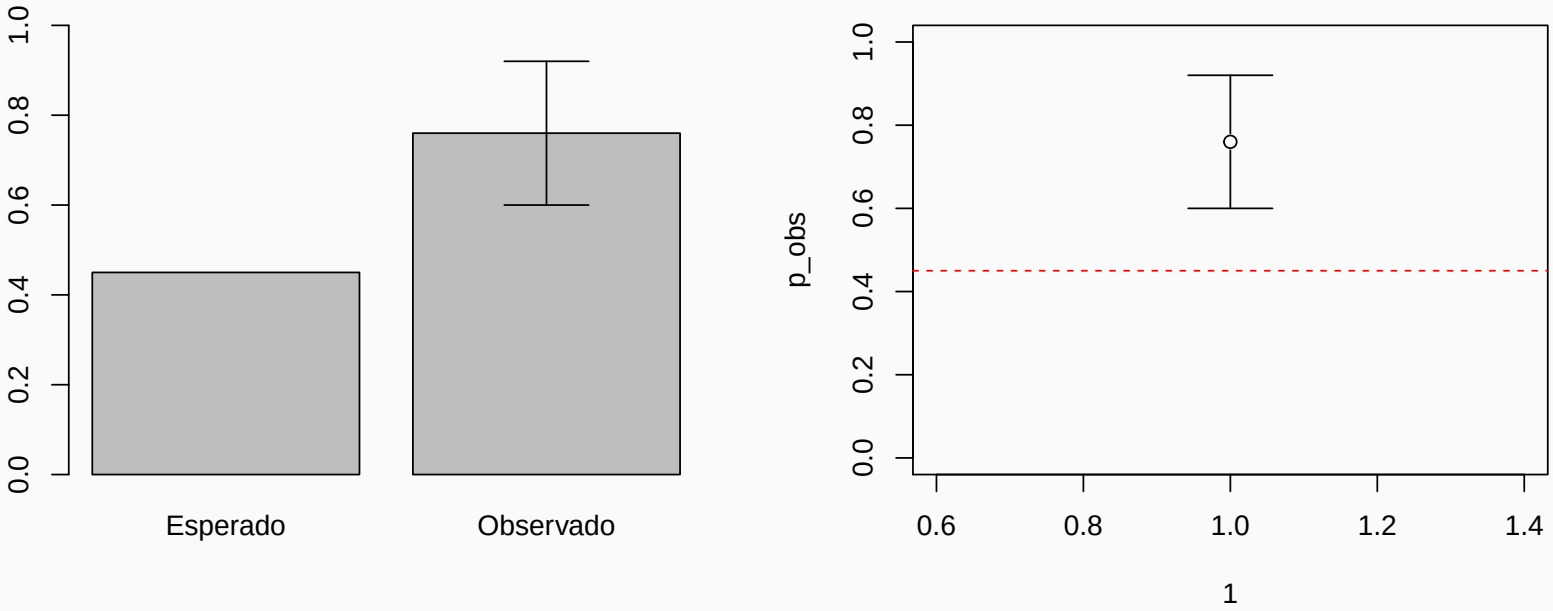
```
## [1] 3.87037
```

Resolução (3)

```
> lower <- qbinom(0.025, n, p_obs) / n
> upper <- qbinom(0.975, n, p_obs) / n
>
> c(lower, upper)
```

```
## [1] 0.60 0.92
```

```
> par(mfrow = c(1,2))
>
> barplot(c("Esperado" = p_exp, "Observado" = p_obs), ylim=c(0, 1))
> arrows(1.9, lower, 1.9, upper, code = 3, angle = 90)
>
> plot(1, p_obs, ylim=c(0,1))
> arrows(1, p_obs + c(-0.02, 0.02), 1, c(lower, upper), angle=90)
> abline(h = p_exp, lty = 'dashed', col = 'red')
```



Resolução (4)

```
> 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 0.65
## 95 percent confidence interval:
##  0.3932420 0.6818508
## sample estimates:
## probability of success
##                   0.54
```

Exercícios

1. 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).
   - $P(15 < X <= 35)$.
   - Quantiles $x_{0.025}$ e $x_{0.975}$.
2. Qual a frequência do sexo e responsta ao tratamento (*response*) na amostra de **cftr**?
3. A amostra apresenta uma proporção similar de homens e mulheres?

Resoluções (1)

```
> 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
```

```
> ## sum(dbinom(16:35, size = 50, p = 0.3))
> pbinom(35, 50, 0.3) - pbinom(15, 50, 0.3)
```

```
## [1] 0.4308216
```

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

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

Resolução (2)

```
> 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
```

Resolução (3)

```
> 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
```

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 está distribuida independentemente da hipoproteinemia?

Avaliando o problema

Testes de Independência

Teste de Fisher

Teste de Qui-quadrado

Exercícios

Resolução

Até a próxima