Problemas — Inferência – Soluções

James R. Hunter

21 de Março de 2017

```
suppressMessages(library(tidyverse))
suppressPackageStartupMessages(library(DescTools))
suppressPackageStartupMessages(library(binom))
suppressPackageStartupMessages(library(knitr))
options(scipen = 10)
```

Usaremos a base de dados `cleveland_heart` que coleciona dados sobre 303 pacientes que sofreram ataques cardíacos nos hospitais de Cleveland, Ohio, EUA. Esses dados vêm de Machine Learning Repository da University of California Irvine. Os dados são de 1988. Esta base de dados fica no formato de RData e pode ser carregada com o comando seguinte: "`load("cleveland_heart.RData")`". O arquivo deve estar na pasta "working directory".

1. Em nossa amostra de Cleveland, homens tem um nível de colesterol total diferente de que as mulheres?

`chol` é a variável para colesterol total e `genero` é para os sexos, codificado como "M" = homens e "F" = mulheres. Siga todas os 4 passos para preparar e executar um teste de hipótese. Mostre alguma análise exploratória.

Passo 1: Formular Hipótese

H_0: d = 0 H_1: d $\neq$ 0

Passo 2: Colecionar Dados

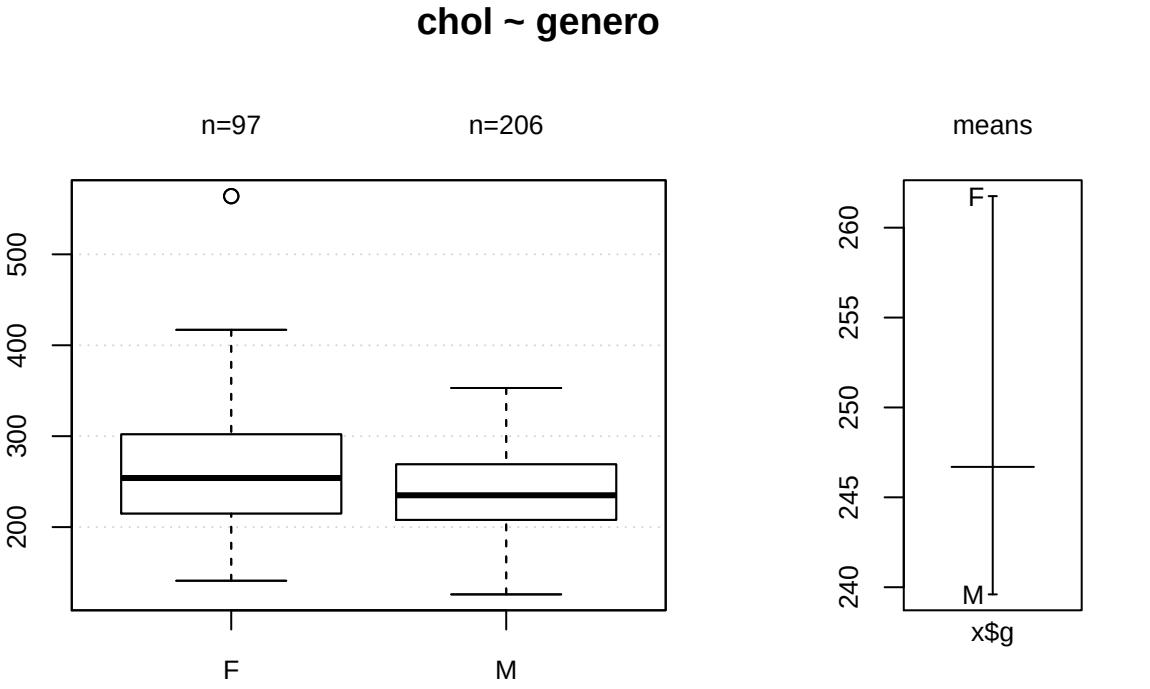
```
load("cleveland_heart.RData")
str(cleveland_heart, give.attr = FALSE)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    303 obs. of  11 variables:
##  $ idade     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ genero    : chr  "M" "M" "M" "M" ...
##  $ tipodor   : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ pressrep  : num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol      : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs       : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ ecgrepouso: num  2 2 2 0 2 0 2 0 2 2 ...
##  $ maxbat    : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang     : num  0 1 1 0 0 0 0 1 0 1 ...
##  $ slope     : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ diagnose  : int  0 2 1 0 0 0 3 0 2 1 ...
```

```
Desc(chol ~ genero, data = cleveland_heart, plotit = TRUE)
```

```
## ------------------------------------------------------------------------------
## chol ~ genero
```

```
##
## Summary:
## n pairs: 303, valid: 303 (100.0%), missings: 0 (0.0%), groups: 2
##
##
##                F        M
## mean     261.753  239.602
## median   254.000  235.000
## sd        64.901   42.650
## IQR       87.000   59.750
## n             97      206
## np        32.013%  67.987%
## NAs            0        0
## 0s             0        0
##
## Kruskal-Wallis rank sum test:
##   Kruskal-Wallis chi-squared = 7.1997, df = 1, p-value = 0.007291
```



/2017-03-21

Passo 3: Executar teste

```
cholTest <- t.test(chol ~ genero, data = cleveland_heart, alternative = "two.sided")
cholTest
```

```
##
##  Welch Two Sample t-test
##
## data:  chol by genero
```

```
## t = 3.0643, df = 136.37, p-value = 0.002631
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.855795 36.445477
## sample estimates:
## mean in group F mean in group M
##        261.7526        239.6019
```

Passo 4 – Interpretar Resultado

Rejeitar H_0: diferença entre as médias não é 0; médias parecem ser diferentes.

2. O valor limítrofe para pressão arterial sistólico é normalmente dado como 140.

Calcule a média da pressão de nossa amostra (variável ***pressrep****) e conte se esta média é significativamente abaixo de 140.*

Passo 1: Formular Hipótese

$H_0 : \mu = 0$

$H_1 : \mu < 0$

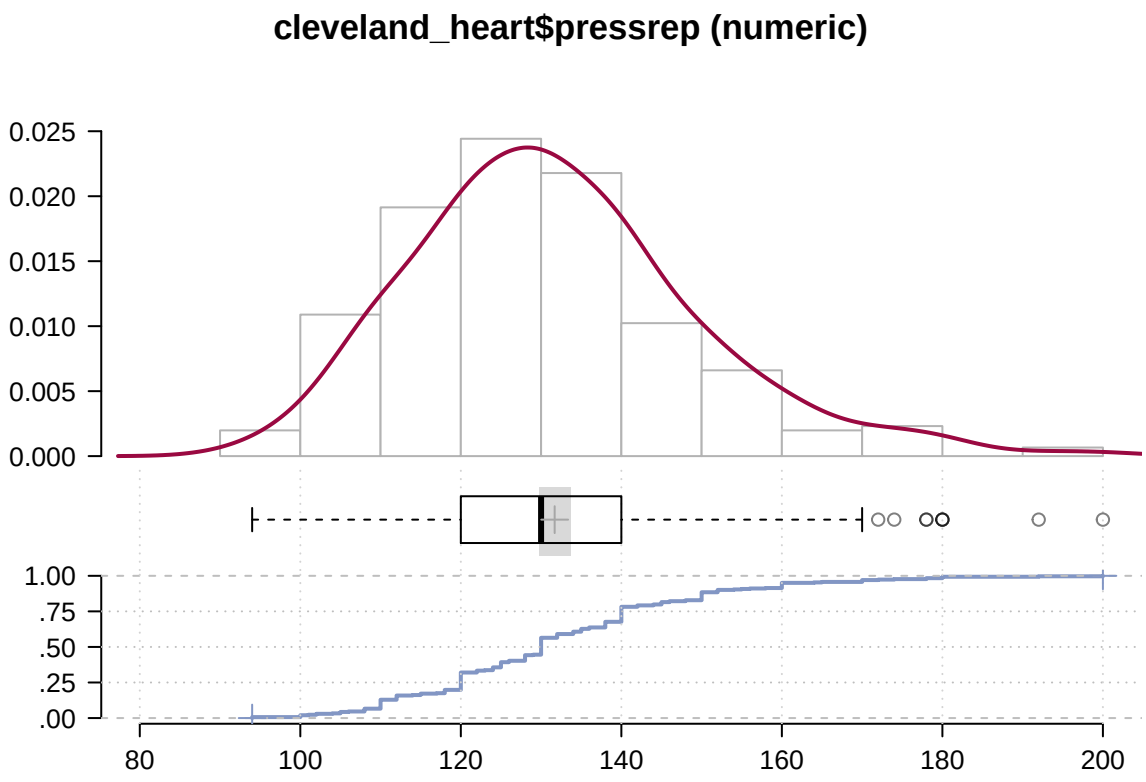
Passo 2: Colecionar Dados

```
glimpse(cleveland_heart$pressrep)
```

```
##  num [1:303] 145 160 120 130 130 120 140 120 130 140 ...
```

```
Desc(cleveland_heart$pressrep)
```

```
## ------------------------------------------------------------------------
## cleveland_heart$pressrep (numeric)
##
##   length       n     NAs  unique      0s    mean  meanCI
##      303     303       0      50       0  131.69  129.70
##           100.0%    0.0%           0.0%          133.68
##
##      .05     .10     .25  median     .75     .90     .95
##   108.00  110.00  120.00  130.00  140.00  152.00  160.00
##
##    range      sd   vcoef     mad     IQR    skew    kurt
##   106.00   17.60    0.13   14.83   20.00    0.70    0.82
##
## lowest : 94.0 (2), 100.0 (4), 101.0, 102.0 (2), 104.0
## highest: 174.0, 178.0 (2), 180.0 (3), 192.0, 200.0
```



Passo 3: Executar teste

```
pressTest <- t.test(cleveland_heart$pressrep, mu = 140, alternative = "less")
pressTest
```

```
##
##  One Sample t-test
##
## data:  cleveland_heart$pressrep
## t = -8.2192, df = 302, p-value = 3.054e-15
## alternative hypothesis: true mean is less than 140
## 95 percent confidence interval:
##     -Inf 133.358
## sample estimates:
## mean of x
##  131.6898
```

Passo 4: Interpretação

Rejeitar H_0. Pressão parece de ser menos que 140.

3. Mais Bagunça com Moedas – Proporções, Monte Carlo e IC

Esta vez, estamos trabalhando com uma moeda de R$1 que você acha injusta, ou seja, a probabilidade de CARA não é igual à probabilidade de COROA. Faça uma simulação de jogar a moeda 20.000 vezes e ver qual é a proporção de CARAS. O valor médio cai dentro de um Intervalo de Confiança 95%? O que é sua conclusão? A moeda é justa? Pode usar o pacote `binom` *para calcular o IC.*

Nosso hipótese nula (H_0): proporção = .5

```
set.seed(1)
p <- 0.5 # proporção igual para CARA CAROA
n <- 20000; k <- 1
tiras <- rbinom(n, k, p)
(caras <- sum(tiras)) ## número de CARAS em 20000
```

```
## [1] 9959
```

```
binom.confint(sum(tiras), n, conf.level = 0.95, methods = "asymptotic")
```

```
##       method    x     n    mean     lower     upper
## 1 asymptotic 9959 20000 0.49795 0.4910205 0.5048795
```

Interpretação: Nosso média (49.775%) cai dentro do intervalo de confiança (49.08 - 50.47%). Moeda parece justa.

4. Voltando a Cleveland – Açúcar no Sangue

O que é a proporção dos pacientes que tem açúcar no sangue acima de limite de 120 (`fbs`)? Mostre para cada gênero. Use `table()` e `prop.table()` para fazer os cálculos.

```
tabfbs <- with(cleveland_heart, table(fbs, genero))
tabfbs
```

```
##    genero
## fbs   F   M
##   0  85 173
##   1  12  33
```

```
prop.table(tabfbs)
```

```
##    genero
## fbs          F          M
##   0 0.28052805 0.57095710
##   1 0.03960396 0.10891089
```

5. Idade Tem a Ver com Hipertrofia Ventricular Esquerda

Divide a amostra em duas partes: pessoas que mostraram hipertrofia ventricular esquerda ou não. Esta condição fica em variável `ecgrepouso`. Esta variável tem três valores:

- *0: normal*
- *1: ST-T segmento abnormal > 0.05 mV*
- *2: hipertrofia ventricular esquerda (HVE)*

A idade de quem tem HVE é significativamente maior da que a idade das pessoas que não têm?

NB: Esta pergunta tem é mais desafiante que as anteriores.

```
cleveland_heart <- cleveland_heart %>% mutate(hve = (ecgrepouso == 2))
table(cleveland_heart$hve)
```

```
##
## FALSE  TRUE
##   155   148
```

```
hveid <- t.test(idade ~ hve, data = cleveland_heart, alternative = "greater")
hveid
```

```
##
##  Welch Two Sample t-test
##
## data:  idade by hve
## t = -2.4267, df = 300.87, p-value = 0.9921
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.1945     Inf
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            53.21935            55.71622
```

Apesar que a média das idades daqueles que mostram HVE seja acima da média daqueles que não mostram a condição, o valor-p indica que a diferença não é significativamente mais alta. A hipótese nula que são iguais não pode ser rejeitada.

Mas, se você faz o teste "two-sided", a diferença fica significativa (p = 0.01582).

6. Crédito Extra

Uma amostra de 12 mulheres participam num drug trial de um novo anti-concepcional oral. Elas têm a pressão sistólico medida antes e depois da administração do remédio. Os médicos querem saber se a pressão arterial tem uma diferença significativa depois de tomar a pilula. Os dados ficam no arquivo `RData` *"syspressmulh.RData" e na tela abaixo. Siga todos os 4 passos para construção e interpretação do teste de hipótese.*

```
load("syspressmulh.RData")
kable(syspressmulh, caption = "Pressão Sistólica em 12 Mulheres Antes e Depois")
```

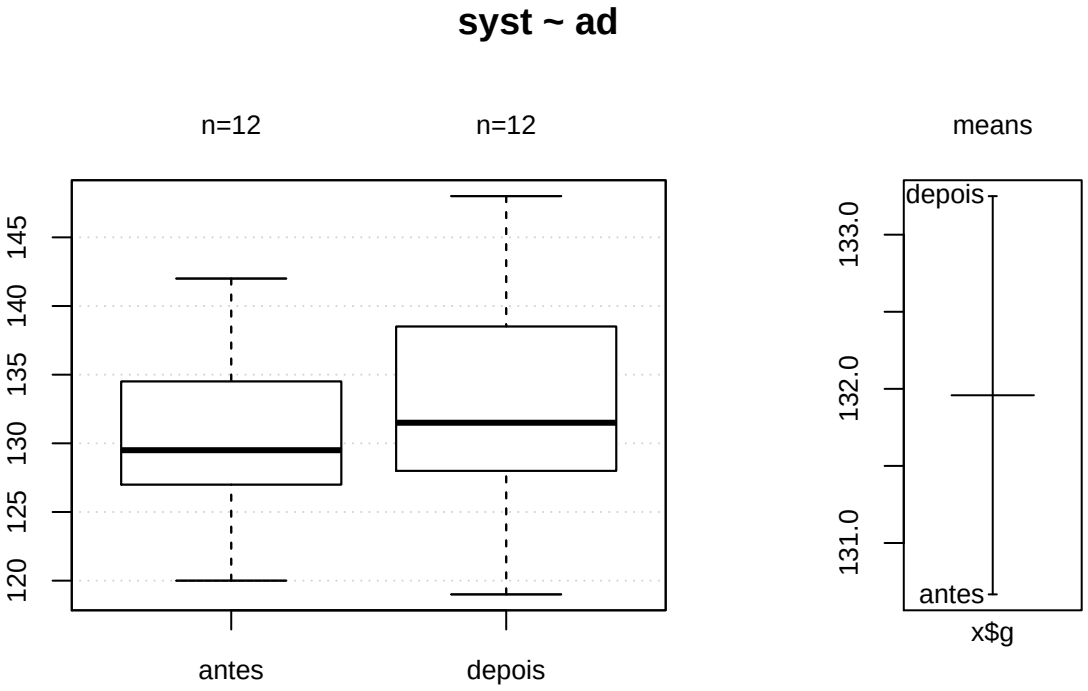
Table 1: Pressão Sistólica em 12 Mulheres Antes e Depois

PacNo	antes	depois
1	122	127
2	126	128
3	132	140
4	120	119
5	142	145
6	130	130
7	142	148
8	137	135
9	128	129
10	132	137
11	128	128
12	129	133

```
## precisa colocar dados numa forma key/value para calcular o teste-t
syspresslong <- syspressmulh %>% gather(ad, syst, -PacNo)
Desc(syst ~ ad, data = syspresslong)
```

```
## ------------------------------------------------------------------------
## syst ~ ad
```

```
##
## Summary:
## n pairs: 24, valid: 24 (100.0%), missings: 0 (0.0%), groups: 2
##
##
##            antes   depois
## mean     130.667  133.250
## median   129.500  131.500
## sd         6.933    8.226
## IQR        5.750    9.750
## n             12       12
## np       50.000%  50.000%
## NAs            0        0
## 0s             0        0
##
## Kruskal-Wallis rank sum test:
##   Kruskal-Wallis chi-squared = 0.52425, df = 1, p-value = 0.469
```




```
acoTtest <- t.test(syst ~ ad, data = syspresslong, paired = TRUE, alternative = "two.sided")
acoTtest
```

```
##
##  Paired t-test
##
## data:  syst by ad
## t = -2.8976, df = 11, p-value = 0.01451
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.5455745 -0.6210921
## sample estimates:
## mean of the differences
##               -2.583333
```

Interpretação

Hipótese nula que a diferença entre antes e depois seja 0 deve ser rejeitada.

Anotações

1. Ou mandem os exercícios para mim antes da próxima aula ou leve-os para a aula.
2. Trabalhem em seus grupos. Grupos podem submeter os problemas juntos. É mais fácil de dominar materiais novos quando tem ajuda dos colegas.