

Bioestatística - Teste de Hipóteses

Eric Kauati Saito

2022-07-04

Relembrando

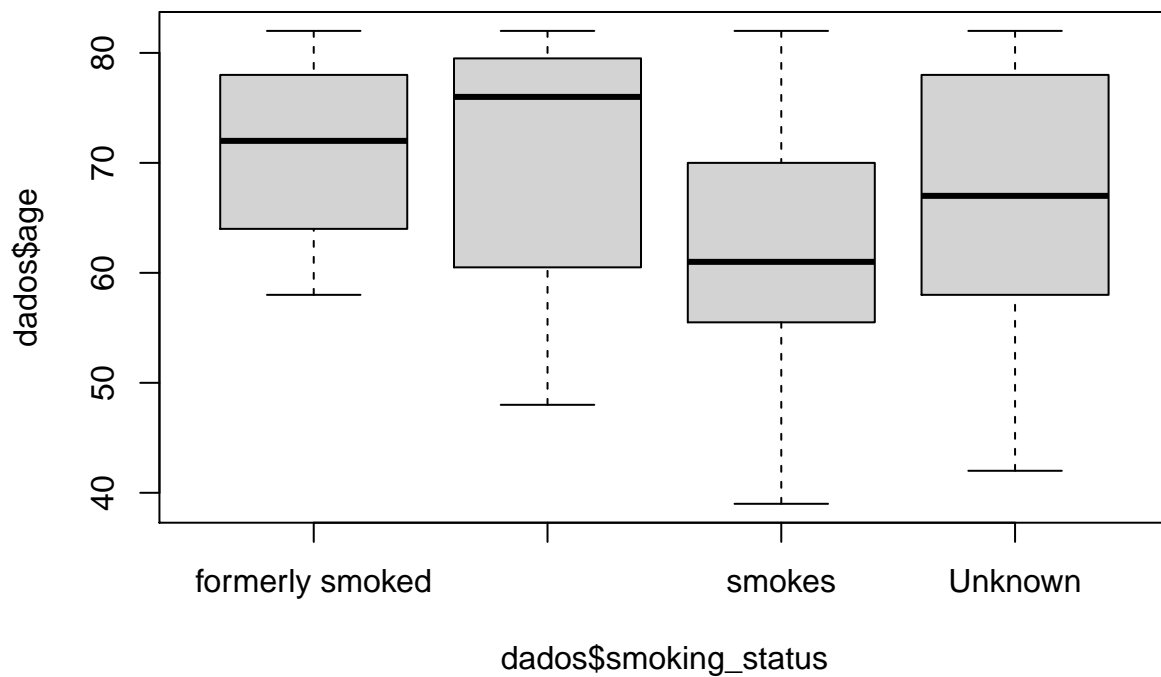
Carregando os dados:

```
dados = read.table(file="D:\\MACAE UFRJ\\2022\\R\\dados_pratica.csv", sep=";", header=TRUE)
```

Vamos tentar comparar a idade dos voluntários e as categoria de tabagismo.

Primeiro vamos tentar ver o boxplot, para isso façamos:

```
boxplot(dados$age ~ dados$smoking_status)
```



Relembrando que ao fazer dados\$age, estamos vendo a coluna da idade ('age') e dados\$smoking_status estamos vendo as categorias de tabagismo.

(dados\$age ~ dados\$smoking_status): “estamos vendo a idade (age) para cada categoria de tabagismo (smoking_status)”

Vejam que tem uma categoria “Unknown”, possivelmente pessoas que não quiseram responder a essa pergunta, ou outros motivos. Vamos remover dos nossos dados para que não influencie nos nossos resultados:

```
dados_remov = dados[!dados$smoking_status=="Unknown",]
```

Vejam dados_remov, não deve ter ninguém com smoking_status “Unknown”. Se prestarem atenção, é bem parecido com a prática anterior, quando estávamos interessados em pegar apenas as pessoas que eram fumantes (“smokes”), só que tem um “!” dentro dos colchetes na frente da expressão.

“!” é não (not), ou seja, não pegue os que smoking_status “Unknown”.

Vamos separar outro conjunto de dados para outras análises:

```
n_fumante = dados[dados$smoking_status=='never smoked',] #Pessoas que nunca fumaram
fumante = dados[dados$smoking_status=='smokes',] #Pessoas que fumam
```

Testes Estatísticos

Intervalo de Confiança

O intervalo de confiança para o nível médio de glicose de fumantes, com nível de confiança de 90%:

```
t.test(fumante$avg_glucose_level, conf.level = 0.9) #Intervalo de confiança para 90%
```

```
##
## One Sample t-test
##
## data: fumante$avg_glucose_level
## t = 10.471, df = 10, p-value = 1.04e-06
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
## 137.2954 194.7719
## sample estimates:
## mean of x
## 166.0336
```

Vejam que aparece a média amostral e o intervalo de confiança.

Obs.: Se você prestar atenção, ele fez um teste de hipótese. Mas como não colocamos mais nenhum parâmetro, ele testou para uma média igual a zero. Não queremos isso, pode ignorar.

Testes de Hipótese

Teste t

teste t para duas amostras, comparando o nível médio de glicose de fumantes e de não fumantes:

```
t.test(fumante$avg_glucose_level, n_fumante$avg_glucose_level, alternative = "two.sided")
```

```
##
## Welch Two Sample t-test
##
## data: fumante$avg_glucose_level and n_fumante$avg_glucose_level
## t = 1.2599, df = 23.305, p-value = 0.2202
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -16.58496 68.35484
## sample estimates:
## mean of x mean of y
## 166.0336 140.1487
```

Teve diferença? Nesse caso a hipótese nula foi dizer que as médias são iguais e a alternativa foi dizer que são diferentes. (teste bilateral) Caso vocês estejam trabalhando com teste unilateral, tem que mudar a opção da hipótese alternativa: `alternative = "less"` `alternative = "greater"`

Wilcoxon rank sum test

Caso fosse necessário um teste não-paramétrico:

```
wilcox.test(fumante$avg_glucose_level, n_fumante$avg_glucose_level, alternative = "two.sided")
```

```
##
## Wilcoxon rank sum exact test
##
## data: fumante$avg_glucose_level and n_fumante$avg_glucose_level
## W = 152, p-value = 0.363
## alternative hypothesis: true location shift is not equal to 0
```

Pareado

Teste pareado (`paired = TRUE`):

```
#t.test(fumante$avg_glucose_level, n_fumante$avg_glucose_level, paired = TRUE)
#wilcox.test(fumante$avg_glucose_level, n_fumante$avg_glucose_level, paired = TRUE)
```

Nesse caso não conseguimos rodar, ele vai dar erro. Isso porque não tem o mesmo número de observações, se fossem dados realmente pareados, teriam que ter o mesmo número de observações. E os dados teriam que estar na ordem correta.

ANOVA

ANOVA

“a idade (age) para cada categoria de tabagismo (smoking_status)”:

```
oneway.test(age ~ smoking_status, data=dados_remov, var.equal=T)
```

```
##  
## One-way analysis of means  
##  
## data: age and smoking_status  
## F = 2.5561, num df = 2, denom df = 46, p-value = 0.08859
```

```
#F = FALSE -> variancia diferentes  
#T = TRUE -> variancia igual
```

Para um nível de significância de $\alpha = 10\%$, como analisamos esse resultado? E agora, o que precisamos fazer?

post-hoc

Teste post-hoc:

```
model <- aov(age ~ smoking_status, data=dados_remov)  
TukeyHSD(model, conf.level=.95)
```

```
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##  
## Fit: aov(formula = age ~ smoking_status, data = dados_remov)  
##  
## $smoking_status  
##
```

	diff	lwr	upr	p adj
## never smoked-formerly smoked	-1.353623	-10.33653	7.629286	0.9293320
## smokes-formerly smoked	-9.357576	-20.10191	1.386758	0.0990571
## smokes-never smoked	-8.003953	-17.92629	1.918384	0.1354379