Bioestatística com SPSS

Notas de Apoio

Mª Rosário Ramos
Pedro Serranho

UNIVERSIDADE DE COIMBRA
UNIVERSIDADE ABERTA
EDUCAÇÃO A DISTÂNCIA
DISTANCE EDUCATION

Conteúdo

Notas Introdutórias para a frequência do curso

Este texto pretende guiar o estudante durante este curso, sobre a aplicações da Bioestatística e utilização do SPSS nesse âmbito. Para um recurso múltimédia sobre a motivação destas aplicações, por favor visite:

A Universidade Aberta disponibiliza licenças de SPSS para os seus estudantes e formandos. As instruções de instalação do software e a forma de requerer o número de licença devem ser consultadas aqui:

http://www.univ-ab.pt/spss/

Os exemplos deste texto foram realizados com recurso ao IBM SPSS®.

Serão cedidos ao longo do curso vários recursos de estudo, sendo que este texto servirá de ligação entre a matéria teórica e os vários recursos cedidos. Todos os recursos serão cedidos no espaço da plataforma online de elearning, sendo que os recursos multimédia serão listados no endereço:

http://www.univ-ab.pt/ pserranho/BioestatisticaSPSS/

Além deste texto, podem também considerar outra bibliografia sobre a utilização do SPSS em Bioestatística, como por exemplo o livro [1]

Capítulo 1

Estatística Descritiva

Neste capítulo vamo-nos centrar na Estatística mais simples, isto é, em como usar a estatística para caracterizar um conjunto de dados de forma adequada aos fins que procuramos. A esta área da Estatística chama-se **Estatística Descritiva**. Este capítulo servirá de base para o resto do texto, onde, como objetivo último, faremos **Estatística Inferencial**, uma área da Estatística onde o objetivo é tirar conclusões sobre uma população a partir do estudo de uma amostra.

Uma apresentação sobre o **método estatístico** pode ser visualizada aqui:

Para uma motivação sobre o uso da estatística descritiva em aplicações da Bioestatística, visite o link seguinte:

1.1 Dados e Variáveis

Comecemos pelo princípio. Num estudo estatístico temos variáveis de interesse, para as quais recolhemos dados de uma amostra, pelo que é importante distinguir entre o conceito de variável e o conceito de dados.

A **variável** é um conceito abstrato e não uma realização. É uma característica de interesse que pode tomar um conjunto de valores diferentes para cada indivíduo ou objeto da população. Geralmente supõe-se que a variável é **aleatória** cuja distribuição de probabilidade pode ser desconhecida, conhecida por estudos empíricos anteriores ou aproximada por uma distribuição teórica conhecida (voltaremos a esta assunto na secção 1.8, mais adiante).

A variável em estudo é selecionada a partir do **objetivo de estudo** previamente definido. Num estudo, é também comum considerar outras variáveis que possam influenciar a variável de estudo, variáveis essas que se designam por **covariáveis**.

Exemplo em Bioestatística 1.1. Como exemplo, supomos que temos como objetivo estudar o número de batimentos cardíacos médio, em repouso, da população adulta em Portugal. A variável aleatória de interesse neste caso seria o número

de batimentos cardíacos por minuto, em repouso. Como covariáveis, poderíamos considerar o peso, a idade ou o sexo do sujeito, uma vez que se espera que possam influenciar a variável de interesse. Outras covariáveis como a região de residência ou o emprego podem também ser interessantes para o estudo.

Definido o objetivo, a variável em estudo e as covariáveis, há que definir uma **amostra** adequada. A amostra deve ser representativa da população, no sentido em que todos os estratos e grupos da população devem estar representados na amostra de forma proporcional à da população. Um forma de recolher uma amostra adequada pode ser por **amostragem aleatória**, isto é, todos os elementos da população têm o mesma probabilidade de serem escolhidos, escolhendo-se depois aleatoriamente os elementos da amostra. Embora num mundo ideal esta opção seja muito boa, na prática são raras as vezes em que se pode optar por ela. Por exemplo, seria difícil escolher aleatoriamente de entre a população adulta portuguesa uma amostra, uma vez que seria preciso ter a lista de todos os elementos, de os escolher aleatoriamente e de esperar que todos os escolhidos estivessem dispostos a entrar no estudo. Assim, outras formas de amostragem têm de ser consideradas, como por exemplo, a **amostragem por estratos**, em que a amostra é escolhida de forma a que os vários estratos de uma população estejam na mesma proporção na amostra recolhida.

Por outro lado, os **dados** recolhidos são as realizações dessas variáveis na amostra escolhida. É sobre os dados recolhidos que se obtêm estatísticas e que se faz a estatística descritiva com vista à caracterização dos mesmos. Em resumo, os dados são uma série de factos sobre os quais se extraem conclusões.

1.2 Classificação de variáveis

Para determinar que tipo de estatísticas se podem calcular para cada variável, é importante determinar a classificação da variável em causa.

As variáveis podem ser classificadas sobre vários aspetos, mas dividem-se em duas grandes classes:

- **variáveis qualitativas** - os seus valores não são numéricos. São geralmente nomes ou designações, que representam uma "qualidade"do indivíduo ou do objeto.

- **variáveis quantitativas** - os seus valores têm representação e significado numérico;

As variáveis qualitativas podem ainda ser classificadas como

- **ordinais** - os valores têm uma ordem associada com significado para a variável;

- **nominais** - os valores não têm uma ordem associada;

Exemplo 1.2 (Variáveis qualitativas nominais). Como exemplos de variáveis quantitativas nominais temos o nome, o género (feminino, masculino), a localidade de residência ou a nacionalidade, entre outros.

Exemplo 1.3 (Variáveis qualitativas ordinais). Como exemplos de variáveis quantitativas ordinais temos a escolaridade (ensino primário, ensino básico, ensino secundário, ensino superior), a classe de índice de massa corporal (magro, saudável, excesso de peso, obeso) ou o alerta da proteção civil em determinada localidade (verde, azul, amarelo, laranja, vermelho).

As variáveis quantitativas por seu lado, podem ser classificadas quanto à continuidade, nomeadamente em

- **contínuas** - a variável pode assumir qualquer valor num dado intervalo;
- **discretas** - a variável apenas pode assumir um conjunto finito ou infinito numerável de valores, pelo que entre o mínimo e o máximo existem valores que a variável não pode assumir;

As variáveis quantitativas podem também ser classificadas quanto à escala, nomeadamente em

- **escala de razão** - O zero tem significado, o que implica que a razão entre dois valores tem significado;
- **escala de intervalos** - O zero não tem significado, o que implica que a razão entre dois valores não tem significado, apenas se podendo dar significado à amplitude entre dois valores;

Exemplo 1.4 (Variáveis quantitativas). Apresentamos a classificação das seguintes variáveis quantitativas:

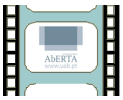
- Número de filhos - é um variável discreta, pois apenas assume valores não-negativos inteiros (não pode, por exemplo, assumir o valor 0.5). Está em escala de razão, porque o zero tem significado (significa a não existência de filhos).
- Tempo - é uma variável contínua, uma vez que pode assumir qualquer valor positivo. Está em escala de razão, porque o zero tem significado (significa o tempo inicial, ou seja, não ter passado qualquer instante).
- Idade - sendo uma medida de tempo (desde o nascimento), pode ser classificada como quantitativa contínua numa escala de intervalos. No entanto, uma vez que é uma variável que é representada usualmente com números inteiros (o número de anos desde o nascimento), pode ser considerada como uma variável discreta.
- Temperatura - é uma variável contínua (pode assumir qualquer valor) em escala de intervalos. Na realidade, a temperatura zero não indica a ausência de calor. A temperatura de 0° centígrados foi definida por convenção como a temperatura de congelação da água.

Estabelecer a correta classificação da variável é importante para definir o tratamento estatístico adequado. Por exemplo, como veremos adiante, não faz sentido calcular a média para variáveis qualitativas, pois estas não são representadas por números. Assim é importante dar esta informação ao SPSS, uma vez que se a classificação estiver correta, o SPSS apenas permite obter estatísticas que estejam de acordo com a variável considerada.

Como fazer em SPSS? 1.5 (Classificação de variáveis). Aberto o SPSS no ficheiro de dados, devemos escolher `Variable View` no canto inferior esquerdo. Na coluna `Measure`, temos as opções `Scale` para uma variável quantitativa, `Ordinal` para uma variável qualitativa ordinal e `Nominal` para uma variável qualitativa nominal. Mais ainda, tratando-se de uma variável qualitativa que esteja codificada em valores numéricos, devemos introduzir na coluna `Values` as legendas para cada um dos valores codificados.

Um vídeo sobre o ambiente SPSS e como definir a classificação da variável, pode ser consultado aqui:

Uma apresentação sobre classificação de variáveis pode ser visualizada aqui:



Muitas vezes é também necessário fazer **transformação de variáveis**. Esta necessidade por ocorrer para efeitos de mudança de escala dos valores, por exemplo para escala logarítmica, ou para construção de novas variáveis em função dos valores de outras já recolhidas.

Exemplo em Bioestatística 1.6. Tendo recolhidos os pesos e alturas dos sujeitos da amostra, podemos obter a variável índice de massa corporal (IMC) através da expressão

$$\text{IMC} = \frac{\text{peso}}{\text{altura}^2}.$$

Um vídeo sobre como fazer transformações de variáveis em SPSS, pode ser consultado aqui:

1.3 Tabelas de frequência

A **tabela de frequências** é uma tabela em que se sistematiza a descrição dos dados recolhidos, ou seja, o conjunto das observações.

As frequências consideradas são geralmente a

- **frequência absoluta simples** - representa o número de observações na amostra que são iguais a determinado valor da variável;
- **frequência absoluta acumulada** - representa o número de observações na amostra que são menores ou iguais a determinado valor da variável;
- **frequência absoluta simples** - representa a proporção (em relação ao número total de observações) de observações na amostra que são iguais a determinado valor da variável;

- **frequência absoluta acumulada** - representa a proporção (em relação ao número total de observações) de observações na amostra que são menores ou iguais a determinado valor da variável;

As tabelas de frequências são usadas para variáveis qualitativas ou quantitativas discretas, cujos valores podem ser descriminados. Para variáveis quantitativas contínuas, a tabelas de frequência terá de considerar classes de intervalos para a variável, por forma a contabilizar o número de valores recolhidos em cada intervalo.

De notar também que as frequências acumuladas não podem ser usadas para variáveis quantitativas nominais, uma vez que não existindo uma ordem, não existe o conceito de "menor".

Uma apresentação sobre tabelas de frequência pode ser visualizada aqui:

Como fazer em SPSS? 1.7 (Tabelas de Frequências). Aberto o SPSS no ficheiro de output, devemos selecionar

```
Analyze->Descriptive Statistics->Frequencies...
```

devendo depois na janela selecionar as variáveis de interesse e arrastá-las para o campo Variable(s). Nessa janela, o botão `Statistics...` permite adicionar mais estatísticas a calcular.

Uma visualização sobre como fazer tabelas de frequências e gráficos em SPSS, pode ser encontrada aqui:

As tabelas de frequências são úteis quando existem poucos valores (ou classes de valores) dos dados. De outra forma, a sua leitura fica confusa, devendo ser utilizado um gráfico para representar os valores recolhidos.

1.4 Representação gráfica

Como vimos na secção anterior, a representação de dados por tabelas de frequência, embora seja exata, pode ser pouco adequada. Nomeadamente, se tivermos dados contínuos ou dados discretos com um número elevado de valores, o uso de tabelas de frequência torna os dados pouco legíveis. Nesse caso, uma hipótese para a representação de dados pode ser a utilização de um gráfico adequado. Nas próximas linhas vamos enumerar alguns tipos de gráficos e, em geral, qual o tipo de dados a que estes se aplicam. Para a escolha do tipo de gráfico, é também importante conhecer o público alvo a quem vai ser apresentado o estudo estatístico, por forma a que este último possa entender os resultados obtidos. De notar que nos focaremos nos gráficos que consideramos mais importantes (o que é por si só uma avaliação subjetiva), sendo que esta apresentação não englobará a miríade de possibilidades para representação gráfica de dados.

O tipo de gráfico que é talvez mais utilizado é o **gráfico de barras**[1]. Neste, cada barra tem o comprimento proporcional à frequência de cada valor da variável. É geralmente utilizado para variáveis qualitativas ou quantitativas discretas, com poucos valores possíveis.

Quando queremos salientar a proporção entre os vários valores (ou classes) de uma variável, optamos geralmente por um **gráfico circular**. Neste, cada "fatia"do gráfico é proporcional, em área, à frequência de cada valor. Mais uma vez, a sua leitura só é aconselhável se:

- existirem poucos valores; ou se,
- existiram valores cuja frequência relativa seja relevante;

pois de outra forma (se existirem muitos valores e todos com frequências relativas muito pequenas), não é possível distinguir entre os vários valores no gráfico circular.

Exemplo em Bioestatística 1.8. Na figura 1.1 temos a representação dos valores recolhidos para a dor no peito, tanto em gráfico de barras (à esquerda) como em gráfico circular (à direita).

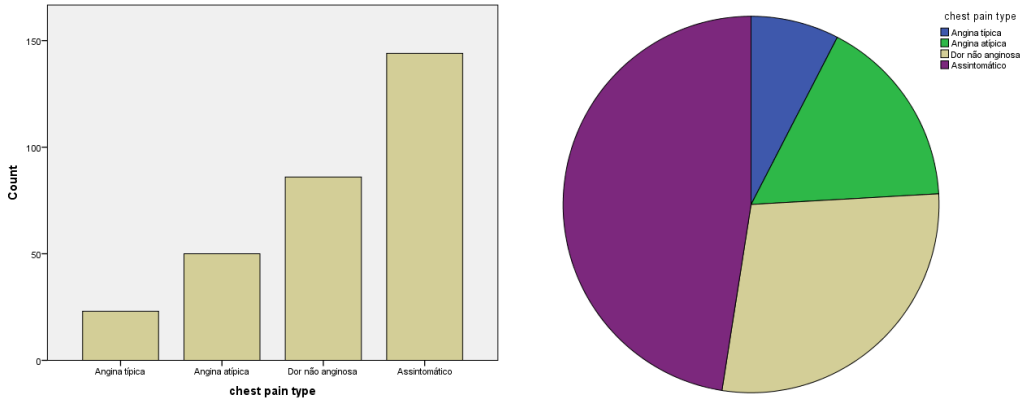


Figura 1.1: Gráfico de barras (à esquerda) e gráfico circular (à direita) dos valores recolhidos para a dor no peito.

Tal como referimos, o gráfico de barras, indica-nos a frequência de cada valor, mas o gráfico circular, indica claramente que o valor "Assimtomático"corresponde a praticamente 50% dos sujeitos da amostra.

Por outro lado, caso tenhamos dados contínuos ou discretos com muitos valores, o mais indicado é representá-los por um **histograma**. Aqui, os dados são primeiro agrupados em classes e é depois apresentada a frequência de cada classe. Em alternativa, pode ser apresentado o **polígono de frequências**, em que em vez de serem apresentadas as barras, são unidos os centros das barras por segmentos de reta.

Exemplo em Bioestatística 1.9. Neste exemplo utilizamos os valores da variável frequência cardíaca máxima no instante inicial, no ficheiro `BaseDados_Notas.sav`.

[1]No âmbito deste texto, chamaremos gráficos de barras quer estas estejam na horizontal ou na vertical, sendo que neste último alguma bibliografia considera a designação de gráficos de colunas.

Na figura 1.2 temos a representação dos valores recolhidos para a frequência cardíaca máxima em determinado período de tempo, tanto em histograma (à esquerda) como em polígono de frequências (à direita).

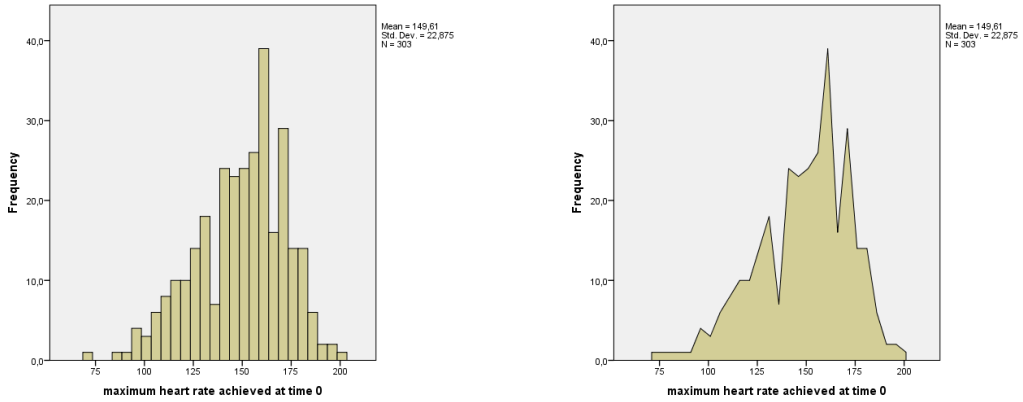


Figura 1.2: Histograma (à esquerda) e polígono de frequências (à direita) dos valores recolhidos para a frequência cardíaca máxima em determinado período de tempo.

O Histograma e o polígono de frequências perdem no entanto a informação de cada valor, devido ao agrupamento destes por classes. Pelo histograma podemos saber que, por exemplo, existem 6 valores no intervalo [90,100[, mas não sabemos quantos são iguais a $90, 91, \ldots, 98$ ou 99. Uma forma de ter uma visualização gráfica e não perder esta informação é utilizar um **diagrama de caule-e-folhas**. Neste caso, o que se faz é considerar uma listagem dos elementos na classe [90,100[utilizando para o efeito o último dígito de cada valor no intervalo. Assim, a linha

$$056679$$

corresponde aos valores reais

$$90, 95, 96, 96, 97, 99.$$

Exemplo em Bioestatística 1.10. Continuando no contexto e dados do exemplo 1.9 temos a representação dos valores recolhidos para a frequência cardíaca máxima em determinado período de tempo em diagrama de caule-e-folhas dado na figura 1.3. Por exemplo, a partir do diagrama de caule-e-folhas é possível ver que existem 4 valores no intervalo [190,200[com valores 190, 192, 194 e 195.

Podemos verificar que em termos gráficos temos a mesma informação no diagrama de caule-e-folhas que o polígono de frequências (a menos de uma troca de eixos), mas temos ainda a informação em cada classe de qual o valor exato, através da listagem dos últimos dígitos. Desta forma, o diagrama de caule-e-folhas permite ter mais informação que um histograma ou polígono de frequências, se bem que o seu uso não esteja tão difundido na sociedade. Desta forma, consoante o público alvo do estudo estatístico, deve ser ponderado se esta é uma boa opção.

Para finalizar esta secção sobre representação gráfica (e voltando a salientar que existem muitas outras possibilidades que não serão aqui abordadas), vamos considerar o gráfico de linhas. Este é especialmente útil para variáveis contínuas que variam ao longo do tempo, nomeadamente no estudo de séries temporais.

```
maximum heart rate achieved at time 0 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1,00 Extremes    (=<71)
     1,00        8 .  8
     6,00        9 .  056679
    10,00       10 .  3355568899
    16,00       11 .  1112234445556678
    25,00       12 .  000122223345555556666789
    26,00       13 .  000011112222222334667888 99
    47,00       14 .  00000011122222233333334444444555566667777788899
    52,00       15 .  000000011112222222233344444555566666677777888888999 9
    57,00       16 .  0000000001111112222222222333333333445555566678888899999 9
    42,00       17 .  00000111122222223333333444445557888889999 9
    15,00       18 .  001122222456678
     4,00       19 .  0245
     1,00       20 .  2

 Stem width:        10
 Each leaf:       1 case(s)
```

Figura 1.3: Diagrama de Caule-e-Folhas dos valores recolhidos para a frequência cardíaca máxima em determinado período de tempo.

Exemplo em Bioestatística 1.11. Para estudar a evolução do peso de um sujeito que participa num programa de emagrecimento, fez-se o gráfico de linhas da figura 1.4 com a evolução do seu peso ao longo das semanas do programa.

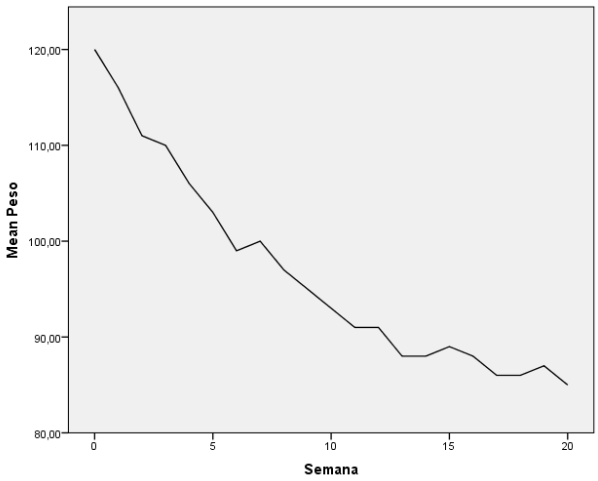
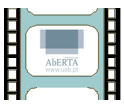


Figura 1.4: Gráfico de linhas da evolução do peso de um sujeito, ao longo das semanas em que esteve inserido num programa de emagrecimento.

Informação complementar sobre representação gráfica pode ser visualizada aqui:



Como fazer em SPSS? 1.12 (Gráficos). Aberto o SPSS no ficheiro de output, devemos selecionar

`Graphs->Chart Builder...`

devendo depois selecionar o tipo de gráfico a usar e quais as variáveis a considerar em cada eixo. Se se tratar de um gráfico de frequências, bastará escolher a variável a considerar no eixo horizontal. Alternativamente, pode escolher

`Graphs->Legacy Dialogs`

e escolher de entre a lista de tipos de gráficos que lhe aparece.

Recordamos que um vídeo sobre como fazer tabelas de frequências e gráficos em SPSS, pode ser encontrada aqui:

1.5 Medidas de tendência central

Outra forma de caracterizar os dados é a utilização de **medidas de tendência central**. Uma medida de tendência central é um valor em torno do qual se encontram distribuídos os valores da variável. Existem várias medidas de tendência central, mas vamo-nos forcar em três: a moda, a mediana e a média (aritmética).

A **moda** é o valor com maior frequência. Pode ser utilizada para qualquer tipo de variável, mas é especialmente útil para caraterizar variáveis nominais, ordinais ou contínuas discretas, especialmente se existirem poucos valores possíveis. A moda determina o valor com maior incidência, pelo que é um indicador importante para a distribuição da variável.

Outra medida de tendência central que é geralmente utilizada é a **mediana**. A mediana é o valor tal que 50% dos valores são inferiores e 50% são superiores. Desta forma, pela definição de mediana, esta não pode ser aplicada a variáveis nominais, uma vez que não existe ordem. Se tivermos uma amostra de dimensão n com valores amostrais $x_1, x_2, \ldots, x_n$, e considerarmos os respetivos valores ordenados $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, do menor para o maior, a mediana é dada por

$$\text{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases},$$

isto é, é o valor central se n é ímpar ou a média dos dois valores centrais se n é par.

A **média** (aritmética) é definida por

$$\bar{x} = \frac{x_1 + x_2 + x_n}{n}. \tag{1.1}$$

pelo que, como implica uma soma, só se pode aplicar a dados quantitativos. É o valor tal que a soma dos desvio aos valores amostrais é nula,isto é,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

e que minimiza a soma dos quadrados dos desvios aos valores amostrais, isto é, $\bar{x}$ minimiza o valor

$$\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

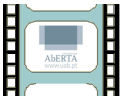
Todas estas medidas caracterizam o centro amostral dos valores, ainda que de formas diferentes. Por exemplo, a mediana não é afetada por valores extremos,

isto é por valores que estejam muito distantes do valor central. Por outro lado, no cálculo da média os valores extremos podem ter uma grande influência. Assim, a comparação da média com a mediana pode inclusivé servir para estudar a assimetria da distribuição dos valores. Em geral, a mediana ser superior à média é um indicador de que existem valores extremos à esquerda e logo que a cauda esquerda da distribuição é mais longa. Por outro lado, valores de média, moda e mediana muito semelhantes é um indicador de uma distribuição praticamente simétrica.

No entanto, para uma análise mais rigorosa, é importante introduzir outras medidas de localização, como veremos na secção seguinte.

Informação complementar sobre medidas de tendência central pode ser visualizada aqui:

Se por outro lado considerar dados agrupados, deve proceder conforme indicado aqui:



Como fazer em SPSS? 1.13 (Medidas de tendência central e outras). Aberto o SPSS no ficheiro de output, devemos selecionar

`Analyze->Descriptive Statistics->Descriptives...`

devendo depois na janela selecionar as variáveis de interesse e arrastá-las para o campo Variable(s). O botão `Options...` permite escolher quais as estatísticas a calcular, incluindo medidas de localização e dispersão que estudaremos nas próximas secções.

Alternativamente, pode usar

`Analyze->Descriptive Statistics->Explore...`

que faz uma análise exploratória da variável, indicando o resultado de várias medidas. Nesse caso, deve colocar as variáveis em estudo no campo `Dependent List` e caso queira fazer uma análise por grupos, colocar a ou as variáveis de grupo no campo `Factor`.

Um vídeo sobre como obter medidas descritivas em SPSS, que engloba os conceitos abordados até ao final da secção 1.7, pode ser encontrada aqui:

SPSS

1.6 Outras medidas de localização

A medidas de tendência central apenas nos dão indicação do centro da distribuição dos valores e uma ideia da simetria da distribuição. No entanto não dão indicação da localização dos restantes valores. Para aferir com maior precisão essa localização, podemos considerar as seguintes **medidas de localização**:

- **Quartis**: Os quartis dividem os valores em quatro partes. O primeiro quartil, denotado por Q1, é o valor tal que 25% dos valores são inferiores e 75% dos valores são superiores. O segundo quartil, coincide com a mediana. O terceiro quartil, denotado por Q3, é o valor tal que 75% dos valores são inferiores e 25% dos valores são superiores. Uma forma de calcular o primeiro quartil é considerar a mediana da primeira sub-amostra, isto é, os valores compreendidos entre o mínimo e a mediana, enquanto que o terceiro quartil pode ser calculado como a mediana da segunda sub-amostra, isto é, os valores compreendidos entre a mediana e o máximo.

- **Decis**: Os decis dividem os valores em dez partes. O primeiro decil, denotado por D1, é o valor tal que 10% dos valores são inferiores e 90% dos valores são superiores; o segundo decil, denotado por D2, é o valor tal que 20% dos valores são inferiores e 80% dos valores são superiores; e assim sucessivamente. É claro que o quinto decil coincide com a mediana.

- **Percentis**: Os percentis dividem os valores em cem partes. O percentil 1, denotado por P1, é o valor tal que 1% dos valores são inferiores e 99% dos valores são superiores; Assim, o percentil 10 (P10) coincide com o 1º decil (D1), O percentil 25 (P25) coincide com o 1º Quartil (Q1) e o percentil 50 (P50) coincide com a mediana, entre outras igualdades possíveis.

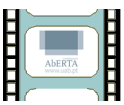
De notar que as medidas anteriores só podem ser aplicadas se a variável tiver uma ordem associada, pelo que terá de ser qualitativa ordinal ou quantitativa. Conhecendo os quartis, o máximo e o mínimo dos valores de uma variável, podemos construir o chamado **diagrama de extremos e quartis**, que permite obter informação visual da distribuição dos valores da variável entre o seu mínimo e máximo e em torno da mediana.

Exemplo em Bioestatística 1.14. Continuando no contexto e dados do exemplo 1.9 temos os valores extremos e quartis dos valores recolhidos para a frequência cardíaca máxima inicial dados por

$$\text{Min} = 71, \quad Q1 = 133, \quad \text{Med} = 153, \quad Q3 = 166, \quad \text{Max} = 202.$$

Assim, obtemos o diagrama de extremos e quartis na figura 1.5. De notar que neste exemplo, o SPSS considera o valor mínimo como valor outlier[2], indicando que este se encontra listado na linha 246. Assim, considera a barra horizontal referente ao mínimo no valor mais pequeno seguinte, nomeadamente 88. Pelo diagrama, é possível ver que a cauda esquerda é ligeiramente maior, uma vez que a distância entre a mediana e o Q1 e o valor mínimo é ligeiramente superior à distância entre a mediana e o Q3 e máximo, respetivamente.

Está disponível um recursos multimédia sobre medidas de localização aqui:



Recordamos que pode ser encontrado no link seguinte um vídeo sobre como obter medidas descritivas em SPSS, em particular, diagramas de extremos e quartis:

[2]Veremos na secção seguinte, como se classificam os outliers.

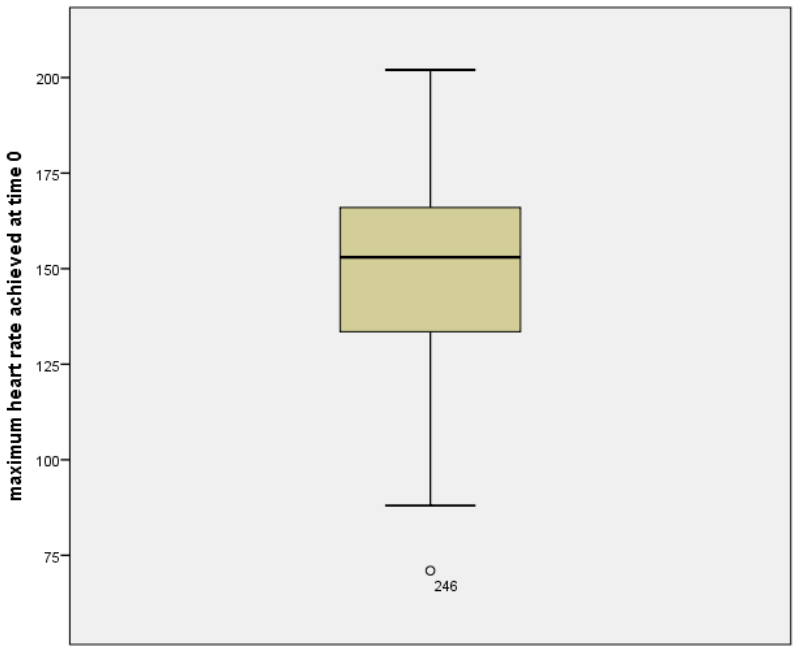


Figura 1.5: Diagrama de extremos e quartis dos valores recolhidos para a frequência cardíaca máxima em determinado período de tempo.

SPSS

1.7 Medidas de dispersão

Além das medidas de tendência central e de outras medidas de localização, importa também estudar a dispersão da dos valores da amostra em torno do valor central. Desta forma, introduzindo as **medidas de dispersão** podemos estudar a variabilidade da variável. Uma maior dispersão, indica uma maior variabilidade, enquanto que uma menor dispersão indica uma menor variabilidade, sugerindo que os valores se encontram concentrados em torno do valor central.

Para variáveis quantitativas, uma medida de dispersão muito utilizada é a **variância amostral**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{1.2}$$

onde x_i são os valores amostrais e $\bar{x}$ é a média definida em (1.1). Como é fácil de notar, quanto maior a dispersão, maior o valor da variância. No sentido inverso, no caso limite em que os valores x_i são todos iguais (e logo iguais à média), a variância é nula.

De notar também que como o termo $(x_i - \bar{x})$ aparece ao quadrado, a variância não é apresentada nas mesmas unidades que os valores da amostra. Desta forma, para que se possam comparar com os valores recolhidos, é comum utilizar o **desvio padrão amostral**

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1.3}$$

O desvio padrão é assim uma medida de dispersão que apresenta valores nas mesmas unidades que os valores da amostra. Assim, o desvio padrão é uma **medida de dispersão absoluta**, isto é, o seu valor depende da ordem de grandeza dos valores recolhidos, pelo que para saber se a dispersão é grande ou pequena teremos de comparar o valor do desvio padrão com a ordem de grandeza dos dados.

Exemplo 1.15 (Aplicação de medidas absolutas e relativas - parte I). Para ilustrar o conceito de medida de dispersão absoluta, consideremos um mesmo desvio padrão de uma unidade em dois estudos diferentes. Por exemplo, um desvio padrão de 1 metro na recolha das alturas de sujeitos, indica uma variabilidade enorme. Na realidade, se assumir que um adulto terá uma altura em torno de 1,70m, um desvio padrão de 1 metro é muitíssimo grande. Por outro lado, um desvio padrão de 1 Kg na recolha de do peso de sujeitos indica uma dispersão pequena. Se pensarmos que um adulto tem peso a rondar os $70Kg$, a variabilidade de 1 Kg pode ser considerada pequena.

Desta forma, é importante ter também **medidas de dispersão relativa**. É com este fim que introduzimos o **coeficiente de variação**

$$CV = \frac{s}{\bar{x}} \tag{1.4}$$

que nos indica a percentagem de dispersão, sendo uma medida de dispersão relativa (em relação à média).

Exemplo 1.16 (Aplicação de medidas absolutas e relativas - parte II). Voltando ao exemplo 1.15, teríamos para a altura e peso, os coeficientes de variação

$$CV_a = \frac{1}{1.70} = 0.588, \quad CV_p = \frac{1}{70} = 0.0143$$

indicando que em termos relativos, a dispersão da altura (58,8%) é muito maior que a dispersão do peso (1.43%).

As medidas de dispersão que vimos até aqui só podem ser aplicadas para variáveis quantitativas, uma vez que implicam operações aritméticas entre os valores amostrais. Outras medidas de dispersão que podem ser usadas para variáveis quantitativas como para variáveis qualitativas ordinais são a **amplitude**

$$\text{Amp} = \text{Max} - \text{Min} \tag{1.5}$$

definida pela diferença entre os valores máximo e mínimo recolhidos e a **amplitude inter-quartil**

$$\text{AIQ} = Q_3 - Q_1 \tag{1.6}$$

definida pela diferença entre o terceiro e primeiro quartil. De notar que as expressões (1.5) e (1.6) anteriores não podem ser calculadas explicitamente se as variáveis ordinais não assumirem valores numéricos, no entanto, a Amplitude e Amplitude Inter-quartil podem ser utilizadas neste contexto, uma vez que podem balizar os valores da variável ordenada.

Exemplo 1.17. Consideremos a variável ordinal "Classe de Peso", que pode assumir os valores

- Magro;

- Normal;
- Excesso de Peso;
- Obeso;

e suponhamos que para determinada amostra recolhemos as medidas de localização seguintes

$$\text{Min} = \text{Magro};\ Q_1 = \text{Normal};\ Q_3 = \text{Excesso de Peso};\ \text{Max} = \text{Obeso}.$$

Podemos então afirmar que a amplitude interquartil varia de Excesso de peso a Normal, abarcando uma classe de valores, enquanto que a Amplitude dos valores é toda a gama de valores possíveis.

Recordamos que pode ser encontrado no link seguinte um vídeo sobre como obter medidas descritivas em SPSS, em particular, sobre medidas de dispersão:

Exemplo em Bioestatística 1.18. Continuando no contexto e dados do exemplo 1.9, temos as medidas de dispersão

$$s^2 = 523.266, \quad s = 22.875, \quad \text{CV} = 0.153, \quad \text{AIQ} = 33, \quad \text{Amp} = 131.$$

1.7.1 Valores Extremos e Outliers

Algumas das medidas de dispersão que vimos até agora são muito influenciadas pelos **valores extremos** (máximo e mínimo) ou **outliers**, isto é, valores que se afastam muito da tendência central. Geralmente, considera-se um valor como **outlier**, quando este dista da média mais de três desvios-padrão. No entanto, o fator multiplicativo pode ser diferente de três, consoante a literatura e a dimensão da amostra. Os outliers merecem a nossa especial atenção por duas razões principais:

(a) O valor pode ser um erro na base de dados, pelo que deve ser revisto. Caso se trate de um erro comprovado[3], o valor deve ser eliminado.

(b) Caso não haja evidência de erro no dados e caso o número de outliers seja relevante, o uso da média, variância, desvio padrão e amplitude deve ser descartado, uma vez que todas estas medidas são influenciadas pelos outliers, desviando a média nesse sentido e aumentando a dispersão global dada pela variância e desvio-padrão e, claro, a amplitude. Nestes casos, as medidas de tendência central e dispersão a usar devem ser, respetivamente, a mediana e a amplitude inter-quartil, uma vez que estas não são influenciadas pelos outliers.

[3]Muitas vezes, é difícil comprovar num estudo estatístico que houve um erro na introdução dos dados, até porque podem ser técnicos diferentes os que fazem a recolha dos dados e os que os tratam estatisticamente. No entanto, é possível em alguns casos perceber que se trata de um erro de introdução, por análise do contexto da recolha e das variáveis. Por exemplo, uma idade superior a 120 anos deverá estar errada, assim como uma idade superior a 18 anos deverá estar errada num estudo sobre crianças.

1.8 Distribuições de probabilidade teóricas mais conhecidas

Antes de iniciarmos o estudo de testes de hipóteses no capítulo seguinte, convém rever o conceito de **variável aleatória**, que pode rever aqui

e algumas **leis de distribuições teóricas** para variáveis contínuas. Estas leis permitem saber a probabilidade de uma variável contínua assumir valores entre a e b dados. Esta probabilidade é dada pelo integral (ou seja, pela área sob o gráfico) da função densidade de probabilidade associada à distribuição. Consoante o caso, é preciso depois escolher a lei de distribuição que melhor aproxima a distribuição da variável.

Comecemos então por enumerar algumas leis de distribuição mais importantes.

1.8.1 Distribuição Uniforme

A **distribuição uniforme** é uma distribuição em que todos os valores têm igual probabilidade de ocorrer, dentro de um intervalo $[a, b]$. Desta forma, diz que a variável aleatória X tem distribuição uniforme, e denota-se por

$$X \sim U(a, b)$$

se a função de probabilidade de X for constante e igual a

$$f(x) = \frac{1}{b-a}.$$

Esta distribuição tem pouco interesse para as aplicações biomédicas, mas é incluída nesta apresentação dada a sua simplicidade. Mais aspetos sobre a distribuição uniforme podem ser visualizados aqui:

1.8.2 Distribuição Normal

A **distribuição Normal** é uma distribuição que serve como modelo para variáveis aleatórias representativas de populações. Denota-se por

$$X \sim N(\mu, \sigma)$$

uma variável aleatória X com distribuição normal de média populacional μ e desvio padrão populacional σ. A distribuição normal é simétrica em relação à média a assume a forma de uma curva gaussiana. Podemos normalizar a variável X com distribuição norma, subtraindo a média e dividindo pelo desvio

padrão, obtendo assim uma variável aleatória $Z = \frac{X-\mu}{\sigma}$ com distribuição normal standard[4]

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

A distribuição normal é simétrica e assume uma função de densidade de probabilidade cujo gráfico é uma curva gaussiana centrada na média populacional.

A distribuição normal está muito bem estudada e em particular sabe-se que

$$P(\mu - \sigma \leq X \geq \mu + \sigma) \approx 0.68,$$

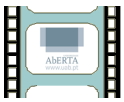
isto é, a probabilidade de um valor da variável aleatória estar a distância da média inferior a um desvio padrão é cerca de 68%. Da mesma forma, a probabilidade de um valor da variável assumir um valor que dista menos de 2 ou 3 desvios-padrão da média é, respetivamente de, 95% e 99%, ou seja,

$$P(\mu - 2\sigma \leq X \geq \mu + 2\sigma) \approx 0.95, \quad P(\mu - 3\sigma \leq X \geq \mu + 3\sigma) \approx 0.99.$$

Consegue-se também mostrar que, sob certas condições, a média de uma amostra de dimensão n proveniente de uma população com distribuição normal, tem distribuição normal de parâmetro

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Mais informações sobre a distribuição normal podem ser visualizados aqui:



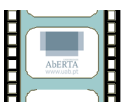
1.8.3 Distribuição Qui-quadrado

A **distribuição qui-quadrado** surge da soma de quadrados de variáveis normais standard. Assim, a variável $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ em que $Z_1, Z_2, \ldots, Z_n$ são variáveis aleatórias e independentes com distribuição normal padrão. Denota-se então por

$$X \sim \chi_n^2$$

uma variável aleatória com distribuição qui-quadrado com n graus de liberdade. Note-se que, por construção, uma variável com distribuição qui-quadrado apenas pode assumir valores não-negativos. Além disso, a distribuição não é simétrica.

Mais detalhes sobre a distribuição qui-quadrado podem ser visualizados aqui:



[4]Chama-se normal standard a uma distribuição normal de média nula e desvio padrão unitário

1.8.4 Distribuição t-Student

A **distribuição t-Student** é uma distribuição que serve como modelo testar a média populacional de variáveis aleatórias representativas de populações com distribuições normais, em que o desvio padrão populacional σ é desconhecido. Denota-se por

$$X \sim t_n$$

uma variável aleatória X com distribuição t-student com n graus de liberdade. À medida que n aumenta, a distribuição t-student tende para a distribuição normal standard, pelo que é comum para n grande aproximar a distribuição de uma variável t-student pela distribuição normal standard.

Mais detalhes sobre a distribuição t-student podem ser visualizados aqui:

1.8.5 Distribuição F-Snedcor

A **distribuição F-Snedcor** tem origem na razão entre duas variáveis com distribuição qui-quadrado. Na realidade, Se X_1 e X_2 são variáveis aleatórias independentes com distribuição qui-quadrado com n_1 e n_2 graus de liberdade, respetivamente, então a variável $F = \frac{X_1/n_1}{X_2/n_2}$ tem distribuição F-Snedcor, denotando-se

$$F \sim F_{n_1,n_2}.$$

De notar que por construção, uma variável com distribuição F-Snedcor só pode assumir valores não-negativos. Como veremos mais adiante, a distribuição F-Snedcor está relacionada com o teste de igualdade de variâncias entre duas ou mais populações.

Mais detalhes sobre a distribuição F-Snedcor podem ser visualizados aqui:

1.8.6 Distribuição Exponencial

A **distribuição exponencial** está relacionada com tempos de espera ou com tempos de vida de componentes elétricos. Esta variável é importante como modelo para alguns modelos de Sobrevivência.

Diz-se que X tem distribuição exponencial com parâmetro λ, denotando-se por

$$X \sim Exp(\lambda)$$

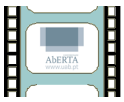
se a sua função densidade de probabilidade desta variável for

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

Uma variável com distribuição exponencial só assume valores não negativos e consegue-se mostrar que uma variável com distribuição exponencial tem média

(no sentido de valor esperado) $\frac{1}{\lambda}$ e variância $\frac{1}{\lambda^2}$. Quer isto dizer que quanto maior o λ, menor o tempo de espera médio (ou tempo de sobrevivência) associado à variável X e menor a variância.

Mais detalhes sobre a distribuição exponencial podem ser visualizados aqui:



1.9 Descrição de dados

Dedicamos este capítulo à forma correta de descrever os dados, tomando como base a distribuição de probabilidade que está por trás da amostra em causa. Se considerarmos a distribuição normal, podemos caracterizá-la pela média e desvio padrão. É comum representar uma amostra (proveniente de uma população com distribuição normal) com média 5.3 e desvio padrão 2.7 na forma 5.5 ± 2.7 (média $\pm$ desvio-padrão). Se a distribuição for, de facto normal, estas duas grandezas definem a distribuição da população.

No entanto, é infelizmente também comum (embora errado!) que se usem estas duas grandezas para caracterizar populações com distribuições não normais, caso em que estes dados podem não fazer sentido.

Mais adiante abordaremos testes estatísticos para testar a normalidade da distribuição da população, usando o **teste de Kolmogorov-Smirnov** ou o **teste de Shapiro-Wilk**. Para já, indicamos que uma forma intuitiva de verificar de determinada variável tem distribuição normal é fazer o seu histograma (ou polígono de frequências) e verificar se este se aproxima de uma curva gaussiana.

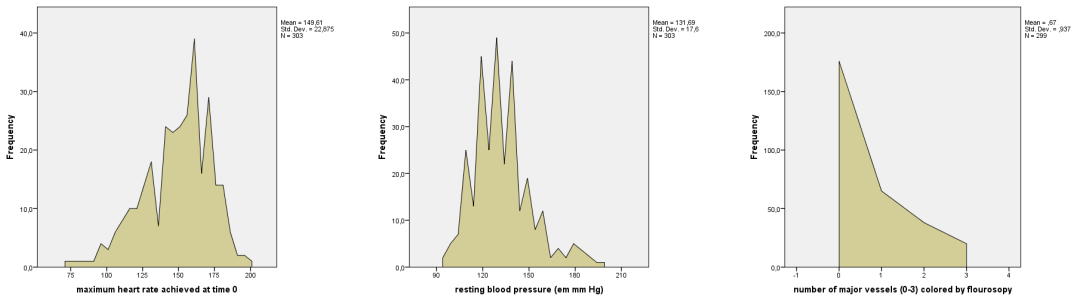


Figura 1.6: Exemplos de histogramas para averiguação intuitiva da distribuição normal da variável.

Será fácil ver na figura 1.6 que a variável número de grandes vasos (à direita) não tem distribuição normal[5]. Por outro as variáveis batimento cardíaco máximo (à esquerda) e pressão arterial em repouso (ao meio), apesar de terem caudas mais longas à esquerda e direita, respetivamente, têm distribuições aproximadas à normal.

[5]Neste caso particular, nem era preciso usar o histograma para concluir a não normalidade, uma vez que se trata de uma variável discreta que só assume 4 valores, de 0 a 3. Desta forma, nunca poderia ter distribuição normal, uma vez que uma variável com distribuição normal terá de ser contínua (ou ser discreta mas assumindo muitos valores distintos, podendo ser assumida como contínua).

Capítulo 2

Testes de Hipóteses

Neste capítulo vamos abordar o conceito de **teste de hipóteses** e a sua implementação. O teste de hipóteses serve para averiguar se determinada hipótese para a população é plausível, sob o pressuposto aleatoriedade e independência da amostra. Este tem obviamente aplicações relevantes na Bioestatística, quando queremos evidenciar diferenças entre grupos, como por exemplo controlo *vs.* patologia ou controlo *vs.* tratamento). Para uma motivação sobre o uso de testes de hipóteses em aplicações da Bioestatística, visite o link seguinte:

O teste de hipóteses é constituído por duas hipóteses: a **hipótese nula** H_0 e a **hipótese alternativa** H_1. As duas hipóteses H_0 e H_1 devem ser complementares[1]. A hipótese alternativa H_1 é também chamada de **hipótese de interesse**, uma vez que esta é a hipótese que queremos sustentar. Assim, se quisermos testar se a média populacional μ (a letra grega lê-se miú) é superior a 10 (dada uma amostra), o teste de hipóteses a considerar é

$$\begin{cases} H_0: & \mu \leq 10, \\ H_1: & \mu > 10. \end{cases}$$

Da mesma forma, se quisermos testar de as médias populacionais μ_1 e μ_2 são distintas nos grupos 1 e 2 da população, respetivamente, o teste a considerar é

$$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2. \end{cases}$$

Veremos de seguida como testar de uma forma probabilística se a hipótese de interesse é verdadeira.

2.1 Erros do Tipo I e do Tipo II

Num contexto estatístico, nunca poderemos mostrar se um determinado resultado é verdadeiro ou falso. Antes, testamos se um determinado resultado é plausível (tem probabilidade elevada de ocorrência), assumindo que a nossa hipótese

[1] Duas hipóteses dizem-se complementares se não se intersectam e cobrem todo o domínio do parâmetro em questão.

é verdadeira. Dito de outra forma, testamos se é plausível que a amostra obtida seja aleatória, dada a hipótese nula H_0. Ao valor dessa probabilidade chama-se **p-value** ou **valor-p** (valor de prova). Assim, num teste estatístico, teremos sempre associados dois tipos de erro: o **erro de Tipo I**, que ocorre quando rejeitamos a hipótese nula H_0 sendo ela verdadeira; e o **erro de Tipo II**, que ocorre quando não rejeitamos a hipótese nula H_0 sendo ela falsa.

Nota 2.1. Na prática, só existe uma forma de evitar erros do tipo I: nunca rejeitar a hipótese nula H_0. De forma semelhante, só existe uma forma de evitar erros do tipo II: rejeitar sempre a hipótese nula H_0. Como é obvio, qualquer das abordagens anteriores é inútil, pelo que num contexto de teste estatístico temos de ter associado uma margem de erro que estamos dispostos a correr.

Geralmente, o controlo do erro é feito através do erro do tipo I, chamando **significância** α (a letra grega lê-se alfa) à probabilidade de erro de tipo I que estamos dispostos a cometer. A significância de um teste é então a probabilidade de cometer um erro do tipo I, ou seja, de rejeitar a hipótese nula H_0, sendo esta hipótese verdadeira na população. Por outro lado, chama-se **potência** do teste à probabilidade de não cometer um erro do tipo II. Por norma, quando a significância do teste aumenta, a sua potência diminui, e vice-versa.

Geralmente, o controlo do erro é feito pela significância, uma vez que como veremos adiante, esta é mais fácil de controlar. Assim, por norma, rejeita-se a hipótese nula se o **p-value** (que, recordamos, é a probabilidade da amostra recolhida ter sido obtida de forma aleatória, caso a hipótese nula seja verdadeira) for inferior à significância escolhida. Daí que caso o resultado do teste seja rejeitar a hipótese nula H_0, o resultado se diga **significativo**.

Normalmente fixamos a significância em 5%, o que significa vamos rejeitar a hipótese nula, sendo esta verdadeira (e logo, erradamente), no máximo em 5% dos casos. Voltando um pouco atrás, o facto de se controlar a significância fixando valores relativamente baixos (de 1%, 5% ou , no máximo, 10%) faz com que a hipótese alternativa seja a hipótese de interesse: queremos com alguma certeza garantir que a hipótese alternativa H_1 seja verdadeira quando rejeitado o teste, "sobrando"apenas 5% para erros (de rejeitar hipótese nula e esta ser verdadeira).

Nota 2.2 (p-value). O conceito de significância justifica também o cuidado necessário a aplicar testes estatísticos a elevado número de variáveis de um conjunto de dados, uma vez que (para uma significância de 5%) em 5% dos casos vamos obter resultados significativos devido a erros do tipo I. Suponhamos que foram recolhidas 20 variáveis para dois grupos de sujeitos: um grupo com patologia e um grupo de controlo. Se testarmos diferenças na distribuição de cada uma das variáveis é possível que em 1 das 20 (correspondente a 5%) os resultados sejam significativos por um erro do tipo I. Desta forma, em vez de correr testes em massa para um grande conjunto de dados, convém limitar através de critérios de especialista na área quais as variáveis que devem ser testadas. Assim, o procedimento correto é pedir a um especialista que selecione quais as variáveis que considera mais relevantes para distinguir entre os grupos e depois testar essas variáveis. A licação em massa de testes estatísticos provoca erros do tipo I e logo conclusões erradas.

Nota 2.3 (Resultados espúrios). Um outro conceito importante é o de **resultado**

espúrio, em que apesar de existir evidência estatística de diferenças nos dados, estes não são explicados pela associação entre variáveis. Por exemplo, poder-se-ia concluir que países com maior número de televisões *per capita* são os países com maior taxa de doenças cardíacas. Uma conclusão (errada) seria que ver televisão implica maior taxa de doença cardíaca, gerada pelo facto de os resultados serem espúrios. De facto, o que acontece é que os países com maior número de televisões *per capita* são os países mais ricos e logo, devido a uma sobre-alimentação, têm maiores problemas cardíacos. Assim, um terceiro factor escondido relaciona-se com as duas variáveis e justifica a variação das mesmas, ainda que o número de televisões no país não tenha influência direta na taxa de doenças cardiacas.

2.2 Amostras Emparelhadas ou Independentes?

Antes de avançarmos para os testes adequados em cada contexto de investigação, convém referir brevemente as técnicas de amostragem e se as amostras são emparelhadas ou independentes

Existem várias formas de amostragem. O processo de amostragem, que não será focado em detalhe neste texto, pode influenciar fortemente os resultados obtidos, tornando-os enviesados. Um exemplo claro é a **amostragem por conveniência**, em que a amostra é obtida por conveniência do inquiridor, seja por mais fácil acesso ou por mais baixo custo, o que por isso pode enviesar totalmente os resultados por não ser representativa da população em estudo.

Exemplo 2.4 (Amostra por conveniência). Imaginemos que se pretendia fazer um estudo sobre a idade média da população portuguesa. Por conveniência minha, perguntava a idade a todos os meus amigos e fazia a média desses valores, obtendo dessa forma uma estimativa para média da idade da população portuguesa. A minha estimativa seria claramente enviesada, partindo do princípio que os meus amigos têm sensivelmente a minha idade.

Assim, é importante a amostra ser representativa da população, ou seja, ser um espelho da população em termos das características/variáveis em estudo e do seu peso nesta. A forma desejável de o fazer é por **amostragem aleatória simples**, isto é, os elementos da população têm todos a mesma probabilidade de ser escolhidos para a amostra. Isto implica uma dimensão de amostra alargada para garantir a representividade, assim como custos financieros e de tempo mais elevados. Assim, uma alternativa é fazer **amostragem estratificada**. Nesse caso, definem-se vários estratos na população com relevância para o estudo, por exemplo, sexo masculino e feminino, intervalos de idade ou distrito de residência, entre outros estratos possíveis. Este critério é frequentemente utilizado quando se pretende recolher uma amostra em que a proporção de cada estrato seja semelhante à proporção de cada estrato na população. Dessa forma, é também garantida a representatividade da amostra.

Estando garantida a representatividade da amostra, há que escolher qual o teste estatístico a utilizar para responder à questão de investigação. Para isso e no que concerne a comparação entre grupos[2], é importante perceber se estamos num cenário de **amostras independentes** ou num cenário de **amostras emparelhadas**.

[2]Este é geralmente o caso em aplicações bioestatística, isto é, saber se existe diferenças signi-

As amostras dizem-se **independentes**, quando os dados (unidades amostrais) de um grupo são independentes dos restantes. Em aplicações da Bioestatística, este contexto encontra-se, por exemplo, quando se recolhem dados num determinado momento sobre sujeitos de diferentes grupos distintos, ou seja, se um sujeito faz parte de um grupo, não faz parte de nenhum dos outros. Um exemplo é o caso em que se quer comparar o valor de um variável entre um grupo de controlo e um grupo com patologia.

Exemplo em Bioestatística 2.5. Pretende-se testar se existem diferenças significativas entre os valores de glicémia em jejum entre diabéticos e controlos (sem diabetes). Para o efeito, foram recolhidos valores de glicémia em jejum para uma amostra de 15 sujeitos com diabetes e 17 sujeitos sem diabetes (grupo de controlo), tendo-se aplicado um teste apropriado para o teste de hipóteses de igualdade de médias de valores de glicémia para amostras independentes, formulado como:

$$\begin{cases} H_0: & \mu_{\text{diabetes}} = \mu_{\text{controlo}}, \\ H_1: & \mu_{\text{diabetes}} \neq \mu_{\text{controlo}}. \end{cases}$$

Por outro lado, diz-se que as amostras são **emparelhadas** se os dados são dependentes entre grupos. Em geral, este cenário aparece em estudos longitudinais, observações repetidas ao longo do tempo, para o mesmo grupo de sujeitos, em que interessa avaliar alterações ao longo do tempo, ou comparar situações do tipo "Antes e Depois". Neste caso, tem de ser levado em linha de conta a resposta de cada sujeito, pelo que este efeito não deve ser descurado, uma vez que é recolhido o valor observado de determinadas variáveis de interesse por sujeito em vários instantes de tempo. O objetivo deste tipo de teste é frequentemente testar a eficácia (ou o efeito) de um determinado tratamento ao longo do tempo.

Exemplo em Bioestatística 2.6. Pretende-se testar um novo tratamento para a diabetes. Para o efeito, foram recolhidos valores de glicémia em jejum pra um grupo de 33 pessoas com diabetes no instante basal (antes de iniciar o tratamento) e um mês após o início do tratamento. Assim aplicou-se um teste apropriado para avaliar a igualdade de médias dos valores de glicémia para amostras emparelhadas (mesmos sujeitos, avaliados em dois momentos):

$$\begin{cases} H_0: & \mu_{1\,\text{mês}} \geq \mu_{\text{basal}}, \\ H_1: & \mu_{1\,\text{mês}} < \mu_{\text{basal}}. \end{cases}$$

De notar que a hipótese de interesse, (no caso, o tratamento ter efeito na redução dos níveis médios de glicémia) deve ser colocada como hipótese alternativa, em H_1.

O contexto de amostras independentes ou emparelhadas implica utilizar testes diferentes, adequados ao tipo de amostras. Enquanto no primeiro caso se pretende saber se a média se altera (se difere) em dois grupos de sujeitos diferentes, no segundo o que se pretende saber é se o valor da medição difere por sujeito, para dois ou mais momentos. Assim, aplicar um teste para amostras independentes sobre amostras que são na realidade emparelhadas (sob algum critério), ou o inverso, leva inevitavelmente a conclusões erradas.

ficativas entre um grupo de controlo e grupos com patologia, por exemplo, com vários tipos de tratamento

Nota 2.7. Algo que é evidente do exposto anteriormente é que para o caso de amostras independentes o tamanho (dimensão) das amostras dos grupos pode diferir (embora não tenha que diferir), enquanto que no caso de amostras emparelhadas a dimensão de amostras por grupo tem de ser igual (com excepção de casos de dados omissos).

2.3 Testes de Normalidade

Um outro aspeto a considerar na escolha do teste a aplicar em cada caso específico é se temos, ou não algum conhecimento *a priori* sobre a distribuição da variável, nomeadamente se a distribuição da mesma é normal (rever secção 1.8.2, se necessário). Caso a distribuição seja normal, podemos optar por um **teste paramétrico**, que é geralmente mais **potente**, isto é, para uma mesma significância, apresenta menor probabilidade de erros do tipo II (maior capacidade de rejeitar a hipótese nula quando ela é efetivamente falsa). Caso contrário, não tendo essa informação *a priori*, teremos que optar por um **teste não paramétrico**, que apesar de ser menos potente, não tem (em geral) restrições para a sua aplicação.

Na prática raramente sabemos se a população segue uma distribuição normal, pelo que temos de utilizar um **teste de normalidade**, cujas hipóteses são

$$\begin{cases} H_0: & \text{A distribuição da variável é normal,} \\ H_1: & \text{A distribuição da variável é não normal.} \end{cases}$$

É claro que, pela sua formulação, o teste falha menos quando a hipótese nula é rejeitada (ou seja, a distribuição não é normal), uma vez que nesse caso temos a hipótese de interesse (ou alternativa). Assim o teste é mais útil para mostar que um teste paramétrico não pode ser utilizado, do que para mostrar que pode. Como exemplos de teste de normalidade temos o **teste de Kolmogorov-Smirnov** e o de **Shapiro-Wilk**, ambos disponíveis no SPSS. O teste de Shapiro-Wilk é por norma mais potente, embora o de Kolmogorov apresente resultados muito semelhantes para amostras de dimensão elevada.

Como fazer em SPSS? 2.8 (Teste de Normalidade). Pode consultar a tabela 2.1 para saber como fazer em SPSS.

Uma alternativa amplamente utilizada pelos investigadores é considerar o **teorema do limite central**, que grosso modo, nos diz que a média (soma) dos valores de uma amostra obtida por amostragem independente e identicamente distribuída segue uma distribuição aproximadamente normal, quando a dimensão da amostra é suficientemente grande. Assim, é comum assumir o pressuposto de normalidade para amostras de dimensão grande.

2.4 Testes Paramétricos

O **teste paramétrico** tem na sua base uma hipótese, uma suspeita, sobre um parâmetro da população (seja o valor médio, uma proporção, p.ex.uma prevalência). Assume-se que o tipo de distribuição da população é conhecido, sendo desconhecidos apenas os valores dos parâmetros da mesma. Assim, o que se coloca

em teste é se o parâmetro populacional pode assumir um ou mais valores hipotéticos, dada uma amostra (sob o pressuposto de ser uma amostra aleatória simples). Por norma, os testes paramétricos clássicos mais usuais assumem que a distribuição da variável é normal em cada grupo, devendo apenas ser aplicado caso esta suposição seja verosímil. Caso contrário, deve-se optar por um teste não paramétrico, tal como descrito na secção 2.6.

2.4.1 Teste Paramétrico para a média (uma amostra)

No caso em que se tem uma única amostra e que se supõe que esta provém de uma população com distribuição normal, pode-se testar se a média desta pode a assumir um dado valor μ_0. É expectável que tanto a diferença entre a média amostral e a dimensão da amostra tenham influência nesta decisão. Por outro lado, é também expectável que o desvio padrão da população tenha influência, uma vez que uma maior dispersão permitiria uma maior diferença entre a amostra amostral e a populacional. De facto, consegue-se mostrar que a média $\bar{X}$ de uma amostra de dimensão n tem (nas condições acima descritas) uma distribuição da forma

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

e logo temos

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

No entanto, na prática o desvio padrão populacional é normalmente desconhecido, pelo que a fórmula anterior não é aplicável. No entanto, consegue-se também mostrar que

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{2.1}$$

em que S é o desvio padrão amostral (1.3). Note-se que neste caso a variável T pode ser calculada através dos valores amostrais, adicionando a hipótese $\mu = \mu_0$ (substituindo assim o valor de μ pelo de μ_0 que se quer testar) e tem uma distribuição t-student com $n-1$ graus de liberdade.

Temos então três tipos de hipóteses a considerar, como veremos de seguida.

Teste bilateral

Dado o teste de hipóteses **bilateral**

$$\begin{cases} H_0: & \mu = \mu_0, \\ H_1: & \mu \neq \mu_0 \end{cases}$$

aplicamos a estatística de teste (2.1) para averiguar se devemos rejeitar a hipótese nula de igualdade e aceitar a hipótese alternativa de que a média difere significativamente do valor em teste (diferir significa aqui ser superior ou inferior ao valor em teste, e por isso se denomina bilateral). Assim, temos regiões de rejeição da hipótese nula à esquerda e à direita, dada uma significância α, cada uma correspondendo a uma probabilidade de $\alpha/2$ (metade de α). Os valores correspondentes a essas probabilidades podem ser obtidos, uma vez que se assume

que a distribuição de T segue uma distribuição t-student. Caso T esteja numa região de rejeição, rejeita-se a hipótese nula H_0, caso contrário, não existe evidência estatística para rejeitar a hipótese nula. Isto é ilustrado através da figura 2.1.

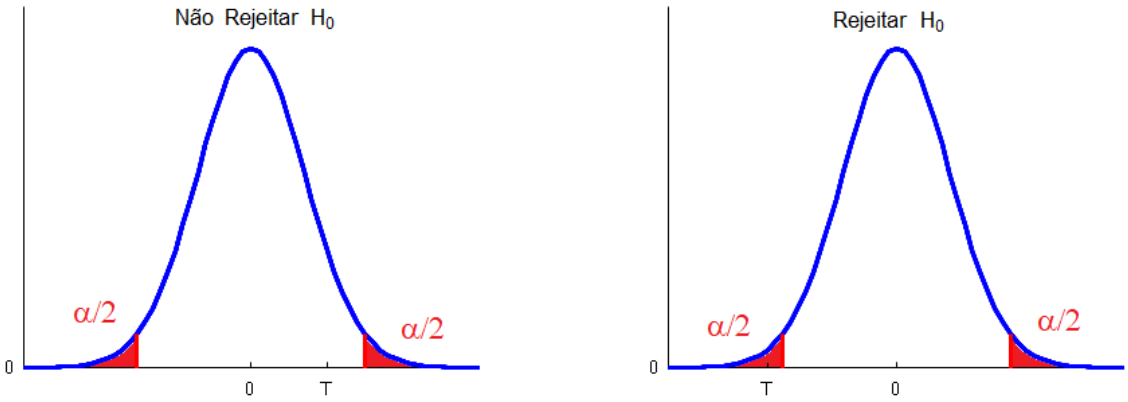


Figura 2.1: Caso Bilateral: teste-T para a média.

Outra forma de o fazer, é calcular o valor-p correspondente à estatística T. Esse valor pode ser obtido no SPSS. Assim, caso o valor-p seja inferior ao nível de significância α previamente fixado, rejeitamos a hipótese nula. Convém nesta altura relembrar a nota 2.2.

Teste unilateral à esquerda

Dado o teste de hipóteses **unilateral esquerdo** formulado por

$$\begin{cases} H_0: \quad \mu \geq \mu_0, \\ H_1: \quad \mu < \mu_0 \end{cases}$$

aplicamos da mesma forma a estatística de teste (2.1) para averiguar se devemos rejeitar a hipótese nula. No entanto como interessa testar para valores à esquerda de μ_0, temos apenas uma região de rejeição à esquerda, correspondendo a uma probabilidade de α (nível de significância). Caso T esteja contida na região de rejeição, rejeita-se a hipótese nula H_0, caso contrário, não existe evidência estatística para rejeitar a hipótese nula. Isto é ilustrado através da figura 2.2.

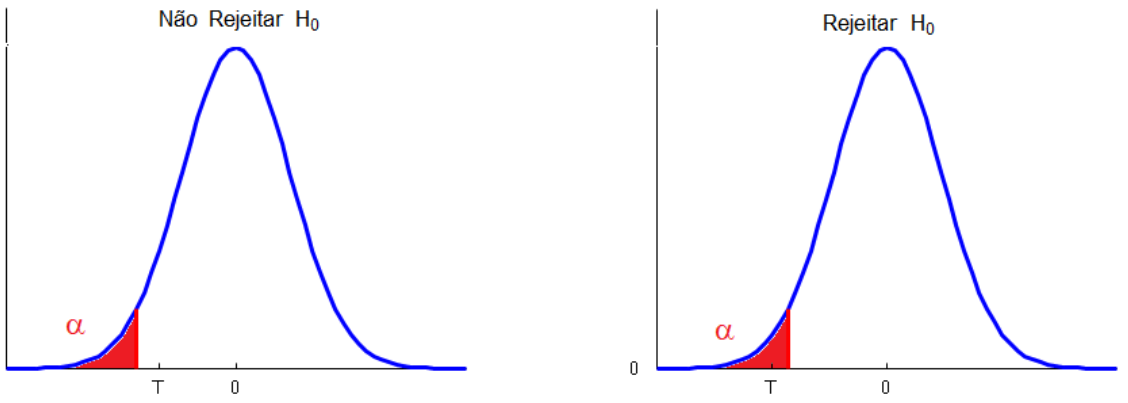


Figura 2.2: Caso Unilateral Esquerdo: teste-T para a média.

Da mesma forma, podemos obter o valor-p unilateral[3] correspondente à estatística T com o SPSS. Caso o valor-p seja inferior à significância, rejeitamos a hipótese nula.

Teste unilateral à direita

Dado o teste de hipóteses **unilateral direito**

$$\begin{cases} H_0: & \mu \leq \mu_0, \\ H_1: & \mu > \mu_0 \end{cases}$$

aplicamos da mesma forma a estatística de teste (2.1) para averiguar se devemos rejeitar a hipótese nula. A região de rejeição encontra-se agora à direita. Caso T esteja na região de rejeição, rejeita-se a hipótese nula H_0, caso contrário, não existe evidência estatística para rejeitar a hipótese nula. Isto é ilustrado através da figura 2.3.

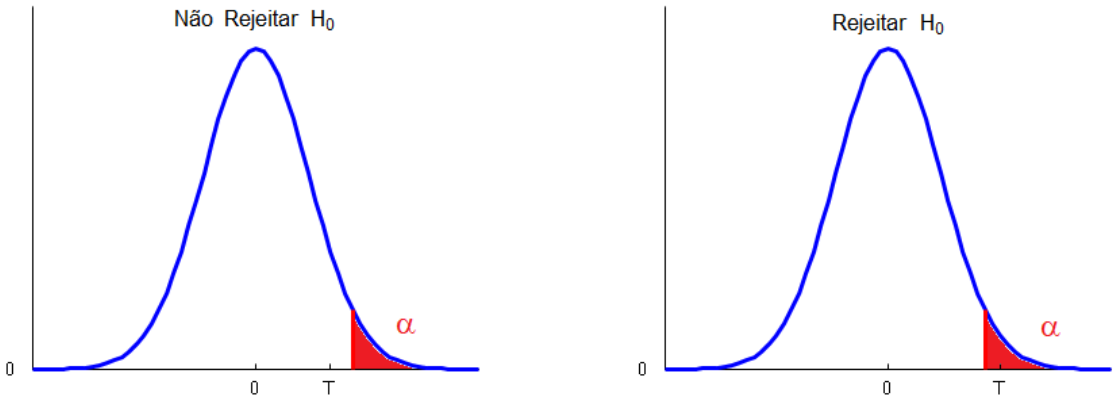


Figura 2.3: Caso Unilateral Direito: teste-T para a média.

Da mesma forma, podemos obter o valor-p unilateral correspondente à estatística T com o SPSS. Caso o valor-p seja inferior à significância, rejeitamos a hipótese nula.

Nota 2.9. Nas secções que se seguem relativas às situações para 2 ou mais amostras, podemos aplicar também teste bilaterais ou unilaterais, sendo o seu funcionamento semelhante (rejeita-se a hipótese nula, caso o valor-p seja inferior à significância, devendo-se considerar o valor-p unilateral (*one-tail*) ou bilateral (*two-tailed*). Assim, faremos a apresentação apenas para o caso bilateral.

2.4.2 Intervalo de Confiança para a média

É muito comum apresentar uma estimativa de um parâmetro desconhecido utilizando um intervalo de valores plausíveis (uma região) atribuindo em certo grau de confiança (probabilidade de estar correto). A estatística de teste (2.1) permite também determinar o **intervalo de confiança** para a média populacional, no caso em que a população tenha distribuição normal. Por **intervalo de confiança** denomina-se um intervalo tal que se pode esperar que o valor da média

[3] O SPSS permite obter o valor-p bilateral (*two-tailed*) e unilateral (*one-tail*).

populacional pertença a esse intervalo, com determinado grau de confiança. Assim, temos de (2.1)

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

que o intervalo de confiança $1-\alpha$ é dado por

$$I.C_{1-\alpha} = \left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

em que $t_{n-1,1-\alpha/2} = -t_{n-1,\alpha/2}$ uma vez que a distribuição t-student é simétrica, sendo estes valores definidos pela distribuição t-student (com $n-1$ graus de liberdade) de forma a satisfazer

$$P(T \leq t_{n-1,\alpha/2}) = \alpha/2, \qquad P(T \leq t_{n-1,1-\alpha/2}) = 1-\alpha/2,$$

para $T \sim t_{n-1}$.

Nota 2.10. Nos casos de uma amostra, testar se $\mu = \mu_0$ com significância α é equivalente a verificar se μ_0 pertence ao intervalo de confiança para a média, com confiança $1-\alpha$. No entanto, no caso de comparação de médias (que veremos adiante), testar se $\mu_1 = \mu_2$ com significância α **não** é equivalente a verificar se existe interseção entre os intervalo de confiança (a $1-\alpha$) para cada média.

Nota 2.11. Embora não o venhamos a ilustrar, nas secções que se seguem podemos também definir intervalos de confiança para as grandezas em questão, sendo estes sempre baseados na estatística de teste da variável em estudo.

2.4.3 Testes Paramétricos para igualdade de médias para amostras independentes

No caso das amostras serem independentes e é do interesse do investigador comparar as duas populações, há que medir se a diferença entre as médias é significativa. Partindo do facto de se assumir que as populações envolvidas têm (aproximadamente) distribuição normal, pode-se chegar à fórmula da estatística de teste para cada caso. Nesta secção seremos pouco profundos na descrição dessas estatísticas de teste, uma vez que no âmbito deste curso se espera que em cada situação prática estas sejam calculadas com recurso a SPSS. Na secção 2.7 serão revistos todos os testes de forma sistemática, para mais fácil consulta.

Testes Paramétricos para igualdade de médias de 2 amostras independentes

No caso geral, temos o teste de hipóteses para verificar se a diferença entre as médias populacionais μ_1 e μ_2 (das duas populações consideradas) é igual a d_μ dado por

$$\begin{cases} H_0: \ \mu_1 - \mu_2 = d_\mu, \\ H_1: \ \mu_1 - \mu_2 \neq d_\mu \end{cases}$$

mas o caso de maior interesse é geralmente testar se as médias das duas populações são iguais, ou seja, o caso $d_\mu = 0$. Nesse caso temos o teste de hipótese

$$\begin{cases} H_0: \ \mu_1 = \mu_2, \\ H_1: \ \mu_1 \neq \mu_2. \end{cases} \tag{2.2}$$

É expectável que tanto a diferença entre a médias, o desvio padrão das variáveis e a dimensão de cada amostra tenham influência nesta decisão, tal como no caso de uma amostra. A estatística de teste a usar difere no entanto, consoante os casos. Por exemplo, caso as variâncias das duas populações sejam idênticas, a estatística de teste pode simplificar-se. No caso mais geral, sem nenhuma suposição (além de que as distribuições de cada variável é normal), esta é dada por

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu$$

em que os graus de liberdade ν são o inteiro mais próximo de

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Como geralmente, caso as amostras sejam de dimensão grande, pode-se tomar a aproximação

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1).$$

Consoante o valor de T, obtém-se o respetivo valor-p, que comparado com a previamente definida significância, permite rejeitar ou não a hipótese nula.

Como fazer em SPSS? 2.12 (Teste-T para amostras independentes). Remetemos para a tabela 2.2 a aplicação em SPSS.

Nota 2.13 (Teste unilateral ou bilateral). Deve-se também referir que no contexto de bioestatística, a decisão entre usar um teste bilateral ou unilateral pode ter influência no resultado, nomeadamente na tese que se quer demonstrar. Por exemplo, se se quiser demonstrar que determinado tratamento tem efeito no aumento de determinado valor, deve-se usar a hipótese alternativa $H_1 : \mu_{\text{tratamento}} > \mu_{\text{controlo}}$, ou seja, o teste unilateral direito

$$\begin{cases} H_0 : & \mu_{\text{tratamento}} \leq \mu_{\text{controlo}}, \\ H_1 : & \mu_{\text{tratamento}} > \mu_{\text{controlo}}. \end{cases}$$

Assim, será mais fácil demonstrar que a média desse valor no grupo de tratamento é maior que no de controlo, uma vez que se concentra toda a significância na região de rejeição, que fica assim na zona da nossa hipótese de interesse.

Testes Paramétricos para igualdade de médias de 3 ou mais amostras independentes

Para comparar 3 ou mais médias (3 ou mais grupos), deve-se utilizar testes específicos para 3 ou mais amostras, uma vez que a utilização de testes (para 2 amostras) de comparação entre as amostras 2-a-2 aumenta o erro do tipo I global (contabilizando todas as comparações). Assim, para k amostras, temos as hipóteses

$$\begin{cases} H_0 : & \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1 : & \text{Existem } i \text{ e } j \text{ tais que } \mu_i \neq \mu_j. \end{cases} \tag{2.3}$$

De notar que a hipótese alternativa é apenas que existem pelo menos duas das médias difiram. Assim, não garante que todas as médias sejam diferentes, nem quais as médias que são diferentes. Para determinar quais as médias diferentes, teremos de optar por um **teste de comparação múltipla**, que abordaremos ainda nesta secção 2.4.3.

Para o teste de hipóteses (2.3) opta-se geralmente pela **ANOVA**, que é a abreviatura usual para *Analysis of Variance*, que tem este nome dado que na sua implementação as médias com base na dispersão dos valores observados, em relação a estas, dentro e entre grupos. O teste ANOVA tem dois pressupostos:

- **Normalidade**: as amostras provêm de populações com distribuição normal;
- **Homocedasticidade**: as amostras provêm de populações com iguais variâncias;

Para inspecionar a Normalidade, vimos já os testes da secção 2.3. Para testar a Homocedasticidade, utilizaremos os testes considerados na secção 2.5.

Nota 2.14 (Robustez da ANOVA). Embora a ANOVA parte dos pressupostos de Normalidade e Homocedasticidade, esta é robusta quanto à falha destes pressupostos. Assim, os resultados de um teste ANOVA são fiáveis, mesmo que estes princípios sejam ligeiramente violados.

O princípio da ANOVA é o de comparar a variação dentro de cada grupo (que se supõe semelhante, devido ao pressuposto da homocedasticidade) com a variação da amostra global. Assim uma variação global muito maior que a de cada grupo, pressupõe que as médias entre os grupos são diferentes, tornando o teste significativo. Remetemos para [2, 3], por exemplo, para pormenores sobre a estatística de teste e a base teórica.

Como fazer em SPSS? 2.15 (Teste ANOVA). Remetemos para a tabela 2.2 a aplicação em SPSS.

Testes de comparação múltipla

Como referimos anteriormente, a comparação das amostras 2-a-2 no caso de comparação de k amostras aumenta o erro do tipo I. Assim, quando o teste ANOVA é significativo, devemos escolher um teste de comparação múltipla para ser aplicado *a posteriori* para determinar entre que grupos existem diferenças[4]. É importante portanto que os testes controlem o erro de tipo I cometido.

Existem vários **testes de comparação múltipla**, entre os quais o teste de **Tukey**, **Scheffe**, **Duncan** ou **Bonferroni**, entre outros. Estes permitem determinar entre que grupos existem diferenças, quando o teste ANOVA foi significativo.

O teste de **Tukey** é aquele, de entre os métodos de uma fase, apresenta intervalos mais pequenos. Assim, o teste de Tukey tem maior facilidade em encontrar diferenças significativas. Este também tem vantagem em ser aplicado em casos equilibrados (ou seja, com amostras de dimensão igual entre grupos). O teste

[4]Recordamos que um teste ANOVA com resultado significativo apenas indica que existem pelo menos duas médias de grupo diferentes, mas não indica quantas nem quais.

de **Duncan** deve também ser aplicado em casos equilibrados e apresenta geralmente resultados mais discriminatórios que o de Tukey. Por outro lado,em casos não equilibrados, o teste de **Scheffe** produz intervalos de confiança maiores, pelo que o seu uso apresenta menos diferenças significativas. O teste de **Bonferroni** é um teste que pode ser utilizado em casos equilibrados ou não equilibrados e que apresenta em geral resultados menos conservadores que o de Scheffe.

Como fazer em SPSS? 2.16 (Teste Comparação múltipla). Remetemos para a tabela 2.2 a aplicação em SPSS.

2.4.4 Testes Paramétricos para igualdade de médias para amostras emparelhadas

No contexto da Bioestatística, testes para amostras emparelhadas têm grande relevância para estudos longitudinais (nos quais interessa a evolução temporal). Por exemplo, se quisermos avaliar o efeito de um tratamento inovador para a diabetes ao longo do tempo, deve-se comparar os níveis de glicémia para um mesmo grupo de sujeitos em tratamento ao longo do tempo, comparando-o, por ventura, a um grupo de controlo (sem tratamento, cou com um tratamento convencional). Assim, as características da amostra são diferentes do caso independente.

Exemplo em Bioestatística 2.17. Considere-se que se recolheram os níveis de glicémia para um grupo de sujeitos ao longo do tempo (instante basal e após 1 e 3 meses), conforme a tabela seguinte:

Sujeitos	A	B	C	D	E	F	G	H
Basal	250	190	198	204	232	189	214	200
1 mês	232	190	201	203	233	188	204	202
3 mês	204	172	171	165	195	150	178	167

No contexto do exemplo anterior, o que importa não é se a média dos valores de glicémia diminui ao longo do tempo, mas sim se os valores de glicémia por sujeito diminuem. Assim, um teste para amostras independentes não é adequado. Antes, deve-se estudar a diferenças de valores entre momentos para cada sujeito, do que os valores em si. Comecemos então por considerar o caso de 2 amostras emparelhadas.

Testes Paramétricos para igualdade de médias de 2 amostras emparelhadas

No caso de duas amostras emparelhadas, as hipóteses a testar são

$$\begin{cases} H_0: & \mu_D = D_0, \\ H_1: & \mu_D \neq D_0 \end{cases}$$

em que μ_D é a média populacional das diferenças e D_0 é um valor de teste para essa diferença. Mais uma vez, o caso de maior interesse é testar se a diferença é nula, ou seja, se

$$\begin{cases} H_0: & \mu_D = 0, \\ H_1: & \mu_D \neq 0 \end{cases}$$

podendo este ser substituído por um teste unilateral, quando adequado.

Nota 2.18 (Teste bilateral ou unilateral). O teste bilateral, deve ser considerado quando se quer mostrar que há diferenças, mas não existe vantagem (ou conhecimento prévio) em demonstrar se a diferença é positiva ou negativa. O teste unilateral, deve ser considerado em casos em que se pretende mostrar que a diferença é positiva ou é negativa. Por exemplo, para o exemplo 2.17 de um tratamento para a diabetes, como se pretende demonstrar que o valor de glicémia diminui (por exemplo, entre os instante basal e os 3 meses), dever-se-ia considerar o teste de hipóteses unilateral esquerdo

$$\begin{cases} H_0: & \mu_D \geq 0, \\ H_1: & \mu_D < 0. \end{cases}$$

A estatística de teste para amostras emparelhadas (provenientes de populações normais) é dada por

$$T_D = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

em que $\bar{D}$ e S_D são as média e desvio-padrão da amostra das diferenças (obtidas pelas diferenças de valores por sujeito entre os momentos 1 e 2). De notar que isto equivale portanto a um teste-T para uma amostra (2.1), definida pelas diferenças anteriores.

Como fazer em SPSS? 2.19 (Teste-T para Amostras emparelhadas). Remetemos para a tabela 2.3 a aplicação em SPSS.

Testes Paramétricos para igualdade de médias de 3 ou mais emparelhadas

Tal como caso de amostras independentes, temos de optar por um teste adequado para 3 ou mais amostras, nomeadamente pelo teste **ANOVA de medidas repetidas**. A aplicação deste teste implica normalidade e **esfericidade**. A esfericidade é o equivalente à homocedasticidade (homogeneidade/igualdade de variâncias), mas em vez de ser aplicada Às variáveis em si, é aplicada aos conjuntos de diferenças, entre diferentes instantes. Na prática, pode ser usado o **teste de Mauchly**, que pode ser obtido no SPSS.

Assim, a ANOVA de medidas repetidas testa a seguinte hipótese

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1: & \text{Existem } i \text{ e } j \text{ tais que } \mu_i \neq \mu_j. \end{cases} \tag{2.4}$$

sendo que as amostras são dependentes. Assim, testa se as diferenças por sujeito são nulas entre os vários instantes, ou, em alternativa, se existem pelo menos dois instantes em que a diferença não é nula.

Como fazer em SPSS? 2.20 (Teste ANOVA de medidas repetidas). Para a sua aplicação em SPSS, consulte a tabela 2.3.

2.5 Testes para igualdade de variâncias

Por vezes interessa comparar a variabilidade (desvio-pdrão, variância) de duas populações, se têm ou não dispersão idêntica, nomeadamente quando esta igualdade é um requisito para outras análises. Quando se assume o pressuposto de

que as amostras provêm de populações normais e queremos testar a igualdade de variâncias das mesmas, podemos o denominado **teste-F**. Dadas as hipóteses

$$\begin{cases} H_0: & \sigma_1 = \sigma_2, \\ H_1: & \sigma_1 \neq \sigma_2 \end{cases}$$

temos a estatística de teste

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1,n_2}$$

com distribuição F-Snedcor.

Como vimos anteriormente, testar se as variâncias de populações são ou não iguais a partir das respetivas amostras tem interesse não só por si, mas também para aferir se alguns testes podem ser aplicados, como por exemplo a ANOVA. Assim, neste capítulo temos obrigatoriamente de considerar testes para a hipótese de homocedasticidade, para 3 ou mais amostras.

O **teste de Levene**, disponível em SPSS, pode ser aplicado para 2 ou mais amostras e não exige normalidade das distribuições. Assim, permite testar

$$\begin{cases} H_0: & \sigma_1 = \sigma_2 = \cdots = \sigma_k, \\ H_1: & \text{Existem } i, j \text{ tal que } \sigma_i \neq \sigma_j. \end{cases}$$

Como fazer em SPSS? 2.21 (Teste Levene). Remetemos para a tabela 2.1 a aplicação deste teste em SPSS.

Outros exemplos de testes não paramétricos de igualdade de variâncias que não abordamos neste texto são os testes de Barlett ou de Cochran.

2.6 Testes Não Paramétricos

Os testes **testes não paramétricos** são testes estatísticos que na generalidade exigem menos condições de aplicabilidade (pressupostos) do que os testes paramétricos. O teste não paramétrico não é tão exigente sobre o tipo e distribuição dos dados. Os dados podem estar numa escala ordinal (por exemplo desde 1 ? concordo completamente até 7 ? discordo completamente). Também existe uma grande variedade de testes não paramétricos para dados que estão classificados em classes (escala nominal).

Nota 2.22 (Qual a necessidade de usar testes paramétricos?). Pode-se então colocar a questão de qual a necessidade de usar testes paramétricos, uma vez que poder-se-ia aplicar sempre testes não paramétricos. Uma maior aplicabilidade dos testes não paramétricos tem no entanto um preço. Por exemplo, se para um mesmo objetivo se verificarem as condições de aplicabilidade de um teste paramétrico, a decisão do teste paramétrico é mais potente (maior probabilidade da decisão ser correta, ao rejeitar uma hipótese nula). Portanto a resposta é simples e reside no seguinte: os testes paramétricos são mais fiáveis, sendo portanto (quando em condições de aplicação) preferíveis aos não-paramétricos.

Um exemplo desta diferença é já dado no argumento que se segue. Enquanto que um teste paramétrico aplicado a amostras provenientes de populações normais, assume essa distribuição e logo testa a veracidade de uma hipótes sobre

o valor do parâmetro média, num teste não paramétrico, não havendo essa suposição de uma lei de distribuição, a abordagem tem de ser diferente. Assim, o que se faz é uma comparação dos *rankings* (posições) dos valores das duas amostras. Assim, ordenando os valores das duas amostras, verificamos se estes estão distribuídos de forma semelhante em torno por exemplo de um mesmo valor central, ou se por outro lado, os valores da amostra A são predominantemente inferiores aos da amostra B, indicando nesse caso uma medida central diferente. Assim, uma vez que o teste se baseia numa comparação de rankings, os teste não paramétricos comparam a mediana das populações, em vez de comparar as respetivas médias.

2.6.1 Testes Não Paramétricos para igualdade de medianas para amostras independentes

Os testes para igualdade de medianas são baseados na comparação dos *rankings* dos valores das duas amostras, conforme descrito anteriormente.

Exemplo 2.23. Mesmo sem saber nada de testes de hipóteses, é expectável que dadas as duas amostras com valores

Amostra A	1	9	8	5	12	7
Amostra B	11	13	18	14	17	15

se conclua que as medianas das duas populações não são iguais, uma vez que quase todos os valores (com excepção de 1) da amostra A são inferiores ao menor valor da amostra B. Por outro lado, se tivermos os valores

Amostra A	11	19	18	15	12	17
Amostra B	15	13	18	14	17	15

já seja mais complicado decidir se de facto as medianas são iguais ou não.

Os testes não paramétricos baseiam-se em estatísticas para comparação de rankings, para confirmar (com determinado grau de significância) se as medianas são diferentes entre amostras. Tal como no caso paramétrico, é importante distinguir entre casos de comparação de 2 amostras ou de 3 ou mais amostras, para não agravar o erro do tipo I.

De referir também que no se se segue não iremos detalhar as estatísticas de teste envolvidas. Antes vamos, apenas referenciar quais os testes que se podem aplicar em cad caso e remeter para o capítulo 2.7 para um resumo e a sua aplicação usando o SPSS.

Testes Não Paramétricos para igualdade de medianas de 2 amostras independentes

Para o caso de duas amostras independentes e as hipóteses

$$\begin{cases} H_0: & \mathrm{med}_1 = \mathrm{med}_2, \\ H_1: & \mathrm{med}_1 \neq \mathrm{med}_2 \end{cases}$$

utiliza-se geralmente o **teste de Mann-Whitney**.

Como fazer em SPSS? 2.24 (Teste de Mann-Whitney). Remetemos para a tabela 2.4 para a sua aplicação usando o SPSS.

Testes Não Paramétricos para igualdade de medianas de 3 ou mais amostras independentes

No caso de três ou mais amostras independentes usa-se, geralmente, o **Kruskal-Wallis**, para testar as hipóteses

$$\begin{cases} H_0: & \text{med}_1 = \text{med}_2 = \cdots = \text{med}_k, \\ H_1: & \text{med}_i \neq \text{med}_j \text{ para algum i,j.} \end{cases}$$

Como fazer em SPSS? 2.25 (Teste de Kruskal-Wallis). Remetemos para a tabela 2.4 para a sua aplicação usando o SPSS.

2.6.2 Testes Não Paramétricos para igualdade de medianas para amostras emparelhadas

No caso de amostras emparelhadas, há que considerar as diferenças dos valores e os *rankings* das diferenças em valor absoluto. Se este ranking for semelhante para valores (de diferenças) positivos e negativos, as amostras têm distribuição semelhante. Caso contrário, existem diferenças nas distribuições.

Testes Não Paramétricos para igualdade de medianas de 2 amostras emparelhadas

Para o caso de duas amostras emparelhadas e as hipóteses

$$\begin{cases} H_0: & \text{med}_1 = \text{med}_2, \\ H_1: & \text{med}_1 \neq \text{med}_2 \end{cases}$$

podemos usar o **teste de Wilcoxon** ou o **teste dos Sinais**. O **teste de McNemar** aplica-se ao caso específico em que as variáveis são dicotómicas, ou seja, apenas assumem o valor 0 ou 1. Este serve para aferir mudanças de estado (de 0 para 1, ou de 1 para 0).

Como fazer em SPSS? 2.26 (Teste de Wilcoxon). Remetemos para a tabela 2.5 para a sua aplicação usando o SPSS.

Testes Não Paramétricos para igualdade de medianas de 3 ou mais emparelhadas

No caso de três ou mais amostras emparelhadas usa-se, geralmente, o **teste de Friedman**, para testar as hipóteses

$$\begin{cases} H_0: & \text{med}_1 = \text{med}_2 = \cdots = \text{med}_k, \\ H_1: & \text{med}_i \neq \text{med}_j \text{ para algum i,j.} \end{cases}$$

Como fazer em SPSS? 2.27 (Teste de Friedman). Remetemos para a tabela 2.5 para a sua aplicação usando o SPSS.

2.6.3 Teste de independência do Qui-quadrado

Além de testes de igualdade de médias e variâncias, podemos também estudar a associação de duas variáveis de outra forma. O **teste de independência do Qui-quadrado** pode ser aplicado a variáveis nominais[5] e parte da tabela de contingência das duas variáveis. Assim, este parte da tabela

		Variável X				Total
		x_1	x_2	$\dots$	x_p	
Variável Y	y_1	O_{11}	O_{12}	$\dots$	O_{1p}	$O_{1\bullet}$
	y_2	O_{21}	O_{22}	$\dots$	O_{2p}	$O_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	y_q	O_{q1}	O_{q2}	$\dots$	O_{qp}	$O_{q\bullet}$
Total		$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet p}$	n

em que O_{ij} é o número de sujeitos observados que satisfazem $Y = y_i$ e $X = x_j$, $n_{\bullet j}$ é o número de sujeitos observados que satisfaz $X = x_j$ e $n_{i\bullet}$ é o número de sujeitos observados que satisfaz $Y = y_j$ e n é o número total de observações de sujeitos.

Dado o teste de hipóteses

$$\begin{cases} H_0: & X \text{ e } Y \text{ são independentes,} \\ H_1: & X \text{ e } Y \text{ são dependentes} \end{cases}$$

a estatística de teste do teste de independência do teste do Qui-quadrado é dada por

$$\chi = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{p-1,q-1}$$

em que o valor esperado E_{ij} para cada célula da tabela é dado por

$$E_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}.$$

Nota 2.28. O teste de independência do Qui-quadrado apenas deve ser utlizado se no máximo 20% dos valores esperados são menores que 5. Noutros casos, devemos seguir outro tipo de testes, como por exemplo o **teste de Fisher**[6] (ver [1]).

Como fazer em SPSS? 2.29 (Teste de Qui-quadrado e de Fisher). Remetemos para a tabela 2.6 para a sua aplicação usando o SPSS.

Exemplo em Bioestatística 2.30. Pretende-se saber se as variáveis género e diagnóstico são independentes, para aferir se o risco é maior no sexo masculino ou feminino. Aplicando o teste Qui-quadrado de independência ao ficheiro `BaseDados_Notas.sav`, obteve-se então um p-value $p < 0.001$, pelo que se conclui que são dependentes.

Nota 2.31 (p-value nunca é nulo). Por vezes o resultado indicado no output SPSS é $p = 0.000$. No entanto, como o p-*value* representa a probabilidade de obtermos a

[5] Veremos adiante que para variáveis ordinais ou quantitativas, estudar a correlação dará mais informação que a associação entre variáveis

[6] O teste de Fisher apenas pode ser aplicado para variáveis dicotómicas.

amostra observada dada a hipótese nula, este nunca será zero, pois representaria um fenómeno impossível. É portanto comum escrever-se nestes casos que $p < 0.001$, como no exemplo anterior.

Uma questão que fica por responder após o teste anterior é qual a variação do risco consoante estamos numa ou noutra classe.

2.6.4 *Odds Ratio* e Risco Relativo

No exemplo 2.30 fica claro que a taxa de diagnósticos é diferente para sexo masculino e feminino, mas não é evidente quanto varia o risco de um caso para outro. Esta informação tem obviamente relevância médica, pelo que é importante conseguir medi-la.

Para o efeito, podemos medir o **risco relativo** e o ***Odds-Ratio***[7], também chamado de **razão de chances**.

Suponhamos então que temos dois grupos de exposição (E-exposto e NE-não exposto) e dois diagnósticos (D- Diagnóstico Positivo, que corresponde ao evento em estudo, e ND-diagnóstico negativo).

O risco de diagnóstico positivo no grupo de exposição é dado pela probbilidade condicionada $P(D|E)$, ou seja, é a razão entre o número de casos com diagnóstico positivo dentro do grupo de exposição e o número de casos total do grupo de exposição.

O risco relativo é dada pela razão entre os riscos dos grupos de exposição e de não exposição, ou seja,

$$RR = \frac{P(D|E)}{P(D|NE)}.$$

Assim, o risco relativo tem a ver com a razão de duas probabilidades, com base em dados passados.

O *Odds-ratio* não tem a ver com probabilidades, mas sim com chances. Neste sentido, está relacionado com eventos futuros que ainda não ocorreram. Assim, a chance para o grupo de exposição é dada pela razão entre a probabilidade de ocorrer o evento no grupo e a probabilidade de não ocorrer, ou seja, é dada pela razão entre as probabilidades condicionadas

$$C_E = \frac{P(D|E)}{P(ND|E)}.$$

Assim, o *Odds-Ratio* é pela razão entre as cahances de cada grupo, ou seja, por

$$OR = \frac{C_E}{C_{NE}} = \frac{\frac{P(D|E)}{P(ND|E)}}{\frac{P(D|NE)}{P(ND|NE)}} = \frac{P(D|E) \times P(ND|NE)}{P(ND|E) \times P(D|NE)}.$$

Geralmente, os valores do *Odds-Ratio* e do Risco Relativo não são muito diferentes.

Como fazer em SPSS? 2.32 (Risco relativo e *Odds-Ratio*). Remetemos para a tabela 2.6 para a sua aplicação usando o SPSS. O SPSS permite obter também os intervalos de confiança para estes valores. De notar que o SPSS considera como

[7] É importante notar que para calcular o risco relativo e Odds-Ratio poderemos apenas considerar variáveis dicotómicas, ou seja, tabelas de contingência 2×2.

evento (diagnóstico positivo) como a primeira coluna, e como grupo de exposição a primeira linha, pelo que as variáveis devem ser introduzidas (ou recalculadas) de forma a satisfazer essa condição.

Ilustraremos estas duas grandezas com um exemplo, seguindo o exemplo 2.30.

Exemplo em Bioestatística 2.33 (Continuação do exemplo 2.30). Temos a tabela de contingência

		Diagnóstico		Total
		Paciente	Controlo	
Género	Feminino	25	72	97
	Masculino	114	92	206
Total		139	164	303

Vamos considerar que o género feminino é o grupo de exposição e o masculino é o de não exposição, embora no caso se pudesse optar pelo contrário (o que levará a resultados diferentes para o risco relatico, mas iguais para o *Odds-ratio*).

Assim, a probabilidade de ter diagnóstico positivo, dentro do grupo do género feminino, é dado por

$$P(D|F) = \frac{25}{97} = 0.2577$$

enquanto que a probabilidade de ter diagnóstico positivo, dentro do grupo do género masculino, é dada por

$$P(D|M) = \frac{114}{206} = 0.5534.$$

Assim, o risco relativo entre os géneros feminino e masculino é dado pela razão das probabilidade condicionadas anteriores, ou seja,

$$RR_F = \frac{25/97}{114/206} = \frac{0.2577}{0.5534} = 0.466.$$

Da mesma forma, temos para o género masculino que o risco relativo é

$$RR_M = \frac{114/206}{25/97} = 2.1472.$$

O *Odds-ratio* é a razão das chances, pelo sendo as chances para cada género dadas por

$$C_F = \frac{P(D|F)}{P(ND|F)} = \frac{25/97}{72/97} = \frac{25}{72} = 0.3472,$$
$$C_M = \frac{P(D|M)}{P(ND|M)} = \frac{114/206}{92/206} = \frac{114}{92} = 1.2391,$$

temos o *Odds-ratio* entre géneros feminino e masculino dado por

$$OR = \frac{C_F}{C_M} = 0.2802$$

pelo que a chance de ter diagnóstico positivo é 0.28 vezes menor nas mulheres, ou dito de outra forma, por cada mulher diagnosticada há cerca de 3.57 homens diagnosticados (uma vez que $3.5687 = 1/0.2802$).

Exercício 2.34. Confirme os resultados anteriores utilizando o SPSS. Talvez seja necessário criar uma nova variável de forma a que o evento (diagnóstico positivo) possa estar na primeira linha, conforme indicado na descrição 2.32 de como fazer em SPSS.

2.7 Tabelas de Resumo de Testes de Hipóteses

Nesta secção apresentamos um resumo deste capítulo, indicando os principais testes de hipóteses a utilizar em cada caso e as instruções de como os executar em SPSS.

Apesar dos links estarem nas tabelas seguintes, deixamos aqui também a listagem dos links para os vídeos sobre a utilização do SPSS neste contexto:

- Testes de Normalidade:

- Testes para média de uma amostra:

- Testes para médias de duas amostras independentes:

- Testes para médias de três ou mais amostras independentes:

- Testes para médias de duas amostras emparelhadas:

- Testes para médias de três ou mais amostras emparelhadas:

- Teste de independência do Qui-quadrado e *Odds-Ratio*:

Tabela 2.1: Teste não-paramétricos de Normalidade e Homocedasticidade. Clique no logo SPSS uma visualização de como fazer em SPSS.

Hipóteses	Teste	Aplicabilidade	SPSS
$\begin{cases} H_0: & \text{A distribuição é normal,} \\ H_1: & \text{A distribuição é não normal} \end{cases}$	Shapiro-Wilk	Sem condições	Analisar ⇝ Estatística Descritiva ⇝ Explorar ⇝ Gráficos • ☑ Teste de normalidade com gráficos ⇝ Continuar • Escolher variáveis ⇝ Ok
	Kolmogorov - Smirnov	Amostras Grandes	(igual ao anterior ou:) Analisar ⇝ Testes não paramétricos ⇝ Caixas de Diálogo Legadas ⇝ K-S de uma amostra • Escolher variáveis • ☑ Dist. Normal ⇝ Ok
$\begin{cases} H_0: & \sigma_1 = \sigma_2 = \cdots = \sigma_k, \\ H_1: & \text{Existem } i, j \text{ tal que } \sigma_i \neq \sigma_j. \end{cases}$	Levene	Sem condições	(Ver ANOVA:) Analisar ⇝ Comparar Médias ⇝ Análise Variância Unidirecional • Escolher variável • Escolher Fator (variável de Grupo) ⇝ Opções • ☑ Teste Homogeneidade Variâncias ⇝ Continuar ⇝ Ok

Tabela 2.2: Testes paramétricos de igualdade de médias para amostras independentes. Clique no logo SPSS uma visualização de como fazer em SPSS.

	Hipóteses	Teste	Aplicabilidade	SPSS
1 amostra	$\begin{cases} H_0: & \mu = \mu_0, \\ H_1: & \mu \neq \mu_0 \end{cases}$	Teste-T	Normalidade	Analisar ⇝ Comparar Médias ⇝ Uma amostra Teste-T • Escolher variável • Escolher valor de teste (μ_0) ⇝ Ok
2 amostras	$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2 \end{cases}$	Teste-T	Normalidade	Analisar ⇝ Comparar Médias ⇝ Amostras Independentes Teste-T • Escolher variável • Escolher variável de Grupo ⇝ Ok
$k \geq 3$ amostras	$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1: & \mu_i \neq \mu_j \text{ para algum i,j} \end{cases}$	ANOVA	Normalidade Homocedasticidade	(Ver teste de Levene): Analisar ⇝ Comparar Médias ⇝ Análise Variância Unidirecional • Escolher variável • Escolher Fator (variável de Grupo) ⇝ Ok
C.Múltipla	$\begin{cases} H_0: & \mu_i = \mu_j, \\ H_1: & \mu_i \neq \mu_j \end{cases}$	Tukey	Equilibrado, Conservador	Seguir 5 passos iniciais da ANOVA e ⇝ Posterori • ☑ do teste pretendido ⇝ Continuar
		Duncan	Equilibrado, Não Conservador	
		Scheffe	Não Equilibrado, Conservador	
		Bonferroni	Não Equilibrado, Não Conservador	

Tabela 2.3: Testes paramétricos de igualdade de médias para amostras emparelhadas. Clique no logo SPSS uma visualização de como fazer em SPSS.

	Hipóteses	Teste	Aplicabilidade	SPSS
2 amostras	$\begin{cases} H_0: & \mu_D = 0, \\ H_1: & \mu_D \neq 0 \end{cases}$	Teste-T	Normalidade	Analisar ⇝ Comparar Médias ⇝ Amostras Pareadas Teste-T • Escolher variável 1 e 2 ⇝ Ok
$k \geq 3$ amostras	$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1: & \mu_i \neq \mu_j \text{ para algum i,j} \end{cases}$	ANOVA Medidas Repetidas	Normalidade Esfericidade	Analisar ⇝ Modelo Linear Geral ⇝ Medidas Repetidas • Escolher nº níveis ⇝ Incluir ⇝ Definir • Escolher variáveis ⇝ Ok

Tabela 2.4: Testes não paramétricos de igualdade de medianas para amostras independentes. Clique no logo SPSS uma visualização de como fazer em SPSS.

	Hipóteses	Teste	SPSS
2 amostras	$\begin{cases} H_0: & \text{med}_1 = \text{med}_2, \\ H_1: & \text{med}_1 \neq \text{med}_2 \end{cases}$	Mann-Whitney	Analisar ⇝ Testes não paramétricos ⇝ Caixas de Diálogo Legadas ⇝ 2 Amostras independentes • ☑ Mann-Whitney • Escolher par de variáveis ⇝ Ok
$k >$ 3 amostras	$\begin{cases} H_0: & \text{med}_1 = \text{med}_2 = \cdots = \text{med}_k, \\ H_1: & \text{med}_i \neq \text{med}_j \text{ para algum i,j} \end{cases}$	Kruskall-Wallis	Analisar ⇝ Testes não paramétricos ⇝ Caixas de Diálogo Legadas ⇝ K Amostras independentes • ☑ Kruskall-Wallis • Escolher k variáveis ⇝ Ok

Tabela 2.5: Testes não paramétricos de igualdade de medianas para amostras emparelhadas. Clique no logo SPSS uma visualização de como fazer em SPSS.

	Hipóteses	Teste	SPSS
2 amostras	$\begin{cases} H_0: & \text{med}_1 = \text{med}_2, \\ H_1: & \text{med}_1 \neq \text{med}_2 \end{cases}$	Wilcoxon	Analisar ⇝ Testes não paramétricos ⇝ Caixas de Diálogo Legadas ⇝ 2 Amostras relacionadas • ☑ Wilcoxon • Escolher par de variáveis ⇝ Ok
		Sinais	Igual ao anterior excepto: • ☑ Sinais
$k \geq 3$ amostras	$\begin{cases} H_0: & \text{med}_1 = \text{med}_2 = \cdots = \text{med}_k, \\ H_1: & \text{med}_i \neq \text{med}_j \text{ para algum i,j} \end{cases}$	Friedman	Analisar ⇝ Testes não paramétricos ⇝ Caixas de Diálogo Legadas ⇝ K Amostras relacionadas • ☑ Friedman • Escolher k variáveis ⇝ Ok

Tabela 2.6: Teste de independência do Qui-quadrado, Fisher e cálculo do Risco relativo e *Odds-Ratio*, para variáveis qualitativas. Clique no logo SPSS uma visualização de como fazer em SPSS.

Hipóteses	Teste	SPSS
$\begin{cases} H_0: & X \text{ e } Y \text{ são independentes,} \\ H_1: & X \text{ e } Y \text{ são dependentes} \end{cases}$	Qui-quadrado	Analisar ⇝ Estatística Descritiva ⇝ Tabela de referência cruzada • Escolher variável linha e coluna ⇝ Estatística • ☑ Qui-quadrado ⇝ Continuar ⇝ Ok
Não aplicável	Risco relativo, Odds-Ratio	Igual ao anterior até Estatísticas • ☑ Risco

Capítulo 3

Correlação e Regressão

Neste capítulo vamos abordar o conceito de **correlação**, como medida de associação entre duas variáveis quantitativas ou qualitativas ordinais. Desta forma podemos quantificar essa associação através de um **coeficiente de correlação** apropriado.

Nos casos em que essa associação é forte e significativa, faz sentido procurar um modelo, uma equação, que relacione as duas variáveis entre si. Nesse sentido, será introduzido o modelo de **regressão**.

Neste capitulo vamos abordar estes dois conceitos, com maior detalhe. Para uma motivação sobre a aplicação destes conceitos em Bioestatística, clique no link seguinte:

3.1 Correlação

O objetivo da **correlação** é quantificar a relação linear entre duas variáveis X e Y, nomeadamente através do cálculo de um **coeficiente de correlação**, normalmente denotado por R.

Desta forma, deve-se iniciar a análise por uma exploração dos dados, e averiguar graficamente através de um **gráfico de dispersão** (como na figura 3.1) se existe alguma tendência (um padrão) linear, ou seja, se a nuvem de pontos se concentra em torno de uma reta de declive não nulo, sendo este positivo (relação linear positiva) ou negativo (relação linear negativa).

No entanto, a análise anterior é bastante subjetiva e insuficiente para avaliar a força da relação entre as duas características, pelo que se pretende quantificar de forma objetiva essa correlação, com interpretação simples. Nessa perspetiva, pretende-se um coeficiente R que tenha valores entre -1 e 1 (um valor padronizado), tal que o sinal de R determine se a relação linear é positiva ou negativa, isto é, se o aumento do valor de uma variável está associado diretamente com o aumento na outra variável (**correlação positiva**) ou se é o oposto, se está associado a um decréscimo da outra variável. (**correlação negativa**). Assim, se

- R é próximo de 1, a relação linear é positiva forte, isto é, quando o valor de uma variável aumenta, existe uma forte tendência ao valor da outra variável aumentar linearmente (uma proporção direta);

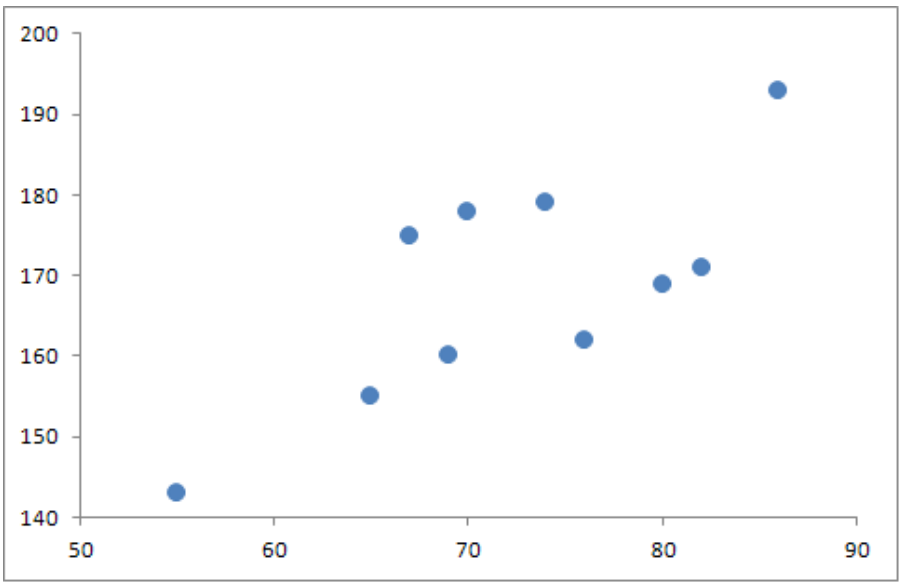


Figura 3.1: Representação gráfica dos dados de altura e peso, para dez sujeitos.

- R é próximo de -1, a relação linear é negativa forte, isto é, quando o valor de uma variável aumenta, existe uma forte tendência ao valor da outra variável diminuir linearmente;
- R é próximo de 0, não existe relação linear entre as duas variáveis, isto é, quando o valor de uma variável aumenta, existe uma forte tendência ao valor da outra variável diminuir linearmente;

Nota 3.1 (Correlação forte, moderada ou fraca). Na bibliografia, a qualificação da correlação consoante o valor do coeficiente R varia bastante. No entanto, por norma considera-se para os seguintes intervalos do valor absoluto |R| que

- se $|R| > 0.7$, a correlação é **forte**;
- se $0.3 < |R| \leq 0.7$, a correlação é **moderada**;
- se $|R| \leq 0.3$, a correlação é **fraca**;

Assim, quando a correlação é fraca, assume-se que as variáveis não estão correlacionadas. Quando é forte, assume-se que a variação de uma variável tem uma forte influencia (positiva ou negativa, consoante o sinal de R) na variação da segunda variável. Quando a correlação é moderada, a influência é também moderada.

Nota 3.2 (Normalização). Note-se também que o valor de R é **normalizado**, no sentido em que varia sempre entre -1 e 1, independentemente da magnitude das grandezas dos valores das variáveis X e Y. Desta forma, o coeficiente de correlação pode ser comparável entre dois pares de variáveis, mesmo que estes tenham grandezas muito distintas.

Nota 3.3 (Simetria). Outra característica desejável no coeficiente de correlação é a **simetria**, isto é, que o coeficiente de correlação ente X e Y seja igual ao coeficiente de correlação entre Y e X. Assim, os coeficientes de correlação apresentados satisfazem também esta condição.

Nas próximas secções vamos estudar em detalhe qual o coeficiente de correlação apropriado em cada caso e como verificar se este é significativo, ou seja, se tem relevância estatística dada a dimensão da amostra.

Como fazer em SPSS? 3.4 (Correlação). Em qualquer caso e como as expressões respetivas para o cálculo destas grandezas são complexas, pode clicar no link seguinte para visualizar um vídeo sobre como obter coeficientes de correlação em SPSS. Aconselhamos no entanto que percorra primeiro o texto seguinte desta secção sobre correlação e só depois visualize a forma de calcular os coeficientes em SPSS:

3.1.1 Correlação Paramétrica - Coeficiente de Pearson

Comecemos por considerar duas variáveis X e Y quantitativas, provenientes de distribuições normais. Neste caso, temos o **coeficiente de correlação de Pearson** dado por

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}. \tag{3.1}$$

Consegue mostrar que este coeficiente satisfaz todas as propriedades descritas anteriormente. Em particular, consegue-se mostrar que se X e Y tiverem uma relação linear perfeita, isto é, se

$$Y = \beta_0 + \beta_1 X$$

para constantes β_0 e $\beta_1 \neq 0$, então $R = \text{sgn}(a)$, isto é, R é 1 ou -1, consoante o sinal[1] de β.

Significância do Coeficiente de Pearson

O valor de R indica-nos se a correlação é positiva ou negativa e se é forte, moderada ou fraca. No entanto, para sabermos se o seu valor tem significado estatístico dada uma amostra, ou seja, se este valor pode ser extrapolado, teremos de testar se este é significativamente diferente de zero para a dimensão da amostra recolhida. Assim, para o teste de hipóteses que assume na hipótese nula que não existe correlação

$$\begin{cases} H_0: & R = 0, \\ H_1: & R \neq 0 \end{cases} \tag{3.2}$$

temos a estatística de teste

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}. \tag{3.3}$$

Note-se portanto que o facto de o valor do coeficiente R ser significativo depende apenas do valor de R e da dimensão da amostra n. Além disso, o valor tende a ser

[1] É comum denotar o sinal de uma constante β por $\text{sgn}(\beta)$.

mais significativo quanto maior o seu valor absoluto (uma vez $|R|$ a tender para 1 implica que $\sqrt{1-R^2}$ no denominador tende para zero e logo a estatística T tende para $\pm\infty$) ou quanto maior for a dimensão da amostra (uma vez que o valor absoluto de T aumenta quando n aumenta). Também pode ocorrer a situação em que valor de correlação fraca ser signaficativa, p. ex. $R = 0,2$ com p-value=0,003. A conclusão é que a correlação existe, pode ser extrapolada para uma população ou grupo, contudo a relação é fraca, ou seja, os valores de uma variável não têm grande impacto na evolução dos valores da outra.

3.1.2 Correlação Não Paramétrica - Coeficiente de Spearman

No caso de variáveis quatitativas que não provêm de distribuições normais ou de variáveis numa escala ordinais, o coeficiente de Pearson 3.1 não é aplicável, perde um pouco da sua potência. Desta forma, opta-se geralmente por um coeficiente alternativo, não paramétrico, que exige menos pressupostos de validade que é o **coeficiente de correlação de Spearman** dado por

$$R_{\text{Spearman}} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

em que D_i é a diferença entre os *rankings* amostrais[2].

Nota 3.5 (Aplicação a variáveis qualitativas ordinais). O coeficiente de Spearman pode ser aplicado a variáveis qualitativas ordinais, uma vez que apenas depende dos *rankings* amostrais. Assim, estando estabelecida uma ordem para os valores da variáveis, pode-se calcular o coeficiente de correlação de Spearman, ainda que a variável não seja quantitativa.

A interpretação do coeficiente de Spearman é semelhante à do coeficiente de Pearson, de acordo com o seu sinal e a nota 3.1.

Nota 3.6 (variáveis não métricas). Existem situações para as quais a interpretação do coeficiente pode ser diferente da regra usual. Por exemplo, no domínio da psicologia, as variáveis psicológicas ou comportamentais são mais difíceis de medir com exatidão, e consideram-se muitas vezes escalas de pontos (1-não gosto nada.....7-gosto muito). Assim, é comum considerar que um coeficiente $|R| > 0.5$ calculado sobre este tipo de medições já é considerado moderado-forte. Recomenda-se uma pesquisa e leitura prévia (artigos, etc) sobre o tipo de variáveis em estudo, por forma a fazer uma interpretação que seja adequada a cada situação concreta.

Significância do Coeficiente de Spearman

Da mesma forma, é necessário verificar se o coeficiente de Spearman é significativo, isto é, se a amostra tem informação suficiente para garantir que o valor R_{Spearman} é significativo.

[2]Sendo x_i o j-ésimo menor valor da amostra para a variável X (e logo com *ranking* k) e y_i o k-ésimo menor valor da amostra para a variável Y (e logo com *ranking* k) então $D_i = j - k$.

Uma hipótese é usar a mesma estatística de teste (3.3) (com R_{Spearman} no lugar de R) para o teste de hipóteses

$$\begin{cases} H_0: & R_{\text{Spearman}} = 0, \\ H_1: & R_{\text{Spearman}} \neq 0. \end{cases} \tag{3.4}$$

No entanto, alguns autores indicam que é mais indicado neste caso usar a estatística de teste aproximada (para uma amostra suficientemente grande)

$$Z = \sqrt{\frac{n-3}{1-06}} F\left(R_{\text{Spearman}}\right) \sim N(0,1)$$

que segue aproximadamente uma distribuição normal, usando a transformada de Fisher

$$F(r) = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

3.2 Regressão Linear Simples

O coeficiente de correlação permite-nos quantificar quão linearmente associadas estão duas variáveis, mas não permite obter um equação mais completa sobre a forma como estão correlacionadas.

Dessa forma, introduzimos o conceito de **Regressão linear simples** entre duas variáveis quantitativas X e Y. Este consiste em assumir que as variáveis X e Y estão relacionadas na forma linear

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{3.5}$$

em que β_1 é o declive da reta e corresponde à variação esperada em Y, dada a variação de uma unidade em X, β_0 é a ordenada na origem e representa o valor de Y para X nulo e o erro ε (a letra grega lê-se epsilon) se assume com distribuição normal e média nula.

Nota 3.7 (Variável dependente e independente). A relação (3.5) define também qual a **variável dependente** e a **variável independente**. No caso, a variável dependente (tal como o nome indica) depende de outras, sendo influenciada pelo seu valor. No caso da relação linear (3.5) a variável independente é X e a dependente é Y, uma vez que cada valor de X define o respetivo valor de Y.

Devido à presença de erro, a troca da variável dependente e independente leva a resultados diferentes, que não são simplesmente a inversão do declive da reta. Assim, a definição da variável dependente e independente influencia o modelo de regressão linear, pelo que esta escolha deve ser bem ponderada, no sentido de se colocar como variável dependente aquela que se quer que seja influenciada pela outra.

Assim, assume-se que os valores amostrais satisfazem

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{3.6}$$

pelo que se pretende estimar os valores β_0 e β_1 que determinam a componente determinística do modelo

$$Y = \beta_0 + \beta_1 X \tag{3.7}$$

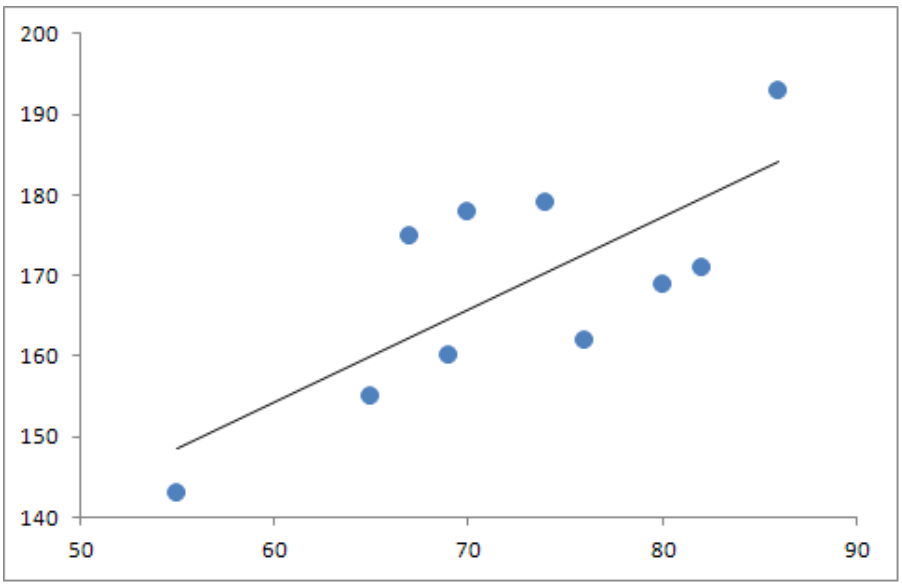


Figura 3.2: Modelo linear para os dados na figura 3.1.

de forma a obter o erro mínimo. Em regressão, geralmente procura-se os estimadores que minimizam a soma dos valores ε_i^2, ou seja, o erro quadrático médio. Dessa forma, obtem-se o estimador para o declive da reta

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$

e para a ordenada na origem

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

obtendo-se assim o modelo de regressão linear

$$\hat{y} = \beta_0 + \beta_1 x.$$

Nota 3.8 (Ajuste do modelo). A qualidade do **ajuste** do modelo linear pode-se medir através de R^2, em que R é o valor do coeficiente de correlação de Pearson. De facto, consegue-se mostrar que R^2 é a percentagem de variação linear de Y explicada pela variação linear X, por outras palavras, o quanto da variação dos valores da variável dependente está relacionado com a variável independente. Daí que também se chame **taxa de explicação** ao valor de R^2. Assim, quanto maior o valor de R^2 (ou do coeficiente de correlação R), maior a qualidade do ajuste do modelo linear.

Nota 3.9 (Relevância do modelo linear). O modelo linear dá-nos informação relevante sobre a associação entre as duas variáveis X e Y e de que forma a variável independente Y influencia a variável dependente Y. A saber:

- O valor de β_1 define qual a variação esperada para Y, dada a variação de uma unidade em X;
- Da mesma forma, se X aumenta em k unidades, espera-se que o aumento de Y seja de $\beta_1 k$ unidades.

- Dado um valor de x, podemos obter o valor previsto para y, por $\hat{y} = \beta_0 + \beta_1 x$. Desta forma, o modelo linear permite fazer predição, para casos em que os dados do par (x, y) são desconhecidos.

Como fazer em SPSS? 3.10 (Regressão Linear Simples). Para visualizar como obter os parâmetros para o modelo de regressão linear simples, por favor clique em:

3.2.1 Inferência sobre o modelo de regressão

Além de estimação pontual para o valor de y, dado um valor concreto para x (ver nota 3.9), podemos fazer outro tipo de inferência sobre o modelo de regressão.

Embora não seja muito relevante obter intervalos de confiança para a ordenada na origem β_0, uma vez que esta constante não influência a variação de y em função da variação de x, existem outros parâmetros para os quais a inferência assume outra relevância em termos de intervalos de confiança ou testes de hipóteses. Entre estes estão o valor do declive β_1 e o valor da previsão para y, que consideraremos de seguida.

Inferência sobre β_1

Assume relevância saber se o declive da reta (e logo o aumento expectável para y quando x aumenta uma unidade) é um determinado valor β_{10}, ou seja se não é nulo. Temos então o teste de hipóteses para esse fim

$$\begin{cases} H_0: & \beta_1 = \beta_{10}, \\ H_1: & \beta_1 \neq \beta_{10} \end{cases}$$

que assume especial relevância no caso $\beta_{10} = 0$, em que se obtém

$$\begin{cases} H_0: & \beta_1 = 0, \\ H_1: & \beta_1 \neq 0. \end{cases} \tag{3.8}$$

Neste caso, testamos se a reta não é horizontal, ou dito de outra forma, se as variáveis estão associadas linearmente de forma significativa.

Nota 3.11 (Modelo significativo). Em particular, o modelo de regressão linear apenas tem significado se o teste de hipóteses (3.8) for significativo, isto é, se existir evidência estatística de que $\beta_1 \neq 0$. Caso contrário, não existe evidência de relação linear entre as duas variáveis.

Temos assim a estatística de teste

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_\varepsilon / \sqrt{S_{xx}}} \sim t_{n-2}$$

em que a variância dos desvios é dada por

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

e

$$S_{xx} = \sum x_i^2 - n\bar{x}^2.$$

No caso $\beta 10 = 0$, a anterior estatística simplifica-se para

$$T = \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\hat{\sigma}_\varepsilon} \sim t_{n-2}. \tag{3.9}$$

A estatística T anterior permite também obter intervalos de confiança para o valor de β_1, nomeadamente através de

$$IC_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-2,1-\alpha/2} \frac{\hat{\sigma}_\varepsilon}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{n-2,1-\alpha/2} \frac{\hat{\sigma}_\varepsilon}{\sqrt{S_{xx}}} \right].$$

Como fazer em SPSS? 3.12 (Intervalo de confiança para β_1). Para fazer o intervalo de confiança para β_1 (e os valores dos coeficientes de regressão) basta seguir os passos para obter o modelo linear, com as seguintes opções:

`Analisar`
⇝ `Regressão`
⇝ `Linear...`
⇝ `Estatísticas...`
• `Coeficientes de Regressão:` ☑ `Estimativas`
• `Coeficientes de Regressão:` ☑ `Intervalos de Confiança`
⇝ `Continuar`
⇝`Ok`

Inferência sobre a previsão y_0

Um outro aspeto que é importante é conseguir balizar uma previsão y_0, dado um valor x_0, ou, dito de outra forma, conseguir testar para um dado x_0 o valor respetivo $y_0 = y_{00}$ é plausível. Desta forma, devemos considerar o teste de hipóteses

$$\begin{cases} H_0: \ y_0 = y_{00}, \\ H_1: \ y_0 \neq y_{00} \end{cases}$$

que pode ser verificado usando a estatística de teste

$$T = \frac{\hat{y}_0 - y_{00}}{\hat{\sigma}_\varepsilon / \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

Da mesma forma, complementando a estimação pontual da forma

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

podemos obter estimação por intervalos em que o valor de y_0 pertence ao intervalo

$$y_0 \in \left[\hat{y}_0 - t_{n-2,1-\alpha/2} \frac{\hat{\sigma}_\varepsilon}{\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}, \hat{y}_0 + t_{n-2,1-\alpha/2} \frac{\hat{\sigma}_\varepsilon}{\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \right] \tag{3.10}$$

com $1 - \alpha$ de confiança.

Como fazer em SPSS? 3.13 (Intervalo de confiança para a predição). Para obter o valor da predição e o intervalo de confiança para os valores x na amostra, há que seguir os passos seguintes:

Analisar
⇝ Regressão
⇝ Linear...
⇝ Salvar...
• Valores preditos: ☑ Não padronizados
• Intervalos de predição: ☑ Média
• Intervalos de predição: ☑ Individual
⇝ Continuar
⇝Ok

obtendo-se assim 5 novas variáveis no ficheiro de dados. A primeira corresponde à previsão $\hat{y}$ para cada valor de x nos dados. As duas seguintes correspondem aos limites do intervalos de confiança para a média $\bar{y}$ para cada valor de x nos dados. As duas últimas colunas são os limites do intervalos de confiança (3.10) para y para cada valor de x nos dados.

Caso se pretende o intervalo de confiança da predição para um valor de x que não esteja nos dados, este pode ser acrescentado À coluna de x, mas sem colocar o correspondente valor de y, para não influenciar os dados.

Exemplo em Bioestatística 3.14. Vamos estudar a correlação e regressão linear entre a frequência cardíaca nos instantes 0 e 1, utilizando a base de dados usual `BaseDados_Notas.sav`. Temos assim o gráfico de dispersão na figura 3.3, que ilustra uma correlação potencialmente elevada.

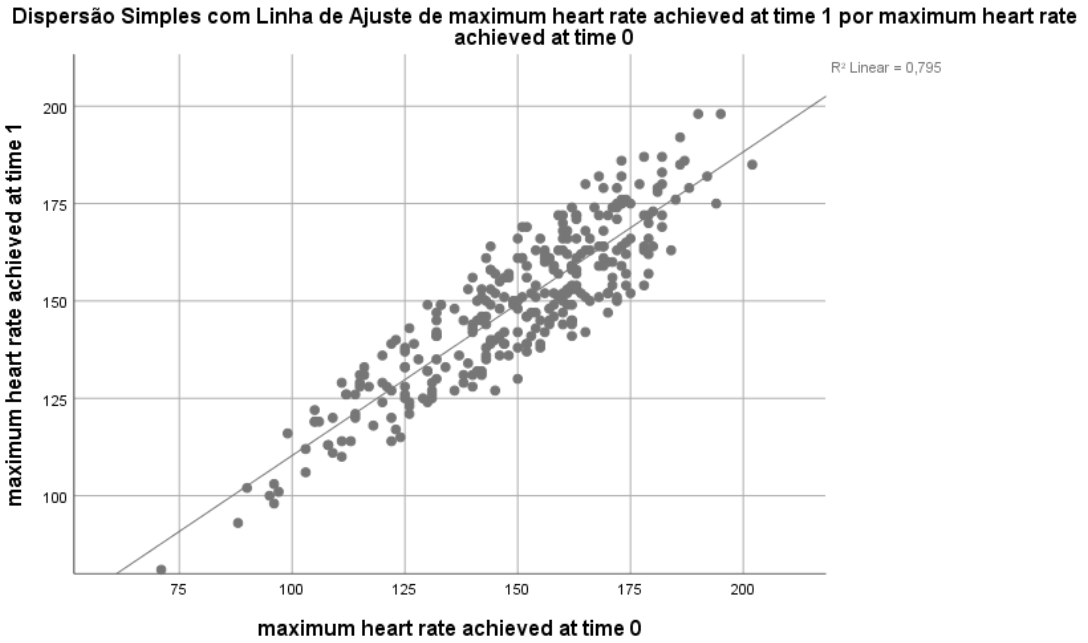


Figura 3.3: Gráfico de dispersão da frequência cardíaca nos instantes 0 e 1 e respetiva reta de regressão linear simples.

Isso é comprovado pelo valor do coeficiente de correlação $R = 0.891$ significativo ($p < 0.001$), indicando uma correlação positiva forte e logo, uma tendência

forte para que o aumento do valor de uma das variáveis implique o aumento do valor da outra.

Uma vez que a correlação é elevada e significativa, faz sentido obter a reta de regressão linear dada por

$$\hat{y} = -2.253 + 1.019x$$

em que x e y são as frequências cardíacas nos instantes 0 e 1, respetivamente. Da reta de regressão concluímos que em média, o aumento de uma unidade na frequência cardiaca no instante 0 implica em média um aumento de 1.019 no valor da frequência cardíaca no instante 1.

Temos também a predição para o valor da frequência cardiaca no instante 1 em $x_0 = 148$ (corresponde ao instante 0) dado por $\hat{y}_0 = 157.76$ e o intervalo de confiança a 95% para este valor $[137.29, 178.24]$.

3.3 Relação entre Correlação e Regressão Linear

O **coeficiente de correlação de Pearson** e o modelo de **regressão linear simples** estão altamente relacionados, como veremos nesta secção.

Na realidade, a partir das expressões para o coeficiente de correlação (3.1) e para o estimador para o declive da reta $\hat{\beta}_1$ em (3.9), podemos obter a relação

$$\hat{\beta}_1 = R\frac{s_y}{s_x}$$

entre os dois, em que s_x e s_y são os desvios padrão amostrais de X e Y, respetivamente. A relação anterior salienta o que foi mencionado na nota 3.2, uma vez que enquanto R não depende das magnitudes que as variações de x provocam em y (uma vez que o coeficiente é normalizado entre -1 e 1), o valor de $\hat{\beta}_1$ quantifica essas magnitudes de variações que a variável independente provoca na dependente.

Outro aspeto importante na relação entre estes dois conceitos já havia sido aflorado na nota 3.11, nomeadamente que o modelo linear apenas faz sentido se houver correlação linear entre as variáveis. De facto, consegue-se mostrar que a estatística de teste (3.3) para testar se $R \neq 0$ é equivalente à estatística de teste (3.9) para testar se $\beta_1 \neq 0$. Por outras palavras, a correlação é significativa se e só se o modelo linear é significativo, pelo que ambos os conceitos (embora nos dêem informação complementar) estão amplamente ligados.

Finalmente, como referido na nota 3.8, a qualidade do ajuste do modelo de regressão linear é aferido com base no coeficiente de correlação de Pearson R, nomeadamente através do cálculo da **taxa de explicação** R^2.

3.4 Regressão Linear Múltipla

A **regressão linear múltipla** é uma extensão da regressão linear simples para o caso em que existem várias **variáveis independentes** $X_1, X_2, \ldots, X_k$ a influenciar a **variável dependente** Y.

Assumindo então que a relação é linear, partimos do modelo de regressão linear múltiplo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

assumindo os erros ε com distribuição normal e média nula. De forma semelhante à regressão linear simples, pretende-se obter estimativas $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ para os parâmetros do modelo, por forma a se poder aferir (à semelhança da nota 3.9)

- que variáveis independentes influenciam a variável dependente Y;
- em que medida a variação das variáveis independentes significativas[3] influencia a variação da variável dependente;
- qual a estimativa prevista para y, dados os valores das variáveis independentes $x_1, x_2, \ldots, x_k$.

Nota 3.15. No contexto de regressão múltipla é comum denominar as variáveis independentes por **covariáveis**.

Nota 3.16 (Covariáveis relevantes). Existem dois casos em que algumas covariáveis devem ser retiradas do modelo de regressão múltipla, a saber:

- Caso algumas das covariáveis não sejam significativas[3];
- Caso duas ou mais covariáveis estejam fortemente correlacionadas entre si. Nesse caso, apenas 1 das vcoariáveis (correlacionadas) deve ser considerada no modelo, devendo ser descartadas as menos significativas, de acordo com o procedimento *stepwise* seguinte.

Nota 3.17. *Stepwise* Existem vários procedimentos para fazer a eliminação de covariáveis não significativas (não correlacionadas significativamente), mas vamos apenas referir o procedimento de **redução de variáveis por *stepwise***. Neste caso, o procedimento testa a significância de todas as covariáveis do modelo (ou, no caso de covariáveis correlacionadas, a significância destas). De seguida, retira-se a covariável com menor significância, que menos contribui para explicar a dependente (destas, no caso de covariáveis correlacionadas) e volta-se a testar o modelo. Repete-se o processo até um dos critérios de paragem seguintes:

- Todas as covariáveis no modelo são significativas;
- O modelo após retirar uma das covariáveis difere significativamente do modelo com essa covariável.

Como fazer em SPSS? 3.18 (Regressão Linear Múltipla). Para visualizar como obter os parâmetros para o modelo de regressão linear múltipla seguir os passos seguintes: `Analisar`
`⇝ Regressão`
`⇝ Linear...`
`⇝ Estatísticas...`
• `Coeficientes de Regressão:` ☑ `Estimativas`
• `Coeficientes de Regressão:` ☑ `Intervalos de Confiança`
☑ `Ajuste do Modelo`
`⇝ Gráficos...`

[3] Abuso de linguagem, uma vez que por variáveis independentes significativas se entende as variáveis independentes cujos respetivos coeficientes β_i do modelo são significativos (para a hipótese alternativa $\beta_i \neq 0$).

• Gráficos de resíduos padronizados: ☑ Histograma
⇝ Continuar
⇝Ok

Para mais detalhes, por favor clique em:

3.5 Outros Modelos de Regressão

Por vezes a relação entre duas (ou mais variáveis) não é linear, casos em que é necessário optar por outro modelo de regressão, por forma a determinar os parâmetros que melhor se adequam aos dados. Nesta secção veremos outros modelos, como por exemplo o polinomial e o exponencial. Nestes casos, o gráfico de dispersão poderá dar algum indício sobre a regressão a usar, uma vez que a forma da nuvem de pontos deverá corresponder a uma forma possível do modelo de regressão.

Na secção 3.6 abordaremos a regressão logística, que é aplicável a variáveis qualitativas dicotómicas.

Como fazer em SPSS? 3.19 (Outros modelos de Regressão). Para visualizar como obter os parâmetros para outros modelos de regressão, por favor clique em:

3.5.1 Regressão Polinomial

Na **Regressão Polinomial** assumimos que a relação entre as variáveis dependente Y e dependente X é da forma

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m + \varepsilon$$

ou seja, é um polinómio de grau m. Assim, o objetivo é obter os coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ que melhor se adaptam aos dados por forma a termos o modelo de regressão polinomial

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m.$$

No caso múltiplo, em que existem mais de uma variáveis independentes $X_1, X_2, \ldots, X_k$, temos o modelo

$$\hat{y} = \sum_{|\alpha|<m} \beta_\alpha x_1^{\alpha_1} \times x_2^{\alpha_2} \times \cdots \times x_k^{\alpha_k}$$

em que para o vetor $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ com componentes α_i não negativas, temos

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_k.$$

Exemplo em Bioestatística 3.20. Um exemplo de regressão polinomial é a relação entre a altura x e o peso y numa população de peso normal. Assim, como a expressão do Indicie de Massa Corporal (IMC) é dado

$$IMC = \frac{\text{peso}}{\text{altura}} = \frac{y}{x^2}$$

para uma população de peso normal com IMC a rondar 25, teremos

$$y = 25x^2.$$

Assim, o modelo de regressão polinomial de grau 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

para uma população de peso normal deveria obter β_2 perto de 25 e β_0 e β_1 próximos de zero.

3.5.2 Regressão Exponencial

O modelo exponencial assume que a relação entre as variáveis X e Y é da forma

$$\hat{y} = e^{\beta_0 + \beta_1 x} = ae^{bx},$$

com $a = e^{\beta_0}$ e $b = \beta_1$.

Nota 3.21 (Relação entre regressão linear e exponencial). É possível, através de uma transformação de variável, obter os coeficientes da regressão exponencial por regressão linear. Assim, aplicando o logaritmo à expressão anterior, temos

$$\underbrace{\ln(y)}_{\hat{y}} = \underbrace{\ln(a)}_{\beta_0} + \underbrace{b}_{\beta_1} x$$

ou seja, se aplicarmos regressão linear aos dados $\Big(x_i, \ln(y_i)\Big)$ e obtermos os respetivos coeficientes β_0, β_1 de regressão linear, podemos obter os coeficientes de regressão exponencial por

$$a = e^{\beta_0}, \quad b = \beta_1.$$

De notar no entanto, que neste caso o erro que se assume nos dados (x_i, y_i) não tem distribuição normal, devido à aplicação do logaritmo.

Para o caso de regressão exponencial múltipla com variáveis independentes $X_1, X_2, \ldots, X_k$, temos o modelo

$$\hat{y} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}.$$

3.6 Regressão Logística

A **Regressão Logística** é utilizada para modelar a probabilidade de um indivíduo/objeto pertencer a uma das duas categorias de uma variável qualitativa binária/dicotómica e tem alguma relação com o *Odds-ratio* abordado na secção 2.6.4.

Por norma, nos estudos de investigação clínica, as duas categorias da variável dependente são denominadas por sucesso e insucesso, sendo que o sucesso é o acontecimento de interesse (ter diabetes, ter cancro, etc.). Em termos das leis de probabilidades, assume-se que Y_i é uma variável com **distribuição binomial** de parâmetros

$$Y_i \sim B(p_i, n_i)$$

representando o número de sucessos em n_i tentativas, em que a probabilidade p_i de exito é desconhecida, mas a dimensão n_i das provas de Bernoulli é conhecida.

Para mais informação sobre a **distribuição binomial** clique no link seguinte:

A Regressão Logística permite identificar variáveis significativas para a classificação numa das duas categorias (sucesso/insucesso), ou seja, permite identificar fatores de risco. Estas variáveis independentes podem ser quantitativas ou qualitativas (também chamadas fatores). Por exemplo, para identificar fatores de risco para o desenvolvimento de diabetes poderiamos considerar a variável contínua "consumo,em gramas de açucares"e como fator poderiamos considerar "faz/não faz, exercício físico regular", ou "Presença/Ausência da doença na família".

Supõe-se então que os vários ensaios $i = 1, 2, \ldots, m$ são influenciados pelas covariáveis $X_1, X_2, \ldots, X_k$. A regressão logística parte do princípio que

$$\hat{p}_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}$$

o que no caso de apenas uma covariável X se reduz a

$$\hat{p}_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_i}}.$$

A ligação entre a variável dependente e as dependentes é uma equação diferente do modelo de regressão linear na sua base, mas equivale a fazer regressão linear para o logaritmo do respetivo ***Odds-Ratio***. De facto, consegue-se mostrar que o *Odds-Ratio* (razão de chances) pra cada prova é dado por

$$\frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}}$$

e logo temos para os valores **logit** respetivos

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

Assim, obtendo os parâmetros $\beta_0, \beta_1, \ldots, \beta_k$ por regressão linear múltipla para o valores anteriores, obtemos a regressão logística para as probabilidades p_i.

Como fazer em SPSS? 3.22 (Regressão Logística). Para visualizar como obter os parâmetros para o modelo de regressão logística seguir os passos seguintes:

`Analisar`
⇝ `Regressão`

⇝ Logística binária...
⇝ Selecionar a variável dependente (binária) e as covariáveis...
⇝ Categórico..(das anteriores selecionar as qualitativas)..
⇝ Opções...
• estatísticas e gráficos: ☑ grupos de classificação
☑ Qualidade do ajuste H-L
☑ I.C. para exp(β)
⇝ Continuar
⇝Ok

Para mais detalhes e exemplos, por favor clique em:

Capítulo 4

Introdução à Análise de Sobrevivência

A **Análise de Sobrevivência** (também usado o termo Sobrevida) é uma área da Estatística que adquiriu a sua designação devido à sua forte aplicação a estudos de sobrevivência de sujeitos a uma determinada doença ou causa de morte. De uma forma mais geral a Análise de Sobrevivência é a área que se debruça sobre o estudo do tempo até à ocorrência de determinado evento de interesse. Como já foi dito, a área de aplicação é geralmente a sobrevivência a determinada doença, mas existem atualmente outras aplicações, desde o tempo de vida de componentes elétricos (evento de interesse pode ser a falha do componente) até ao tempo de incumprimento de crédito bancário, por exemplo.

Desta forma, na Análise de Sobrevivência a variável de interesse é contínua, uma vez que é o tempo até à ocorrência do evento de interesse. O evento de interesse é geralmente denominado neste contexto por **falha**, sendo que a variável de interesse é portanto geralmente denominada por **tempo de sobrevivência** . Esta pode ser dada em minutos, horas, dias, meses ou anos e como exemplo, o tempo de sobrevivência pode ser, entre outros, o tempo:

- de sobrevivência a determinada doença;
- até à cura de determinada doença;
- da cura até à recidiva, em especial em estudos de cancro;
- até à alta hospitalar, para determinada doença;
- de vida de um determinado componente elétrico;
- até ao incumprimento após crédito bancário;

Exercício 4.1. Pesquise alguns exemplos específicos de contextos para estudos de Análise de Sobrevivência e partilhe-os nos fóruns de turma.

Assim, o objetivo principal da análise de sobrevivência é estimar grandezas associadas com o tempo de sobrevivência, muitas delas baseadas da **Função de Sobrevivência**, que definimos de seguida. A função de sobrevivência no instante t indica a probabilidade de sobrevivência até ao instante t, pelo que a partir dela se pode sumariar os seguintes como os objetivos principais da análise de sobrevivência:

- Estimar a função de sobrevivência a partir de dados recolhidos;
- Comparar funções de sobrevivência entre grupos (a uma menor função de sobrevivência corresponde uma maior mortalidade);
- Estudar a influência de covariáveis de interesse na função de sobrevivência (por exemplo, determinar se o aumento do valor de determinada covariável faz aumentar ou diminuir a função de sobrevivência no intervalo de estudo).

Para uma motivação sobre a aplicação de Análise de Sobrevivência em Bioestatística, clique no link seguinte:

Como veremos na secção seguinte a partir da função de sobrevivência consegue-se caracterizar a forma como a população sobrevive à doença ao longo do tempo.

4.1 Função de Sobrevivência

A **Função de Sobrevivência** é geralmente denotada por $S(t)$, em que t é o tempo. A função de sobrevivência no instante t indica a probabilidade da falha ainda não ter ocorrido, ou seja, a percentagem de sujeitos na população que ainda sobrevivem no instante t. Assim, sendo T a variável aleatória tempo de sobrevivência[1], temos a definição de função de sobrevivência dada por

$$S(t) = P(T \geq t).$$

É fácil verificar que a função de sobrevivência define a distribuição de probabilidade para a variável aleatória tempo de sobrevivência T. Na realidade, a função distribuição de probabilidade de T é definida por

$$F(t) = P(T \leq t),$$

logo esta é dada por

$$F(t) = 1 - S(t), \tag{4.1}$$

de onde sai que a função densidade de probabilidade associada é dada por

$$f(t) = -S'(t). \tag{4.2}$$

A função de sobrevivência permite também tirar conclusões sobre a sobrevivência a determinada doença ao longo do tempo. Permite, por exemplo, determinar algumas medidas de tendência central para o tempo de vida, como o **tempo de vida mediano** designado por t_{mediano}, que é definido pelo instante de tempo em que a função de sobrevivência é igual a 50%, ou seja,

$$S\left(t_{\text{mediano}}\right) = 0.5. \tag{4.3}$$

[1]Como habitualmente, consideramos T maiúsculo para designar a variável aleatória e t minúsculo para designar um determinado valor dessa variável.

Temos também o **tempo de vida médio** denotado por t_m dado a partir da função de sobrevivência por

$$t_m = \int_0^\infty S(t)\, dt. \tag{4.4}$$

A função de sobrevivência permite também definir grandezas como a vida média residual a partir do instante t, denotada por $\mathrm{vmr(t)}$, que é definida como o tempo de vida médio restante para os sujeitos com idade t e é definida por

$$\mathrm{vmr(t)} = \frac{1}{\mathrm{S(t)}} \int_{\mathrm{t}}^\infty \mathrm{S(u)du}. \tag{4.5}$$

consegue-se verificar que verificar que $\mathrm{vmr(0)} = \mathrm{t_m}$. A função de sobrevivência está também fortemente relacionada com a taxa de falha, que explanaremos de seguida.

4.2 Taxa de Falha

A taxa de variação da falha num intervalo de tempo é a razão entre a probabilidade de uma falha ocorrer nesse intervalo de tempo e a amplitude do intervalo, condicionada ao facto da falha não ter acontecido antes. Considerando um instante t, a **taxa de falha** $\lambda(t)$ nesse instante é obtida fazendo o limite do intervalo anterior em torno de t para uma amplitude nula.

A taxa de falha é um indicador da variação instantânea da sobrevivência. Instantes com taxa de falhas maiores indicam maior decrescimento da função de sobrevivência do que instantes com taxas de falha menores. Da mesma forma, se a taxa de falha for crescente, isto indica que à medida que o tempo avança a probabilidade instantânea de morrer aumenta, o que traduz geralmente o efeito do envelhecimento. Além disso,consegue-se mostrar que

$$\lambda(t) = -\frac{S'(t)}{S(t)}. \tag{4.6}$$

Consegue-se também mostrar que

$$S(t) = e^{-\int_0^t \lambda(u)\, du} \tag{4.7}$$

e que

$$\lambda(t) = -(\ln S(t))'. \tag{4.8}$$

Pelas expressões anteriores, em especial pela expressão (4.7), vemos que a informação da taxa de falha chega muito suavizada à função de sobrevivência, por via da aplicação do integral, que tende a suavizar variações pontuais bruscas. Assim, a taxa de falha tende a dar maior informação que a função de sobrevivência, uma vez que funções de sobrevivência muito próximas podem ser originadas por taxas de falha muito distintas.

Uma outra função que aparece muitas vezes na literatura é a **taxa de falha acumulada**, que é definida por

$$\Lambda(t) = \int_0^t \lambda(u)\, du. \tag{4.9}$$

Esta função é mais fácil de ser estimada, pelo que, embora não tenha um significado direto, permite obter estimativa da taxa de falha a partir dela. É também fácil de perceber que a partir de (4.7) se obtem

$$S(t) = e^{-\Lambda(t)}. \tag{4.10}$$

ou em alternativa

$$\Lambda(t) = -\ln S(t). \tag{4.11}$$

Notamos que o conhecimento da função de sobrevivência S permite obter a taxa de falha λ através de (4.8), a função densidade de probabilidade f através de (4.2), a função densidade de probabilidade (4.1), a taxa de falha acumulada a partir de (4.11), assim como os tempo médio (4.4) e mediano de vida (4.3) e a vida média residual (4.5). Assim, nos próximos capítulos dedicaremo-nos a formas de estimar a função de sobrevivência. Antes, no entanto, vamo-nos dedicar a alguns aspetos da montagem da recolha de dados e modelos, que são importantes no contexto de análise de sobrevivência.

4.3 Tempo de Sobrevivência

Para todas as grandezas e funções que temos considerado neste texto, é importante definir o tempo de sobrevivência de forma inequívoca para cada sujeito, de forma a que o tempo de sobrevivência tenha o significado de interesse para o estudo a ser feito. Neste texto consideraremos apenas um evento de interesse, isto é, a falha originada por uma causa de interesse. Por exemplo, para estudar a morte a determinado tipo de cancro, não se pode considerar uma falha a morte devido a outro factor como por exemplo um acidente de trabalho.

Além do instante de falha, para a definição do tempo de sobrevivência é de crucial importância a definição do instante de início de contagem desse tempo. Ao contrário dos estudos estatísticos usuais, geralmente o instante de início de contagem do tempo de sobrevivência não coincide com o início da recolha dos dados. Por exemplo, suponhamos que estudamos a sobrevivência a determinado cancro, utilizando os dados dos pacientes que se encontram inscritos em determinado instituto entre as datas X e Y. Supomos agora que um dos paciente falece (ou seja, ocorre uma falha) na data Z, que se encontra no período do estudo entre X e Y. Qual será então o tempo de sobrevivência a considerar? Ingenuamente o leitor poderia considerar que seria o tempo que decorre entre X (o início do estudo) e Z (a data de falha). Na realidade este tempo não tem significado estatístico para o estudo, pois a data de início de estudo X é pouco relevante para a progressão da doença no sujeito em causa. Assim, a data de início a considerar deverá, neste caso, ser a data de diagnóstico (ou, em alternativa, a data do início do tratamento), uma vez que o que tem relevância para o estudo é o tempo de sobrevivência após desta data. Dessa forma, para todos os sujeitos no estudo, o tempo de sobrevivência seria o tempo entre a data de diagnóstico e a data de morte (falha), que é de facto o tempo com interesse para o estudo de sobrevivência.

O exemplo anterior levanta outras questões que provavelmente já assolam a mente do leitor:

(a) O que fazer com indivíduos que no período de estudo não falecem, ou seja, não ocorre a falha?

(b) Como entrar no estudo com fatores que influenciam o tempo de sobrevivência?

Começamos pela questão (b), porque na realidade não a vamos responder para já. Apenas a mencionaremos na secção 4.5, sendo que a aprofundaremos mais adiante no texto, nomeadamente quando iniciarmos o estudo de estimativas para a função de sobrevivência. Num estudo de sobrevivência a cancro, obviamente que o estadio do cancro na data de diagnóstico, o género e a idade do sujeito (entre outros fatores) podem influenciar o tempo de sobrevivência. Dessa forma, é necessário considerar estes fatores como covariáveis no modelo, ou seja, como variáveis independentes que poderão influenciar a variável (dependente) de interesse: o tempo de sobrevivência. Trataremos desse aspeto a partir do próximo capítulo.

Quanto à questão (a), o leitor provavelmente poderia sugerir (mais uma vez de forma ingénua) que os dados relativos a sujeitos em que a falha não ocorreu poderiam simplesmente ser eliminados do estudo. No entanto, é fácil de ver que se perde informação se se proceder dessa forma. Se sabemos para determinado sujeito que a falha não ocorreu no período de observação, sabemos que o seu tempo de sobrevivência é superior a determinado tempo t, pelo que essa informação não deve ser desprezada no modelo. Esta é outra grande diferença da Análise de Sobrevivência em relação a outros estudos estatísticos: permite utilizar dados censurados, isto é, dados em que a falha não ocorre no período de observação.

4.4 Censura e Truncamento

Entende-se por **censura** a observação parcial da resposta. Sempre que para determinado sujeito não ocorre a falha até ao tempo t de observação, dizemos que em t ocorreu uma censura. Em presença de censura, o que sabemos é que o tempo de sobrevivência é superior a t, tratando-se neste caso de uma **censura à direita**. A censura à direita pode ter várias origens, como por exemplo, a saída de um sujeito do estudo por ter mudado a região de residência, a cura do paciente ou a morte devida a outra causa, que não a doença em estudo. Existem contextos no entanto em que temos **censura à esquerda**, isto é, em que quando a censura ocorre sabemos que o tempo de sobrevivência é inferior ao censurado. Por exemplo, o caso em que se estuda em determinada localidade em que idade as crianças começam a caminhar. Podem existir crianças que já sabem caminhar e em que não se saiba a data precisa em que isso aconteceu, pelo nesse caso a censura é à esquerda.

Assim, cada i-ésimo sujeito é caracterizado pelo tempo t_i e por uma variável dicotómica δ_i que assume os valores

$$\delta_i = \begin{cases} 1, & \text{em } t_i \text{ ocorre uma falha} \\ 0, & \text{em } t_i \text{ ocorre uma censura} \end{cases} \tag{4.12}$$

Desta forma, o conjunto de dados base em análise de sobrevivência são pares $(t_i, \delta_i), \quad i = 1, 2, \ldots, n$, que formam o conjunto de n pacientes observados, em que em alguns ocorreram falhas e noutros censuras.

Um outro fenómeno que é geralmente confundido com a censura é o **truncamento** , embora seja um conceito totalmente diferente. O truncamento é o fenómeno que impede indivíduos de fazerem parte da amostra e logo do estudo. Por exemplo, se para estudar a taxa de sobrevivência a determinada doença em determinada localidade eu recorrer aos dados do recenseamento eleitoral na junta de freguesia, apenas considero os sujeitos maiores de idade. Dessa forma, existe uma truncatura à esquerda, uma vez que os menores de idade são excluídos do estudo.

4.5 Covariáveis

Juntamente com o par (T_i, δ_i) definido anteriormente, podem (e devem) ser recolhidos para cada sujeito os valores de outras covariáveis de interesse, que se espera poder influenciar o tempo de sobrevivência. Ao incluir covariáveis, obtemos diferenças na probabilidade de sobrevivência consoante o valor da covariável e assim podemos testar se existem diferenças significativas entre valores diferentes e grupos de interesse. No contexto de estudos de sobrevivência a doenças, salientamos as seguintes covariáveis, que têm geralmente um efeito significativo e que portanto devem ser consideradas:

- **Coorte**: Em estudos de sobrevivência a doenças, a função de sobrevivência é fortemente afetada pelo horizonte temporal em que é feito o estudo. É evidente que a taxa de sobrevivência ao cancro tem vindo a aumentar ao longo dos anos, devido aos avanços da medicina. Assim, caso tenhamos dados transversais a várias épocas, é importante introduzir no modelo uma variável coorte que fixa o periodo a que relata. Por exemplo, se tivermos dados de sobrevivência a cancro recolhidos em períodos de observação entre 1981 e 1984, depois entre 1990 e 1992 e entre 1996 e 1999, será indicado considerar no modelo uma variável coorte com valores 0, 1 e 2, consoante o período em que o sujeito foi observado, uma vez que é esperado que a função de sobrevivência (a probabilidade de sobreviver mais x tempo) seja diferente consoante o coorte considerado. Os Coortes podem ser décadas, séculos ou períodos definidos no tempo e permitem ao modelo dar estimativas diferentes consoante o coorte considerado.

- **Grupo de Controlo vs. Grupo de Estudo:** A análise de sobrevivência tem aplicações para mostrar que determinado(s) tratamento(s) inovador(es) tem melhores resultados que o tratamento convencional. Assim, no desenho do estudo é importante considerar e recolher dados de sujeitos de vários grupos diferentes e posteriormente comparáveis. É usual considerar um grupo de controlo (sujeitos sem tratamento tomando um placebo, ou com o tratamento convencional) e o grupo de estudo (ou grupo experimental) (sujeito ao tratamento A, B e/ou C). Assim, a variável Grupo deve ser considerada no modelo com valores 0 para o grupo de controlo (tipicamente) e 1,2,3,... para os grupos sujeitos ao tratamento A, B, C, etc...

- **Outras covariáveis:** Consoante a doença e o objetivo do estudo importa recolher dados de outros fatores que possam influenciar (ou condicionar) o tempo de sobrevivência. Por exemplo, em geral, a idade do paciente na data de diagnóstico é importante para o modelo. A idade do sujeito influencia a progressão da maioria das doenças. O género do sujeito é uma variável de interesse em muitas situações, uma vez que influencia não só a taxa de incidência, mas também a progressão da doença. Em estudos de cancro, os estadio do cancro (tumor) no momento do diagnóstico influencia fortemente a função de sobrevivência, pelo que o estadio deve ser uma variável considerada no modelo.

Estamos agora em condições de iniciarmos o estudo de métodos para a estimativa da função de sobrevivência. Dado o contexto breve deste curso, abordaremos apenas estimadores não-paramétricos, assunto ao qual nos dedicaremos nas próximas linhas.

4.6 Estimadores Não paramétricos

Nas situações práticas, em contexto de investigação médica e outros, a lei (função) de sobrevivência não é conhecida, pelo que se torna necessário recorrer a estimativas, que são obtidas através de estimadores desenvolvidos, com boas propriedades de estimação. Os estimadores não paramétricos não assumem qualquer modelo de distribuição para a associação entre as covariáveis de interesse nem para a taxa de falha acumulada. Assim, estes são aplicáveis na maioria dos casos em que não existe informação para o modelo próprio da função de sobrevivência.

De notar que os estimadores têm de ter a capacidade de lidar com censuras, pois de outra forma não seriam aplicáveis ao contexto de Análise de Sobrevivência. No entanto, e como motivação, vamos começar por considerar o caso sem censuras, isto é, o caso em que para todos os sujeitos ocorreu uma falha no período de observação.

Exemplo 4.2. Suponhamos que recolhemos os dados de falhas ocorridas nos vários instantes de tempo tabelados abaixo, para o total de 168 sujeitos:

t_i	Nº de falhas	Nº de sobreviventes
0	-	168
1	5	163
2	7	156
3	10	146
4	25	121
5	52	69
6	35	34
7	19	15
8	15	0

Neste caso sem censuras, temos então como possibilidade para estimativa

para a função de sobrevivência

$$\hat{S}(t_i^+) = \frac{\text{nº de sujeitos vivos no instante } t_i}{\text{nº de sujeitos inicialmente no estudo}} \tag{4.13}$$

em que a notação t^+ indica que a estimativa é válida entre o instante t_i e o seguinte t_{i+1}. Por exemplo, para $t = 3$ teríamos a estimativa

$$\hat{S}(3) = \frac{136}{168} = 0.86905$$

ou seja, cerca de 86.9% dos sujeitos sobrevive ao instante $t = 3$. Da mesma forma, se pretendessemos estimar a taxa de falha, teríamos, por exemplo a estimativa

$$\hat{\lambda}([t_i, t_{i+1}[) = \frac{\text{nº de falhas entre } t_i \text{ e } t_{i+1}}{\text{nº de sujeitos vivos no instante } t_i}$$

pelo que para o intervalo $[3, 4[$ teríamos

$$\hat{\lambda}([3, 4[) = \frac{25}{146} = 0.17123.$$

Completando a tabela, teríamos as estimativas

t_i	Nº de falhas	Nº de sobreviventes	$\hat{S}(t_i^+)$	$\hat{\lambda}([t_i, t_{i+1}[)$
0	—	168	1	0.02976
1	5	163	0.97024	0.04294
2	7	156	0.92857	0.06410
3	10	146	0.86905	0.17123
4	25	121	0.72024	0.42975
5	52	69	0.41071	0.50725
6	35	34	0.20238	0.55882
7	19	15	0.08929	1
8	15	0	0	—

De notar que quando não há censuras, a função de sobrevivência é nula após o último instante medido, uma vez que para todos os sujeitos existe uma falha no período observado. Mais ainda, a taxa de falha é 1 no último intervalo medido, uma vez que todos os que ainda sobreviviam acabam por ter uma falha nesse período, devido à razão anterior. No entanto, este é problema no contexto geral de análise de sobrevivência, uma vez que teremos de ter estimativas que possam permitir a sobrevivência de alguns sujeitos, o que não acontece no caso das censuras. Assim fica explícito que a eliminação das censuras para um estudo de Análise de sobrevivência desvirtua totalmente os resultados.

Assim, iniciamos o estudo de técnicas que permitam considerar censuras, começando pelo estimador de Kaplan-Meier para a função de sobrevivência

4.6.1 Estimador de Kaplan-Meier

O **estimador de Kaplan-Meier** foi sugerido por Kaplan e Meier em 1958 [4]. Na realidade, o estimador de Kaplan-Meier é a estimativa de máxima verosimilhança da função de sobrevivência, mas pode ser visto como uma adaptação

da estimativa empírica considerada em (4.13), sendo que para a contabilidade do número de sujeitos vivos no instante t_i são contabilizados também os sujeitos que tiveram uma censura num instante superior. Na realidade, o estimador de Kaplan-Meier é uma função em escada em que o início dos degraus da escada coincídem com os instantes onde ocorrem falhas. Assim, o primeiro passo para a estimativa de Kaplan-Meier é dividir o intervalo em n sub-intervalos, com limites nos pontos de falha. A partir daí, começamos por estimar a função de sobrevivência no primeiro instante de falha t_1 como a razão entre o número de sujeitos em risco no instante t_1 sobre o número se sujeitos inicialmente no estudo. No segundo instante de falha t_2 consideramos a relação

$$P(T \geq t_2) = P(T \geq t_1, T \geq t_2) = P(T \geq t_1).P(T \geq t_2 | t \geq 1)$$

ou seja, a probabilidade do tempo de sobrevivência ser superior ou igual a t_2 é igual à do tempo de sobrevivência ser superior ou igual a t_1 vezes a probabilidade do tempo de sobrevivência ser superior ou igual a t_2 condicionada ao tempo ser superior a t_1. Dito em linguagem corrente, para a probabilidade de um sujeito sobreviver até t_2 é igual à probabilidade de sobreviver até t_1 vezes a probabilidade de sobreviver até t_2 sabendo que sobreviveu até t_1. Este processo é depois repetido sucessivamente para se ter as estimativas sucessivas nos vários instantes de falha.

Vejamos um exemplo, antes de explicitarmos a expressão geral.

Exemplo 4.3. Consideramos os dados recolhidos e ordenados de tempos de falha seguintes, em que o sinal + indica uma censura em vez de uma falha:

$$1, 3+, 4, 5, 5+, 6+, 7+, 8, 8, 9+, 10, 11+, 13+ .$$

Começamos por estabelecer os extremos dos intervalos t_i a considerar para a estimativa de Kaplan-Meier, ou seja,

$$t_0 = 0 \text{ (instante inicial)}, \; t_1 = 1, \; t_2 = 4, \; t_3 = 5, \; t_4 = 8, \; t_5 = 10.$$

Em seguida, para cada intervalo $[t_i, t_{i+1}[$ é preciso definir o número n_i de sujeitos em risco (ou seja, vivos) e o número f_i de falhas no intervalo, dando origem à tabela seguinte:

t_i	Intervalo	n_i	f_i
0	[0,1[	13	0
1	[1,4[	13	1
4	[4,5[	11	1
5	[5,8[	10	1
8	[8,10[	6	2
10	[10, 13[	3	1

A probabilidade de sobrevivência inicial é 1, logo consideramos que

$$S(0^+) = P(T \geq 0) = 1,$$

ou seja, a probabilidade de sobreviver ao instante 0 é 1.

De seguida, temos que no instante $t = 1$ ocorre 1 falha em 13 sujeitos em risco. Assim, a estimativa para o sujeito não sobreviver ao instante $t = 1$ é $1/13$

(número de falhas sobre número de sujeitos em risco), pelo que a estimativa para sobreviver ao instante 1 é dada por

$$\hat{S}(1^+) = (1 - 1/13) = 0.92308.$$

No instante seguinte $t = 4$, ocorre 1 falha em 11 sujeitos em risco. A estimativa para o sujeito não sobreviver até ao instante $t = 4$, condicionada ao sujeito estar vivo no instante anterior $t = 1$ é então dada por

$$P(T \leq 4|T \geq 1) = 1/11 = 0.090909,$$

pelo que

$$P(T \geq 4|T \geq 1) = 1 - 1/11 = 0.90909.$$

Assim, a estimativa da função de sobrevivência é dada por

$$\hat{S}(4^+) = \underbrace{P(T \geq 1)}_{\hat{S}(1^+)}.P(T \geq 4|T \geq 1) = 0.92308 \times 0.90909 = 0.83916.$$

De forma semelhante, a estimativa para a probabilidade de sobrevivência ao instante $t = 5$, condicionada ao sujeito estar vivo no instante $t = 4$ é

$$P(T \geq 5|T \geq 4) = 1 - 1/10 = 0.9$$

pelo que a estimativa para o sujeito estar vivo no instante $t = 5$ é dada por

$$\hat{S}(5^+) = \hat{S}(4^+).P(T \geq 5|T \geq 4) = 0.83916 \times 0.9 = 0.75524.$$

Repetindo o processo, temos a estimativa de Kaplan-Meier para a função de sobrevivência dada na tabela seguinte:

Tabela 4.1: Estimativa de Kaplan Meier para a função de sobrevivência do exemplo 4.3.

t_i	Intervalo	n_i	f_i	$\hat{S}(t_i^+)$
0	[0,1[	13	0	1
1	[1,4[	13	1	0.92308
4	[4,5[	11	1	0.83916
5	[5,8[	10	1	0.75524
8	[8,10[	6	2	0.50350
10	[10, 13[	3	1	0.33566

Vamos então definir formalmente o **estimador de Kaplan-Meier** seguinte, que em [4] foi demonstrado ser o estimador de máxima verosimilhança da função de sobrevivência S. Sejam para $i = 1, 2, \ldots, n$

- t_i os instantes em que foram observadas falhas na amostra e
- n_i os número de sujeitos em risco no instante t_i;
- f_i o número de falhas ocorridas no intervalo $[t_i, t_{i+1}]$;

Então, o **estimador de Kaplan-Meier** para a função de sobrevivência é dado pela função em escada

$$\hat{S}(t) = \prod_{i:t \geq t_i} \left(1 - \frac{f_i}{n_i}\right), \tag{4.14}$$

em que $\prod$ significa o produtório dos vários valores em i.

Nota 4.4 (Normalidade e Consistência). Como nota, referimos que é possível mostrar (ver [5]) que a distribuição do estimador de Kaplan-Meier converge assintoticamente para uma distribuição normal e que este é fracamente consistente, ou seja, para qualquer $\varepsilon > 0$ quando a dimensão n da amostra tende para infinito, temos

$$\lim_{n \to \infty} P(|\hat{(S)} - S| < \varepsilon) = 1.$$

Felizmente, hoje é possível utilizar recursos computacionais para obter este tipo de estimativa. Ainda que a quantidade de dados utilizada no exemplo 4.3 para fins ilustrativos é relativamente pequena, em casos de dados recolhidos para estudos de clínicos a quantidade de dados a tratar torna inviável o recurso a cálculo manual para estimar a função de sobrevivência. Isto é ilustrado no exemplo 4.5.

Exemplo em Bioestatística 4.5. Vamos então confirmar os valores obtidos da estimativa de Kaplan-Meier na tabela para o exemplo 4.3 com recurso ao SPSS. Além disso vamos calcular o tempo de vida médio e mediano e traçar o gráfico da função de sobrevivência e da taxa de risco acumulado.

Para o efeito, escolhemos

`⇝ Analisar`
`⇝ Sobrevivência`
`⇝ Kaplan-Meier`
- `Escolher variável tempo`
- `Escolher variável censura (Status)`

`⇝ Definir evento...`
- `Indicar o valor único correspondente à falha (geralmente 1)`

`⇝ Continuar`
`⇝ Opções`
- `Estatísticas ☑ Tabela de Sobrevivência`
- `Estatísticas ☑ Sobrevivência de média e mediana`
- `Gráficos ☑ Sobrevivência`
- `Gráficos ☑ Risco`

`⇝ Continuar`
`⇝ Continuar`
`⇝ Ok`

Temos então o tempo médio de vida $t_m = 8.888$ e o tempo mediano de vida $t_{\text{mediano}} = 10$, com os gráficos de sobrevivência e risco acumulado da figura 4.1.

De notar que como em qualquer estimador, é importante ter noção da sua variância, para assim estimar um intervalo de confiança adequado. Como o estimador tem uma distribuição assintótica normal (ver nota 4.4), esperamos que o

intervalo de confiança (a $(1-\alpha)$ de confiança) seja dado em cada instante t por

$$IC_{1-\alpha}(t) = \left[\hat{S}(t) + z_{\alpha/2}\sqrt{\text{Var}\left(\hat{S}(t)\right)}, \hat{S}(t) + z_{1-\alpha/2}\sqrt{\text{Var}\left(\hat{S}(t)\right)}\right],$$

em que z_α é o percentil α da distribuição normal. Assim, falta obter a variância do estimador.

Teorema 4.6 (Fórmula de Greenwood)**.**
A variância assintótica do estimador de máxima verosimilhança de Kaplan-Meier é dado pela fórmula de Greenwood

$$\widehat{\text{Var}}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{i:t\geq t_i} \frac{f_i}{n_i(ni - f_i)}. \tag{4.15}$$

Demonstração. A prova vem das propriedades do estimador de máxima verosimilhança e pode ser detalhada em [6].

Exemplo em Bioestatística 4.7. Determine o Intervalo de Confiança a 95% segundo a fórmula de Greenwood para a função de sobrevivência do exemplo 4.3.

Resposta.

O SPSS não determina automaticamente o intervalo de confiança, apenas calcula o erro padrão $\sqrt{\text{Var}\left(\hat{S}(t)\right)}$. A partir deste, teremos de construir o intervalo de confiança a 95%, que para $z_{0.975} = -z_{0.025} = 1.96$, é dado na tabela seguinte:

t_i	n_i	f_i	$\hat{S}(t_i^+)$	erro padrão	IC
1	13	1	0.923	0.0739	[0.778,1.000]
4	11	1	0.839	0.1045	[0.634,1.000]
5	10	1	0.755	0.1232	[0.514,0.997]
8	6	2	0.503	0.1669	[0.176,0.831]
10	3	1	0.336	0.1765	[0.000,0.682]

De notar que pela fórmula de Greenwood, o intervalo de confiança pode ter limites superiores a 1 ou inferiores a zero, o que no contexto de uma função de sobrevivência não faz sentido[2].

[2]Por norma, os valores superiores a 1 são apresentados como 1 e os inferiores a 0 como 0.

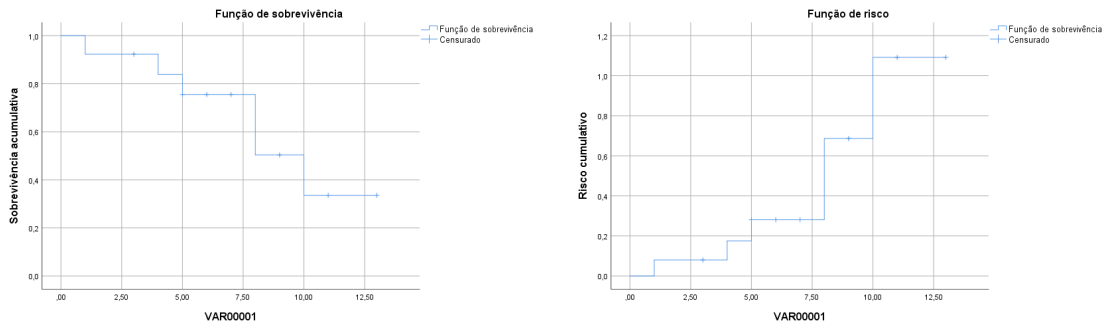


Figura 4.1: Gráficos da função de sobrevivência e risco acumulado referentes ao exemplo 4.5.

Desta forma, foi sugerida em 1980 uma transformação logarítmica da função de sobrevivência em [6] por Kalbfleish e Prentice, que não abordaremos neste texto.

4.6.2 Comparação entre grupos

Uma das aplicações principais da análise de sobrevivência é a comparação entre grupos, isto é, determinar se existem diferenças estatisticamente significativas entre as funções de sobrevivência de dois ou mais grupos distintos da população. Os grupos podem ser várias regiões do país, várias faixas etárias, grupos de controlo e de tratamento, entre outros. Nesta secção veremos como podemos aplicar o uso de R para obter estimativas para diferentes grupos e como podemos testar se existem diferenças estatisticamente significativas.

Comecemos por considerar o seguinte exemplo, que utilizaremos para ilustrar as técnicas utilizadas nas seguintes linhas.

Exemplo 4.8. Sujeitos com determinada doença foram separados aleatoriamente em 3 grupos distintos. Ao grupo de controlo foi administrado um placebo. Ao grupo do tratamento 1, foi administrado um medicamento em duas doses diárias. Ao grupo de tratamento 2 foi administrado a mesma dosagem diária de medicamento mas em 3 doses diárias. Foi registado para cada um dos sujeitos o tempo em dias até obterem alta hospitalar. Consideramos os dados recolhidos e ordenados de tempos de falha seguintes, em que o sinal + indica uma censura em vez de uma falha:

Controlo	Tratamento 1	Tratamento 2
5, 8, 8+, 9, 10, 10+, 14	4, 5, 5+, 6, 8, 9, 9+, 10+, 11, 12+	5+, 5, 6, 6+, 7, 8, 8+, 10+

Esta base de dados é cedida no ficheiro `Sobrevivencia48.sav` na plataforma.

Exercício SPSS 4.9. Determine a estimativa de Kaplan-Meier da função de sobrevivência para cada um dos três grupos do exemplo 4.8.

Resposta.

Repetindo o procedimento do exemplo 4.5 e acrescentando o Factor Tratamento, temos o gráfico na figura 4.2.

A comparação na figura 4.2 ilustra-nos que aparentemente não existe grande diferença entre os vários grupos. No entanto, é necessário fazer um teste estatístico adequado para a verificação desta hipótese, o que trataremos de seguida.

Teste Logrank

O **teste de logrank** permite testar a hipótese nula de comparação de funções de sobrevivência entre p grupos dada por

$$\begin{cases} S_1(t) = S_2(t) = \cdots = S_p(t), & \text{no intervalo de observação,} \\ S_i(t) \neq S_j(t) & \text{para algum par } (i,j), \text{ no intervalo de observação,} \end{cases}$$

Este teste é talvez o mais usado no contexto de análise de sobrevivência, uma vez que permite testar igualmente em todo o intervalo, em vez de fazer a comparação assintótica para t grande. No entanto, como hipótese para a aplicação do

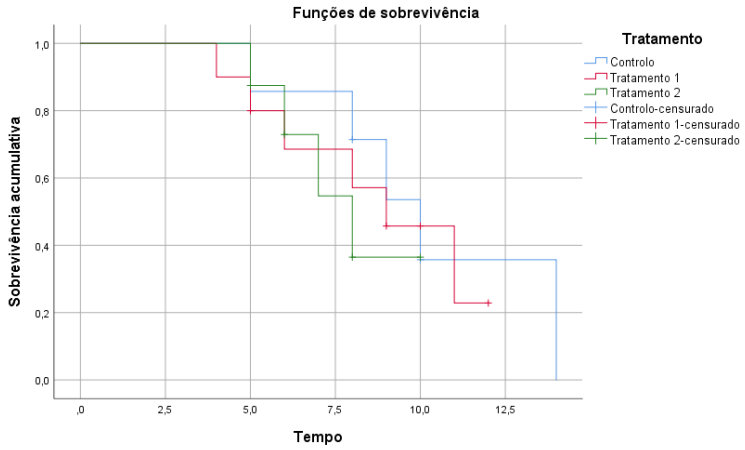


Figura 4.2: Estimativa de Kaplan Meier para a função de sobrevivência em cada grupo do exemplo 4.8.

teste supõe-se que a razão entre as taxa de falha em cada grupo é sensivelmente constante ao longo do tempo.

O teste de logrank parte da listagem ordenada de todos os tempos de falha em todos os grupos dada por

$$t_1 < t_2 < \cdots < t_m.$$

Em cada instante t_k, podemos construir a tabela seguinte, em que se assinalam por grupo i e no total o número de falhas f_{ik}, f_k e o número de não falhas dados por $n_{ik} - f_{ik}, n_k - f_k$, respetivamente, no instante t_k, em que n_{ik} é o número de sobreviventes no grupo i e n_k é o número total de sobreviventes no instante t_k:

	Grupo 1	Grupo 2	...	Grupo p	Total
Falha	f_{1k}	f_{2k}	...	f_{pk}	f_k
Não Falha	$n_{1k} - f_{1k}$	$n_{2k} - f_{2k}$	...	$n_{pk} - f_{pk}$	n_k-f_k
Total	n_{1k}	n_{2k}	...	n_{pk}	n_k

A partir da tabela de contingência anterior, podemos estabelecer o valor esperado para o número de falhas no grupo i no instante j no caso da hipótese nula (de funções de sobrevivência iguais), que seria dado por $e_{ik} = \frac{f_k n_{ik}}{n_k}$. Notando agora que a distribuição conjunta da variável $(f_{2k}, f_{3k}, \ldots, f_{pk})$ é uma hipergeométrica multivariada, consegue-se obter a estatística de teste

$$T = e' V^{-1} e \sim \chi^2_{p-1}$$

que tem distribuição qui-quadrado com $p-1$ graus de liberdade e em que V é a matriz de covariância das variáveis $(f_{2k}, f_{3k}, \ldots, f_{pk})$ e o vetor e é simplesmente a soma em todos os instantes k da diferença entre o valor observado e o valor esperado para o número de falhas dado por

$$e = \left(\sum_{k=1}^{m} f2k - e_{2k}, \sum_{k=1}^{m} f3k - e_{3k}, \ldots, \sum_{k=1}^{m} fpk - e_{pk} \right).$$

Exercício SPSS 4.10. Determine se existe diferença significativas entre as funções de sobrevivência dos três grupos do exemplo 4.8.

Resposta.

Para aplicar o teste de logrank, acrescentamos aos comandos do exemplo 4.5 a opção do teste logrank. Assim, temos com as opções:

```
   ⇝ Analisar
⇝ Sobrevivência
⇝ Kaplan-Meier
• Escolher variável tempo
• Escolher variável censura (Status)
⇝ Definir evento...
• Indicar o valor único correspondente à falha (geralmente 1)
⇝ Continuar
• Escolher variável de grupo (Factor)
⇝ Compara factor...
• Estatísticas de teste ☑ Posição log
⇝ Continuar
⇝ Opções
• Estatísticas ☑ Tabela de Sobrevivência
• Estatísticas ☑ Sobrevivência de média e mediana
• Gráficos ☑ Sobrevivência
• Gráficos ☑ Risco
⇝ Continuar
⇝ Continuar
⇝ Ok
```

um *p-value* de $p = 0.795$, pelo que não existe evidência estatística de diferenças de sobrevivência entre grupos.

Um vídeo sobre como obter o modelo de sobrevivência de Kaplan-Meier em SPSS é dado no link seguinte:

SPSS

4.7 Outros modelos de sobrevivência

O modelo não-paramétrico de Kaplan-Meier permite apenas englobar covariáveis nominais ou em estratos, uma vez que para cada um dos seus valores é estimada uma função de sobrevivência de forma independente. Para incorporar outro tipo de variáveis (por exemplo, quantitativas) teremos de optar por outros modelos.

Nesta secção vamos abordar outros tipos de modelos possíveis, embora não sejam detalhadas as suas aplicações em SPSS e condições de aplicabilidade. Dados os limites de tempo deste curso introdutório, servem apenas para ilustrar outras possibilidades de modelos de sobrevivência.

4.7.1 Modelos de Regressão paramétricos

Os modelos de regressão permitem a introdução de covariáveis de interesse, que se espera possam ter influência na variável tempo de sobrevivência e consequen-

temente na função de sobrevivência.

A abordagem que trataremos nesta secção, passa por considerar a introdução de covariáveis (quantitativas ou não) em modelos paramétricos. Para isso, vamos considerar que a alguns parâmetros da distribuição da variável tempo de falha T associada ao modelo são afetados pela covariável de interesse X, estabelecendo-se nessa relação um modelo de correlação. A regressão linear não é geralmente apropriada para o efeito, ou seja, considerar que

$$\theta = \beta_0 + \beta_1 x + \varepsilon$$

em que o erro ε teria uma distribuição normal, não é adequado, uma vez que em geral não se traduz geralmente num modelo válido em análise de sobrevivência.

É no entanto importante verificar a adequação deste tipo de modelos aos dados em causa, uma vez que o modelo escolhido pode não ser adequado. Ao contrário dos modelos não-paramétricos, os modelos de regressão paramétricos têm menos adaptabilidade, uma vez que partem de um pressuposto para a distribuição.

Regressão Exponencial

Consideramos que a taxa de falha λ se mantém constante ao longo do tempo de sobrevivência T, mas depende da covariável X. Dessa forma, assumimos uma regressão exponencial da forma

$$\lambda(t|x) = \varepsilon e^{-\beta_0 - \beta_1 x} \tag{4.16}$$

em que o erro ε tem distribuição exponencial com média unitária.

Assim, a função de sobrevivência para o modelo exponencial, tem a expressão

$$S(t|x) = \exp\left(-\frac{t}{e^{\beta_0+\beta_1 x}}\right), \quad t \geq 0. \tag{4.17}$$

A determinação dos coeficientes β_0 e β_1 define a função de sobrevivência pelo modelo linear, sendo que estes podem ser obtidos pelo método da máxima verosimilhança, tomando em conta as censuras.

De notar que no caso com m covariáveis, ou seja, em que se pode considerar $X = (X_1, X_2, \ldots, X_m)$, o caso anterior é generalizado por

$$S(t|x) = \exp\left(-\frac{t}{e^{\beta(x)}}\right), \quad t \geq 0, \tag{4.18}$$

em que

$$\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{4.19}$$

Regressão de Weibull

Na regressão de Weibull assumimos que o parâmetro de escala α da distribuição de Weibull depende da covariável de interesse X, ou seja, tomamos a regressão exponencial para este parâmetro

$$\alpha = \varepsilon e^{\beta(x)},$$

com a mesma notação para $\beta(x)$ considerada em (4.19) no caso de considerar m covariáveis. Consegue-se mostrar que isto se traduz na função de sobrevivência

$$S(t|x) = \exp\left(-\left(\frac{t}{e^{\beta(x)}}\right)^{\gamma}\right), \quad t \geq 0, \tag{4.20}$$

o que por sua vez se traduz numa taxa de falha dada por

$$\lambda(t|x) = \frac{\gamma}{e^{\gamma\beta(x)}} t^{\gamma-1}, \quad t \geq 0. \tag{4.21}$$

Regressão de Log-normal

Para a regressão log-normal, consideramos a regressão linear da média μ da variável $Y = \ln(T)$ em função das covariáveis, isto é,

$$\mu = \beta(x) + \varepsilon, \tag{4.22}$$

em que $\beta(x)$ é definido por (4.19) e ε tem uma distribuição normal com média nula, o que se traduz no modelo de regressão log-normal para a função de sobrevivência dado por

$$S(t|x) = \Phi\left(\frac{\beta(x) - \ln t}{\sigma}\right). \tag{4.23}$$

em que Φ é a função de distribuição normal padrão. Assim, assumimos que as covariáveis apenas afetam a média, mas não o desvio padrão.

4.7.2 Modelo semi-paramétrico de Cox

O modelo de Cox [7] é um modelo semi-paramétrico, ou seja, é um modelo que em parte tem a flexibilidade de um modelo não-paramétrico, mas por outro lado envolve a estimação de alguns parâmetros. Estes parâmetros estão relacionados com hipóteses para o modelo que precisam de ser satisfeitas, sendo um modelo com condições de aplicação, ainda que menos forte que nos modelos paramétricos de regressão usuais.

O modelo de Cox standard parte do princípio dos riscos proporcionais que esclarecemos de seguida.

4.7.3 Modelo de Cox

O **modelo de Cox** é também designado por modelo de riscos proporcionais, uma vez que assume que a razão entre as taxas de risco entre dois estados de uma covariável é constante. Dito de outra forma, para a covariável X_i, e como habitualmente assumindo regressão exponencial, temos

$$\frac{\lambda(t|x_1, x_2, \ldots, x_i + x, \ldots, x_m)}{\lambda(t|x_1, x_2, \ldots, x_i, \ldots, x_m)} = e^{\beta_i x}. \tag{4.24}$$

Sendo esta a única hipótese a considerar, o modelo de Cox assume que a taxa de falha para os valores $x_1, x_2, \ldots, x_m$ das m covariáveis $X_1, X_2, \ldots, X_m$ é dada por

$$\lambda(t|x_1, x_2, \ldots, x_m) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m}. \tag{4.25}$$

em que $\lambda_0(t)$ é denominada a taxa de falha basal e corresponde à taxa de falha quando todas as covariáveis são nulas.

Assim, a estimativa do modelo de Cox passa por estimar a função $\lambda_0(t)$ (componente não paramétrica) e os parâmetros $\beta_1, \beta_2, \ldots, \beta_m$ (componente paramétrica). A partir da taxa de falha, podemos estimar a função de sobrevivência pelo modelo de Cox a partir da relação

$$S(t|x_1, x_2, \ldots, x_m) = [S_0(t)]^{\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}. \tag{4.26}$$

em que $S_0(t) = e^{-\int_0^t \lambda_0(u)du}$, ou seja, a função de sobrevivência basal.

É fácil verificar que o modelo de Cox é mais versátil que os modelos de regressão paramétricos referidos no capítulo anterior. Na realidade, enquanto estes assumem o comportamento de λ_0, no modelo de Cox este comportamento é livre *a priori* e estimado *a posteriori* consoante os dados recolhidos. Assim, se o modelo paramétrico adequado for desconhecido, o uso do modelo de Cox é apropriado, o que faz deste modelo muito popular em estudos de sobrevivência.

No entanto, é importante notar que este modelo apenas pode ser utilizado caso os riscos sejam proporcionais, o que pode ser estudado a partir dos resíduos de Schoenfeld.

4.7.4 Modelo de Cox estratificado

Quando a hipótese de taxas de falha proporcionais não é verificada para determinada variável, não se pode aplicar o modelo de Cox conforme apresentado no capítulo 4.7.3. Nesse caso, uma possibilidade pode ser a de estratificar a variável em causa e encontrar um modelo de Cox em cada estrato. De alguma forma, a ideia é semelhante à avançada para a aplicação de métodos não paramétricos a variáveis contínuas, sendo que as causas para a estratificação são diferentes. No caso dos métodos não-paramétricos, estratificava-se a variável contínua por forma a obter um modelo não-paramétrico em cada estrato, uma vez que no caso de modelos não paramétricos não é possível introduzir covariáveis diretamente no modelo. No caso do modelo de Cox, a estratificação ocorre para se obter um modelo diferente de Cox em cada estrato e assim eliminar a necessidade de taxas de falha proporcionais entre estratos. Mais detalhes sobre este modelo de Cox estratificado podem ser obtidos, por exemplo, em [8, 9].

4.7.5 Modelo de Cox com covariáveis dependentes do tempo

O modelo de Cox que ilustrámos neste capítulo assume entre outras hipóteses que as covariáveis não se alteram no tempo. No caso de se alterarem, este não pode ser aplicado. Note-se que a suposição de não variação das covariáveis pode ser limitativa em alguns casos. Por exemplo o estado civil de um sujeito pode-se alterar durante um estudo, assim como por exemplo o seu nível de glicémia em jejum ou o peso do paciente. Nesse contexto já não faz sentido falar-se em risco proporcionais, pois sendo a taxa da falha agora dada por

$$\lambda(t|x_1(t), x_2(t), \ldots, x_i(t), \ldots, x_m(t)) = \lambda_0(t)e^{\beta_1 x_1(t) + \beta_2 x_2(t) + \cdots + \beta_m x_m(t)} \tag{4.27}$$

a razão das taxas de falha não é constante ao longo do tempo. Este modelo é portanto ainda mais versátil que o standard e é uma generalização importante

do método de Cox. Mais detalhes deste tipo de modelos podem ser encontrados por exemplo em [8, 9].

4.7.6 Modelos de tempo de vida acelerado

Na apresentação feita nas secções 4.7.1 a 4.7.3, foi sempre considerado que a presença de covariáveis significativas influenciaria a taxa de falha de forma proporcional. No entanto, nem sempre será o caso, na prática. Um outro tipo de modelos que é possível de utilizar é o chamado modelo de tempo de sobrevivência acelerado. Neste caso, a presença de covariáveis acelera ou diminui o tempo de sobrevivência na função de sobrevivência basal, ou seja, considera-se um modelo da forma

$$S(t|\boldsymbol{x}) = S_0(tg(\boldsymbol{x}))$$

em que a função g altera a escala de tempo em função do valor das covariáveis $\boldsymbol{x}$. Mais detalhes deste tipo de modelos podem ser encontrados por exemplo em [8, 9].

4.7.7 Modelos *first hitting time*

Uma nova porta que se vem abrindo para o estudo de sobrevivência é o de assumir que existe um modelo estocástico subjacente à falha. Assim, assume-se que o sujeito inicia com um valor positivo da variável Y, que corresponde a um estado saudável, e que a falha ocorre quando a variável estocástica associada Y assume o valor nulo. Esta variável estocástica é impossível de medir na realidade, mas assume-se que está subjacente ao processo. Assim, modelar o tempo de falha é equivalente a modelar o processo estocástico da variável Y. Estas técnicas estão associadas a modelos de *first hitting time* (FHT) e regressão *threshold* (TR), que podem ser consultados com mais detalhe em [10, 11].

Índice

Links

Bibliografia

[1] Miguel Patrício, Marisa Loureiro, and Francisco Caramelo. *Bioestatística com SPSS*. Plátano Editora, 2017.

[2] Teresa Oliveira. *Estatística Aplicada*. Universidade Aberta, 1994.

[3] Elizabeth Reis, Paulo Melo, Rosa Andrade, and Teresa Calapez. *Estatística Aplicada*, volume 2. Edições Sílabo, 2016.

[4] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):pp. 457–481, 1958.

[5] N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, 2(3):437–453, 05 1974.

[6] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc., 2002.

[7] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[8] Enrico Antonio Colosimo and Suely Ruiz Giolo. *Anï¿½lise de Sobrevivï¿½ncia Aplicada*. Editora Blücher, 2006.

[9] David G. Kleinbaum and Michel Klein. *Survival Analysis, a Self-Learning Text*. Springer, 2nd edition edition, 2005.

[10] Mei-Ling Ting Lee. *Lifetime Models and Risk Assessment*. John Wiley & Sons, Ltd, 2014.

[11] D. Stogiannis, C. Caroni, C. E. Anagnostopoulos, and I. K. Toumpoulis. Comparing first hitting time and proportional hazards regression models. *Journal of Applied Statistics*, 38(7):1483–1492, 2011.