Estatística para Ciência de Dados
Resolução do trabalho 07

Eduardo Façanha Dutra

Conteúdo

1 Enunciado

Questão 1

Assista aos vídeos 3, 4 e 5, do prof. Jacob. Em seguida, execute o programa contido no arquivo "pref-ABC-teste-qui-quadrado.R", usando o RStudio. Pede-se: a) Explique todos os testes de proporção realizados. b) Reporte os resultados para as hipóteses contidas no programa.

Questão 2

Nesta questão, você deve descobrir as condições que sua amostra precisa ter para você aplicar um certo teste estatístico. Sugiro você consultar as possíveis condições usando a tabela de análise. Nos dois itens pedidos a seguir, você irá responder as perguntas formuladas pelo sistema, para chegar aos testes que se pede em cada item. Para isto, considere que o sistema está acessível em https://www.socscistatistics.com/tests/what_stats_test_wizard.aspx Pede-se: a) O sistema deve sugerir teste Qui-quadrado. Serve para comparar proporções, isto é, a distribuição de diversos acontecimentos (preferências, aprendizado, etc.) em diferentes amostras, a fim de avaliar se as amostras diferem significativamente quanto às proporções desses acontecimentos.

- Você acertou a resposta em quantas tentativas?
- Quais as respostas você deu para cada uma das perguntas formuladas quando você acertou o teste solicitado?
- Dê um exemplo, em que esta técnica seria aplicada, considerando as suas respostas corretas.

b) O sistema deve sugerir o teste estatístico T-student, e como alternativa a ele, o Man-whitney U test.

- Você acertou a resposta em quantas tentativas?
- Quais as respostas você deu para cada uma das perguntas formuladas quando você acertou o teste solicitado?
- Dê um exemplo, em que esta técnica seria aplicada, considerando as suas respostas corretas.

2 Resolução

2.1 Questão 1

```
a) Explique todos os testes de proporção realizados.
b) Reporte os resultados para as hipóteses contidas no programa.
```

A explicação dos resultados se dará logo após a explicação dos testes realizados.

2.1.1 Teste chi-quadrado

O primeiro teste realizado é o Chi-quadrado para proporções, onde será avaliado se existe diferença na preferência entre sites.

```
#carregamento das bibliotecas necessárias
library("csv")
library("Rcpp")

# leitura dos dados em arquivo
prefsABC = read.csv("Dados/prefsABC.csv")

#visualização em tabela dos dados recém carregados
#View(prefsABC)

#conversão da coluna Subject para representação de dados categóricos
prefsABC$Subject = factor(prefsABC$Subject)

#visualização de algumas estatísticas dos dados carregados
#summary(prefsABC)

#formatação dos dados para apresentação da contagem de observações por categoria
prfs = xtabs( ~ Pref, data = prefsABC)

#visualização da contagem das observações
# prfs

#Realização do teste qui-quadrado
teste = chisq.test(prfs)
teste
```

```
##
##  Chi-squared test for given probabilities
##
## data:  prfs
## X-squared = 13.3, df = 2, p-value = 0.001294
```

O teste acima realizado é um teste de proporções, cuja a hipótese nula é de que as frequências das variáveis observadas em relação ao total são iguais, que, no caso particular estudado, se traduz em conjecturar que as preferências entre os sites A, B e C são iguais.

No caso estudado, o teste é realizado comparando as frequências observadas com as frequências esperadas, como visto abaixo:

```
teste$expected
```

```
##  A  B  C
## 20 20 20
```

A fórmula utilizada para calcular a estatística do teste é:

$$\chi^2 = \sum_{k=A}^{C} \frac{(o_k - e_k)^2}{e_k}$$

De acordo com o resultado do teste, pode-se afirmar que, com os dados obtidos, a hipótese de que a preferência entre os sites A, B e C pode ser rejeitada a um nível de significância estatística de 0,05.

Uma ilustração mais detalhada do teste é mostrada na figura a seguir.

Chi-Square Calculator for Goodness of Fit

Success!

The data below should be self-explanatory. The only thing to note is that if you want to redo the calculation, you should press the "Restart Calculation" button (rather than using your browser back button).

The Chi^2 value is: 13.313

	Observed	*Expected*	*Difference*	*Difference Sq.*	*Diff. Sq. / Exp Fr.*
A	8	0.33 (19.98)	-11.98	143.52	7.18
B	21	0.33 (19.98)	1.02	1.04	0.05
C	31	0.33 (19.98)	11.02	121.44	6.08
					13.313

The Chi^2 value is 13.313. The *p*-value is .00129. The result is significant at $p < .05$.

2.1.2 Teste multinomial

O teste multinomial aplicado visa comparar a frequência dos dados observados em cada categoria (site de preferência) com uma distribuição multinomial com frequências teóricas à escolha do pesquisador.

O teste informa o quão raro seriam os resultados observados obtidos caso fossem obtidos de uma distribuição multinomial. A distribuição multinomial é utilizada para modelar uma variável aleatória que representam eventos com probabilidade fixa e independentes entre si.

A hipótese nula do teste é de que as variáveis observadas seguem possuem frequências iguais às predeterminadas pelo pesquisador. No nosso caso estudado a hipótese nula pode ser ser expressada como a assunção de que a proporção de escolha entre os sites A, B e C são iguais.

Portanto, o teste é realizado como mostrado no bloco a seguir, passando como parâmetros os dados estudados e assumindo iguais chances de escolha de cada site: `A= 1/3, B= 1/3, C= 1/3`. Ao predeterminar chances iguais de escolha para cada opção, assume-se, implicitamente, que as variáveis são independentes entre si e que a escolha de um site em particular é devido apenas a uma aleatoriedade.

```
# multinomial test
library(XNomial)
xmulti(prfs, c(1/3, 1/3, 1/3), statName="Prob")
```

```
##
## P value (Prob) = 0.0008024
```

Os resultados obtidos no teste nos mostram que a frequência de escolha de cada site divergem bastante de uma distribuição multinomial equilibrada(`P(A) = P(B) = P(C)`) devido ao valor P calculado se encontrar muito abaixo de `0.05`, limiar de significância adotado.

Pode-se interpretar o resultado, alternativamente, como, caso a escolha do site seja aleatória, e não haja preferência entre eles, um resultado tão ou mais raro quanto o resultado encontrado será observado apenas `0.08%` das vezes. Como as chances são muito baixas é provável que a escolha dos sites não sejam aleatórias e que pelo menos um site pode possuir características de causem a preferência de um site sobre outro.

2.1.3 Testes binomiais

Os testes binomiais aplicados em seguida são utilizados para comparar o percentual de escolha de um site isoladamente contra as outras opções. Semelhante ao teste multinomial, é escolhido um percentual de referência de `1/3`, que corresponde a uma escolha aleatória, sem preferências, entre 3 opções.

A hipótese nula de cada teste individual é de que a probabilidade de escolha do teste é igual a 1/3.

```
# teste site A
aa = binom.test(sum(prefsABC$Pref == "A"), nrow(prefsABC), p=1/3)
# teste site B
bb = binom.test(sum(prefsABC$Pref == "B"), nrow(prefsABC), p=1/3)
# teste site C
cc = binom.test(sum(prefsABC$Pref == "C"), nrow(prefsABC), p=1/3)
aa$p.value
```

```
## [1] 0.000553318
```

```
bb$p.value
```

```
## [1] 0.7852017
```

```
cc$p.value
```

```
## [1] 0.00372349
```

Os resultados de cada teste nos informa que as proporções de escolhas dos sites A e C estão divergentes da probabilidade de 1/3, pois o p-value de cada teste está muito abaixo de `0.05`.

O site A foi escolhido abaixo do esperado, pois seu intervalo de confiança de 95% (6% a 25%) está abaixo da probabilidade de referência (33%).

O site C foi escolhido acima do esperado, pois seu intervalo de confiança de 95% (38% a 65%) está acima da probabilidade de referência (33%)

Já o site B, escolhido por 21 das 60 pessoas, não diverge significamente da probabilidade de referência de 1/3, que está contida no intervalo de confiança do teste realizado (23% a 48%).

```
p.adjust(c(aa$p.value, bb$p.value, cc$p.value), method="holm")
```

```
## [1] 0.001659954 0.785201685 0.007446980
```

```
#cada valor p é multiplicado pelo seu índice de ordenação
c(aa$p.value*3,bb$p.value*1, cc$p.value*2)
```

```
## [1] 0.001659954 0.785201685 0.007446980
```

Após a realização dos testes, os p-values foram ajustados como uma forma de verificar se não há erro do tipo 1, rejeição de hipótese nula verdadeira, em cada um dos testes aplicados. No método de Holm os valores p são ordenados do maior para o menor e multiplicados pelo respectivo índice de ordenação(menor valor-p por 3 e o maior valor-p por 1).

Mesmo após as correções dos p-values de cada teste a conclusão sobre os testes binomiais permanece inalterada, pois nenhum deles ultrapassou o limiar de `0.05`.

2.1.4 Teste qui-quadrado ABCSex

O teste aplicado a seguir tem como objetivo avaliar se a escolha de site é afetada pelo gênero do participante.

Diferentemente do teste chi-quadrado anterior não será avaliado se as proporções entre A, B e C são iguais, mas se as proporções das escolhas feitas por pessoas de cada gênero dadas as proporções de A, B e C observadas são semelhantes.

Portanto, a hipótese nula é que a escolha do site é independente do gênero do participante

```
# leitura dos dados em arquivo da preferências do sites por gênero
prefsABCsex <- read.csv("Dados/prefsABCsex.csv")

#conversão da coluna Subject para representação de dados categóricos
prefsABCsex$Subject=factor(prefsABCsex$Subject)

#visualização dos dados
#View(prefsABCsex)
#summary(prefsABCsex)

#formatação dos dados para apresentação da contagem de observações por categoria
prfsABCSex = xtabs( ~ Pref + Sex, data = prefsABCsex)

t = chisq.test(prfsABCSex)
```

```
## Warning in chisq.test(prfsABCSex): Chi-squared approximation may be incorrect
```

```
t
```

```
##
##  Pearson's Chi-squared test
##
## data:  prfsABCSex
## X-squared = 6.9111, df = 2, p-value = 0.03157
```

Com o p-value obtido do teste menor que 0.05, pode-se rejeitar a hipótese nula de que as escolhas de site são independente de gênero.

2.1.5 Teste de proporções G test

Teste utilizado com o mesmo objetivo do chi-quadrado, para avaliar se há diferença nas proporções de escolha de site por pessoas de cada gênero. Utiliza-se a mesma tabela do chi-quadrado. O teste geralmente é utilizado para amostras maiores que 200.

```
library(RVAideMemoire)
```

```
## *** Package RVAideMemoire v 0.9-77 ***
```

```
# G-test
G.test(prfsABCSex)
```

```
##
##  G-test
##
## data:  prfsABCSex
## G = 7.0744, df = 2, p-value = 0.02909
```

O valor p obtido nos leva às mesmas conclusões do teste chi-quadrado anterior.

2.1.6 Teste exato de Fisher

O teste exato de Fisher é utilizado para saber o quão provavel é de se obter um resultado tão ou mais atípico quanto o observado, dadas as proporções apresentadas. Normalmente é utilizado para amostras com poucas observações.

A hipótese nula do teste é de que as chances de escolha dos sites são iguais (odds ratio = 1).

```
# Fisher's exact test
fisher.test(prfsABCSex)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  prfsABCSex
## p-value = 0.03261
## alternative hypothesis: two.sided
```

O resultado do teste permite rejeitar a hipótese nula de que a razão das chances de escolha são iguais a 1.

2.1.7 Testes binomiais ABC para o sexo Masculino

Os testes binomiais aplicados em seguida são utilizados para comparar o percentual de escolha de um site por participantes apenas do sexo masculino isoladamente contra as outras opções. Semelhante ao teste multinomial, é escolhido um percentual de referência de `1/3`, que corresponde a uma escolha aleatória, sem preferências, entre 3 opções.

A hipótese nula de cada teste individual é de que a probabilidade de escolha do teste é igual a 1/3.

```
# Teste binomial de escolha do site A por sujeitos do sexo masculino
ma = binom.test(sum(prefsABCsex[prefsABCsex$Sex == "M", ]$Pref == "A"),
                nrow(prefsABCsex[prefsABCsex$Sex == "M", ]),
                p = 1/3)

# Teste binomial de escolha do site B por sujeitos do sexo masculino
mb = binom.test(sum(prefsABCsex[prefsABCsex$Sex == "M", ]$Pref == "B"),
                nrow(prefsABCsex[prefsABCsex$Sex == "M", ]),
                p = 1/3)

# Teste binomial de escolha do site C por sujeitos do sexo masculino
mc = binom.test(sum(prefsABCsex[prefsABCsex$Sex == "M", ]$Pref == "C"),
                nrow(prefsABCsex[prefsABCsex$Sex == "M", ]),
                p = 1/3)

ma$p.value
```

```
## [1] 0.05473678
```

```
mb$p.value
```

```
## [1] 0.1266222
```

```
mc$p.value
```

```
## [1] 0.0004322513
```

Para o site A não se rejeita a hipótese nula, pois o p-value(0.055) está acima do valor de `0.05`.

Para o site B não se rejeita a hipótese nula, pois o p-value(0.127) está acima do valor de `0.05`.

Para o site C se rejeita a hipótese nula, pois o p-value(0.0004) está muito abaixo do valor de `0.05`.

```
p.adjust(c(ma$p.value, mb$p.value, mc$p.value), method="holm")
```

```
## [1] 0.109473564 0.126622172 0.001296754
```

```
#cada valor p é multiplicado pelo seu índice de ordenação
c(ma$p.value*2,mb$p.value*1, mc$p.value*3)
```

```
## [1] 0.109473564 0.126622172 0.001296754
```

Após a realização dos testes, os p-values foram ajustados como uma forma de verificar se não há erro do tipo 1, rejeição de hipótese nula verdadeira, em cada um dos testes aplicados. No método de Holm os valores p são ordenados do maior para o menor e multiplicados pelo respectivo índice de ordenação(menor valor-p por 3 e o maior valor-p por 1).

Mesmo após as correções dos p-values de cada teste a conclusão sobre os testes binomiais permanece inalterada, pois nenhum deles ultrapassou o limiar de `0.05`.

2.1.8 Testes binomiais ABC para o sexo Feminino

Os testes aplicados para o sexo masculino são agora aplicados para o sexo feminino.

```
fa = binom.test(sum(prefsABCsex[prefsABCsex$Sex == "F", ]$Pref == "A"),
                nrow(prefsABCsex[prefsABCsex$Sex == "F", ]),
                p = 1/3)
fb = binom.test(sum(prefsABCsex[prefsABCsex$Sex == "F", ]$Pref == "B"),
                nrow(prefsABCsex[prefsABCsex$Sex == "F", ]),
                p = 1/3)
fc = binom.test(sum(prefsABCsex[prefsABCsex$Sex == "F", ]$Pref == "C"),
                nrow(prefsABCsex[prefsABCsex$Sex == "F", ]),
                p = 1/3)

fa$p.value
```

```
## [1] 0.009010912
```

```
fb$p.value
```

```
## [1] 0.0472391
```

```
fc$p.value
```

```
## [1] 0.6939695
```

Para o site A se rejeita a hipótese nula, pois o p-value(0.009) está muito abaixo do valor de `0.05`.

Para o site B se rejeita a hipótese nula, pois o p-value(0.047) está abaixo do valor de `0.05`.

Para o site C não se rejeita a hipótese nula, pois o p-value(0.694) está acima do valor de `0.05`.

```
p.adjust(c(fa$p.value, fb$p.value, fc$p.value), method="holm")
```

```
## [1] 0.02703274 0.09447821 0.69396951
```

```
c(fa$p.value*3,fb$p.value*1, fc$p.value*1)
```

```
## [1] 0.02703274 0.04723910 0.69396951
```

Após a realização dos testes, os p-values foram ajustados como uma forma de verificar se não há erro do tipo 1, rejeição de hipótese nula verdadeira, em cada um dos testes aplicados. No método de Holm os valores p são ordenados do maior para o menor e multiplicados pelo respectivo índice de ordenação(menor valor-p por 3 e o maior valor-p por 1).

Mesmo após as correções dos p-values de cada teste a conclusão sobre os testes binomiais permanece inalterada, pois nenhum deles ultrapassou o limiar de `0.05`.

2.2 Questão 2

2.2.1 a)

- Foram realizadas 2 tentativas;
- Nominal > One nominal variable;
- Realizar um teste em uma população para aferir se, por exemplo, a proporção de mortes por Covid-19 em um país segue ou diverge da tendência mundial.

2.2.2 b)

- Foram realizadas 3 tentativas;
- Interval/Ratio > Differences between populations > No, I'm working with more than one sample (or treatment condition) > Independent-measures > Two samples (treatments);
- Realizar um teste em uma população para aferir se, por exemplo, há eficácia de um medicamento para determinada doença entre dois grupos.