DEPARTAMENTO DE MATEMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Estatística Computacional

André Couto
António Santos
Diogo Santiago
Diogo Reis

Trabalho realizado no âmbito de **Estatística Computacional**
Disciplina da Licenciatura em Matemática

Introdução

Foi realizado um estudo com o objetivo de identificar fatores de risco para o desenvolvimento de insuficiência renal aguda após cirurgia cardíaca. Formaram-se dois grupos de pacientes submetidos a cirurgia cardíaca, sendo um deles constituído por 42 pacientes que desenvolveram insuficiência renal aguda após a cirurgia e o outro constituído por 42 pacientes que não desenvolveram insuficiência renal aguda após a cirurgia. A informação recolhida encontra-se organizada no ficheiro *Renal* segundo as seguintes variáveis:

- *InsufRenal* (desenvolveu insuficiência renal aguda: Sim, Não)
- *Idade* (em anos)
- *DurEstadia* (duração da estadia no hospital, em dias)
- *Sexo* (Masculino, Feminino)
- *Hipert* (historial de hipertensão: Sim, Não)
- *Diabetes* (historial de diabetes mellitus: Sim, Não)
- *FalCardiaca* (historial de falência cardíaca: Sim, Não)
- *TempoBloco* (duração da estadia no bloco operatório, em horas)
- *TempoCEC* (duração da circulação extracorpórea, em horas)
- *NumComp* (número de complicações)

Assim este trabalho tem como objetivo responder às questões pretendidas com a ajuda do software R.

Exercício 1

a)

Em primeiro lugar, com o objetivo de fazer um estudo descritivo da amostra no que diz respeito à variável *TempoBloco* no grupo dos pacientes que desenvolveram insuficiência renal aguda após a cirurgia e no grupo dos que não desenvolveram essa insuficiência, avaliámos as medidas de localização (média, mediana, moda, mínimo, máximo e quartis) e as medidas de dispersão (variância, variância corrigida, desvio padrão, desvio padrão corrigido, coeficiente de variação, amplitude, amplitude interquartis e coeficiente de assimetria).

Assim, obtivemos os seguintes *outputs* para nos ajudar a obter esses valores, sendo que analisamos os dados dos pacientes no seu total, e divididos nos que desenvolveram ou não insuficiência renal.

```
     mean       sd IQR        cv  skewness 0% 25% 50% 75% 100%  n
 4.821429 1.309472   1 0.2715941 0.9493941  2   4   5   5    9 84
```

```
        mean        sd IQR        cv   skewness 0% 25% 50% 75% 100%
Nao 4.357143 0.8136467   1 0.1867386 -0.4395416  2   4 4.5   5    6
Sim 5.285714 1.5386994   2 0.2911053  0.5944859  3   4 5.0   6    9
    TempoBloco:n
Nao           42
Sim           42
```

Figura 1: Parâmetros obtidos em R

```
> discreteCounts(Renal[,"TempoBloco", drop=FALSE])
Distribution of TempoBloco
      Count Percent
2         1    1.19
3         6    7.14
3.5       6    7.14
4        20   23.81
4.5       8    9.52
5        24   28.57
5.5       2    2.38
6         8    9.52
7         4    4.76
8         4    4.76
9         1    1.19
Total    84   99.98
```

```
> discreteCounts(Renal[Renal$InsufRenal == 'Sim'
+   # InsufRenal = Sim
Distribution of TempoBloco
      Count Percent
3         3    7.14
3.5       3    7.14
4         7   16.67
4.5       2    4.76
5        11   26.19
5.5       1    2.38
6         6   14.29
7         4    9.52
8         4    9.52
9         1    2.38
Total    42   99.99
```

```
> discreteCounts(Renal[Renal$InsufRenal == 'Nao'
+   # InsufRenal = Nao
Distribution of TempoBloco
      Count Percent
2         1    2.38
3         3    7.14
3.5       3    7.14
4        13   30.95
4.5       6   14.29
5        13   30.95
5.5       1    2.38
6         2    4.76
Total    42   99.99
```

Figura 2: Parâmetros obtidos em R para determinar a moda

Precisamos ainda de calcular alguns parâmetros e para isso utilizámos algumas fórmulas:

Amplituda da amostra = Máximo-Mínimo;

Moda é a observação ou observações cuja frequência é $\geq$ que as duas adjacentes;

Variância corrigida = $\hat{s}^2$, com $\hat{s}$ o desvio padrão corrigido;

Variância $= s^2 = \frac{n-1}{n}\hat{s}^2$
Desvio Padrão $= s = \sqrt{s^2}$
Posto isto obtemos:

Parâmetros	PIRAC	NPIRAC	Total de Doentes
n	42	42	84
Média	5.285714	4.357143	4.821429
Mediana	5	4.5	5
Moda	4,5,6 e 8	4 e 5	3,4,5,6 e 8
Mínimo	3	2	2
Máximo	9	6	9
$Q_{\frac{1}{4}}$	4	4	4
$Q_{\frac{3}{4}}$	6	5	5
Variância	2.3112245	0.64625855	1.6943036
Variância Corrigida	2.3675958	0.66202095	1.7147169
Desvio Padrão	1.5202712	0.8039021	1.3016542
Desvio Padrão Corrigido	1.5386994	0.8136467	1.309472
Amplitude da Amostra	6	4	7
Amplitude Interquartis	2	1	1
Coeficiente de Assimetria	0.5944859	-0.4395416	0.9493941
Coeficiente de Variação	0.2911053	0.1867386	0.2715941

PIRAC: Pacientes que desenvolveram Insuficiência Renal Aguda após a cirurgia.
NPIRAC:Pacientes que não desenvolveram Insuficiência Renal Aguda após a cirurgia

Quanto à existência de *outliers* temos

- **PIRAC**

  $Q_{\frac{1}{4}} - 1.5IQR = 4 - 1.5 \times 2 = 1$

  $Q_{\frac{3}{4}} + 1.5IQR = 6 + 1.5 \times 2 = 9$

  Concluímos então que existem *outliers* para valores inferiores a 1 e superiores a 9. O máximo e o mínimo da amostra são, respetivamente, 9 e 3, pelo que os dados observados estão no intervalo $[3, 9]$. Assim, não há *outliers* para esta amostra (como poderemos confirmar no diagrama de extremos e quartis mais à frente apresentado).

- **NPIRAC**

  $Q_{\frac{1}{4}} - 1.5IQR = 4 - 1.5 \times 1 = 2.5$

  $Q_{\frac{3}{4}} + 1.5IQR = 5 + 1.5 \times 1 = 6.5$

  Concluímos então que existem *outliers* para valores inferiores a 2.5 e superiores a 6.5. O máximo e o mínimo da amostra são, respetivamente, 6 e 2, pelo que os dados observados estão no intervalo $[2, 6]$. Assim, há *outliers* à esquerda, mas não há à direita (como poderemos confirmar no diagrama de extremos e quartis mais à frente apresentado).

$Q_{\frac{1}{4}} - 3IQR = 4 - 3 \times 1 = 1$

$Q_{\frac{3}{4}} + 3IQR = 5 + 3 \times 1 = 8$

Assim, há *outliers severos* para valores inferiores a 1 e superiores a 8. Como os dados estão no intervalo $[2, 6]$ os *outliers* que existem são moderados.

- **Total de Doentes**

  $Q_{\frac{1}{4}} - 1.5IQR = 4 - 1.5 \times 1 = 2.5$

  $Q_{\frac{3}{4}} - 1.5IQR = 5 - 1.5 \times 1 = 6.5$

  Concluímos então que existem *outliers* para valores inferiores a 2.5 e superiores a 6.5. O máximo e o mínimo da amostra são, respetivamente, 9 e 2, pelo que os dados observados estão no intervalo $[2, 9]$. Assim há *outliers* à esquerda e à direita (como poderemos confirmar no diagrama de extremos e quartis mais à frente apresentado)

  $Q_{\frac{1}{4}} - 3IQR = 4 - 3 \times 1 = 1$

  $Q_{\frac{3}{4}} + 3IQR = 5 + 3 \times 1 = 8$

  Assim, há *outliers* severos para valores inferiores a 1 e superiores a 8. Como os dados estão no intervalo $[2, 9]$ os *outliers* que existem à esquerda são moderados e os *outliers* que existem à direita são moderados no intervalo $[6.5; 8]$ e severos no intervalo $]8, 9]$.

Quanto ao estudo descritivo gráfico podemos confirmar a existência dos *outliers* que vimos em cima, através dos diagramas de extremos e quartis:

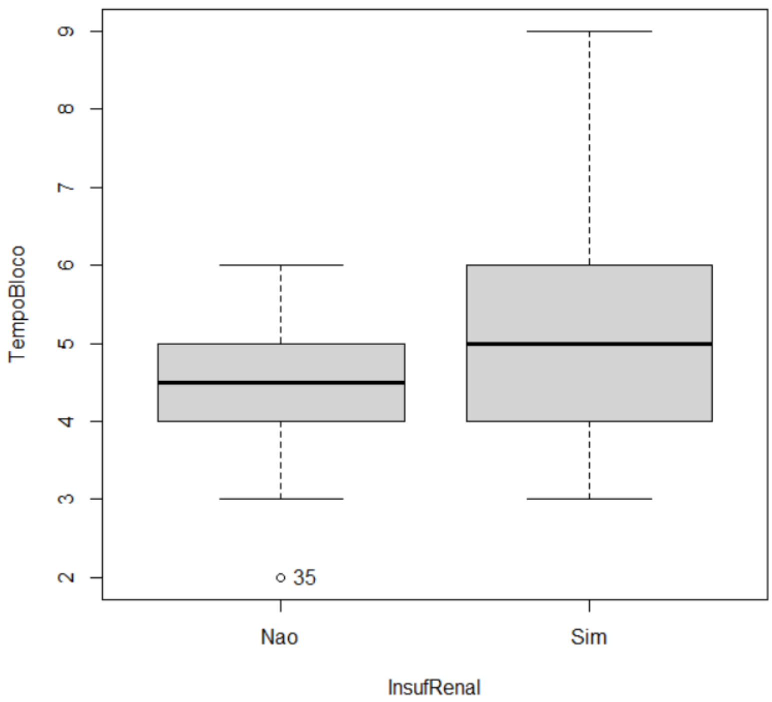

Figura 3: Diagrama de Extremos e Quartis para NPIRAC e PIRAC

Assim confirma-se que 2 é um *outlier* moderado à esquerda para **NPIRAC** e que não existem *outliers* para **PIRAC**.

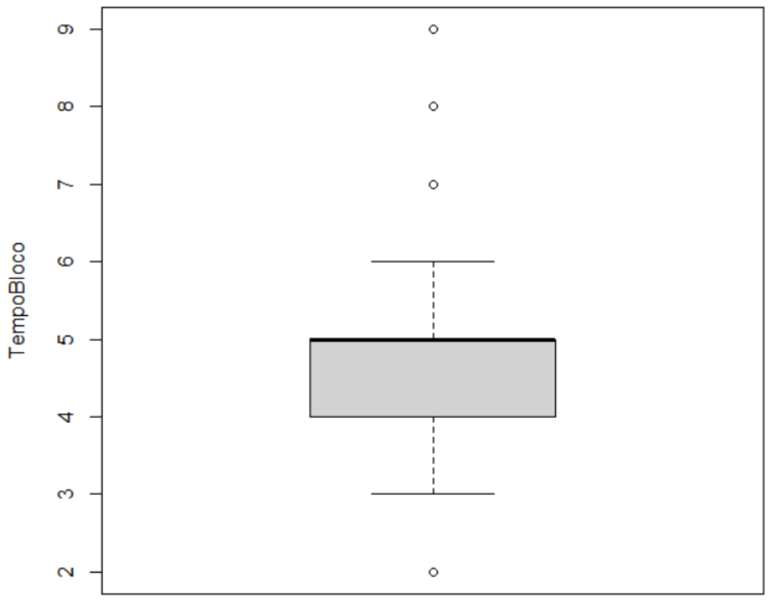

Figura 4: Diagrama de Extremos e Quartis para Total de Doentes

Podemos então concluir que 2 é um *outlier* moderado à esquerda, 7 e 8 são *outliers* moderados à direita e 9 é um *outlier* severo à direita.

Observemos agora os histogramas e os diagramas de Caule e Folhas:

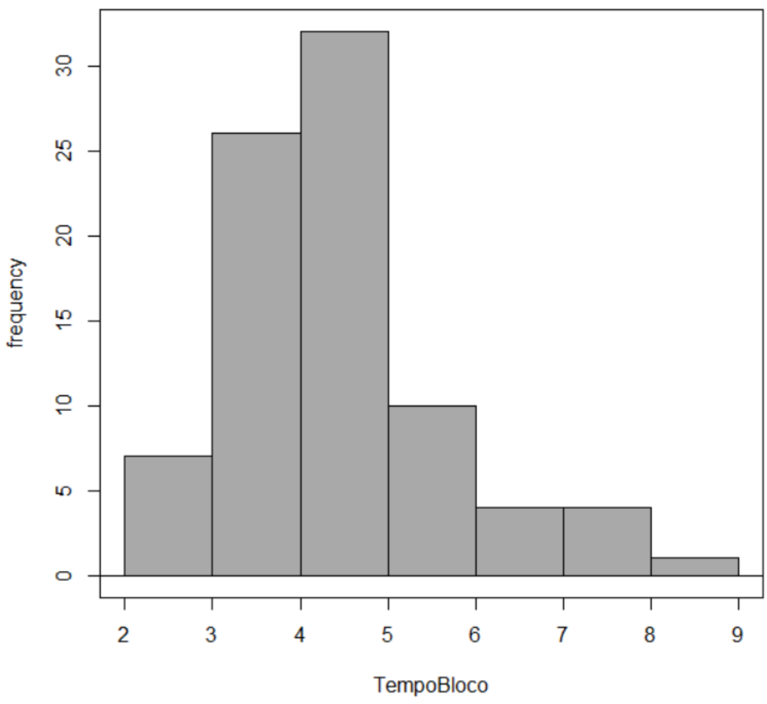

Figura 5: Histograma para Total de Doentes

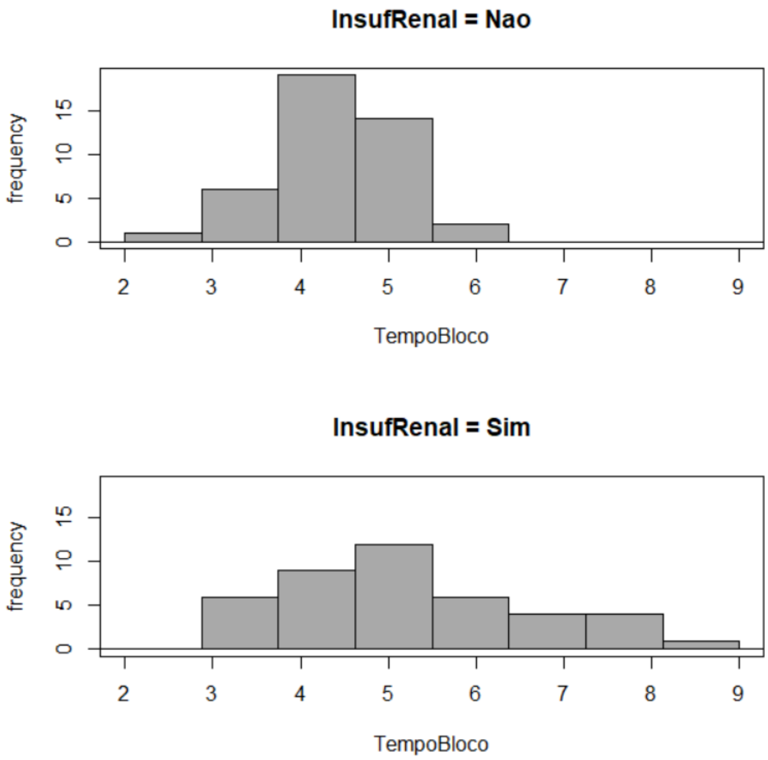

Figura 6: Histogramas para NPIRAC e PIRAC, respetivamente

```
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 84
LO: 2
    7    3* | 000000
          t |
   13     f | 555555
          s |
         3. |
   33    4* | 00000000000000000000
          t |
   41     f | 55555555
          s |
         4. |
  (24)   5* | 000000000000000000000000
          t |
   19     f | 55
          s |
         5. |
   17    6* | 00000000
HI: 7 7 7 7 8 8 8 8 9
```

Figura 7: Diagrama de Caule e Folhas para Total de Doentes

```
  1 | 2: represents 1.2, leaf unit: 0.1
TempoBloco[InsufRenal == "Nao"]
                                TempoBloco[InsufRenal == "Sim"]
LO: 2
   4              000| 3* |000              3
                     |  t |
   7              555|  f |555              6
                     |  s |
                     | 3. |
  20    0000000000000| 4* |0000000         13
                     |  t |
 (6)           555555|  f |55              15
                     |  s |
                     | 4. |
  16    0000000000000| 5* |00000000000   (11)
                     |  t |
   3                5|  f |5               16
                     |  s |
                     | 5. |
   2               00| 6* |000000          15
                          HI: 7 7 7 7 8 8 8 8
                          9
n:                 42     42
```

Figura 8: Diagrama de Caule e Folhas para NPIRAC e PIRAC

Averiguemos agora os níveis de assimetria com base nos gráficos acima e com os cálculos que faremos de seguida.

- **Total de Doentes**

  $\overline{x} = 4.821529 < 5 = Med$, logo existe uma assimetria negativa

  $$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} = 0.2626505406$$

  $|\frac{Skewness}{SES}| = |\frac{0.9493941}{0.2626505406}| = 3.614666461 > 2$, pelo que a assimetria é acentuada como também podemos ver nos gráficos acima já que as observações estão localizadas muito mais à direita.

- **NPIRAC**

  $\overline{x} = 4.357143 < 4.5 = Med$, logo existe uma assimetria negativa

  $$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} = 0.3653606056$$

  $|\frac{Skewness}{SES}| = |\frac{-0.4395416}{0.3653606056}| = 1.203035011 < 2$, pelo que a assimetria não é acentuada como também podemos ver nos gráficos acima já que as observações estão localizadas mais à direita.

- **PIRAC**

  $\overline{x} = 5.285714 > 5 = Med$, logo existe uma assimetria positiva

  $$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} = 0.3653606056$$

  $|\frac{Skewness}{SES}| = |\frac{0.5944859}{0.3653606056}| = 1.627120962 < 2$, pelo que a assimetria não é acentuada como também podemos ver nos gráficos acima já que as observações estão localizadas mais à esquerda.

b)

i) Como a duração da estadia no bloco operatório(*DurEstadia*) é uma variável quantitativa e o grupo a que os doentes pertencem(*InsufRenal*) é uma variável qualitativa vamos calcular o coeficiente *Eta* para determinar o grau de associação:

```
> eta(InsufRenal,DurEstadia)
[1] 0.4177293
```

Figura 9: Output do Coeficiente Eta

Assim como o valor obtido é 0.4177293 e está mais próximo de 0 indica fraca associação.

ii) Como o historial de hipertensão(*Hipert*) e o grupo a que os doentes pertencem(*InsufRenal*) são ambas variáveis qualitativas nominais, vamos calcular o *Coeficiente de Contingência de Pearson* através do R:

```
> library(vcd)

> tab=xtabs(~Hipert+InsufRenal, data = Renal)

> assocstats(tab)
                     X^2 df P(> X^2)
Likelihood Ratio 0.83149  1  0.36184
Pearson          0.82963  1  0.36238

Phi-Coefficient   : 0.099
Contingency Coeff.: 0.099
Cramer's V        : 0.099
```

Figura 10: Output do Coeficiente de Contingência de Pearson

Assim como o valor obtido é 0.099 é bastante próximo de 0 existe muito fraca associação.

Exercício 2

Nos próximos três exercícios vamos utilizar nível de significância 0.05.

Consideremos as variáveis aleatórias:

X = "duração da estadia no bloco operatório dos pacientes que não desenvolvem insuficiência renal aguda após a cirurgia"

Y = "duração da estadia no bloco operatório dos pacientes que desenvolvem insuficiência renal aguda após a cirurgia"

Comecemos por testar a normalidade das variáveis. As hipóteses em teste são:

H_0: X tem distribuição normal vs H_1: X não tem distribuição normal

H_0': Y tem distribuição normal vs H_1': Y não tem distribuição normal

Como a dimensão das amostras X e Y é $42 > 30$, usamos o teste de Lilliefors (Kolmogorov-Smirnov).

Obtivemos os seguintes resultados para os testes de ajustamento das variáveis X e Y.

```
--------
InsufRenal = Nao

       Lilliefors (Kolmogorov-Smirnov) normality test

data:  TempoBloco
D = 0.16622, p-value = 0.005117

--------
InsufRenal = Sim

       Lilliefors (Kolmogorov-Smirnov) normality test

data:  TempoBloco
D = 0.1927, p-value = 0.0004397

--------
```

Figura 11: Teste de ajustamento para X e Y

O p-valor para a mostra X é $0.005117 < 0.05$, logo rejeitamos H_0 pelo que X não tem distribuição normal.

O p-valor para a amostra Y é $0.0004397 < 0.05$, logo rejeitamos H_0' pelo que Y não tem distribuição normal.

Nota: Apesar das variáveis não terem distribuição normal, como têm dimensão $42 > 30$ e não têm assimetria acentuada (provámos em 1. (a)) podemos aplicar o teste T para as médias de X e Y.

Em primeiro lugar é necessário testar a igualdade das variâncias de X e de Y efetuando o teste F. Sejam σ_X^2 e σ_Y^2 as variâncias de X e Y, respetivamente. Queremos testar

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2$$

Neste caso, como não temos normalidade das amostras, para testar a igualdade das variâncias usamos o teste de Levene. Obtivemos o seguinte resultado.

```
> Tapply(TempoBloco ~ InsufRenal, var, na.action=na.omit,
+   data=Renal) # variances by group
      Nao       Sim
0.6620209 2.3675958

> leveneTest(TempoBloco ~ InsufRenal, data=Renal, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value    Pr(>F)
group  1  13.881 0.0003566 ***
      82
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 12: Teste de Levene para igualdade de variâncias

O p-valor, PR(>F), é $0.0003566 < 0.05$ pelo que rejeitamos H_0 ao n.s. 0.05 e não temos igualdade de variâncias.

Sejam então $m_X = E(X)$ e $m_Y = E(Y)$. Queremos verificar se a duração da estadia no bloco operatório dos pacientes que desenvolvem insuficiência renal aguda após a cirurgia é significativamente superior à dos pacientes que não desenvolvem tal insuficiência.

As hipóteses a testar são:

$$H_0 : m_Y - m_X = 0 \text{ vs } H_1 : m_Y - m_X < 0$$

Nota: A hipótese H_1 seria, teoricamente, $m_X - m_Y > 0$. No entanto, pela forma como o R organiza as variáveis vamos efetuar, de modo equivalente, $m_Y - m_X < 0$.

Obtivemos o seguinte resultado.

```
        Welch Two Sample t-test

data:  TempoBloco by InsufRenal
t = -3.4574, df = 62.266, p-value = 0.0004944
alternative hypothesis: true difference in means between group Nao and
95 percent confidence interval:
       -Inf -0.4801291
sample estimates:
mean in group Nao mean in group Sim
         4.357143          5.285714
```

Figura 13: Teste T para diferença de médias

Como p-valor$= 0.0004944 < 0.05$ rejeitamos H_0 ao n.s. 0.05. Assim, concluímos que a duração da estadia no bloco operatório dos pacientes que desenvolveram insuficiência renal aguda, é significativamente superior à dos pacientes que não desenvolveram tal insuficiência.

Exercício 3

a)

Vamos efetuar o teste de independência do Qui-quadrado. As hipóteses a testar são:

H_0 : As variáveis são independentes vs H_1: As variáveis não são independentes

A tabela de contingência vai ser uma tabela 2x2, pelo que vamos efetuar a correção de continuidade de Yates.

Obtivemos assim o seguinte resultado.

```
Frequency table:
          Hipert
InsufRenal Nao Sim
       Nao  17  25
       Sim  13  29

Total percentages:
       Nao  Sim Total
Nao   20.2 29.8    50
Sim   15.5 34.5    50
Total 35.7 64.3   100

        Pearson's Chi-squared test with Yates' continuity correction

data:  .Table
X-squared = 0.46667, df = 1, p-value = 0.4945


Expected counts:
          Hipert
InsufRenal Nao Sim
       Nao  15  27
       Sim  15  27
```

Figura 14: Teste de independência do Qui-quadrado

Como podemos observar não existem valores inferiores a 5 ou a 1, logo podemos considerar o p-valor, que é calculado de forma aproximada.. Assim, como p-valor$=$ $0.4945 > 0.05$, aceitamos H_0 ao n.s. 0.05 e concluímos que não há associação significativa entre as variáveis *InsufRenal* e *Hipert*.

b)

Vamos efetuar o teste de independência do Qui-quadrado. As hipóteses a testar são:

H_0 : As variáveis são independentes vs H_1: As variáveis não são independentes

A tabela de contingência vai ser uma tabela 2x2, pelo que vamos efetuar a correção de continuidade de Yates.

Obtivemos assim o seguinte resultado.

```
Frequency table:
          FalCardiaca
InsufRenal Nao Sim
       Nao  32  10
       Sim  22  20

Total percentages:
       Nao  Sim Total
Nao   38.1 11.9    50
Sim   26.2 23.8    50
Total 64.3 35.7   100

        Pearson's Chi-squared test with Yates' continuity correction

data:  .Table
X-squared = 4.2, df = 1, p-value = 0.04042


Expected counts:
          FalCardiaca
InsufRenal Nao Sim
       Nao  27  15
       Sim  27  15
```

Figura 15: Teste de independência do Qui-quadrado

Como podemos observar não existem valores inferiores a 5 ou a 1, logo podemos considerar o p-valor, que é calculado de forma aproximada. Assim, como p-valor$= 0.04042 < 0.05$, rejeitamos H_0 ao n.s. 0.05 e concluímos que há associação entre as variáveis *InsufRenal* e *FalCardiaca*.

Duas medidas dessa associação são o coeficiente de contingência V de Cramer (que toma valores entre 0 e 1) e o coeficiente de contingência de Pearson (que, no caso de tabelas 2x2, toma valores entre 0 e 0.71, aproximadamente). Obtivemos os seguintes resultados.

```
Contingency Coeff.: 0.241
Cramer's V        : 0.248
```

Figura 16

Estes valores não são muito elevados sendo mais próximos de zero do que do valor máximo que cada um pode atingir. Assim, a associação entre as duas variáveis não é elevada.

Exercício 4

Em primeiro lugar, começámos por criar um subficheiro com a condição $Falcardiaca ==$ "Sim", ficheiro esse que designámos por RenalFalCardiacaSim.

Seja p a proporção de pacientes que têm historial de falência cardíaca que desenvolvem insuficiência renal aguda após cirurgia cardíaca. Queremos testar

$$H_0 : p = 0.6 \text{ vs } H_1 : p > 0.6$$

Obtivemos o seguinte resultado.

```
Frequency counts (test is for first level):
InsufRenal2
Sim Nao
 20  10

        Exact binomial test

data:  rbind(.Table)
number of successes = 20, number of trials = 30, p-value = 0.2915
alternative hypothesis: true probability of success is greater than 0.6
95 percent confidence interval:
 0.5005613 1.0000000
sample estimates:
probability of success
             0.6666667
```

Figura 17

Como p-valor= $0.2915 > 0.05$ aceitamos H_0 ao n.s. 0.05. Assim, não podemos afirmar que a percentagem de utentes que têm historial de falência cardíaca que desenvolvem insuficiência renal aguda após cirurgia cardíaca é significativamente superior a 60%