CC0291 - Estatística Não Paramétrica

Teste de Homogeneidade - 04/05/2023.

Prof. Maurício

1 Teste de Homogeneidade- Duas Populações Independentes:

Vamos apresentar como testar se duas distribuições multinomiais independentes são as mesmas.

Este teste é conhecido na literatura como teste de homogeneidade.

Vamos considerar duas distribuições multinomiais independentes com parâmetros

$$n_j, p_{1j}, p_{2j}, \ldots, p_{kj}, \quad j = 1, 2 \;\; i = 1, 2, \ldots, k,$$

respectivamente.

Sejam

$$X_{ij}, i = 1, 2, \ldots, k \;\; , j = 1, 2,$$

as frequências correspondentes. Se n_1 e n_2 são grandes, a variável aleatória

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(X_{ij} - n_j \, p_{ij})^2}{n_j \, p_{ij}} \sim \chi^2(gl = 2k - 2),$$

é a soma de duas variáveis aleatória independentes

$$Q = \sum_{i=1}^{k} \frac{(X_{i1} - n_1 \, p_{i1})^2}{n_1 \, p_{i1}} + \sum_{i=1}^{k} \frac{(X_{i2} - n_2 \, p_{i2})^2}{n_2 \, p_{i2}} \sim \chi^2(gl = (k-1) + (k-1) = 2k - 2),$$

A nossa hipótese nula fica:

$$H_0 : p_{11} = p_{12} = p_1 \;\; p_{21} = p_{22} = p_2, \ldots, p_{k1} = p_{k2} = p_k,$$

em que

$$p_{1i} = p_{2i} = p_i, \quad i = 1, 2, \ldots, k$$

são desconhecidas e portanto temos que estimar

$$p_1, p_2, \ldots, p_{k-1}$$

a estimativa de p_k sai por diferença já que :

$$\sum_{i=1}^{k} p_i = 1.$$

Assim para calcular as frequências esperadas vão ser estimados

$$m = k - 1$$

parâmetros.

O número de graus de liberdade do teste

$$gl = k - m = 2k - 2 - (k-1) = k - 1.$$

O estimador de máxima verossimilhança de p_i é dado por:

$$\hat{p}_i = \frac{X_{i1} + X_{i2}}{n_1 + n_2}.$$

Seja $E_{ij} = n_j \times \hat{p}_i$ as nossas frequências esperadas. Logo

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(k-1),$$

pois vamos perder $m = k - 1$ graus de liberdade pois estimamos

$$p_1, p_2, \ldots, p_{k-1}.$$

Assim

$$gl = 2k - 2 - (k-1) = k - 1.$$

Vamos apresentar um exemplo.

Exemplo 4 Duas técnicas distintas de ensino foram usadas em dois grupos de estudantes independentemente. Cada grupo contava com 100 estudantes de capacidade bem similares. No final um mesmo exame foi aplicado a cada grupo e aplicada uma nota literal de E, excelente, B, bom, R, regular, sendo I,insuficiente e M,mau.

Os dados tabulados foram:

Nota	E	B	R	I	M	Total
Grupo I	15	25	32	17	11	100
Grupo II	9	18	29	28	16	100

Vamos considerar que estes dados são oriundos de duas multinomiais independentes com $k = 5$ e ao nível de significância de 5% , teste a hipótese que estas duas distribuições são as mesmas e portanto as duas técnicas tem a mesma eficácia.

Temos $k = 5$ classes, o grupo I será representado por $j = 1$ e o grupo II será representado por $j = 2$..

O grau E será $i = 1$, grau B será $i = 2$, grau R será $i = 3$, grau I será $i = 4$ e o grau M será $i = 5$.

Além disso p_{ij}, representa a probabilidade de um estudante do grupo j tirar nota i.

Queremos testar:

$$H_0 : p_{11} = p_{12} = p_1 \ ; \ p_{21} = p_{22} = p_2; \ \ , p_{31} = p_{32} = p_3; p_{41} = p_{42} = p_4, p_{51} = p_{52} = p_5.$$

Temos $n_1 = n_2 = 100$ e $n_1 + n_2 = 200$

Se H_0 é verdade temos:

1. $p_{11} = p_{12} = p_1$ e é estimado por:

$$\hat{p}_1 = \frac{X_{11} + X_{12}}{n_1 + n_2} = \frac{15 + 9}{200} = \frac{12}{100} = 0,12.$$

Assim

$$\hat{E}_{11} = \hat{E}_{12} = n_1 \times \hat{p}_1 = 100 \times \frac{12}{100} = 12.$$

2. $p_{21} = p_{22} = p_2$ e é estimado por:

$$\hat{p}_2 = \frac{X_{21} + X_{X2}}{n_1 + n_2} = \frac{25 + 18}{200} = \frac{21,5}{100} = 0,215.$$

Assim

$$\hat{E}_{21} = \hat{E}_{22} = n_1 \times \hat{p}_2 = 100 \times \frac{21,5}{100} = 21,5.$$

3. $p_{31} = p_{32} = p_3$ e é estimado por:

$$\hat{p}_3 = \frac{X_{31} + X_{32}}{n_1 + n_2} = \frac{32 + 29}{200} = \frac{30,5}{100} = 0,305.$$

Assim

$$\hat{E}_{31} = \hat{E}_{32} = n_1 \times \hat{p}_3 = 100 \times \frac{30,5}{100} = 30,5.$$

4. $p_{41} = p_{42} = p_4$ e é estimado por:

$$\hat{p}_4 = \frac{X_{41} + X_{42}}{n_1 + n_2} = \frac{17 + 28}{200} = \frac{22,5}{100} = 0,225.$$

Assim

$$\hat{E}_{41} = \hat{E}_{42} = n_1 \times \hat{p}_4 = 100 \times \frac{22,5}{100} = 22,5.$$

5. $p_{51} = p_{52} = p_5$ e é estimado por:

$$\hat{p}_5 = \frac{X_{51} + X_{52}}{n_1 + n_2} = \frac{11 + 16}{200} = \frac{13,5}{100} = 0,135.$$

Assim

$$\hat{E}_{51} = \hat{E}_{52} = n_1 \times \hat{p}_5 = 100 \times \frac{13,5}{100} = 13,5.$$

A tabela das frequências esperadas é dada por:

Nota	E	B	R	I	M	Total
Grupo I	12	21,5	30,5	22,5	13,5	100
Grupo II	12	21,5	30,5	22,5	13,5	100

A estatística do teste será:

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{5} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2(4),$$

em que

$$O_{ij} = X_{ij} \quad e \quad \hat{E}_{ij} = n_j \times \hat{p}_j.$$

$$Q = \frac{(15-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(25-21,5)^2}{21,5} + \frac{(18-21,5)^2}{21,5} + \frac{(32-30,5)^2}{30,5} +$$
$$+ \frac{(29-30,5)^2}{30,5} + \frac{(17-22,5)^2}{22,5} + \frac{(28-22,5)^2}{22,5} + \frac{(11-13,5)^2}{13,5} + \frac{(16-13,5)^2}{13,5} = 6,40.$$

A região crítica do teste é dada por:

$$Q_{cal} > c,$$

em que $c = P_{95} = 9,49$ é o percentil de ordem 95 da qui-quadrado com $gl = k - 1 = 4$. O valor da estatística do teste é

Como

$$Q_{cal} = 6,40 < Q_{tab} = 9,49,$$

não há evidências de que H_0 não seja verdadeira.

O nível descritivo do teste é dado por:

$$nd = P(Q \geq Q_{cal}) = 0,1711.$$

Vamos usar o R como calculadora:

```
>
> O_1=c(15,25,32,17,11)
>
> k=length(O_1);k
[1] 5
>
> n_1=sum(O_1);n_1
[1] 100
>
>
> O_2=c(9,18,29,28,16)
>
> n_2=sum(O_2);n_2
[1] 100
>
>
> p_est=(O_1+O_2)/(n_1+n_2);p_est
[1] 0.120 0.215 0.305 0.225 0.135
> sum(p_est)
[1] 1
>
> E_1=n_1*p_est
> E_2=n_2*p_est
>
> D_1=O_1-E_1
>
> D12=D_1^2
> A_1=D12/E_1
> S_1=sum(A_1);S_1
[1] 3.200945
>
> D_2=O_2-E_2
>
> D22=D_2^2
> A_2=D22/E_2
> S_2=sum(A_2);S_2
[1] 3.200945
>
> Q_cal=S_1+S_2;Q_cal
[1] 6.401891
>
> alfa=0.05
> gama=1-alfa;gama
[1] 0.95
```

```
>
>
>
>
> P95=qchisq(gama,k-1);P95
[1] 9.487729
>
> c=round(P95,2);c
[1] 9.49
>
>

> nd=pchisq(Q_cal,k-1,lower.tail=FALSE);nd
[1] 0.171078
>
```

Fazendo direto no R temos usando o chisq.test. Tomando cuidado pois o objeto é uma matriz. Cada linha é uma população multinomial.

```
>
> Grau=rbind(O_1,O_2);Grau
[,1] [,2] [,3] [,4] [,5]
O_1   15   25   32   17   11
O_2    9   18   29   28   16
> rownames(Grau)=c("Grupo I","Grupo II")
> colnames(Grau)=c("E","B","R","I","M")
>    Grau
          E  B  R  I  M
Grupo I  15 25 32 17 11
Grupo II  9 18 29 28 16
>
>
> mod=chisq.test(Grau);mod

Pearson's Chi-squared test

data:  Grau
X-squared = 6.4019, df = 4, p-value = 0.1711

>
```

2 Teste de Homogeneidade- r Populações Independentes:

Vamos estender o teste homogeneidade para comparar se as probabilidades de r distribuições multinomiais independentes são as mesmas.

Vamos considerar que as distribuições multinomiais independentes tenham parâmetros

$$n_j, p_{1j}, p_{2j}, \ldots, p_{cj}, \;\; j = 1, 2, \ldots, r \;\; i = 1, 2, \ldots, c.$$

respectivamente.

Sejam

$$Oij = X_{ij}, i = 1, 2, \ldots, c \;\; , j = 1, 2, \ldots, r$$

as frequências correspondentes.

A nossa hipótese nula fica:

$$H_0 : p_{i1} = p_{i2} = \ldots = p_{ir} = p_{i.}, \;\; i = 1, 2, \ldots c - 1.$$

O estimador de máxima verossimilhança de $p_{i.}$ é dado por:

Seja

$$n = \sum_{j=1}^{r} n_j.$$

$$\hat{p}_{i.} = \frac{\sum_{j=1}^{r} X_{ij}}{n}.$$

A frequências esperadas são estimadas por:

$$\hat{E}_{ij} = n_j \;\times \hat{p}_{i.}, \;\; i = 1, 2, \ldots, c.$$

Se H_0 é verdade então a estatística do teste será:

$$Q = \sum_{j=1}^{r} \sum_{i=1}^{c} \frac{(X_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2((r-1)(c-1)).$$

temos agora $r(k-1)$ parâmetros livres nas r distribuições e estimamos $m = (k-1)$ parâmetros para calcular as frequências esperadas.
Logo

$$gl = r(k-1) - (k-1) = (r-1)(k-1).$$

Exemplo 5: Um estudo do uso do cinto de segurança em táxis abrangeu 77 táxis de Nova York, 129 em Chicago e 72 em Pittsburgh, com os resultados apresentados na tabela a seguir. Um porta-voz do sindicato de motoristas de táxi argumenta que, embora pareça que poucos táxis têm cintos em condições de uso, o tamanho da amostra extraída das três cidades foi de apenas 278, de modo que os resultados não são significativos. Com o nível de significância de 0,05, teste a afirmação de que as três cidades têm a mesma proporção de táxis com cinto de segurança em condições de uso: isto é, teste a homogeneidade das proporções de cintos de segurança em condições de uso nas três cidades.

Condições de Uso	Boas	Não Boas	Total
Nova York	3	74	77
Chicago	42	87	129
Pittsburgh	2	70	72
Total	47	231	278

Com base no artigo "The Phantom Taxi Seat Belt de "Welkon and Reisinger"publicado no "American Journal of Public Health, Vol. 67, n^o11. "

Solução Aqui temos $r = 3$ e $c = 2$.

Temos 3 binomiais pois quando $c = 2$ temos duas classes a saber: sucesso ou fracasso.

Considere p_1 a probabilidade de se encontrar um táxi em Nova York com cinto de segurança em condições de uso.

p_2 a probabilidade de se encontrar um táxi em Chicago com cinto de segurança em condições de uso.

p_3 a probabilidade de se encontrar um táxi em Pittsburgh com cinto de segurança em condições de uso.

Queremos testar:

$$H_0 : p_1 = p_2 = p_3 = p.$$

Na realidade queremos comparar se as probabilidades de três binomiais independentes são iguais.

Na cidade de Nova York foi retirada uma amostra de $n_1 = 77$ táxis e observado que

$0_{11} = 3$ estavam em condições de uso e $0_{12} = 74$ não estavam.

Na cidade de Chicago foi retirada uma amostra de $n_2 = 129$ táxis e observado que

$0_{21} = 42$ estavam em condições de uso e $0_{22} = 87$ não estavam.

Na cidade de Pittsburgh foi retirada uma amostra de $n_3 = 72$ táxis e observado que

$0_{31} = 2$ estavam em condições de uso e $0_{32} = 70$ não estavam. Se H_0 é verdade na realidade temos uma única população e foi retirada uma amostra de tamanho $n = 278$ com um total de sucessos $S = 0_{11} + 0_{21} + 0_{31} = 47$. A estimativa de máxima verossimilhança de p é dada por:

$$\hat{p} = \frac{0_{11} + 0_{21} + 0_{31}}{n} = \frac{47}{278} = 0,169.$$

e

$$\hat{q} = \frac{231}{278} = 0,831.$$

Vamos calcular as frequências esperadas estimadas:

$$\hat{E}_{11} = n_1 \times \hat{p} = 77 \times \frac{47}{278} = \frac{77 \times 47}{278} = \frac{47}{278} = 13,02.$$

$$\hat{E}_{12} = n_1 \times \hat{q} = 77 \times \frac{231}{278} = \frac{77 \times 231}{278} = \frac{47}{278} = 63,98.$$

$$\hat{E}_{21} = n_2 \times \hat{p} = 129 \times \frac{47}{278} = \frac{129 \times 47}{278} = \frac{47}{278} = 21,81$$

$$\hat{E}_{22} = n_2 \times \hat{q} = 129 \times \frac{231}{278} = \frac{129 \times 231}{278} = \frac{47}{278} = 107,19.$$

$$\hat{E}_{31} = n_3 \times \hat{p} = 72 \times \frac{47}{278} = \frac{72 \times 47}{278} = \frac{47}{278} = 12,17.$$

$$\hat{E}_{32} = n_3 \times \hat{q} = 72 \times \frac{231}{278} = \frac{72 \times 231}{278} = \frac{47}{278} = 59,83.$$

A estatística do teste terá $gl = (r-1)(c-1) = 2$ graus de liberdade.

O percentil 95 de $Q \sim \chi^2(2)$ é dado por:

$$Q_{tab} = 5,99.$$

A nossa região crítica é dada por:

Rejeitar H_0 se

$$Q_{cal} > 5,99.$$

```
>
> r=3;c=2
> gl=(r-1)*(c-1);gl
[1] 2
>
> alfa=0.05;gama=1-alfa
```

```
>
> Q_tab=qchisq(gama,gl);Q_tab
[1] 5.991465
> round(Q_tab,2)
[1] 5.99
>
```

Vamos obter o valor da estatística do teste.

```
>
>
>
> O_1=c(3,74);n_1=sum(O_1);n_1
[1] 77
> O_1
[1]  3 74
> O_2=c(42,87);n_2=sum(O_2);n_2
[1] 129
> O_2
[1] 42 87
>
> O_3=c(2,70);n_3=sum(O_3);n_3
[1] 72
>
> O_3
[1]  2 70
> n=n_1+n_2+n_3;n
[1] 278
>
>
> ####Estimativa de p e q=1-p
>
> S=O_1[1]+O_2[1]+O_3[1];S
[1] 47
> p_est=S/n;p_est
[1] 0.1690647
> require(MASS)
> fractions(p_est)
[1] 47/278
>
> q_est=1-p_est
>
> fractions(q_est)
[1] 231/278
>
> ####Estimando as frequências esperadas:
>
>
>
> E11_est=n_1*p_est;E11_est
```

```
[1] 13.01799
>
> fractions(E11_est)
[1] 3619/278
>
>
>
> E12_est=n_1*q_est;E12_est
[1] 63.98201
>
> fractions(E12_est)
[1] 17787/278
>
> E1_est=c(E11_est,E12_est)
>
>
> E21_est=n_2*p_est;E21_est
[1] 21.80935
>
> fractions(E21_est)
[1] 6063/278
>
>
> E22_est=n_2*q_est;E22_est
[1] 107.1906
>
> fractions(E22_est)
[1] 29799/278
>
> E2_est=c(E21_est,E22_est)
>
> E31_est=n_3*p_est;E31_est
[1] 12.17266
>
> fractions(E31_est)
[1] 1692/139
>
>
> E32_est=n_3*q_est;E32_est
[1] 59.82734
>
> fractions(E32_est)
[1] 8316/139
>
> E3_est=c(E31_est,E32_est)
>
>
> O=c(O_1,O_2,O_3);O
[1]  3 74 42 87  2 70
>
> E_est=c(E1_est,E2_est,E3_est)
>
```

```
> E_est
[1]  13.01799  63.98201  21.80935 107.19065  12.17266  59.82734
> D=O-E_est
> D2=D^2
>
> A=D^2/E_est
>
> tab=cbind(O,E_est,D,D2,A);tab
      O     E_est         D       D2         A
[1,]  3  13.01799 -10.01799 100.3600  7.709337
[2,] 74  63.98201  10.01799 100.3600  1.568566
[3,] 42  21.80935  20.19065 407.6622 18.692084
[4,] 87 107.19065 -20.19065 407.6622  3.803151
[5,]  2  12.17266 -10.17266 103.4830  8.501267
[6,] 70  59.82734  10.17266 103.4830  1.729695
>
>
> Q_cal=sum(A);Q_cal
[1] 42.0041
>
>
>
```

Vamos fazer a conclusão:

```
>
> Q_cal >Q_tab
[1] TRUE
```

Assim as proporções nas 3 cidades não podem ser consideradas iguais.

O nível descritivo do teste é dado por:

```
>
> nd=1-pchisq(Q_cal,gl);nd
[1] 7.56703e-10
>
>
```

Para descobrir onde estão estas diferenças vamos olhar os resíduos:

Direto pelo R temos que entrar com os dados na forma matricial. As frequências esperadas estimadas são obtidas também. Veja a saída:

```
>
> taxi=matrix(c(3,42,2,74,87,70),ncol=2)
>
> taxi=cbind(taxi,apply(taxi,1,sum))
>
> taxi=rbind(taxi,apply(taxi,2,sum))
>
```

```
>
>
> colnames(taxi)=c("Sim","Não","Total")
> rownames(taxi)=c("Nova York","Chicago","Pittsburgh", "Total")
>
>  taxi
            Sim Não Total
Nova York    3  74    77
Chicago     42  87   129
Pittsburgh   2  70    72
Total       47 231   278
>
>
> mod=chisq.test(taxi[1:3,1:3]);mod

Pearson's Chi-squared test

data:  taxi[1:3, 1:3]
X-squared = 42.004, df = 4, p-value = 1.665e-08

> names(mod)
[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
[7] "expected"  "residuals" "stdres"
> mod$observed
            Sim  Não  Total
Nova York    3  74    77
Chicago     42  87   129
Pittsburgh   2  70    72
> mod$expected
             Sim       Não     Total
Nova York  13.01799  63.98201    77
Chicago    21.80935 107.19065   129
Pittsburgh 12.17266  59.82734    72
>
> mod$res[1:2,1:2]
              Sim       Não
Nova York -2.776569  1.252424
Chicago    4.323434 -1.950167
> mod$stdres[1:2,1:2]
              Sim       Não
Nova York -3.412804  1.926511
Chicago    6.172152 -3.484149
>
```

Vamos comentar!!!!!!