Análise de resíduos e a estatística qui-quadrado de Pearson

Gustavo Almeida Silva

Table 1:

	Cura	Obito	Permanece.internado
>80	57	90	3
<80	180	83	37

a) O que o valor calculado para a estatística qui-quadrado nos fala sobre a hipótese nula? calcular pelo R a estatística qui-quadrado de Pearson e tomar a decisão estatística, explicitando H0, a decisão tomada em relação a H0 e escrevendo a conclusão do teste de hipóteses. Sugestão: após criar a matriz representando a tabela, usar a função chisq.test()

O **Teste Qui-Quadrado de Pearson** é um metodo que busca determinar se existe correlação estatística significante entre 2 variáveis categóricas. O teste possui as seguintes hipóteses

H_0 : Variaveis não possuem correlação significativa

H_1 : Variaveis possuem correlação significativa

```
test = as.table(as.matrix(data)) |>
  chisq.test()

data.frame('Estatistica de Teste'=test$statistic,
           'P-valor'=test$p.value) |>
      kbl(
        booktabs = TRUE,
        escape = FALSE,
        caption = "Resultado Teste Qui-Quadrado"
      ) |>
      kable_classic() |>
      kable_styling(
        font_size = 9,
        latex_options = "HOLD_position"
      )
```

Table 2: Resultado Teste Qui-Quadrado

	Estatistica.de.Teste	P.valor
X-squared	48.39601	0

Assim, o p-valor retornado é baixissimo, ficando do nivel de significancia de 1%. Assim, rejeita-se H_0, ou seja, as variaveis de idade e status possuem correlação significativa

b) Apresentar a tabela acima com todos os três resíduos discutidos em aula, usando como formato de saída aquele considerado padrão do SPSS, obtido pelos argumentos correspondentes da função do R sugerida a seguir. Sugestão: usar a função CrossTable()1 do pacote gmodels do R

O valor esperado para cada celula é dada pela seguinte tabela:

```
test$expected|>
      kbl(
        booktabs = TRUE,
        escape = FALSE,
        caption = "Valor esperado"
      ) |>
      kable_classic() |>
      kable_styling(
        font_size = 9,
        latex_options = "HOLD_position"
      )
```

Table 3: Valor esperado

	Cura	Obito	Permanece.internado
>80	79	57.66667	13.33333
<80	158	115.33333	26.66667

A partir da tabela podemos calcular 3 tipos diferentes de residuos, são eles:

- Resíduos Brutos
  - A frequência observada em cada célula menos a frequência esperada (na hipótese de independência), ou seja:

$$e_{ij} = n_{ij} - E_{ij}$$

Calculando tal residuo:

```
(test$observed-test$expected)|>
      kbl(
        booktabs = TRUE,
        escape = FALSE,
        caption = "Resíduos Brutos"
      ) |>
      kable_classic() |>
      kable_styling(
        font_size = 9,
        latex_options = "HOLD_position"
      )
```

Table 4: Resíduos Brutos

	Cura	Obito	Permanece.internado
>80	-22	32.33333	-10.33333
<80	22	-32.33333	10.33333

- Resíduos de Pearson
  - A frequência observada em cada célula menos a frequência esperada, dividido pela raiz quadrada da frequência esperada, conforme equação abaixo:

$$e_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

```
test$residuals|>
      kbl(
        booktabs = TRUE,
        escape = FALSE,
        caption = "Residuos de Pearson"
      ) |>
      kable_classic() |>
      kable_styling(
        font_size = 9,
        latex_options = "HOLD_position"
      )
```

Table 5: Residuos de Pearson

	Cura	Obito	Permanece.internado
>80	-2.475193	4.257827	-2.829900
<80	1.750226	-3.010739	2.001041

- Resíduos ajustados de Pearson:
  - Como os resíduos de Pearson não apresentam variâncias compatíveis com a distribuição normal padrão, sugere-se um ajuste a este resíduo (ver Haberman, 19734), que consiste em dividir o resíduo de Pearson pelo desvio-padrão de todos os resíduos:

$$e_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{[(1 - \frac{n_{i+}}{N})(1 - \frac{n_{+j}}{N})]}}$$

```
test$stdres|>
      kbl(
        booktabs = TRUE,
        escape = FALSE,
        caption = "Resíduos ajustados de Pearson"
      ) |>
      kable_classic() |>
```

```
      kable_styling(
        font_size = 9,
        latex_options = "HOLD_position"
      )
```

Table 6: Resíduos ajustados de Pearson

	Cura	Obito	Permanece.internado
>80	-4.406271	6.646608	-3.63104
<80	4.406271	-6.646608	3.63104

Para representar os 3 residuos no formato do **SPSS**, podemos utlizar a função *CrossTable()* do pacote **gmodels**:

```
data|>
  as.matrix()|>
  as.table()|>
  gmodels::CrossTable(resid=T,sresid=T,asresid=T,format = 'SPSS')
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |                Residual |
## |            Std Residual |
## |           Adj Std Resid |
## |-------------------------|
##
## Total Observations in Table:  450
##
##              |
##              |                Cura  |               Obito  | Permanece.internado  |           Row To
## -------------|----------------------|----------------------|----------------------|-----------------
##          >80 |                  57  |                  90  |                   3  |               150
##              |               6.127  |              18.129  |               8.008  |
##              |              38.000% |              60.000% |               2.000% |             33.333%
##              |              24.051% |              52.023% |               7.500% |
##              |              12.667% |              20.000% |               0.667% |
##              |            -22.000  |              32.333  |            -10.333  |
##              |              -2.475  |               4.258  |              -2.830  |
##              |              -4.406  |               6.647  |              -3.631  |
## -------------|----------------------|----------------------|----------------------|-----------------
##          <80 |                 180  |                  83  |                  37  |               300
##              |               3.063  |               9.065  |               4.004  |
##              |              60.000% |              27.667% |              12.333% |             66.667%
##              |              75.949% |              47.977% |              92.500% |
##              |              40.000% |              18.444% |               8.222% |
```

```
##              |             22.000  |            -32.333  |             10.333  |
##              |              1.750  |             -3.011  |              2.001  |
##              |              4.406  |             -6.647  |              3.631  |
## -------------|---------------------|---------------------|---------------------|---------
## Column Total |                237  |                173  |                 40  |             450
##              |            52.667% |            38.444% |             8.889% |
## -------------|---------------------|---------------------|---------------------|---------
##
##
```

c) Escolher duas células quaisquer da tabela e escrever, para cada célula escolhida, um resumo analítico sobre os resíduos calculados que contenha uma conclusão sobre o que foi observado.

A primeira célula a ser analisada é a de posição [1,2], que é o numero de óbitos de pessoas com mais de 80 anos. Analisando os residuos é possivel ver que a celula possui um numero de observações maior que o valor esperado, possuindo um residuo ajustado de Pearson de 6.647, isso indica que o numero de óbitos entre pessoas com mais de 80 anos é maior do que o valor esperado calculado via tabela de contigencia

A segunda célula a ser analisada é a de posição [2,1], que é o numero de pessoas com menos 80 curadas . Analisando os residuos é possivel ver que a celula possui um numero de observações maior que o valor esperado, possuindo um residuo ajustado de Pearson de 4.406, isso indica que o numero de pessoas curadas que possuem menos de 80 anos é maior do que o valor esperado calculado via tabela de contigencia