Exercícios em Sala

ANÁLISE DE DADOS CATEGORIZADOS

Tailine J. S. Nonato

28 de março

Descrição da atividade

Estimar a proporção de casas com renda menor que 5 u.m.

Aula de inferências/estimativa de proporção.

```
set.seed(4)
n<-10
amostra <- sample(1:90,size=n, replace = FALSE)
salario<- c(9,2,2,8,3,5,8,3,6,8)
carac <- salario<5

df<-data.frame(amostra,salario,carac)
kable(df,align='c')
```

amostra	salario	carac
75	9	FALSE
51	2	TRUE
3	2	TRUE
71	8	FALSE
44	3	TRUE
58	5	FALSE
89	8	FALSE
56	3	TRUE
30	6	FALSE
62	8	FALSE

Estimativa pontual

```
pia <-sum(df$carac)/10
kable(pia,align='c')
```

x
0.4

Estimativa do intervalo de confiância

```
alpha<- 0.1
sup<- pia - qnorm(1-(alpha/2))*sqrt((pia*(1-pia))/n)
inf<- pia + qnorm(1-(alpha/2))*sqrt((pia*(1-pia))/n)

ic <- data.frame(sup,inf)
kable(round(ic,2),align='c')
```

sup	inf
0.15	0.65

Descrição da atividade

Jornal; preferem certa marca

$$\begin{cases} H_0 : \pi = 0.6 \\ H_1 : \pi > 0.6 \end{cases}$$

Dados

```
n=200
carac=156

prop<- carac/n

kable(prop,align='c')
```

x
0.78

Teste de hipóteses

```
pib<-0.6
alpha<-0.025

z<- (prop-pib)/sqrt((pib*(1-pib))/n)


kable(z, align = 'c')
```

x
5.196152

```
kable(z>qnorm(1-(alpha)),align='c')
```

x
TRUE

```
sup<- pib - qnorm(1-(alpha))*sqrt((pib*(1-pib))/n)
inf<- pib + qnorm(1-(alpha))*sqrt((pib*(1-pib))/n)

ic <- data.frame(sup,inf)
kable(round(ic,2),align='c')
```

sup	inf
0.53	0.67

09 de abril

Descrição da atividade

Calcular o risco relativo pontual e intervalar. Dados de um estudo de caso-controle: aspirina e placebo para prevenção de infarto.

Dados

```
alpha<- 0.05
n1=11034
n2=11037
p1=0.0171
p2=0.0094

p12=p1-p2
kable(p12,align='c')
```

x
0.0077

```
riscorelativo<-p1/p2
kable(riscorelativo,align='c')
```

x
1.819149

```
suplog<- log(p1/p2) - qnorm(1-(alpha/2))*sqrt(((1-p1)/(n1*p1))+((1-p2)/(n2*p2)))
inflog<- log(p1/p2) + qnorm(1-(alpha/2))*sqrt(((1-p1)/(n1*p1))+((1-p2)/(n2*p2)))
iclog<- c(suplog,inflog)
kable(iclog,align='c')
```

x
0.3602717
0.8364658

```
ic<- exp(iclog)
kable(ic,align='c')
```

x
1.433719
2.308195

Agora testando outras proporções:

```
alpha<- 0.05
n1=11034
n2=11037
p1=0.510
p2=0.501

p12=p1-p2
kable(p12,align='c')
```

x
0.009

```
riscorelativo<-p1/p2
kable(riscorelativo,align='c')
```

x
1.017964

```
suplog<- log(p1/p2) - qnorm(1-(alpha/2))*sqrt(((1-p1)/(n1*p1))+((1-p2)/(n2*p2)))
inflog<- log(p1/p2) + qnorm(1-(alpha/2))*sqrt(((1-p1)/(n1*p1))+((1-p2)/(n2*p2)))
iclog<- c(suplog,inflog)
kable(iclog,align='c')
```

x
-0.0082944
0.0439036

```
ic<- exp(iclog)
kable(ic,align='c')
```

x
0.9917399
1.0448817

Calculando Odds Ratio

```
alpha<- 0.05
n11=189
n21=104
n12=10845
n22=10933
p1=0.0171
p2=0.0094

odds1<- p1/(1-p1)
kable(odds1,align='c')
```

x
0.0173975

```
odds2<- p2/(1-p2)
kable(odds2,align='c')
```

x
0.0094892

```
oddsratio<- odds1/odds2
kable(oddsratio,align='c')
```

x
1.8334

```
#ou

oddsratio<- (n11*n22)/(n12*n21)
kable(oddsratio,align='c')
```

x
1.832054

```r
ASE<- sqrt((1/n11)+(1/n12)+(1/n21)+(1/n22))
kable(ASE,align='c')
```

x
0.1228416

```r
suplog<- log(odds1/odds2) + qnorm(alpha/2)*ASE
inflog<- log(odds1/odds2) - qnorm(alpha/2)*ASE
iclog<- c(suplog,inflog)
kable(iclog,align='c')
```

x
0.3654071
0.8469374

```r
ic<- exp(iclog)
kable(ic,align='c')
```

x
1.441100
2.332492

11 de abril

Descrição da atividade

Calcular o risco relativo pontual e intervalar. Dados de um estudo: acredita em vida após a morte e gênero.

Dados

```r
alpha=0.05
n11=435
n21=375
n12=147
n22=134
```

```
n1=582
n2=509

p1<- n11/(n11+n12)
kable(p1,align='c')
```

x
0.7474227

```
p2<- n21/(n21+n22)
kable(p2,align='c')
```

x
0.7367387

```
p12<- p1-p2
kable(p12,align='c')
```

x
0.010684

```
riscorelativo11<- p1/p2
kable(riscorelativo11,align='c')
```

x
1.014502

```
suplog11<- log(p1/p2) + qnorm(alpha/2)*sqrt(((1-p1)/(n1*p1))+((1-p2)/(n2*p2)))
inflog11<- log(p1/p2)- qnorm(alpha/2)*sqrt(((1-p1)/(n1*p1))+((1-p2)/(n2*p2)))
iclog<- c(suplog11,inflog11)
kable(iclog,align='c')
```

x
-0.0557973
0.0845924

```
ic<- exp(iclog)
kable(ic,align='c')
```

x
0.9457309
1.0882734

O risco relativo é de 1.01, com intervalo de confiança de 0.95 a 1.09. Isso significa que a chance de acreditar em vida após a morte é 1.02 vezes maior para mulheres do que para homens, ou seja 1.01% maior.

Risco relativo -> **probabilidade** de um evento ocorrer em um grupo dividido pela probabilidade de um evento ocorrer em outro grupo.

```
odds1<- p1/(1-p1)
kable(odds1,align='c')
```

x
2.959184

```
odds2<- p2/(1-p2)
kable(odds2,align='c')
```

x
2.798507

```
theta<- odds1/odds2
kable(oddsratio,align='c')
```

x
1.832054

```
#ou

theta<- (n11*n22)/(n12*n21)
kable(oddsratio,align='c')
```

x
1.832054

```
ASE<- sqrt((1/n11)+(1/n12)+(1/n21)+(1/n22))
kable(ASE,align='c')
```

x
0.1386756

```
suplog<- log(odds1/odds2) + qnorm(alpha/2)*ASE
inflog<- log(odds1/odds2) - qnorm(alpha/2)*ASE
iclog<- c(suplog,inflog)
kable(iclog,align='c')
```

x
-0.2159720
0.3276264

```
ic<- exp(iclog)
kable(ic,align='c')
```

x
0.8057579
1.3876705

Odds ratio é de 1.22, com intervalo de confiança de 0.89 a 1.67. Isso significa que a chance de acreditar em vida após a morte é 1.22 vezes maior para mulheres do que para homens.

Odds ratio -> **razão de chances** de um evento ocorrer em um grupo dividida pela razão de chances de um evento ocorrer em outro grupo.

16 de abril

Descrição da atividade

Testes de Qui Quadrado (Independência) para os dados de identificação partidária.

Hipóteses:

$$\begin{cases} H_0 : \text{X e Y independentes, ou seja } \pi_{ij} = \pi_{i.}\pi_{.j} \\ H_1 : \text{X e Y dependentes, ou seja } \pi_{ij} \neq \pi_{i.}\pi_{.j} \end{cases}$$

```
alpha<- 0.05
quicrit<- qchisq(1-alpha,df=2)
kable(quicrit,align='c')
```

x
5.991465

```
obs<- matrix(c(279,73,225,165,47,191),nrow=2,byrow=TRUE)
row.names(obs)<- c('Fem','Masc')
colnames(obs)<- c('Dem','Ind','Rep')
kable(obs,align='c')
```

	Dem	Ind	Rep
Fem	279	73	225
Masc	165	47	191

```
pim<- rowSums(obs)/sum(obs)
pjm<- colSums(obs)/sum(obs)
```

```r
qui <- chisq.test(obs)
quis <- data.frame(qui$statistic,qui$p.value)
colnames(quis)<- c('Qui','p-value')
kable(quis,align='c')
```

	Qui	p-value
X-squared	7.009544	0.0300536

```r
residuals <- (obs - qui$expected)/sqrt(qui$expected*(1-pim)*(1-pjm))
residualsf <- qui$residuals
```

Odds ratio

```r
oddsratio<- (obs[1,1]*obs[2,3])/(obs[1,3]*obs[2,1])
kable(oddsratio,align='c')
```

x
1.435394

Ou seja, a chance de se identificar como Democrata (em vez de Republicano) é 1.44 vezes maior para mulheres do que para homens.

```r
oddsratioh <- 1/oddsratio
kable(oddsratioh,align='c')
```

x
0.6966729

Ou seja, a chance de se identificar como Republicano (em vez de Democrata) é 0.7% menor para homens do que para mulheres.

18 de abril

Descrição da atividade

Calcular Razao de Verossimilhança para os dados de identificação partidária.

Hipóteses:

$$\begin{cases} H_0 : \text{X e Y independentes, ou seja } \pi_{ij} = \pi_{i.}\pi_{.j} \\ H_1 : \text{X e Y dependentes, ou seja } \pi_{ij} \neq \pi_{i.}\pi_{.j} \text{ para qualquer } i, j \end{cases}$$

```
G<- 2*sum(obs*log(obs/qui$expected))
kable(G,align='c')
```

x
7.002594

G^2 tem distribuição Qui Quadrado com V graus de liberdade, onde V é o nº de parâmetros sob H_1 - o nº de parâmetros sob H_0,

Sob H_1, $V_1 = (i * j) - 1$

Sob H_0, $V_0 = (i - 1) + (j - 1)$

Logo,

$V = V_1 - V_0$

$V = (i * j) - 1 - [(i - 1) + (j - 1)]$

$V = (j - 1)(i - 1)$

Como $i = 2$ e $j = 3$, então $V = 2$

Sabe-se que para $\alpha = 0.05$, $G_{crit} = 5.99$. Logo, como G^2 é maior que G_{crit}, rejeita-se H_0.

Em amostras grandes, G^2 terá um resultado muito próximo ao de χ^2. Mas em amostras pequenas, G^2 é mais confiável/robusto.

Exercício - Consumo de álcool e mal formação congênita

1. Identifique as variáveis em estudo e classifiqye quanto ao tipo.
2. Identifique a variável resposta e a variável explicativa.
3. Determine a proporção de presença de malformação congênita para cada nível de consumo de álcool e analise os resultados obtidos.
4. Verifique se a presença de malformação congênita está associada ao consumo de álcool das mães a um nível de significância de 5%. e tratando as variáveis como qualitativas nominais e ordinais.

a. Comente a decisão tomada considerando o nível de significância solicitado. A decisão seria a mesma para outro nível de significância? Qual seria sua recomendação?

b. Os pressupostos do teste foram atendidos? O que poderia ser feito?

5. Refaça o teste utilizado no item 4 agregando categorias para contornar o problema indicado no item 4b. Comente o a decisão tomada com relação aos aspectos considerados nos itens 4a e 4b.

6. Os resultados dos testes realizados permitem concluir sobre a existência de tendências na associação entre as variáveis considerando o nível de consumo de álcool? Justifique sua resposta.

7. Construa tabelas 2 x 2 que permitam medir a associação entre presença de mal formação congênita para cada nível de consumo de álcool em relação a ausência de consumo de álcool. Comente os resultados. Eles sugerem alguma tendência?

Solução

```
obs<- matrix(c(17066,48,14464,38,788,5,126,1,37,1),nrow=5,byrow=TRUE)
row.names(obs)<- c('0','<1','1-2','3-5','6+')
colnames(obs)<- c('Ausente','Presente')
kable(obs,align='c')
```

	Ausente	Presente
0	17066	48
<1	14464	38
1-2	788	5
3-5	126	1
6+	37	1

```
expected <- outer(rowSums(obs),colSums(obs))/sum(obs)
kable(expected,align='c')
```

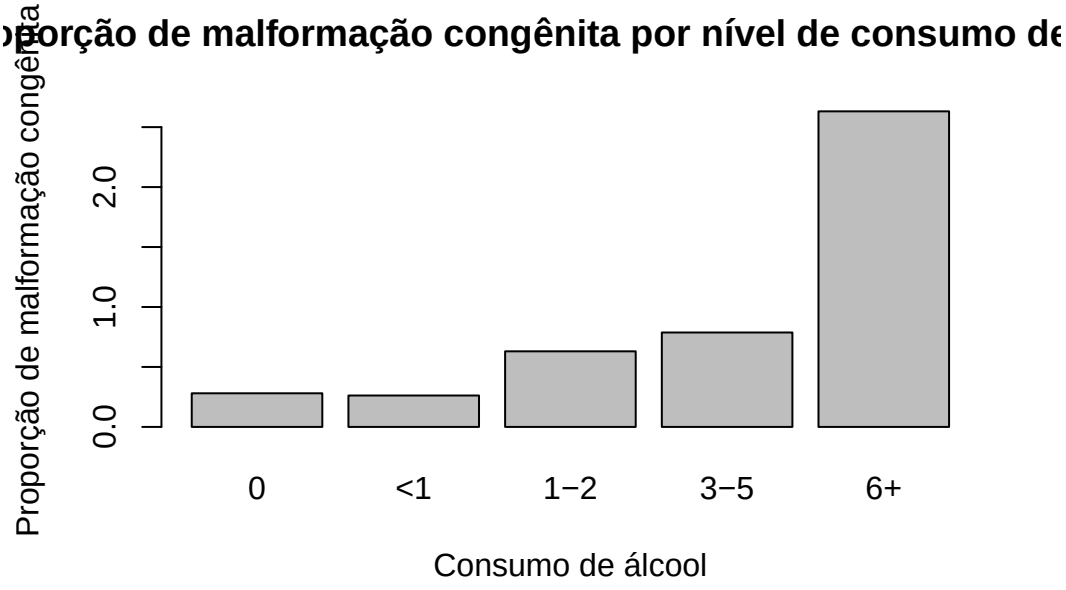
	Ausente	Presente
0	17065.13888	48.8611162
<1	14460.59624	41.4037576
1-2	790.73596	2.2640449
3-5	126.63741	0.3625898
6+	37.89151	0.1084914

1. Consumo de álcool (quantitativa discreta - categorizada em faixas ordinais) e malformação congênita (qualitative nominal).
2. Variável resposta: malformação congênita. Variável explicativa: consumo de álcool.
3. Proporção de malformação congênita para cada nível de consumo de álcool:

```
p_i <- 100*obs[,2]/rowSums(obs)
kable(p_i,align='c')
```

	x
0	0.2804721
<1	0.2620328
1-2	0.6305170
3-5	0.7874016
6+	2.6315789

```
barplot(p_i,main='Proporção de malformação congênita por nível de consumo de álcool',xlab=
```



4. Teste de Qui Quadrado para associação/independência.

Hipóteses:

$$\begin{cases} H_0 : \text{Consumo de álcool e malformação congênita são independentes} \\ H_1 : \text{Consumo de álcool e malformação congênita são dependentes} \end{cases}$$

em termos de proporções:

$$\begin{cases} H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \\ H_1 : \pi_{ij} \neq \pi_{i.}\pi_{.j} \end{cases}$$

```
gl <- (nrow(obs)-1)*(ncol(obs)-1)
qui_crit<- qchisq(1-alpha,df=gl)
kable(qui_crit,align='c')
```

x
9.487729

```
qui <- chisq.test(obs)
qui$residuals
```

```
         Ausente   Presente
0    0.006591843 -0.1231913
<1   0.028305154 -0.5289794
1-2 -0.097295603  1.8183038
3-5 -0.056641925  1.0585496
6+  -0.144828680  2.7066232
```

```
quis <- data.frame(qui$statistic,qui$p.value)
colnames(quis)<- c('Qui','p-value')
kable(quis,align='c')
```

	Qui	p-value
X-squared	12.08205	0.0167514

Com $\alpha = 0.05$,

$\chi^2_{obs} > \chi^2_{crit}$

p-valor $< \alpha$

Logo, rejeita-se H_0.

No entanto, é possível identificar algumas limitações:

- No teste de Qui Quadrado, como a necessidade de amostras grandes (na amostra existem caselas com frequência esperada menor que 5).
- No teste Qui Quadrado não é levada em consideração de ordem entre as categorias (consumo de álcool é uma variável ordinal).
- É possível identificar uma associação, mas não a direção da associação (quanto maior o consumo de álcool, maior a probabilidade de malformação congênita?). O apoio gráfico é útil para identificar a direção da associação.

a. A decisão muda para outros níveis de significância. Para $\alpha = 0.01$, a decisão seria a mesma. Para $\alpha = 0.10$, a decisão seria diferente. A recomendação é que sejam realizados testes de associação com amostras maiores e que sejam consideradas outras técnicas de análise.

b. Os pressupostos do teste Qui Quadrado são: independência entre as categorias, frequências esperadas maiores que 5 e amostras grandes. Uma solução possível é agrupar as categorias de consumo de álcool em três categorias: Zero consumo (0), Baixo consumo (1-2) e Alto consumo(3+). E caso ainda haja frequências esperadas menores que 5, é possível agrupar as categorias em duas categorias: Zero consumo (0) e Consumo (1+).

Realizando o teste de razão de verossimilhança:

```
G<- 2*sum(obs*log(obs/qui$expected))
pvalorG <- 1-pchisq(G,df=gl)
g2 <- data.frame(G,pvalorG)
colnames(g2)<- c('G','p-valor')
kable(g2,align='c')
```

G	p-valor
6.201998	0.1845623

Nesse caso, a decisão seria não rejeitar H_0, já que G^2 é menor que G_{crit}.

Calculando a Odds Ratio para consumo 0 e consumo 1-2:

De forma que sucesso é a presença de malformação congênita e fracasso é a ausência de malformação congênita.

```r
oddsratio<- (obs[1,1]*obs[3,2])/(obs[1,2]*obs[3,1])
kable(oddsratio,align='c')
```

x
2.255975

Assim, a chance de malformação congênita é 2.26 vezes maior para mães que consomem de 1 a 2 doses de álcool por dia do que para mães que não consomem álcool. Ou seja, a chance de malformação congênita é 0.44% menor para que não consomem álcool.

5. Teste de Qui Quadrado para associação/independência com categorias agrupadas.

```r
obs2<- matrix(c(17066,48,14464,38,951,7),nrow=3,byrow=TRUE)
row.names(obs2)<- c('0','<1','1+')
colnames(obs2)<- c('Ausente','Presente')
kable(obs2,align='c')
```

	Ausente	Presente
0	17066	48
<1	14464	38
1+	951	7

```r
expected2 <- outer(rowSums(obs2),colSums(obs2))/sum(obs2)
kable(expected2,align='c')
```

	Ausente	Presente
0	17065.1389	48.861116
<1	14460.5962	41.403758
1+	955.2649	2.735126

Hipóteses:

$$\begin{cases} H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \\ H_1 : \pi_{ij} \neq \pi_{i.}\pi_{.j} \end{cases}$$

```r
gl2 <- (nrow(obs2)-1)*(ncol(obs2)-1)
qui_crit2<- qchisq(1-alpha,df=gl2)
kable(qui_crit2,align='c')
```

x
5.991465

```r
qui2 <- chisq.test(obs2)
quis2 <- data.frame(qui2$statistic,qui2$p.value)
qui2$residuals
```

```
        Ausente   Presente
0   0.006591843 -0.1231913
<1  0.028305154 -0.5289794
1+ -0.137988941  2.5787991
```

```r
colnames(quis2)<- c('Qui','p-value')
kable(quis2,align='c')
```

	Qui	p-value
X-squared	6.965085	0.0307292

Com $\alpha = 0.05$,

$\chi^2_{obs} > \chi^2_{crit}$

p-valor $< \alpha$

Logo, rejeita-se H_0. No entanto, é possível observar que ainda com esse agrupamento, existem frequências esperadas menores que 5 e que H_0 não é rejeitada em todos os níveis de significância. Assim, é possível tentar mais uma vez mas agrupando as categorias em duas categorias: Zero consumo (0) e Consumo (1+).

```r
obs3<- matrix(c(17066,48,15415,45),nrow=2,byrow=TRUE)
row.names(obs3)<- c('Não consome alcool','Consome')
colnames(obs3)<- c('Ausente','Presente')
kable(obs3,align='c')
```

	Ausente	Presente
Não consome alcool	17066	48
Consome	15415	45

```
expected3 <- outer(rowSums(obs3),colSums(obs3))/sum(obs3)
kable(expected3,align='c')
```

	Ausente	Presente
Não consome alcool	17065.14	48.86112
Consome	15415.86	44.13888

Hipóteses:

$$\begin{cases} H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \\ H_1 : \pi_{ij} \neq \pi_{i.}\pi_{.j} \end{cases}$$

```
gl3 <- (nrow(obs3)-1)*(ncol(obs3)-1)
qui_crit3<- qchisq(1-alpha,df=gl3)
kable(qui_crit3,align='c')
```

x
3.841459

```
qui3 <- chisq.test(obs3)
quis3 <- data.frame(qui3$statistic,qui3$p.value)
colnames(quis3)<- c('Qui','p-value')
kable(quis3,align='c')
```

	Qui	p-value
X-squared	0.0056394	0.9401383

Com $\alpha = 0.05$,

$\chi^2_{obs} < \chi^2_{crit}$

p-valor $> \alpha$

Logo, não rejeita-se H_0 para nenhum nível de significância.

30 de abril

Dados: Pena de morte

```
obs <- matrix(c(53,430,15,176),nrow=2,byrow=TRUE)
row.names(obs) <- c('B','N')
colnames(obs) <- c('Sim','Não')
kable(obs, align='c')
```

	Sim	Não
B	53	430
N	15	176

Hipóteses:

$$\begin{cases} H_0 : \text{Veredito de pena de morte independe da raça do réu} \\ H_1 : \text{Existe dependência entre veredito de pena de morte e raça do réu} \end{cases}$$

```
qui <- chisq.test(obs)
kable(qui$expected, align='c')
```

	Sim	Não
B	48.72997	434.27
N	19.27003	171.73

```
res <- data.frame(qui$statistic, qui$p.value)
kable(res,align='c')
```

	qui.statistic	qui.p.value
X-squared	1.144741	0.2846528