notas de estatística

maria joana soares e ricardo severino

maio de 2011

Conteúdo

CONTEÚDO

1

Introdução

Na actual sociedade da informação, existe uma cada vez maior facilidade de acesso e interesse na recolha de grandes quantidades de dados. De facto, a quantidade de dados disponíveis tem crescido a um ritmo impressionante nos últimos tempos.

Uma questão que se coloca, naturalmente, perante esta super abundância de dados é a seguinte: como tirar partido e extrair significado desses dados?

Os meios computacionais existentes actualmente permitem-nos processar, resumir e analisar grandes quantidades de dados de forma muito rápida e eficiente, para deles extrair informação relevante.

Podemos dizer, de um modo informal, que o principal objectivo da estatística é precisamente o de extrair *informação* a partir de dados.

Mais especificamente:

A **Estatística** pode ser definida como o conjunto de instrumentos, procedimentos e técnicas que permitem, de forma adequada, recolher, organizar, explorar, descrever, analisar e interpretar dados.

1.1 Conceitos básicos

- Chama-se **população** ao conjunto de todos os objectos/indivíduos/etc. que têm em comum uma ou mais características sobre as quais temos interesse em efectuar um estudo estatístico.
- A cada elemento da população dá-se o nome de **unidade estatística**.
- Cada uma das características em estudo[a] é chamada **variável**.
- Uma **observação** é o valor que a variável assume numa determinada unidade estatística.

[a] As quais, naturalmente, só nos interessam quando ocorrem em diferentes quantidades ou tipos, isto é, quando não têm um valor constante para todos os elementos da população.

Exemplo 1.1. *Como exemplos de populações, têm-se:*

1. *o conjunto de todos cidadãos portugueses;*
2. *o conjunto de todos os livros existentes na biblioteca da Universidade do Minho;*
3. *o conjunto de todos os parafusos produzidos por uma certa fábrica (num determinado período de tempo);*
4. *o conjunto de todos os possíveis lançamentos de um dado (um número infinito de vezes).*[1]

Em relação com cada uma das populações anteriores, as variáveis a estudar poderiam ser, por exemplo (respectivamente):

1. *sexo, cor dos olhos, altura, peso;*
2. *ano de publicação, tipo de encadernação (hardback, paperback), número de exemplares existentes ;*
3. *tipo de parafuso (de cabeça chata, de cabeça redonda), qualidade do parafuso (defeituoso, não defeituoso);*
4. *número da face obtida (1 a 6), paridade do número da face obtida (par/ímpar);*

Dada a impossibilidade de observar toda uma população – ou devido à sua dimensão, ou pelo facto de a observação poder implicar a sua destruição – ou apenas por razões de economia, comodidade

[1] Trata-se, neste caso, daquilo a que chamamos uma população hipotética, uma vez que não consiste de objectos realmente existentes.

ou tempo, é fundamental recolher um subconjunto, que se pretende que seja representativo, dessa população. Tem-se, então, a seguinte definição.

- Dada uma população relativa a um certo estudo estatístico, chama-se **amostra** a um subconjunto finito dessa população (usado para estudar as características que nos interessam na população).
- A cada observação da variável (ou variáveis) em estudo respeitante a cada unidade estatística pertencente à amostra, chamamos **dado estatístico**.

Nota: Por vezes, chama-se população ao conjunto de potenciais valores que a variável em estudo assumiria na população, por exemplo, fala-se na população das alturas ou dos pesos dos alunos inscritos na Universidade do Minho no ano lectivo 2010/2011; neste caso, chamar-se-á, amostra ao conjunto dos dados estatísticos (i.e. ao conjunto dos valores da variável observados nos diversos elementos do subconjunto da população seleccionado). A amostra é, nesse caso, uma lista de dados.

É fundamental que a amostra seleccionada seja representativa da população. Um dos processos de garantir essa representatividade é fazer **amostragem aleatória simples**, em que todos os elementos da população têm as mesmas hipóteses (i.e. igual probabilidade) de ser incluídos na amostra.[2] Uma amostra mal recolhida, também chamada *viciada, enviesada* ou *tendenciosa*, levará naturalmente a conclusões e previsões distorcidas; ver, e.g. [Ath07, p.14] para um exemplo interessante (previsão dos resultados das eleições americanas de 1936, que opuseram F. D. Roosevelt a A. Landon).

1.2 Classificação das variáveis

Para a descrição dos fenómenos é importante a classificação das variáveis em estudo. As variáveis distinguem-se (quanto à sua natureza) em:

1. variáveis **qualitativas** (também ditas **categóricas** ou **factores**);
2. variáveis **quantitativas**.

As variáveis qualitativas variam em tipo ou qualidade, mas não em quantidade; os seus valores são intrinsecamente não numéricos. Exemplos de variáveis qualitativas, com indicação de seus possíveis valores, são: cor do cabelo (preto, castanho, louro, ruivo); sexo (masculino, feminino);

[2] Este método pode ser de difícil execução em populações de dimensão muito elevada, mas pode sempre recorrer-se a métodos que simulam essa aleatoriedade (como por exemplo, a amostragem estratificada, por grupos, etc.); ver. e.g. [Ath07, p.13] ou [PV08, p. 56].

grupo sanguíneo (A+,B+,AB+,O+,A−,B−,AB−,O−); classificação obtida numa entrevista (insuficiente, suficiente, bom, muito bom); nível de satisfação com a alimentação na cantina universitária (muito insatisfeito, insatisfeito, satisfeito, muito satisfeito); dureza dos minerais na escala de Mohs $(1, 2, \ldots, 10)$.

Note-se que, por exemplo, no caso da dureza dos minerais, o que está em causa é uma característica (a resistência ao risco) a qual é intrinsecamente não numérica, sendo a escala numérica de 1 a 10 uma simples convenção.

As variáveis quantitativas variam em quantidade, sendo os seus valores intrinsecamente numéricos. São exemplos de variáveis quantitativas: altura; peso; idade; número de filhos; número de dentes sãos; temperatura (em graus Celsius), temperatura absoluta (medida em graus Kelvin), hora do dia.

As variáveis quantitativas podem ainda distinguir-se entre:

1. **variáveis discretas**: quando assumem apenas um número finito ou infinito numerável (isto é, contável) de valores; estão associadas, geralmente, a processos de contagem, por exemplo: número de filhos, número de dentes sãos, número de parafusos defeituosos,etc.

2. **variáveis contínuas**: quando podem tomar valores num conjunto infinito não numerável (por exemplo um certo intervalo de números reais); estão, geralmente, associadas a processos de medida: peso, altura, tempo de vida (de um ser humano), temperatura (em graus centígrados) temperatura absoluta (em graus Kelvin), etc.

Note-se que as observações (medições) de uma variável contínua são sempre arredondadas (isto é, são dadas até uma certa precisão), uma vez que não é possível medir uma variável contínua exactamente. Por exemplo, no caso da altura poderemos arredondar os dados até ao cm e usar $175cm$ para alturas tais como $175,123...cm$ ou $174,9235...cm$; no caso do tempo de vida (idade) de um ser humano, é usual medirmos a idade em número inteiro de anos, etc. Assim, os dados amostrais com que trabalharemos são sempre conjuntos discretos (na realidade, conjuntos finitos), ainda que, em certos caso, sejam tratados como dados contínuos, no sentido em são aproximações de números que, teoricamente, podem assumir valores num "contínuo"de pontos.

1.3 Escalas de medida

Existem quatro tipos de escalas de medida que podemos usar quando lidamos com dados estatísticos (resultantes de observações/medições de uma variável).

1. **Nominal** ou **categórica**: usada para as variáveis qualitativas, para distinguir diferentes categorias ou classes de indivíduos, quando a ordem das categorias ou classes (ainda que rotuladas com números) não tem qualquer significado; aos indivíduos da mesma categoria ou classe é atribuído o mesmo *nome* ou *código*; por exemplo, para a variável sexo, a escala poderá ser F para o sexo feminino e M para o sexo masculino (ou, se preferimos , 1 para o sexo feminino e 2 para o sexo masculino). Neste caso não faz sentido efectuar operações aritméticas sobre os "nomes"das classes ou categorias, ainda que estes estejam representados por números. Os métodos estatísticos apropriados para a análise de dados em escala nominal são os que se baseiam em contagens de efectivos de cada classe ou análise de proporções.

2. **Ordinal**: usada para as variáveis qualitativas, para distinguir diferentes categorias ou classes de indivíduos, quando a ordem das categorias ou classes já tem significado; por exemplo, no caso da variável anteriormente referida relativa ao grau de satisfação com a alimentação na cantina, faz sentido considerar o grau de satisfação por *ordem crescente*: 0-muito insatisfeito, 1-insatisfeito, 2-satisfeito, 3-muito satisfeito; o mesmo se passa com a escala de Mohs para a dureza dos minerais. Para os dados em escala ordinal, para além de usarmos métodos baseados em contagens e proporções, também podemos usar métodos baseados nas ordens ou *ranks* das observações; [PV08].

3. **Intervalar**: usada para variáveis quantitativas; trata-se de uma escala numérica em que pode haver um zero, mas em que esse zero é apenas convencional, não significando "ausência"da característica medida; por exemplo, a escala Celsius ou a escala Fahrenheit para medir temperaturas; 0^o Celsius não significa ausência de calor (note-se que 0^oC corresponde a 32^o F); neste tipo de escala faz sentido ordenar as medições (32^oC é uma temperatura superior a 16^oC), fazer operações aritméticas envolvendo somas e diferenças (por exemplo achar a média das temperaturas máximas registadas ao longo de certos dias ou calcular a amplitude térmica diurna), mas não faz sentido fazer divisões ou razões; uma temperatura de 32^oC não é o dobro de 16^oC (basta converter na escala Fahrenheit para perceber porque não faz sentido dizer que uma temperatura de 32^oC é duas vezes uma temperatura de 16^oC).

4. **Absoluta** (ou **de razões**): usada para variáveis quantitativas; trata-se de uma escala numérica em que o zero não é uma simples convenção, mas corresponde, na realidade, a ausência da característica medida. Por exemplo, a escala Kelvin para temperatura ou a escala em kg para o peso. Neste caso, faz sentido, para além das operações aritméticas já referidas para a escala intervalar, fazer também divisões (comparações por quocientes). Um indivíduo com $100\,kg$ pesa, efectivamente, o dobro de um com $50\,kg$ (se converter o peso noutras unidades, a relação será sempre a mesma).

As escalas anteriores estão apresentadas em grau crescente de complexidade. Cada nível de complexidade acrescenta qualquer coisa ao nível anterior:

1. Nominal: nível mais baixo, só nomes têm significados;
2. Ordinal: acrescenta uma ordem aos nomes;
3. Intervalar: acrescenta sentido às diferenças;
4. Absoluta: acrescenta sentido às razões.

Os dados quantitativos podem ser categorizados e, consequentemente, transformados em factores: por exemplo a idade de uma pessoa, variável quantitativa em **escala absoluta**, pode ser convertida num factor com, por exemplo, 5 classes etárias (primeira infância, infância, adolescência, adulto, idoso) tornando-se uma variável qualitativa em **escala ordinal**.

De facto, é sempre possível passar de dados expressos numa escala complexa para uma escala mais simples, mas o contrário não faz sentido.

1.4 Principais fases da análise estatística

As análises estatísticas são, essencialmente, compostas por três fases:

1. recolha de dados – *amostragem* ou *planeamento de experiências*;
2. tratamento inicial de dados, que inclui a sua ordenação, resumo (cálculo de algumas das suas características), apresentação (em tabelas, gráficos, etc.) e exploração desses dados – *estatística descritiva* e *análise exploratória* de dados;
3. indução, a partir do que se verifica numa amostra, para a população de que esta foi extraída – *inferência estatística*.

Neste curso introdutório, debruçar-nos-emos apenas sobre alguns aspectos das fazes 2. e 3. da análise estatística.

2

Análise Inicial dos Dados

2.1 Dados qualitativos

Os dados provenientes de variáveis qualtitativas são geralmente descritos usando tabelas de frequências (absolutas ou relativas) e gráficos, tais como gráficos de barras ou diagramas circulares.

- Uma tabela que lista as diversas categorias e as correspondentes frequências com que ocorrem (isto é, a contagem do número de indivíduos que estão em cada categoria) é chamada uma **tabela de frequências** (absolutas).
- As frequências absolutas podem ser divididas pelo número total de indivíduos para obter as chamadas **frequências relativas** (por vezes expressas em percentagem).

A distribuição do número total de observações pelas diversas categorias diz-se uma **distribuição** de frequências.

Exemplo 2.1. *Foi recolhida uma amostra de 400 estudantes a viver em residência universitária e estes foram inquiridos relativamente ao seu grau de satisfação com as condições de alojamento, na seguinte escala: Muito Insatisfeito, Insatisfeito, Razoavelmente Satisfeito, Satisfeito, Muito Satisfeito. Os resultados indicaram que: 28 escolheram a categoria "Muito Insatisfeito", 60 escolheram a categoria "Insatisfeito", 160 escolheram a categoria "Razoavelmente Satisfeito", 120 escolheram a categoria "Satisfeito"e 32 escolheram a categoria "Muito Satisfeito".*

Uma tabela de frequências (absolutas e relativas) para os dados recolhidos poderá ser apresentada

da forma seguinte:

Opinião de estudantes sobre o alojamento na residência

Grau de satisfação	*Frequência absoluta*	*Frequência relativa*	*Frequência relativa (%)*
Muito Insatisfeito	*28*	0.07	7
Insatisfeito	*60*	0.15	15
Razoavelemente Satisfeito	*160*	0.40	40
Satisfeito	*120*	0.30	30
Muito Satisfeito	*32*	0.08	8
Total	*400*	*1*	100

Estando os dados em escala ordinal, podemos também apresentar na tabela de frequências, as chamadas **frequências acumuladas** (absolutas ou relativas), as quais se obtêm adicionando as frequências.

Exemplo 2.1 (cont.) *Considerando os dados do Exemplo 2.1 , ter-se-ia a seguinte tabela.*

Opinião de estudantes sobre o alojamento na residência

Grau de satisfação	*Frequência absoluta*	*Frequência absoluta acumulada*	*Frequência relativa*	*Frequência relativa acumulada*
Muito Insatisfeito	*28*	*28*	0.07	0.07
Insatisfeito	*60*	88	0.15	0.22
Razoavelmente Satisfeito	*160*	*248*	0.40	0.62
Satisfeito	*120*	368	0.30	0.92
Muito Satisfeito	*32*	400	0.08	1
Total	*400*		*1*	

Na construção de um gráfico de barras ou diagrama circular, deve ter-se em atenção que:

1. as alturas das barras (num gráfico de barras) ou as amplitudes dos diversos sectores (num diagrama circular) são proporcionais às respectivas frequências;

2. num gráfico de barras, estas devem estar igualmente distanciadas umas das outras, e podem ser representadas na horizontal ou na vertical;

3. os diagramas circulares utilizam-se apenas quando o número de categorias em que a variável é classificada é pequeno.

Exemplo 2.1 (cont.) *Na Figura 2.1 apresentam-se dois gráficos de barras e na Figura 2.2 um diagrama circular relativos aos dados do Exemplo 2.1.*

A função do Mathematica apropriada para esboçar gráficos de barras é a função `BarChart`. Para diagramas circulares podemos usar a função `PieChart` ou a função `PieChart3D`.

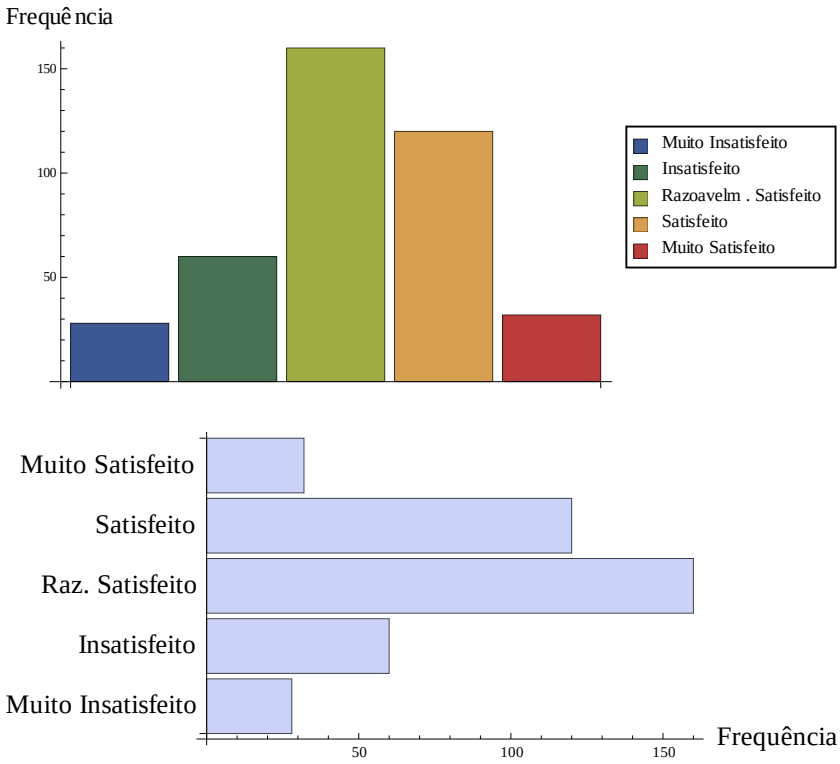

Figura 2.1: Gráficos de barras relativos aos dados do Exemplo 2.1.

2.2 Dados quantitativos

Os dados quantitativos correspondentes a variáveis que não tomem um grande número de valores podem, de modo análogo ao que é feito para variáveis qualitativas, ser apresentados através de tabelas de frequências. Quanto à representação gráfica destes dados, é usual usar gráficos de linhas (mas, por vezes, também gráficos de barras) para os representar.

Os dados provenientes de variáveis que tomam um número pequeno de valores podem ser representados também através de diagramas circulares.

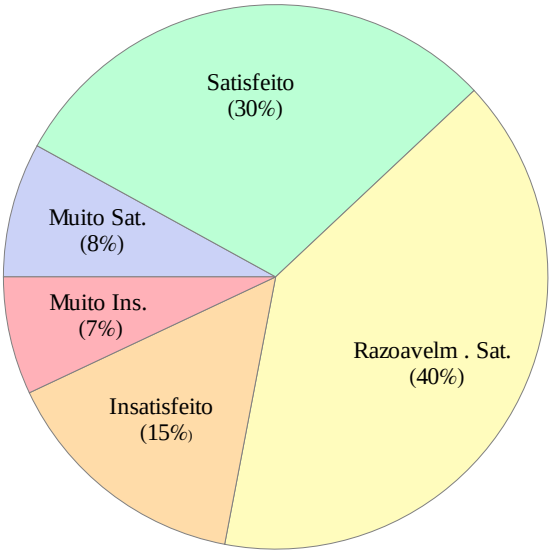

Figura 2.2: Diagrama circular relativo aos dados do Exemplo 2.1.

Exemplo 2.2. *Registou-se o número de irmãos de cada um dos 20 alunos de uma dada turma de um liceu, tendo-se obtido os seguintes dados:*

$$\mathbf{x} = (1, 1, 3, 2, 0, 1, 0, 3, 1, 1, 4, 0, 2, 2, 1, 1, 0, 1, 1, 1)$$

A tabela de frequências correspondente aos dados anteriores é:

Número de irmãos dos alunos de uma turma

Nº irmãos	*Frequência absoluta*	*Frequência relativa (%)*
0	*4*	*20*
1	*10*	*50*
2	*3*	*15*
3	*2*	*10*
4	*1*	*5*
Total	*20*	*100*

Um gráfico de linhas correspondente a estes dados é apresentado na Figura 2.3.

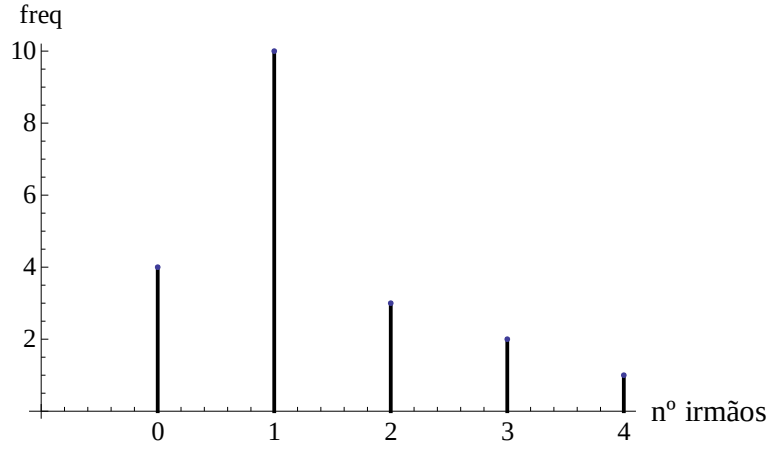

Figura 2.3: Gráfico de linhas relativo aos dados do Exemplo 2.2

No caso de uma variável contínua, ou no caso de uma variável discreta que tome um elevado número de valores, a sua apresentação numa tabela e respectiva representação gráfica pressupõe um agrupamento de dados em **classes**, as quais, salvo casos excepcionais, devem ser intervalos de igual amplitude, designada por **intervalo de classe**. Isto significa que os dados são categorizados. As classes em que os dados são categorizados devem ser exaustivas (isto é, a união das classe deve conter todos os dados) e mutuamente exclusivas (as classes não se podem sobrepor, pelo que cada um dos dados irá pertencer a apenas uma classe).

As primeiras questões que se levantam são as seguintes:

1. Qual deve ser o número de classes a considerar?
2. Qual deve ser a amplitude de cada classe?

Existem diversas regras para calcular o número de classes. A mais usual é a chamada **regra de Sturges** que estabelece que, para uma amostra de dimensão n, o número k de classes a considerar é o menor inteiro tal que

$$2^{k-1} \geq n.^{1}$$

De notar que a regra de Sturges (ou outras regras existentes para a determinação do número de classes) são meramente indicativas. A prática, a experiência e o bom senso são os melhores "guias" para a escolha do número de classes a utilizar.

[1] Este valor de k pode ser obtido pela fórmula $k = \lceil \log_2(n) + 1 \rceil$ onde $\lceil x \rceil$ designa o menor inteiro não inferior a x.

Descrevemos agora a forma mais usual de obter as classes.

Supondo que os dados amostrais que estamos a considerar são $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, vamos representar por $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ a sequência obtida ordenando os dados x_i, por ordem crescente. Chamamos-lhe **amostra ordenada**.

A quantidade

$$R = x_{(n)} - x_{(1)} = \max\{x_i\} - \min\{x_i\}.$$

é chamada **amplitude da amostra**.

Após a escolha do número de classes, k, determinamos o valor $h^* = \frac{R}{k}$ e tomamos um valor h tal que $h > h^*$, o qual será a amplitude de cada classe. O valor de h a utilizar deverá ser razoavelmente simples.

Determinamos, então o valor $\epsilon = kh - R$ e tomamos para primeira classe o intervalo, aberto à direita e fechado à esquerda[2] dado por

$$\left[x_{(1)} - \frac{\epsilon}{2}, x_{(1)} - \frac{\epsilon}{2} + h\right);$$

a classe seguinte será, então,

$$\left[x_{(1)} - \frac{\epsilon}{2} + h, x_{(1)} - \frac{\epsilon}{2} + 2h\right)$$

e assim sucessivamente, sendo a última classe dada por

$$\left[x_{(n)} + \frac{\epsilon}{2} - h, x_{(n)} + \frac{\epsilon}{2}\right).$$

Um exemplo de aplicação do processo acima descrito, é apresentado de seguida.

Exemplo 2.3. *Considere-se a seguinte amostra de* 100 *dados:*

$$\begin{aligned}\mathbf{x} = (&193, 193, 171, 197, 197, 177, 186, 180, 187, 163, 169, 176, 171, 183, 168, 168, 183, 189, 175, 176,\\ &158, 167, 182, 165, 180, 178, 186, 185, 182, 168, 170, 202, 200, 194, 165, 172, 173, 174, 181, 188,\\ &194, 172, 174, 173, 187, 172, 194, 167, 193, 187, 182, 170, 184, 166, 171, 176, 188, 169, 180, 170,\\ &168, 191, 194, 196, 173, 167, 184, 166, 180, 166, 163, 184, 173, 161, 173, 158, 187, 184, 184, 177,\\ &192, 161, 167, 169, 197, 182, 160, 200, 188, 201, 188, 199, 176, 196, 191, 166, 192, 187, 188, 175)\end{aligned}$$

Como $2^6 = 64 < 100$ *e* $2^7 = 128 > 100$, *usando a regra de Sturges, devemos considerar um número de classes* k *tal que* $k - 1 = 7$, *ou seja, devemos usar* $k = 8$. *Temos, além disso* $x_{(n)} = \max_i\{x_i\} =$

[2] É esta a convenção que adoptamos aqui, por ser a que mais se ajusta ao uso do Mathematica, embora sejam possíveis outras escolhas.

202 *e* $x_{(1)} = \min_i\{x_i\} = 188$, *pelo que* $R = 202 - 158 = 44$. *Neste caso, tem-se* $h^* = 44/8 = 5.5$ *e é razoável considerar* $h = 6$. *Então, vem* $\epsilon = 6 \times 8 - 44 = 4$. *Então, a primeira classe será o intervalo*

$$[x_{(1)} - \epsilon/2, x_{(1)} - \epsilon/2 + h) = [158 - 2, 158 - 2 + 6) = [156, 162)$$

e as sete restantes classes serão

$$[162, 168),\ [168, 174),\ [174, 180),\ [180, 186),\ [186, 192),\ [192, 198)\ \textit{e}\ [198, 204).$$

Para construir a tabela de frequências absolutas (e relativas) de cada classe, teríamos de contar os efectivos de cada classe (e dividir pelo número total de elementos). Essa tabela seria a seguinte:

Classe	*Frequência absoluta*	*Frequência relativa*
$[156, 162)$	*5*	*0.05*
$[162, 168)$	*12*	*0.12*
$[168, 174)$	*21*	*0.21*
$[174, 180)$	*11*	*0.11*
$[180, 186)$	*17*	*0.17*
$[186, 192)$	*15*	*0.15*
$[192, 198)$	*14*	*0.14*
$[198, 204)$	*5*	*0.05*
Total	*100*	*1*

O correspondente histograma é apresentado na Figura 2.4.

Por vezes, desenha-se o chamado **polígono de frequências**, unindo por segmentos de recta os pontos médios do topo dos rectângulos; os pontos médios dos rectângulos inicial e final são unidos até aos pontos, situados no eixo dos xx, cujas abcissas são os pontos médios das classes que seriam adjacentes às classes inicial e final; veja-se Figura 2.5.

Os histogramas podem também ser feitos usando frequências relativas no eixo dos yy, em vez de frequências absolutas, o que permite comparar mais facilmente diferentes distribuições; o respectivo polígono de frequências é, nesse caso, o chamado **polígono de frequências relativas**.

O Mathematica dispõe de uma função apropriada para desenhar histogramas: `Histogram`. O seu uso (com escolha adequada de parâmetros) será explorado nas aulas práticas.

2.3 Medidas de localização, de dispersão e de forma

As tabelas de frequências e os gráficos constituem processos de redução de dados. No caso de dados quantitativos, é possível resumir de uma forma mais drástica esses dados, calculando certas

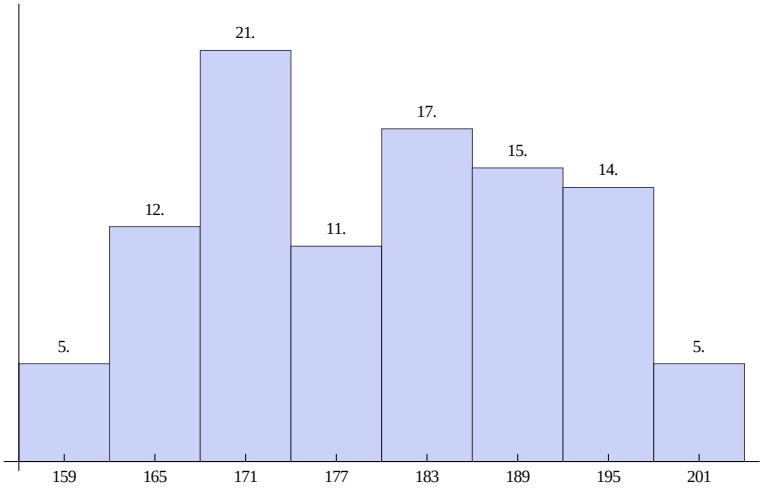

Figura 2.4: Histograma relativo aos dados do Exemplo 2.3

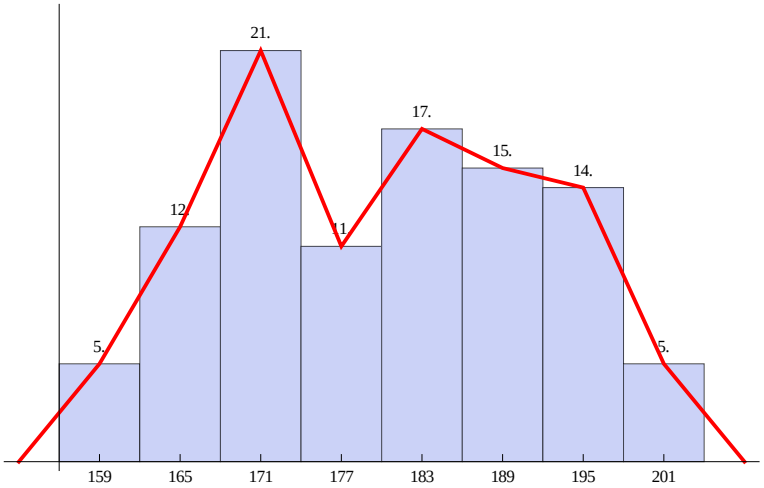

Figura 2.5: Histograma e polígono de frequências relativo aos dados do Exemplo 2.3.

características amostrais.

Chamamos **estatísticas** às características numéricas da amostra e **parâmetros** às características numéricas da população; as características amostrais (empíricas, obtidas a partir dos dados) são estimativas das correspondentes características populacionais, que estudaremos posteriormente.

Três aspectos importantes de um conjunto de dados são a sua localização (ou tendência central),

a sua variabilidade (ou dispersão) e a sua forma.

2.3.1 Medidas de localização

No que se segue, dada uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$, designaremos por $(x_{(1)}, \ldots, x_{(n)})$ a respectiva amostra ordenada por ordem crescente.

Dado uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$, a **média amostral** (ou média da amostra), denotada por $\bar{x}$, é definida por

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{k=1}^{n} x_k. \tag{2.1}$$

O **quantil** p **amostral** (ou quantil p da amostra) definido para $p \in [0, 1]$, que representaremos por q_p é, "grosso modo", o valor que separa os $p \times 100\%$ valores menores da amostra dos $(1-p) \times 100\%$ valores maiores da amostra. Ou seja, o quantil p separa os $p \times 100\%$ primeiros valores da amostra **ordenada** dos $(1-p) \times 100\%$ últimos valores dessa amostra.

É importante salientar que existem diversas formas de definir os quantis, nem sempre coincidentes. Por exemplo, por vezes exige-se que o quantil seja um dos valores da amostra, outras vezes o quantil pode ser um valor situado entre dois valores da amostra (determinado de acordo com um certo critério).

No Mathematica, se usarmos a função `Quantile`, sem especificar quaisquer parâmetros adicionais, para determinar os quantis, estes serão sempre valores da amostra. A função `Quantile` pode, no entanto, ser usada com outros parâmetros para permitir calcular os quantis de acordo com diversas definições.

Alguns quantis especialmente importantes, correspondem aos casos:

- $p = 1/2$, o qual é designado por **mediana amostral**;
- $p = i/4; i = 1, 2, 3$, que são os chamados **quartis amostrais** (respectivamente primeiro quartil, segundo quartil e terceiro quartil);
- $p = i/10; i = 1 \ldots, 9$, que são os chamados **decis amostrais**;
- $p = i/100, i = 1, 2, \ldots, 99$, que são os chamados **percentcis amostrais**.

O Mathematica tem funções próprias para calcular a média, a mediana e os três quartis: `Mean`, `Median` e `Quartiles`, respectivamente. A mediana calculada com a função `Median` do Mathematica corresponde ao uso da fórmula

$$q_{1/2} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ for ímpar} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{se } n \text{ for par} \end{cases} . \tag{2.2}$$

Note-se que, quando n é par, o valor da mediana não é um dos valores da amostra, mas sim uma média dos dois valores "centrais". Isto significa que a mediana assim definida não pode ser calculada com a função `Quantile` com os parâmetros por defeito; para a calcular, se $\mathbf{x}$ designar a "lista"com a amostra, teremos de usar a função `Quantile` com a seguinte escolha de parâmetros : `Quantile`$[\mathbf{x}, 1/2, \{\{1/2, 0\}, \{0, 1\}\}]$.
Uma situação idêntica se passa com os quartis: os quartis dados com a função `Quartiles` só coincidem com os dados pela função `Quantile` se usarmos `Quantile`$[\mathbf{x}, 1/4, \{\{1/2, 0\}, \{0, 1\}\}]$, `Quantile`$[\mathbf{x}, 1/2, \{\{1/2, 0\}, \{0, 1\}\}]$ e `Quantile`$[\mathbf{x}, 3/4, \{\{1/2, 0\}, \{0, 1\}\}]$.
De notar que a função `Quartiles` dá como resultado os três quartis.

A **moda amostral** é o valor (ou valores) da amostra que ocorre com maior frequência. Pode haver uma ou mais modas amostrais. No primeiro caso, dizemos que os dados são **unimodais**, dizendo-se **pluriomodais** no segundo caso.

Todas as medidas referidas acompanham a mudança de localização dos dados, i.e. se forem considerados novos dados obtido por translação

$$y_k = x_k + b,$$

tem-se, por exemplo,

$$\bar{y} = \bar{x} + b,$$

o mesmo se passando com as restantes medidas (moda, mediana, quantis).

Se os dados sofrerem uma transformação linear, $y_k = ax_k + b$, então a média, a moda e a mediana acompanham essa transformação; por exemplo, tem-se

$$\bar{y} = a\bar{x} + b.$$

2.3.2 Medidas de dispersão

No que se segue, continuamos a considerar uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$ e a respectiva amostra ordenada $(x_{(1)}, \ldots, x_{(n)})$.

A **variância amostral** (ou variância da amostra) que representamos por var($\mathbf{x}$) ou s^2 é a média (corrigida) dos quadrados dos desvios dos dados em relação à média amostral, sendo dada pela fórmula

$$\text{var}(\mathbf{x}) = s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2. \tag{2.3}$$

Nota: A razão de considerarmos a média corrigida – isto é, de dividirmos por $n-1$ e não por n – será explicada posteriormente; note-se, no entanto que, para valores de n grandes, o uso de n ou $n-1$ no denominador fará pouca diferença. Por vezes, define-se a variância pela fórmula $\sum_{k=1}^{n}(x_k - \bar{x})^2$, sendo a expressão calculada pela fórmula (2.3) designada por *variância corrigida*.

Note-se que quanto mais concentrados estiverem os dados em torno do valor médio $\bar{x}$, mais pequena será a variância amostral e quanto mais dispersos estiverem, maior será a variância amostral.

Uma vez que a unidade de medida da variância não é a mesma dos dados (por exemplo, se os dados estiverem expressos em cm, a variância virá expressa em cm^2) é usual recorrer-se à sua raiz quadrada, s, que já se expressa na mesma unidade de medida. O valor s é chamado **desvio padrão amostral** (ou desvio padrão da amostra), sendo, portanto, dado pela fórmula

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

A **amplitude amostral** (ou amplitude da amostra), R é, tal como referimos anteriormente, a diferença entre o valor máximo e o valor mínimo da amostra, isto é, é dada por

$$R = x_{(n)} - x_{(1)}. \tag{2.4}$$

A **amplitude inter-quartis** da amostra, AIQ, é a diferença entre o terceiro e o primeiro quartis amostrais, isto é, é dada por

$$AIQ = q_{3/4} - q_{1/4}. \tag{2.5}$$

O **coeficiente de dispersão** da amostra (definido apenas quando a média é positiva) é definido como o quociente entre o desvio padrão e a média, isto é, é dado por

$$\frac{s}{\bar{x}}, \quad \bar{x} > 0. \tag{2.6}$$

Quando o coeficiente de dispersão é expresso em percentagem, obtemos o chamado **coeficiente de variação**, CV:

$$CV = \frac{s}{\bar{x}} \times 100\%, \quad \bar{x} > 0. \tag{2.7}$$

O Mathematica dispõe de funções próprias `Min`, `Max`, `InterQuartileRange`, `Variance`, `StandardDeviation` para calcular, respectivamente o valor mínimo, o valor máximo (e, portanto, facilmente calcular a amplitude da amostra), a amplitude inter-quartis, a variância e o desvio-padrão, respectivamente.

Quando os dados sofrem uma transformação linear, $y_k = ax_k + b$, têm-se as seguintes expressões para a variância e desvio-padrão dos dados transformados:

$$\mathrm{var}(\mathbf{y}) = s_y^2 = a^2 s_x^2 = a^2\,\mathrm{var}(\mathbf{x})$$

e

$$s_y = |a| s_x.$$

Note-se que a constante aditiva (que está associada a uma mudança de localização, ou mudança de origem, dos dados) não afecta a variância nem o desvio padrão; já a constante multiplicativa (associada a uma mudança de escala nos dados) afecta a variância e o desvio-padrão.

Outliers

Designam-se por **outliers** as observações demasiado afastadas dos 50% valores centrais da amostra (ordenada). Mais precisamente, um dado da amostra é considerado um **outlier** se estiver situado fora do intervalo

$$(q_{1/4} - 1.5AIQ, q_{3/4} + 1.5AIQ),$$

em que AIQ é a amplitude inter-quartis, dada pela fórmula (2.5), sendo considerado um **outlier severo** se estiver situado fora do intervalo

$$(q_{1/4} - 3AIQ, q_{3/4} + 3AIQ).$$

Um *outlier* não severo é também chamado um *outlier* **moderado**. Os *outliers* são elementos a que devemos dar especial atenção, porque podem desvirtuar totalmente uma análise estatística. A primeira coisa a fazer é ver se não houve erros de registo; se tal não aconteceu, será aconselhável fazer uma análise com e sem esses dados, de forma a avaliar o efeito que eles têm na análise e na interpretação dos resultados.

Diagramas de caixas-com-bigodes

O chamado **diagrama de caixa-com-bigodes** ou **diagrama de extremos-e-quartis** é um gráfico com duas caixas centrais limitadas pelos três quartis e umas linhas que se prolongam para fora das

caixas até ao valor mínimo e máximo da amostra (ou, se quisermos evidenciar os *outliers*, até ao menor e maior valores que não sejam *outliers*, sendo os *outliers* identificados por símbolos próprios).

O Mathematica tem uma função específica para desenhar diagramas de caixas-com-bigodes – `BoxWhiskerChart` – a qual será usada nas aulas práticas.

Exemplo 2.4. *Considere-se a seguinte amostra*

$$\mathbf{x} = \{1, 700, 800, 1100, 1200, 1000, 900, 1000, 2500\},$$

à qual corresponde a amostra ordenada

$$(1, 700, 800, 900, 1000, 1000, 1100, 1200, 2500).$$

Neste caso, tem-se $\min_i\{x_i\} = 1$ *e* $\max_i\{x_i\} = 2500$*. Os quartis (calculados usando a função* `Quartiles`*) são* $q_{1/4} = 775$*,* $q_{1/2} = 1000$ *e* $q_{3/4} = 1125$*, pelo que a amplitude inter-quartis é* $AIQ = 350$*. Vemos, então, que os outliers são os valores* 1 *(outlier moderado) e* 2500 *(outlier severo). Nas figuras 2.6 e 2.7 são apresentados exemplos de dois diagramas de caixas-com-bigodes, feitos usando a função* `BoxWhiskerChart` *do Mathematica, relativos a esta amostra. No primeiro, não há indicação dos outliers, sendo estes apresentados no segundo. Note-se que o Mathematica usa símbolos diferentes ("pontos"de intensidade de cinzento diferentes) para distinguir os outliers moderados (mais escuros) dos outliers severos (mais claros).*

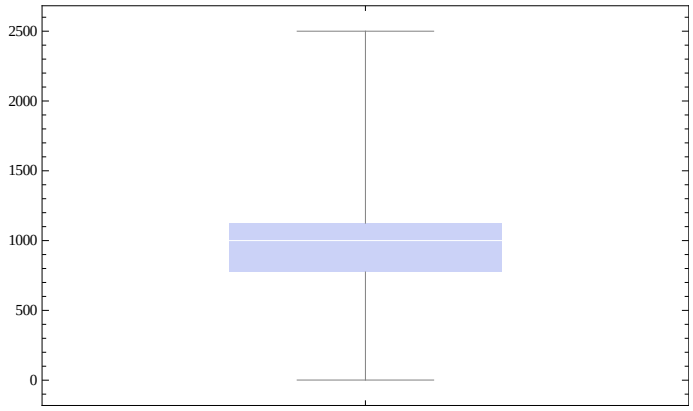

Figura 2.6: Diagrama de caixa-com-bigodes relativo ao Exemplo 2.4 (sem indicação de *outliers*).

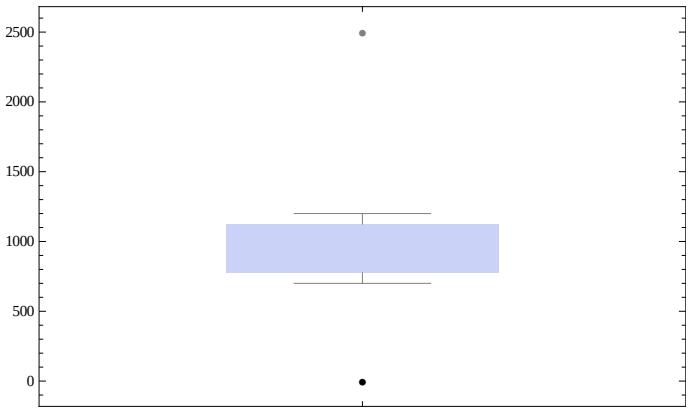

Figura 2.7: Diagrama de caixa-com-bigodes relativo ao Exemplo 2.4 (com indicação de *outliers*).

2.3.3 Medidas de forma

A forma de uma distribuição dos dados é medida em duas perspectivas: assimetria e achatamento.

Começamos por definir o que se entende por **momento central (empírico) de ordem** r, denotado por m_r:

$$m_r = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^r, \quad \text{para } r = 1, 2, \ldots. \tag{2.8}$$

Note-se que $m_1 = 0$ (verifique!) e que m_2 é a variância amostral (não corrigida, ou seja, obtida dividindo por n e não $n-1$).

O **coeficiente de assimetria (empírico)** é dado por

$$b_1 = \frac{m_3}{(\sqrt{m_2})^3}. \tag{2.9}$$

Nota: A divisão por $(\sqrt{m_2})^3$ é feita para uniformizar esta estatística, a qual deixa, assim, de ser expressa em unidades.

O **coeficiente de achatamento** ou **de curtose empírico** é dado por

$$b_2 = \frac{m_4}{m_2^2}. \tag{2.10}$$

O coeficiente de assimetria dá indicação sobre o peso relativo das *caudas* (direita e esquerda); ver Figura 2.8. Note-se que a assimetria pode ser negativa, positiva ou nula. Nas distribuições assimétricas unimodais com cauda direita prolongada, tem-se, geralmente

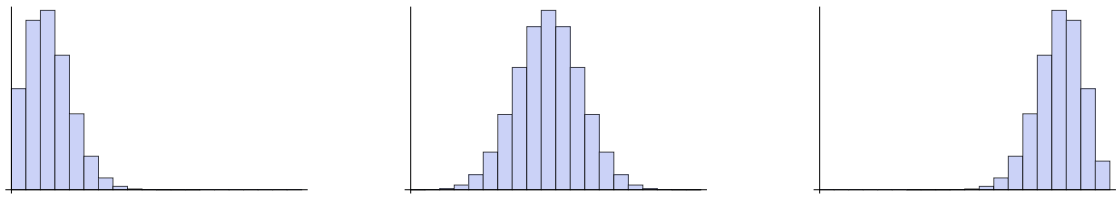

Figura 2.8: Histogramas com diferentes assimetrias: Esquerda: $b_1 > 0$; Centro: $b_1 = 0$; Direita: $b_1 < 0$

$$\text{moda} < \text{mediana} < \text{média (e } b_1 > 0)$$

e, nas com cauda esquerda prolongada, tem-se

$$\text{média} < \text{mediana} < \text{moda (e } b_1 < 0).$$

O coeficiente de achatamento ou curtose dá indicação sobre o maior ou menor achatamento (ou, de outro modo, sobre a maior ou menor concentração à volta do valor médio, ou ainda, se quisermos, sobre a existência de caudas "leves"ou de caudas "pesadas"). Note-se que o achatamento é sempre positivo.

Dizemos que distribuição dos dados é **platicúrtica**[3], **mesocúrtica** ou **leptocúrtica**[4], conforme seja $\beta_1 < 3$, $\beta_1 = 3$ ou $\beta_1 > 3$, respectivamente. Note-se que só faz sentido falar de achatamento para distribuições que sejam (quase) simétricas. [5] [6] [7]

[3] *plati...* – elemento de origem grega de composição de palavras que exprime a ideia de *plano*, *chato*, *dilatado*; do grego *platýs*, "largo", "chato".

[4] *lepto...* – elemento de origem grega de composição de palavras que exprime a ideia de *delgado*, *ténue*, *subtil*; do grego *leptós*, "delgado"

[5] A razão da classificação das distribuições, no que respeita ao achatamento, depender da comparação do coeficiente de curtose com o valor 3, será entendida mais à frente, quando falarmos na chamada *curtose empírica* e mostramos que, para a distribuição normal ou Gaussiana (que o aluno já deve conhecer bem do Ensino Secundário), ela tem precisamente o valor 3; assim, de certo modo, a classificação depende de uma comparação com a curtose da Gaussiana.

[6] Os nomes de *leptocúricas* – no sentido de "delgadas"e platicúrticas – no sentido de "achatadas"foram originalmente escolhidos porque em geral, para a maioria das distribuições simétricas, as leptocúrticas são de facto mais "delgadas"e as platicúrticas mais "achatadas"; veja a Figura 2.9.

[7] Na prática, consideramos como mesocúrticas distribuições para as quais se tenha $\beta_1 \approx 3$, chamando lapticúrticas aquelas para as quais β_2 é claramente inferior a 3 e leptocúrticas aquelas para as quais β_2 é claramente superior a 3.

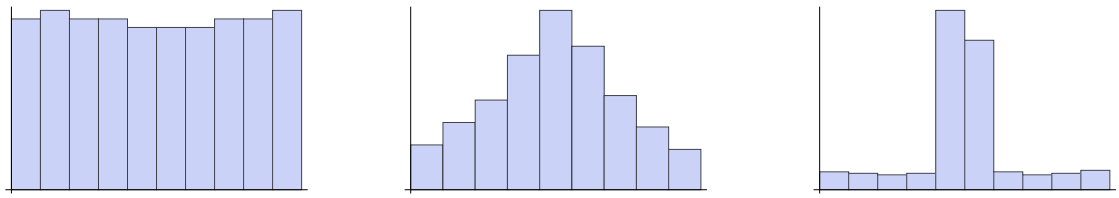

Figura 2.9: Histogramas com diferentes curtoses: Esquerda: $b_2 < 3$; Centro: $b_2 \approx 3$; Direita: $b_2 > 3$.

2.4 Dados bivariados: estudo de relações entre variáveis

Até este momento, temos estado a estudar características de dados correspondentes a uma só variável (dados univariados). Nesta secção, consideramos a relação entre duas variáveis. Por exemplo, é natural pensar que a altura e o peso de um indivíduo estão relacionados ou que o preço de um determinado produto (por exemplo, vinho) e o montante da colheita também estão relacionados.

Quando se fala de uma relação determinística, fala-se numa correspondência biunívoca entre duas variáveis. Por exemplo, o perímetro P de uma circunferência e o raio r da mesma circunferência estão relacionados; a relação que liga essas duas variáveis é definida e inalterável e pode expressar-se pela seguinte fórmula: $P = 2\pi r$.

Quando falamos numa relação estatística, como por exemplo a relação entre o peso e altura de um indivíduo, pode suceder (e sucede) que indivíduos com a mesma altura tenham pesos diferentes, mas, em média, quanto maior é altura de um indivíduo, maior é o seu peso; no caso do preço do vinho, em média, quanto maior é a colheita, menor é o preço. Assim, uma relação estatística entre duas variáveis ocupa-se da variação em média. Os fenómenos não estão ligados de forma determinística, mas a intensidade de um é acompanhada pela intensidade do outro no mesmo sentido (associação positiva) ou no sentido inverso (associação negativa).

Vamos, então, dedicar-nos ao caso em que temos uma amostra de dimensão n constituída por **pares** de observações $(x_k, y_k); k = 1, 2, \ldots, n$, em que a primeira entrada do par é relativa à medição de uma variável x no indivíduo k e a segunda entrada é relativa à medição de uma variável y no mesmo indivíduo k.

2.4.1 Dados quantitativos (em escala intervalar)

Para já, vamos supor que as variáveis em causa são quantitativas e expressas em escala (no mínimo) intervalar.

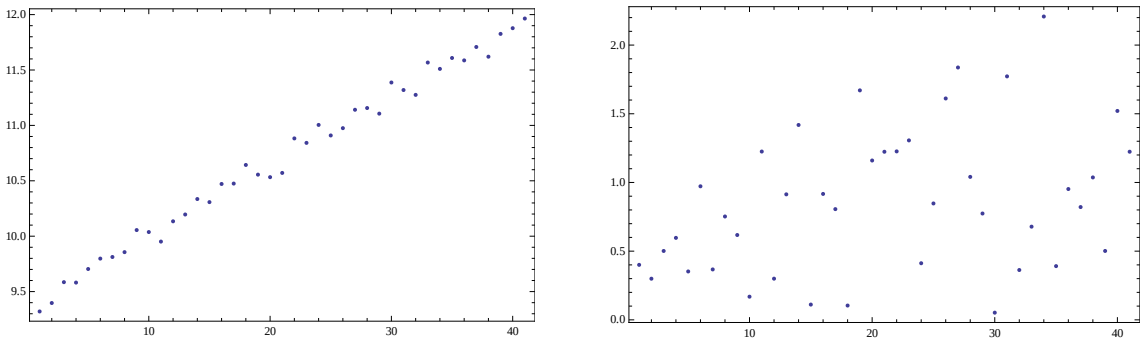

Figura 2.10: Diagramas de dispersão: Esquerda: sugere uma relação linear; Direita: não evidencia qualquer relação

Para uma primeira indicação do tipo de associação eventual entre as variáveis, é conveniente elaborar o chamado **diagrama de dispersão**, o qual é, simplesmente, a representação gráfica dos pontos (x_k, y_k) (como pontos de um plano).

No caso de os pontos do diagrama de dispersão tenderem a colocar-se aproximadamente sobre uma recta, dizemos que as variáveis estão linearmente correlacionadas. Para medir, numericamente, o grau de associação (correlação) linear entre duas variáveis podemos usar uma estatística, conhecida por **coeficiente de correlação amostral de Pearson**, o qual é definido por

$$r = \frac{1}{n-1} \sum_{k=1}^{n} \frac{x_k - \bar{x}}{s_x} \frac{y_k - \bar{y}}{s_y}, \tag{2.11}$$

onde s_x e s_y representam os desvios padrões amostrais das amostras provenientes das variáveis x e y, respectivamente.

Note-se que permutando as duas amostras, isto é, considerando a amostra emparelhada (y_k, x_k), o valor do coeficiente de correlação mantém-se inalterado. Também se mostra facilmente que este coeficiente é invariante para mudanças de localização e escala dos dados.

Pode provar-se que o coeficiente de correlação r satisfaz as seguintes propriedades:

- $-1 \leq r \leq 1$;
- $r = \pm 1$ se e só se os n pontos (x_k, y_k) estiverem sobre uma recta, isto é, se e só se existir uma relação linear perfeita entre x e y;
- se as variáveis não estiverem relacionadas, então $r = 0$;
- se $r = 0$, então não existe relação linear entre as variáveis (podendo, no entanto, existir uma relação **não linear** entre as variáveis).

Em resumo, podemos dizer que r mede o grau de **relação linear** entre as duas variáveis. Quanto mais próximo se encontra r, em valor absoluto, de 1, mais forte é a associação linear entre as variáveis. Se as variáveis não aparentam qualquer padrão ou, havendo padrão, este não for linear, então $r \approx 0$.

É importante salientar que uma grande correlação não significa necessariamente uma relação de causa e efeito; poderá, por exemplo, acontecer que ambas as variáveis estejam positivamente correlacionadas com uma terceira variável (chamada *variável de confundimento*).

2.4.2 Dados em escala ordinal

Quando pelo menos uma das variáveis está apenas em escala ordinal, utiliza-se o chamado **coeficiente de correlação ordinal de Spearman**[8], o qual coincide com o coeficiente de correlação de Pearson, mas aplicado às *ordens* dos dados. Assim, cada par (x_k, y_k) é substituído pelo par $(\text{ord}(x_k), \text{ord}(y_k))$, onde $\text{ord}(x_k)$ representa a ordem da observação x_k na colecção dos dados (com significado análogo para $\text{ord}(y_k)$) e calcula-se o respectivo coeficiente de correlação de Pearson. Pode mostrar-se que tal equivale ao uso da seguinte fórmula

$$r_S = 1 - \frac{6\sum_{k=1}^{n} d_k^2}{n(n^2-1)},$$

onde $d_k = \text{ord}(x_k) - \text{ord}(y_k)$. Tem-se, tal como no caso anterior, $-1 \leq r_S \leq 1$. Além disso, se a ordenação for totalmente concordante, teremos $d_k = 0$, para $k = 1, \ldots, n$ donde virá $r_S = 1$; se a ordenação for totalmente discordante (se uma ordenação for inversa da outra), pode mostrar-se que será $\sum_{k=1}^{n} d_k^2 = \frac{n(n^2-1)}{3}$, donde virá $r_S = -1$.

Exemplo 2.5. *[Kit98, p.177]*

Pediu-se a ambos os membros de um casal que ordenassem 10 determinados factores, na educação dos filhos, do mais importante (10) para o menos importante (1). Os dados recolhidos estão apresentados na tabela seguinte:

[8] Este coeficiente é também aplicável a dados em escala intervalar e absoluta, bastando para tal, convertê-los em escala ordinal.

Factor	*Ord. Marido*	*Ord. Mulher*
1	*6*	*6*
2	*3*	*3*
3	*1*	*2*
4	*7*	*9*
5	*2*	*1*
6	*8*	*7*
7	*4*	*5*
8	*9*	*8*
9	*5*	*4*
10	*10*	*10*

Calculemos o coeficiente de correlação de Spearman relativo a estes dados.

Como os dados estão já na forma de ordens, basta aplicar-lhes directamente a fórmula $r_S = 1 - \frac{6\sum_{k=1}^{n} d_k^2}{n(n^2-1)}$. *Neste caso, tem-se* $d_1 = 0$, $d_2 = 0$, $d_3 = -1$, $d_4 = -2$, $d_5 = 1$, $d_6 = 1$, $d_7 = -1$, $d_8 = 1$, $d_9 = 1$ *e* $d_{10} = 0$, *vindo, então*

$$r_S = 1 - \frac{6 \times (1+4+1+1+1+1+1)}{10 \times 99} = 1 - \frac{60}{990} = 0.939.$$

Como r_S *está bastante próximo de* 1, *podemos, portanto, concluir que existe uma grande concordância entre o casal sobre quais os factores mais relevantes na educação dos seus filhos.*

2.5 Análise de Regressão

Vimos como um diagrama de dispersão é um processo gráfico para detectar, visualmente, relações entre dados bivariados; vimos também como o coeficiente de correlação pode ser usado para medir a associação linear entre duas variáveis. A análise de regressão, que estudaremos agora, fornece-nos ferramentas para descrever, numericamente, relações entre variáveis, de modo a permitir fazer previsões. Ao contrário da correlação, na regressão há que distinguir entre a *variável resposta* ou *dependente* - aleatória - e a *variável independente* ou *variável preditora* (em muitas situações, controlada pelo experimentador) - não aleatória, que supomos medida sem erro. Em geral, designamos a variável dependente por y e a variável independente por x. Como exemplo, consideremos uma experiência laboratorial em que são administradas doses x_k (escolhidas, e portanto, controladas e não aleatórias) de um certo medicamento, em diversos animais, e se medem determinadas respostas y_k da administração dessas doses de medicamento. Neste caso, não faz sentido trocar os papéis de x e y, sendo indispensável distinguir entre a variável independente ou preditora x e a variável dependente ou resposta, y.

O diagrama de dispersão pode evidenciar alguma relação funcional entre x e y. Quando tentamos "encontrar"essa relação funcional, isto é, quando ajustamos um modelo $y = f(x)$, falamos de **regressão de** y **em** x (ou de y sobre x).

2.5.1 Regressão linear simples

No caso da regressão linear simples pressupõe-se haver uma relação de linearidade entre as variáveis x e y. Existe assim um *modelo* da forma $\hat{y} = a + bx$, considerando-se que os valores observados y_k são flutuações amostrais (com erro) em torno dos valores fornecidos pelo modelo – chamados **valores previstos** ou **valores ajustados** de y – isto é dos valores

$$\hat{y}_k = a + bx_k. \tag{2.12}$$

O **erro de predição**, **desvio** ou **resíduo** e_k correspondente à k-ésima observação é a diferença entre o valor observado y_k e o valor previsto $\hat{y}_k$, isto é, é dado por

$$e_k = y_k - \hat{y}_k. \tag{2.13}$$

Um dos critérios mais usados para encontrar a recta $\hat{y} = a + bx$ que "melhor"se ajusta aos dados é usar o chamado processo dos **mínimos quadrados**: neste caso, os valores dos parâmetros a e b que definem a recta $\hat{y} = a + bx$ são determinados de forma a que seja minimizada a soma dos quadrados dos resíduos SSE.

Note-se que o valor absoluto do resíduo $|e_k|$ é a distância entre o ponto (x_k, y_k) e o ponto, $(x_k, \hat{y}_k)$, isto é, o ponto alinhado com este, na vertical, mas situado sobre a recta. Assim, ao minimizarmos a soma total dos quadrados dos desvios, estaremos a minimizar a soma total dos quadrados das distâncias, medidas na vertical, dos pontos (x_k, y_k) à recta $a + bx$; ver Figura 2.11.

Pode mostrar-se que este problema de minimização tem por solução os valores de a e b dados por

$$a = \bar{y} - b\bar{x}, \tag{2.14}$$

$$b = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2} = \frac{\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}}{(n-1)s_x^2}. \tag{2.15}$$

A recta $\hat{y} = a + bx$ assim obtida é chamada **recta de regressão** de y em x.

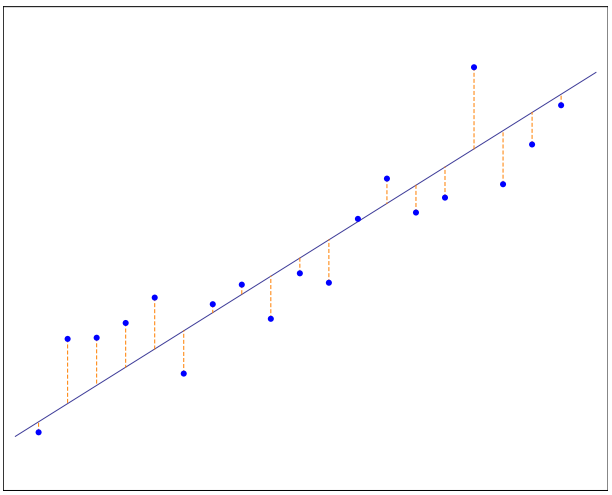

Figura 2.11: Recta de regressão

Propriedades da recta de regressão

Facilmente se verificam as seguintes propriedades.

RR1 A recta de regressão passa pelo ponto $(\bar{x}, \bar{y})$, isto é, tem-se

$$\bar{y} = a + b\bar{x} \tag{2.16}$$

RR2 O valor de b pode ser obtido através de

$$b = r\frac{s_y}{s_x}, \tag{2.17}$$

onde r é dado pela fórmula (2.11) do coeficiente de correlação de Pearson para os dados (x_k, y_k).

RR3

$$\sum_{k=1}^{n} e_k = \sum_{k=1}^{n}(y_k - \hat{y}_k) = 0. \tag{2.18}$$

RR4

$$\sum_{k=1}^{n} y_k = \sum_{k=1}^{n} \hat{y}_k. \tag{2.19}$$

2.5.2 Qualidade do ajustamento

Representação gráfica dos desvios

Tendo encontrado a recta de regressão para os dados, a próxima questão que, naturalmente se levanta, é a de tentar aferir a qualidade do ajustamento obtido.

Uma vez que os resíduos medem a discrepância entre a recta e os dados observados, um gráfico de dispersão dos valores (x_k, e_k) pode ajudar a pôr em evidência as eventuais deficiências de considerarmos $a + bx$ como modelo para a relação entre y e x.

A inferência estatística baseada neste modelo assenta no pressuposto que os erros de ajustamento têm um comportamento aleatório normal, com valor médio nulo, não estão correlacionados e têm variância constante. Por isso, na representação gráfica dos resíduos:

- não devem existir padrões ou tendências, devendo a distribuição dos pontos no plano ter um aspecto aleatório;
- os pontos devem estar dispostos (ao acaso) numa banda horizontal (uma vez que se espera variância constante para os desvios) centrada no eixo dos xx (uma vez que se espera uma média nula para os desvios).

Coeficiente de determinação

Comecemos por observar que, uma vez que a recta de regressão passa pelo ponto $(\bar{x}, \bar{y})$, se tem

$$\hat{y}_k - \bar{y} = a + bx_k - \bar{y} = a + bx_k - (a + b\bar{x}) = b(x_k - \bar{x}).$$

Mas, então, tem-se

$$\begin{aligned}\sum_{k=1}^{n}(y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) &= \sum_{k=1}^{n}\left(y_k - \bar{y} - b(x_k - \bar{x})\right)\left(b(x_k - \bar{x})\right) \\ &= b\left(\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) - b\sum_{k=1}^{n}(x_k - \bar{x})^2\right) = 0,\end{aligned}$$

onde, na última igualdade se usou a expressão (2.15) do coeficiente b. Então, temos

$$\begin{aligned}\sum_{k=1}^{n}(y_k-\bar{y})^2 &= \sum_{k=1}^{n}(y_k-\hat{y}_k+\hat{y}_k-\bar{y})^2\\ &= \sum_{k=1}^{n}(y_k-\hat{y}_k)^2+\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{y})\\ &= \sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2+\sum_{k=1}^{n}(y_k-\hat{y}_k)^2\end{aligned}$$

A equação anterior costuma escrever-se como

$$SST = SSA + SSE \tag{2.20}$$

em que SST, SSA e SSE representam, respectivamente, a soma de quadrados total ($\sum_{k=1}^{n}(y_k-\bar{y})^2$), a soma dos quadrados devida ao ajustamento ($\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2$) e a soma de quadrados devida ao erro ($\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$). Então, tem-se

- no caso de um ajuste perfeito (relação linear perfeita), será $SSE = 0$, donde
$$\frac{SSE}{SST} = 0 \quad \text{e} \quad 1-\frac{SSE}{SST} = 1$$
- num ajustamento totalmente desadequado (ausência total de relação linear), será $SSA = 0$, donde
$$\frac{SSE}{SST} = 1 \quad \text{e} \quad 1-\frac{SSE}{SST} = 0$$
- num ajustamento intermédio (relação linear imperfeita), terse-á $SSA \neq 0$ e $SSE \neq 0$, donde
$$0 < \frac{SSE}{SST} < 1 \quad \text{e} \quad 0 < 1-\frac{SSE}{SST} < 1$$

Mas,

$$\begin{aligned}1-\frac{SSE}{SST} = \frac{SSA}{SST} &= \frac{\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2}{\sum_{k=1}^{n}(y_k-\bar{y})^2}\\ &= \frac{b^2\sum_{k=1}^{n}(x_k-\bar{x})^2}{\sum_{k=1}^{n}(y_k-\bar{y})^2} = b^2\frac{s_x^2}{s_y^2} = r^2,\end{aligned}$$

onde r é o coeficiente amostral de Pearson (veja a fórmula (2.17)) para os dados (x_k, y_k). Assim, r^2, que varia entre 0 e 1, mede o grau de linearidade dos dados e chama-se **coeficiente de determinação**.

Este é tanto maior quanto mais o modelo linear se adequa aos dados. Em resumo, temos a seguinte fórmula para o coeficiente de determinação

$$r^2 = 1 - \frac{SSE}{SST}. \tag{2.21}$$

Dividindo ambos os membros da equação (2.20) por $n-1$, obtém-se

$$s_y^2 = s_{\hat{y}}^2 + s_e^2, \tag{2.22}$$

em que s_y^2 representa a variância total da amostra dos y_k, $s_{\hat{y}}^2$ representa a variância explicada pelo ajustamento (i.e. pela regressão linear de y em x) e s_e^2 representa a variância residual, devida a erro (note-se que $\sum_{k=1}^n (y_k - \hat{y}_k)^2 = \sum_{k=1}^n e_k^2 = \sum_{i=1}^n (e_k - \bar{e})^2$, uma vez que $\bar{e} = 0$, pelo que $\frac{1}{n-1}(y_k - \hat{y}_k)^2$ é de facto a variância dos desvios e_k). Temos também

$$r^2 = \frac{\sum_{k=1}^n (\hat{y}_k - \bar{y})^2}{\sum_{k=1}^n (y_k - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2},$$

o que mostra que r^2 representa a fracção (ou percentagem) da variância total que é devida ao ajustamento.

Exemplo 2.6. *Considere-se a seguinte amostra de pares* (x_k, y_k)*:*

$$\{(0.1, 2.51649), (0.2, 2.64119), (0.3, 2.7158), (0.4, 2.8884), (0.5, 2.99668),$$
$$(0.6, 3.10415), (0.7, 3.19486), (0.8, 3.31053), (0.9, 3.52461), (1., 3.57375),$$
$$(1.1, 3.68104), (1.2, 3.89518), (1.3, 3.98911), (1.4, 4.05582), (1.5, 4.24346),$$
$$(1.6, 4.33153), (1.7, 4.42073), (1.8, 4.55742), (1.9, 4.72277), (2., 4.86406)\}.$$

Na Figura 2.12 apresenta-se o diagrama de dispersão dos dados, o qual evidencia uma relação linear entre x *e* y*.*

Neste caso, calculando a recta de regressão $\hat{y} = a + bx$*, usando as fórmulas (2.14) e (2.15), obtém-se*

$$\hat{y} = 2.37366 + 1.2264x.$$

Na Figura 2.13 apresentam-se novamente os pontos (x_k, y_k)*, sobrepondo, no mesmo gráfico, a recta de regressão acima obtida, sendo patente o "bom"ajustamento da recta aos pontos considerados.*

Na Figura 2.14 estão representados os pontos (x_k, e_k) *onde* $e_k = y_k - \hat{y_k}$*.*

Neste caso, o valor do coeficiente de determinação é $r^2 = 0.9981$ *(valor muito próximo de 1).*

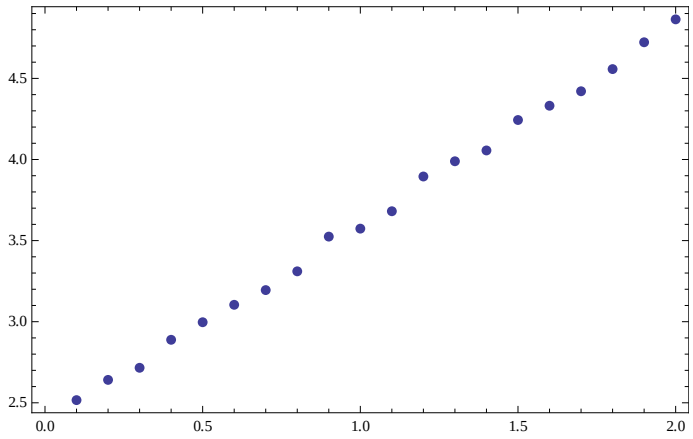

Figura 2.12: Diagrama de dispersão dos dados do Exemplo 2.6.

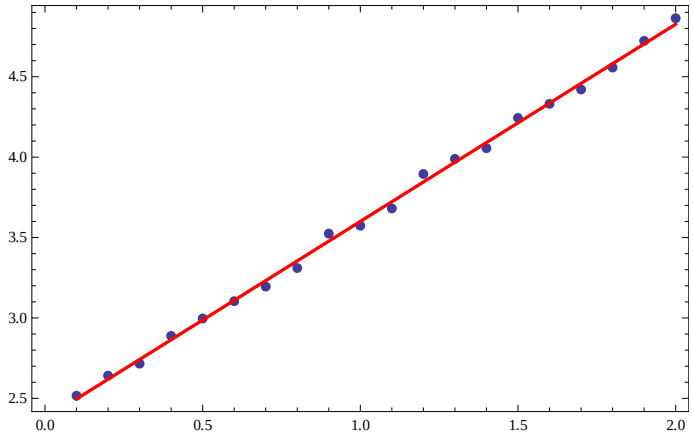

Figura 2.13: Recta de regressão para os dados do Exemplo 2.6.

2.6 Outros modelos

A regressão linear simples $\hat{y} = a + bx$ insere-se no caso mais geral de um **modelo linear**, isto é, de um modelo da forma

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots$$

(Trata-se de um modelo **linear nos parâmetros** $a, b_1, b_2, \ldots$). Nesta fórmula, $x_1, x_2, \ldots$ podem ser diferentes variáveis (teremos um modelo de regressão *linear múltipla*), podem ser funções de uma

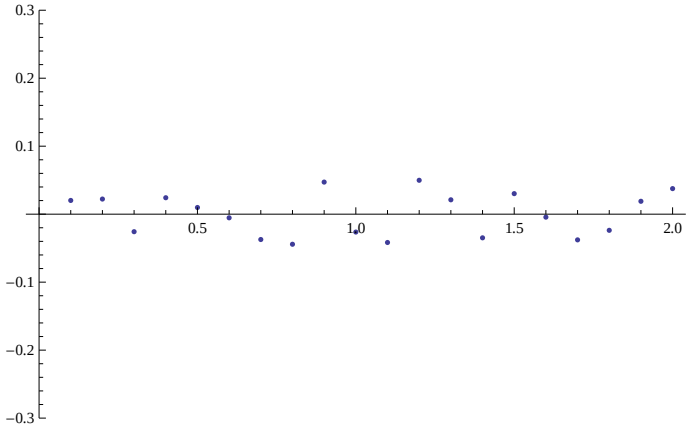

Figura 2.14: Gráfico dos desvios relativos à recta de regressão do Exemplo 2.6.

mesma variável (dizemos então que temos regressão *curvilínea*), por exemplo

$$\hat{y} = a + b_1 x + b_2 x^2,$$

podendo ainda ter-se uma combinação dos dois casos, por exemplo,

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 \cos x_2.$$

Todos estes casos são resolvidos de forma semelhante à regressão linear simples, pelo critério dos mínimos quadrados, sendo os parâmetros $a, b_1, b_2, \ldots$ determinados de forma a minimizar a soma dos quadrados dos desvios $\sum_{k=1}^{n} e_k^2 = \sum_{i=1}^{n} (\hat{y}_k - y_k)^2$. Um caso particularmente importante deste tipo de modelos é o da **regressão polinomial**, em que se procura ajustar aos dados um polinómio de um determinado grau m,

$$\hat{y} = a + b_1 x + b_2 x^2 + \ldots + b_m x^m.$$

Também é possível (embora seja um problema de mais difícil resolução) ajustar um **modelo não linear**, ou seja, um modelo da forma

$$\hat{y} = f(x, a, b_1, b_2, \ldots)$$

em que $a, b_1, b_2, \ldots$ são parâmetros e f é uma função *não linear* desses parâmetros. Por exemplo, um modelo desse tipo será

$$\hat{y} = a e^{bx},$$

(modelo de crescimento/decrescimento) exponencial). Outros modelos deste tipo serão estudados nas aulas práticas. Veremos também que, para grande parte dos modelos, o Mathematica pode ser usado com relativa facilidade para determinar os parâmetros. Há que ter em atenção, no entanto, que no caso de um modelo não linear, a determinação dos parâmetros envolve resolução de equações não lineares, as quais passam pelo uso de métodos iterativos (e, por vezes, necessitam de uma indicação de uma aproximação inicial "razoável"para os parâmetros.)

De seguida apresenta-se o código *Mathematica* que gerou tudo aquilo que compõe cada um dos exemplos deste capítulo.

---

exemplo 2.1

- **dados**

```
dados = {28, 60, 160, 120, 32};
```

- **gráfico de barras**

```
BarChart[dados, ChartStyle → "DarkRainbow",
 ChartLegends → {"Muito Insatisfeito ", "Insatisfeito ",
   "Razoavelm . Satisfeito ", "Satisfeito ", "Muito Satisfeito "},
 AxesLabel → {None, "Frequência"}, ImageSize → 400]
```

- **um segundo gráfico de barras**

```
BarChart[dados, ChartLabels → {"Muito Insatisfeito ", "Insatisfeito ",
   "Razoavelm . Satisfeito ", "Satisfeito ", "Muito Satisfeito "},
 BarOrigin → Left, AxesLabel → {"Frequência"}, ImageSize → 500]
```

- **diagrama circular**

```
PieChart[dados, ChartLabels → {"Muito Sat. \n (8%)",
   "Satisfeito \n (30%)", "Razoavelm . Sat. \n (40%)",
   "Insatisfeito \n (15%)", "Muito Ins.\n (7%)"}, ImageSize → 400]
```

Atente-se na utilização do comando *nova linha*, $\backslash n$, na opção `ChartLabels` do comando `PieChart`, que permitiu escrever duas linhas de informação.

Como alternativa ao comando `PieChart`, experimente usar o comando `PieChart3D` para gerar o gráfico circular.

exemplo 2.2

- **dados**

```
dados = {1, 1, 3, 2, 0, 1, 0, 3, 1, 1, 4, 0, 2, 2, 1, 1, 0, 1, 1, 1};
```

- **tabela de frequências**

```
Union[dados]
```

```
tabelaDados = Table[
  {k, Count[dados, k], N[100 * Count[dados, k] / Length[dados], 3]}, {k, 0, 4}]
```

```
Grid[Prepend[tabelaDados,
  {"Nº irmãos", "freq. absoluta", "freq. relativa (%)"}],
 Frame → All, Spacings → {3, 1}]
```

- **gráfico de linhas**

```
contagens = Tally[dados]
```

```
ListPlot[contagens, AxesOrigin → {-1, 0}, Filling → Axis,
 FillingStyle → Thick, AxesLabel → {"nº irmãos", "freq"}, ImageSize → 600]
```

As contagens efectuadas inicialmente para a construção da tabela de frequências são feitas mediante a informação dos valores distintos presentes nos dados. Por outras palavras, só depois de sabermos que nos dados temos apenas os valores $0, 1, 2, 3$ e 4, é que vamos proceder à contagem do número das respectivas repetições, afirmando que a variável discreta `k` deve tomar valores inteiros de 0 a 4.

Para a obtenção do gráfico de linhas das frequências absolutas dos dados foi utilizado um comando muito específico, `Tally`, que, de uma forma automática, efectua a contagem do número de repetições de cada valor presente nos dados. Observe o aspecto curioso das contagens obtidas pelo comando `Tally` surgirem por ordem do aparecimento dos valores nos dados e não por ordem crescente desses valores.

exemplo 2.3

- **dados**

```
dados = {193, 193, 171, 197, 197, 177, 186, 180, 187, 163, 169, 176, 171, 183, 168,
   168, 183, 189, 175, 176, 158, 167, 182, 165, 180, 178, 186, 185, 182, 168,
   170, 202, 200, 194, 165, 172, 173, 174, 181, 188, 194, 172, 174, 173,
   187, 172, 194, 167, 193, 187, 182, 170, 184, 166, 171, 176, 188, 169,
   180, 170, 168, 191, 194, 196, 173, 167, 184, 166, 180, 166, 163, 184,
   173, 161, 173, 158, 187, 184, 184, 177, 192, 161, 167, 169, 197, 182,
   160, 200, 188, 201, 188, 199, 176, 196, 191, 166, 192, 187, 188, 175};
```

- **agrupar os dados em classes**

```
Length[dados]
```

```
2^6 < Length[dados] < 2^7
```

```
Min[dados]
```

```
Max[dados]
```

```
(Max[dados] - Min[dados]) / 8
```

- **tabela de frequências**

```
freqAbs = BinCounts[dados, {156, 204, 6}]
```

```
Grid[{{"classes", "freq. absolutas", "freq. relativas (%)"},
  {"[156, 162)", freqAbs[[1]], N[100 * freqAbs[[1]] / Length[dados], 3]},
  {"[162, 168)", freqAbs[[2]], N[100 * freqAbs[[2]] / Length[dados], 3]},
  {"[168, 174)", freqAbs[[3]], N[100 * freqAbs[[3]] / Length[dados], 3]},
  {"[174, 180)", freqAbs[[4]], N[100 * freqAbs[[4]] / Length[dados], 3]},
  {"[180, 186)", freqAbs[[5]], N[100 * freqAbs[[5]] / Length[dados], 3]},
  {"[186, 162)", freqAbs[[6]], N[100 * freqAbs[[6]] / Length[dados], 3]},
  {"[156, 162)", freqAbs[[7]], N[100 * freqAbs[[7]] / Length[dados], 3]},
  {"[156, 162)", freqAbs[[8]], N[100 * freqAbs[[8]] / Length[dados], 3]}},
 Frame → All, Spacings → {3, 1}]
```

- **histograma**

```
marcasClasse = Table[159 + 6 (k - 1), {k, 1, 8}]
```

```
graf1 = Histogram[dados, {156, 204, 6}, Automatic,
  LabelingFunction → Above, Ticks → {marcasClasse, None}, ImageSize → 500]
```

- **histograma com polígono de frequências**

```
poligono = Append[Prepend[freqAbs, 0], 0];
abcissas = Append[
   Prepend[marcasClasse, First[marcasClasse] - 6], Last[marcasClasse] + 6];
```

```
graf2 = ListLinePlot[Table[{abcissas[[n]], poligono[[n]]},
    {n, 1, Length[poligono]}], PlotStyle → {Red, Thick}];
```

```
Show[{graf1, graf2}, PlotRange → All, ImageSize → 500]
```

A construção da tabela de frequências foi efectuada com a introdução explícita de todos os seus elementos. Naturalmente que tal forma de proceder só é admissível quando o número de classes não seja muito elevado, caso contrário ter-se-ia que utilizar o comando `Table`.

exemplo 2.4

- **dados**

```
dados = {1, 700, 800, 1100, 1200, 1000, 900, 1000, 2500};
```

- **medidas de localização**

```
N[Mean[dados], 6]
```

```
Median[dados]
```

```
Commonest[dados]
```

```
Quartiles[dados]
```

```
Quantile[dados, 0.1, {{1 / 2, 0}, {0, 1}}]
```

- **medidas de dispersão**

```
Variance[dados]
```

```
N[StandardDeviation[dados], 5]
```

```
Max[dados] - Min[dados]
```

```
InterquartileRange[dados]
```

```
N[StandardDeviation[dados] / Mean[dados], 3]
```

- **diagramas de caixas-com-bigodes**

```
BoxWhiskerChart[dados, ImageSize → 500]
```

```
BoxWhiskerChart[dados, "Outliers", ImageSize → 500]
```

exemplo 2.5

- **coeficiente de Spearman**

```
marido = {6, 3, 1, 7, 2, 8, 4, 9, 5, 10}
```

```
mulher = {6, 3, 2, 9, 1, 7, 5, 8, 4, 10}
```

```
ds = (marido - mulher)^2
```

$$\mathrm{N}\left[1 - \frac{6\,\mathrm{Total}[\mathrm{ds}]}{\mathrm{Length}[\mathrm{ds}]\left(\mathrm{Length}[\mathrm{ds}]^2 - 1\right)}, 4\right]$$

exemplo 2.6

- **dados**

```
dados = {{0.1, 2.51649}, {0.2, 2.64119}, {0.3, 2.71580}, {0.4, 2.88840},
   {0.5, 2.99668}, {0.6, 3.10415}, {0.7, 3.19486}, {0.8, 3.31053},
   {0.9, 3.52461}, {1.0, 3.57375}, {1.1, 3.68104}, {1.2, 3.89518},
   {1.3, 3.98911}, {1.4, 4.05582}, {1.5, 4.24346}, {1.6, 4.33153},
   {1.7, 4.42073}, {1.8, 4.55742}, {1.9, 4.72277}, {2.0, 4.86406}}
```

- **recta de regressão**

```
graf1 =
 ListPlot[dados, Frame → True, Axes → False, PlotStyle → {PointSize[0.01]}]
```

```
recta[t_] = a * x + b /. FindFit[dados, a * x + b, {a, b}, x]
```

```
graf2 = Plot[recta[x], {x, 0.1, 2.0}, PlotStyle → {Red, Thickness[0.005]}];
```

```
Show[{graf1, graf2}, ImageSize → 500]
```

- **desvios relativos à recta de regressão**

```
xxs = Transpose[dados][[1]]
```

```
yys = Transpose[dados][[2]]
```

```
desvios = yys - recta[xxs]
```

```
pontosDesvios = Transpose[{xxs, desvios}]
```

```
ListPlot[pontosDesvios, PlotStyle → {Red, PointSize[0.01]},
 PlotRange → {-0.06, 0.06}, ImageSize → 500]
```

- **quadrado do coeficiente de determinação**

```
Correlation[xxs, yys]^2
```

3

Probabilidade e Variáveis Aleatórias

Já referimos que em geral, é impossível estudar toda uma população, pelo que nos limitamos a estudar uma amostra dessa população. A partir das propriedades da amostra, *inferimos* propriedades da respectiva população. Naturalmente, as decisões ou predições sobre a população baseadas numa amostra dela retirada envolvem sempre um certo grau de incerteza. Para "medir"esta incerteza é necessário o recurso à *probabilidade*.

Neste capítulo começamos por fazer uma muito breve revisão dos principais conceitos e resultados básicos sobre teoria da probabilidade (conceitos e resultados que os alunos já estudaram no ensino secundário, mas que são aqui revistos, essencialmente com o objectivo de uniformizar as notações). De seguida, estudamos as chamadas *variáveis aleatórias*.

3.1 Probabilidade

Chama-se **experiência aleatória** a uma experiência[a] que satisfaça os três requisitos seguintes:

- pode ser repetida em condições análogas;
- conhecemos, à partida, todos os possíveis resultados que podem ocorrer;
- o resultado que obtemos em cada realização da experiência é incerto (variando com as diferentes realizações da experiência).

Por oposição às experiências aleatórias temos as experiências determinísticas, que, quando realizadas nas mesmas condições conduzem aos mesmos resultados.

[a] Isto é, um procedimento que permite a obtenção de observações.

Como exemplo típico de uma experiência aleatória, tem-se a experiência de lançamento de um dado (com as faces numeradas de 1 a 6) com a observação de qual a face que fica virada para cima.

- Ao conjunto de todos os resultados possíveis de uma experiência aleatória, chamamos **espaço amostral** ou **espaço de resultados** associado a essa experiência; este conjunto será denotado por Ω.
- Aos subconjuntos do espaço amostral Ω chamamos **acontecimentos**. Estes conjuntos são, geralmente, designados por letras maiúsculas do início do alfabeto latino, $A, B, C, \ldots$.
- O conjunto Ω (sendo um subconjunto de si próprio) é um acontecimento, chamado **acontecimento universal**.
- Ao conjunto vazio $\emptyset$ (que é um subconjunto de Ω) chamamos **acontecimento nulo**.
- Um subconjunto de Ω formado apenas por um elemento $\{\omega\}$ é chamado **acontecimento elementar**.[a]
- Ao efectuar uma realização da experiência, dizemos que que um determinado acontecimento A ocorreu ou se realizou, se o resultado da experiência foi um dos elementos de A.

[a] Por vezes, por uma questão de simplicidade, denotamos o acontecimento elementar $\{\omega\}$, simplesmente por ω.

No caso do exemplo anterior do lançamento de um dado, o espaço amostral seria o conjunto $\Omega = \{1, 2, 3, 4, 5, 6\}$. Um possível acontecimento seria o conjunto $A = \{1, 3, 5\}$ (que poderíamos descrever como "saída de face com um número ímpar"). Este acontecimento ocorreria se, ao lançar o dado, saísse, por exemplo, o número 3.

3.1.1 Definição axiomática de probabilidade

Naturalmente, gostaríamos de associar a cada acontecimento A uma "probabilidade"', que, de algum modo, medisse o "grau de certeza/incerteza"da sua ocorrência.

Numa abordagem *empírica* de probabilidade, poderíamos tentar definir a probabilidade de A do seguinte modo: repetir-se-ia a experiência n vezes (nas mesmas condições) e contar-se-ia o número m de ocorrências de A nestas n repetições da experiência; o quociente $f_n(A) = \frac{m}{n}$ é chamada *frequência relativa* de ocorrência de A. Este valor $f_n(A)$ deve ser uma aproximação para a probabilidade de A, devendo essa aproximação ser tanto melhor quanto maior for o valor de n; se observarmos, empiricamente, que $f_n(A)$ tende para um certo valor p_A quando n cresce, será natural considerar esse valor p_A como a "probabilidade de A". Esta definição frequencista de probabilidade é, no entanto, pouco rigorosa do ponto de vista matemático (de facto, o conceito de "tende para"que usámos

aqui, um pouco informalmente, não coincide com o conceito matemático associado a um limite, usado vulgarmente em análise). Na prática, contudo, interpretamos geralmente a probabilidade em termos de frequências relativas; assim, quando dizemos que um determinado acontecimento tem probabilidade 0.5, esperamos que, numa repetição da experiência em causa um grande número de vezes, A deva ocorrer aproximadamente em metade das repetições.

Note-se que a frequência relativa $f_n(A)$ satisfaz os seguintes propriedades: $f_n(A) \geq 0$ e $f_n(\Omega) = 1$; além disso, se A e B forem acontecimentos disjuntos (i.e. tais que $A \cap B = \emptyset$), teremos $f_n(A \cup B) = f_n(A) + f_n(B)$. De facto, estas três propriedades básicas das frequências relativas inspiraram a *definição axiomática* de medida de probabilidade (introduzida por Kolmogorov, 1933) e que apresentamos de seguida.

Chama-se **medida de probabilidade** a uma função que a cada acontecimento $A \subseteq \Omega$ associa um valor real, $P(A)$, designado por **probabilidade de** A, e que satisfaz os seguintes axiomas:

P1 $P(A) \geq 0$

P2 $P(\Omega) = 1$

P3 Se $A_1, A_2, \ldots$ são acontecimentos disjuntos dois a dois (isto é, se $A_i \cap A_j = \emptyset$, para $i \neq j$), então $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots$

Nota: Quando Ω é um conjunto infinito não numerável, existem alguns subconjuntos "patológicos" de Ω aos quais não é possível associar uma probabilidade; estes conjuntos são ditos não-probabilizáveis ou não-mensuráveis; em teoria elementar da probabilidade, ignoramos estes conjuntos. De aqui em diante, quando nos referirmos a um acontecimento, isto é, a um subconjunto de Ω, assumimos sempre que esse conjunto é probabilizável.

Da definição axiomática, resultam as seguintes propriedades importantes, válidas para quaisquer $A, B \subseteq \Omega$:

Prop P1 $P(\bar{A}) = 1 - P(A)$, onde $\bar{A}$ designa o complementar de A (em Ω), isto é, $\bar{A} = \{\omega \in \Omega : \omega \notin A\}$.

Prop P2 $P(\emptyset) = 0$.

Prop P3 $P(A) \leq 1$.

Prop P4 Se $A \subseteq B$, então $P(A) \leq P(B)$.

Prop P5 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

3.1.2 Definição clássica de probabilidade

No caso de Ω ser finito, com n elementos, um caso muito particular de medida de probabilidade consiste em atribuir a cada acontecimento elementar a probabilidade $\frac{1}{n}$ e a cada acontecimento A a probabilidade

$$P(A) = \frac{\text{número de elementos de } A}{n}.$$

Este quociente também se costuma escrever na forma

$$P(A) = \frac{\text{número de casos favoráveis a } A}{\text{número de casos possíveis}}.$$

Dizemos então que estamos no caso de resultados *igualmente prováveis*. Por exemplo, no caso do lançamento do dado que temos vindo a referir, esta atribuição de probabilidade corresponde ao caso de o dado ser *equilibrado* ou *perfeito* para o qual a probabilidade de sair cada face é $\frac{1}{6}$. Este caso de resultados elementares igualmente prováveis corresponde à primeira definição formal de probabilidade (definição clássica de probabilidade), introduzida por Pierre Simon de Laplace, em 1812, na sua obra *Théorie Analytique des Probabilités*.

3.2 Probabilidade condicional; acontecimentos independentes

Após a realização de uma dada experiência aleatória, suponhamos que ocorreu um determinado acontecimento B, tal que $P(B) > 0$. Um dado acontecimento A terá agora uma nova probabilidade associada (eventualmente diferente da que tinha inicialmente) chamada **probabilidade de A condicional a B** que denotaremos por $P(A|B)$, e que é definida por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.1}$$

Por exemplo, no caso do lançamento do dado, suponhamos que sabemos que ocorreu o acontecimento "saída de face far" ($B = \{2, 4, 6\}$) e que pretendemos saber a probabilidade de ter saído a face com o número 2, isto é $A = \{2\}$. Então

$$P(A|B) = \frac{P(\{2\} \cap \{2, 4, 6\})}{P(\{2, 4, 6\})} = \frac{P(\{2\})}{P(\{2, 4, 6\})} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

A fórmula (3.2) pode também ser escrita como

$$P(A \cap B) = P(A|B)P(B). \tag{3.2}$$

Se a probabilidade inicial $P(A)$ e a probabilidade condicional $P(A|B)$ coincidirem, dizemos que A é **independente** de B.

Nesse caso, tem-se

$$P(A \cap B) = P(A)P(B),$$

donde se segue (supondo que $P(A) > 0$) que

$$\frac{P(A \cap B)}{P(A)} = P(B)$$

o que significa que $P(B|A) = P(B)$, pelo que B também será independente de A, fazendo, portanto, mais sentido dizer simplesmente que A e B são **acontecimentos independentes** (um do outro). Em resumo, dizemos que dois acontecimentos A e B são independentes se

$$P(A \cap B) = P(A)P(B). \tag{3.3}$$

Dizemos que n acontecimentos $A_1, \ldots, A_n$ são mutuamente independentes se, para qualquer escolha de r (com $r = 2, 3, \ldots, n$) desses acontecimentos a probabilidade da intersecção destes for igual ao produto das respectivas probabilidades, isto é, tivermos

$$\begin{aligned}
&P(A_i \cap A_j) = P(A_i)P(A_j), \forall i, j \in \{1, \ldots, n\},\ i \neq j,\\
&P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k), \forall i, j, k \in \{1, \ldots, n\}\ \ (i, j, k \text{ distintos}),\\
&\ldots\\
&P(A_1 \cap A_2 \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n).
\end{aligned}$$

3.3 Varáveis aleatórias

Ao efectuar experiências aleatórias, os resultados podem ser descritos de várias formas. Por exemplo, na experiência do lançamento de uma moeda, os acontecimentos podem ser descritos por palavras, como "saída de cara"ou "saída de escudo"[1], ou usando, o símbolo C para "saída de cara", e E para "saída de escudo", pelos dois símbolos C e E; ao retirar uma carta de um baralho, os diversos resultados podem ser descritos por uma mistura de números e palavras (e.g. "sair o 10 de espadas"), no lançamento de um dado com as faces numeradas, os resultados podem ser descritos simplesmente pelos números $1, 2, 3, 4, 5, 6$.

Do ponto de vista de matemático, seria bem mais simples se todos os resultados das experiências estivessem associados a números (reais). Tal pode ser feito, através da atribuição de um *código*

[1] Ou Euro...

numérico para os diversos resultados da experiência. Por exemplo, no caso do lançamento da moeda, bastaria associar o código 0 para "saída de cara" ou "C" e 1 para "saída de escudo" ou "E"; no caso do baralho, bastaria numerar as diversas cartas do baralho de 1 a 52 e os diversos resultados passariam a estar associados a cada um desses números; no caso do dado, temos já números associados ao diversos resultados da experiência.

No fundo, o que pretende, é definir uma função de Ω (espaço amostral original) em $\mathbb{R}$. Tal função é chamada uma **variável aleatória**. Mais precisamente, tem-se a seguinte definição:

Chama-se **variável aleatória** (v.a.)[a] a toda a função X do espaço amostral Ω em $\mathbb{R}$, isto é

$$\begin{aligned} X :&\Omega \to \mathbb{R} \\ &\omega \mapsto X(\omega) \end{aligned}$$

[a] Associada a uma dada experiência aleatória, à qual corresponde uma certo espaço amostral Ω.

No caso anteriormente considerado do lançamento da moeda, a função X seria definida por $X(\text{C}) = 0$ e $X(\text{E}) = 1$.

Ao definirmos uma variável aleatória estamos, no fundo, a substituir o espaço amostral Ω por um novo espaço (mais familiar em matemática), o conjunto $\mathbb{R}$. Se o espaço amostral for já $\mathbb{R}$ (ou um seu subconjunto), não será necessária a introdução de qualquer função adicional, ou, se assim preferimos, podemos associar-lhe a v.a. *identidade* isto é, a aplicação de Ω em $\mathbb{R}$ tal que $X(\omega) = \omega$.

As variáveis aleatórias são vulgarmente designadas por letras maiúsculas do final do alfabeto: X, Y, etc.

Até este momento, definimos probabilidade para *acontecimentos*, isto é, para subconjuntos do espaço amostral. No entanto, estamos muitas vezes interessados em fazer afirmações envolvendo probabilidades de (valores de) *variáveis aleatórias*, por exemplo, dizer que "a probabilidade de a v.a. X assumir valores (ou "estar") entre a e b é 0.5". Tal é feito, transformando as afirmações acerca dos valores de X em afirmações acerca de subconjuntos de Ω, do seguinte modo.

Dado $B \subseteq \mathbb{R}$, definimos a "probabilidade de X assumir valores em B", e escrevemos $P(X \in B)$, como sendo a probabilidade do conjunto formado pelos elementos de Ω cuja imagem por X está em B,[2] isto é,

$$P(X \in B) = P\Big(\{\omega \in \Omega : X(\omega) \in B\}\Big). \tag{3.4}$$

[2] Este conjunto chama-se a imagem inversa de B por X e denota-se por $X^{-1}(B)$.

Exemplo 3.1. *Consideremos a experiência aleatória que consiste no lançamento de duas moedas equilibradas e seja X a v.a. que representa o "número de caras que ocorrem". Neste caso, o espaço amostral seria, por exemplo, $\Omega = \{(C,C),(C,E),(E,C),(E,E)\}$ e $X : \Omega \to \mathbb{R}$ dada por $X(C,C) = 2$, $X(C,E) = 1$, $X(E,C) = 1$ e $X(E,E) = 0$. Sendo as moedas equilibradas, cada um dos resultados possíveis, (C,C), (C,E), (E,C) e (E,E) tem probabilidade $\frac{1}{4}$ e portanto, tem-se, por exemplo*

$$\begin{aligned}
P(X=0) &= P\Big(\{\omega \in \Omega : X(\omega) = 0\}\Big) = P\Big(\{(E,E)\}\Big) = \frac{1}{4},\\
P(X=1) &= P\Big(\{\omega \in \Omega : X(\omega) = 1\}\Big) = P\Big(\{(E,C),(C,E)\}\Big) = \frac{1}{2},\\
P(1 \le X < 4) &= P\Big(X \in [1,4)\Big) = P\Big(\{\omega \in \Omega : X(\omega) \in [1,4)\}\Big)\\
&= P\Big(\{\omega : X(\omega) = 1 \text{ ou } X(\omega) = 2 \text{ ou } X(\omega) = 3\}\Big)\\
&= P\Big(\{(C,C),(E,C),(C,E)\}\Big) = \frac{3}{4}.
\end{aligned}$$

3.3.1 Variáveis discretas

Uma v.a. diz-se **discreta** se o seu contradomínio, também chamado **suporte** da v.a., for um conjunto discreto, i.e. for um conjunto $\{x_1,\ldots,x_n\}$ (finito) ou um conjunto $\{x_1,x_2,x_3\ldots\}$ (infinito, mas numerável) de pontos. No que se segue, para designar um qualquer deste tipo de conjuntos, usaremos uma notação compacta $\{x_k : k \in \mathbb{K}\}$, onde $\mathbb{K}$ designa o conjunto de todos os naturais $\mathbb{K} = \{1,2,3,\ldots\}$ ou uma sua parte inicial $\mathbb{K} = \{1,2,\ldots,n\}$.

As variáveis aleatórias discretas ocorrem em muitos problemas, especialmente em casos em que estamos interessados em contar o número de vezes que qualquer coisa acontece. Por exemplo, se X representa o número de parafusos defeituosos de um conjunto contendo 1000 parafusos, então X é uma v.a. cujos valores possíveis são $1,2,\ldots,1000$.

Função massa de probabilidade (f.m.p.)

Quando temos uma v.a. discreta, interessa-nos não só conhecer os valores que ela pode tomar, mas também a probabilidade com que assume cada um desses valores. Designa-se por **função massa de probabilidade** (f.m.p.) de uma v.a. discreta X a função que associa a cada elemento do contradomínio dessa v.a. a respectiva probabilidade. Sendo $\{x_k : k \in \mathbb{K}\}$ o contradomínio da v.a.

X, é usual designarmos por p_k a probabilidade do elemento x_k, isto é,

$$p_k = P(X = x_k).$$

É frequente explicitar-se a f.m.p. de uma v.a. discreta na forma

$$X : \begin{cases} x_k, & k \in \mathbb{K} \\ p_k \end{cases}$$

onde na primeira linha se indicam os pontos que constituem o suporte da variável e na segunda linha a probabilidade de cada um desses pontos. A f.m.p. pode também ser descrita na forma $\{(x_k, p_k) : k \in \mathbb{K}\}$ ou, se não houver dúvidas na identificação do seu suporte, simplesmente por $\{p_k : k \in \mathbb{K}\}$.

Exemplo 3.1 (cont.) *No caso do Exemplo 3.1, do lançamento das duas moedas equilibradas, a v.a. X associada ao número total de caras, teria a seguinte f.m.p.:*

$$X : \begin{cases} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{cases}$$

As f.m.p. podem também representar-se graficamente de modo idêntico aos diagramas de linhas referidos no capítulo anterior.

Exemplo 3.1 (cont.) *No caso do Exemplo 3.1, a representação gráfica da f.m.p. seria a apresentada na Figura 3.1.*

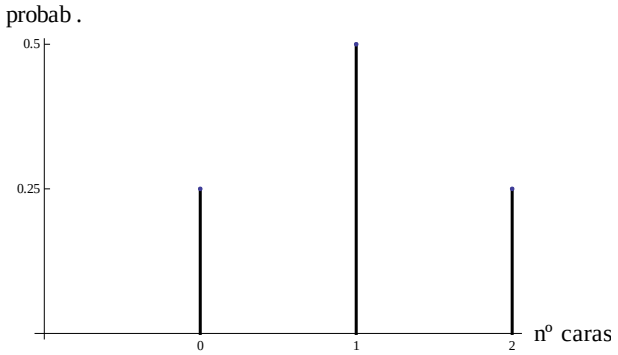

Figura 3.1: Função massa de probabilidade do Exemplo 3.1.

Uma f.m.p. deve satisfazer as seguintes propriedades:

FMP1 $\forall k \in \mathbb{K},\ p_k \geq 0.$

FMP2 $\sum_{k \in \mathbb{K}} p_k = 1.$

Além disso, facilmente se mostra que

$$P(X \in B) = \sum_{k\,:\,x_k \in B} p_k. \tag{3.5}$$

Exemplo 3.1 (cont.) *Assim, tem-se, relativamente à f.m.p. do Exemplo 3.1,*

$$P(0 \leq X < 2) = P(X = 0) + P(X = 1) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Função de distribuição (f.d.)

A v.a. X também fica completamente caracterizada pelo conhecimento da chamada **função de distribuição** (f.d.), a qual é definida, para todo o $x \in \mathbb{R}$, por

$$F_X(x) = P(X \leq x). \tag{3.6}$$

No caso que estamos a considerar, de uma v.a. discreta com f.m.p. $\{(x_k, p_k)\}$ (e onde supomos que os x_k estão ordenados por ordem crescente), a respectiva f.d. será dada por

$$F_X(x) = \sum_{k\,:\,x_k \leq x} p_k. \tag{3.7}$$

Exemplo 3.1 (cont.) *No caso do Exemplo 3.1, ter-se-ia*

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{4}, & 0 \leq x < 1, \\ \frac{3}{4}, & 1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

A função F_X é uma função em escada, não decrescente, com descontinuidades nos pontos x_k (valores do suporte de X), sendo contínua à direita e com limite finito à esquerda de cada um desses pontos. Além disso, tem-se

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{e} \quad \lim_{x \to +\infty} F_X(x) = 1.$$

Vimos que, a partir da f.m.p. podemos construir a f.d. de uma v.a. discreta X. Reciprocamente, se tivermos uma v.a. discreta da qual conheçamos a f.d., podemos reconstruir a respectiva f.m.p.

(os x_k's são os pontos de descontinuidade de F e p_k é o valor do salto no ponto x_k, i.e. $p_k = F(x_k) - F(x_k^-)$.

Logo, uma v.a. discreta fica caracterizada quer através da sua f.m.p., quer através da sua f.d.

Note-se também que temos

$$P(a < X \leq b) = F_X(b) - F_X(a), \quad a, b \in \mathbb{R}. \tag{3.8}$$

Exemplo 3.1 (cont.) *A função de distribuição da v.a. discreta do Exemplo 3.1 é apresentada na Figura 3.2.*

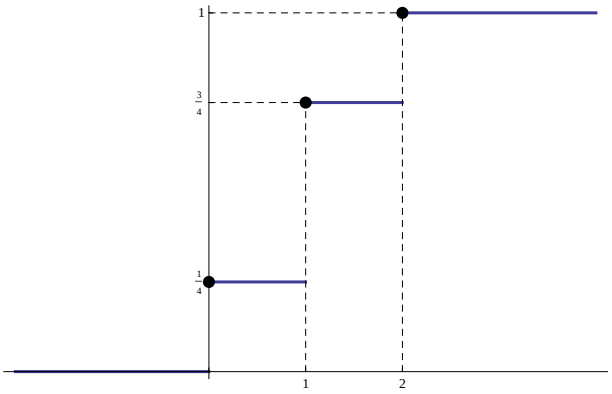

Figura 3.2: Função de distribuição do Exemplo 3.1.

3.3.2 Variáveis contínuas

Função densidade de probabilidade (f.d.p.)

Consideremos, por exemplo, a v.a. X que representa a dose de um determinado medicamento que deve ser dada a um certo doente até que este reaja positivamente. Neste caso, os valores possíveis desta v.a. são todos os valores do intervalo $(0, \infty)$ de $\mathbb{R}$, o qual é um conjunto não numerável. Naturalmente, neste caso não é possível definir a f.m.p da maneira que fizemos para o caso de uma variável discreta. No entanto, poderá existir uma função f que desempenhe uma papel semelhante ao da f.m.p para o caso discreto. Tal função, chamada **função densidade de probabilidade** (f.d.p.), deverá satisfazer as seguintes propriedades:

FDP1 $f_X(x) \geq 0, \quad \forall x$

FDP2 $\displaystyle\int_{-\infty}^{\infty} f_X(x)dx = 1$

Note-se que as condições anteriores equivalem a afirmar que a área da figura plana limitada pela "curva" f_X e pelo eixo dos xx é igual a 1.

Neste caso, o cálculo de probabilidades efectuar-se-á através da fórmula

$$P(X \in B) = \int_B f_X(x)dx, \quad \forall B \subset \mathbb{R}. \tag{3.9}$$

Em particular, se B for um intervalo da forma $[a, b], (a, b], (a, b]$ ou (a, b), tem-se

$$P(X \in B) = \int_a^b f_X(x)dx. \tag{3.10}$$

Neste caso de existência de uma função com as propriedades FDP1 e FDP2 e para a qual tenhamos (3.9), dizemos que a variável em causa é contínua.[3]

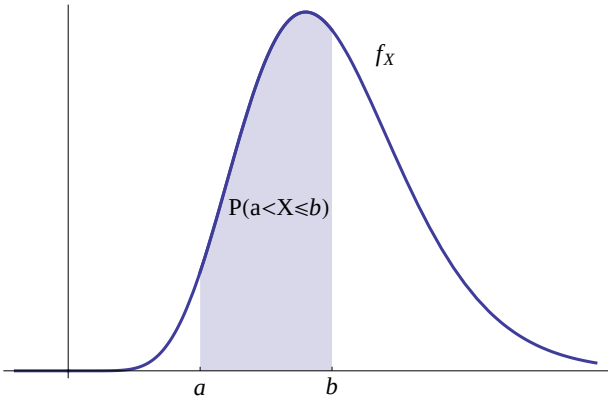

Figura 3.3: Interpretação geométrica de $P(a < X \leq b)$

Note-se que, de (3.9) se segue que

$$P(X = x) = \int_x^x f_X(t)dt = 0,$$

ou seja, *a probabilidade de uma v.a. contínua assumir um valor particular é zero*.

No entanto, se considerarmos um intervalo infinitesimal $(x, x + \delta x]$ com $\delta x \approx 0$, podemos dizer que

$$P(X \in (x, x + \delta x]) = \int_x^{x+\delta x} f_X(t)dt \approx f_X(x)\delta x,$$

[3]Devemos referir que existem v.a.'s que não são nem discretas nem contínuas, tendo um comportamento misto. No entanto, neste curso não estudaremos este tipo de variáveis.

se aproximarmos o integral $\int_x^{x+\delta x} f_X(t)dt$, que corresponde à área abaixo da curva f_X entre x e $x+\delta x$, pelo valor $f_X(x)\delta x$, que corresponde à área do rectângulo que tem por base δx e por altura $f_X(x)$; ver Figura 3.4.

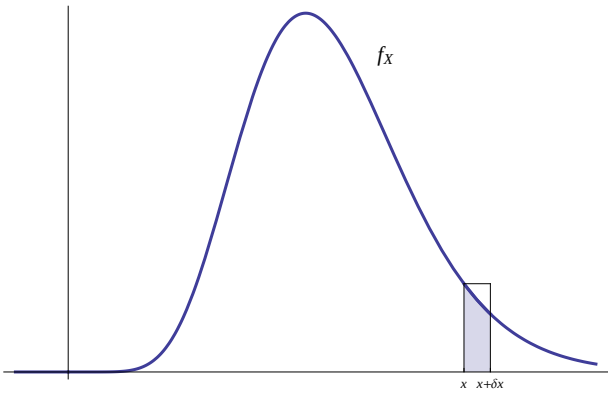

Figura 3.4: Interpretação probabilística da f.d.p.

Exemplo 3.2. *Seja X a v.a. cuja f.d.p. é dada por*

$$f_X(x) = \begin{cases} \frac{3}{8}(4x - 2x^2), & 0 < x < 2, \\ 0, & \textit{restantes valores de } x \end{cases}$$

O gráfico desta função é apresentado na Figura 3.5. Note-se que temos, de facto, $f_X(x) \geq 0$ para

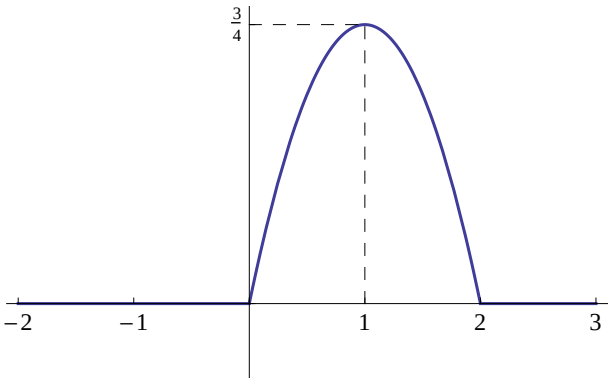

Figura 3.5: Função densidade de probabilidade do Exemplo 3.2

todo o $x \in \mathbb{R}$ *(uma vez que, para* $0 < x < 2$*, se tem* $\frac{3}{8}(4x - 2x^2) > 0$*) e também*

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^2 \frac{3}{8}(4x - x^2)dx$$

$$= \frac{3}{8}\left[2x^2 - \frac{2x^3}{3}\right]_{x=0}^{x=2} = 1.$$

Neste caso, tem-se, por exemplo,

$$P(X > 1) = \int_1^{\infty} f_X(x)dx = \frac{3}{8}\int_1^2 (4x - 2x^2)dx = \frac{1}{2}.$$

Função de distribuição (f.d.)

Tal como para o caso discreto, também no caso de uma v.a. contínua é possível definir a função de distribuição, a qual é dada, para $x \in \mathbb{R}$, por

$$F_X(x) = P(X \leq x).$$

Se X tem f.d.p. f_X, então

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{3.11}$$

Graficamente, o valor de $F_X(x)$ é a área da figura compreendida entre o eixo das abcissas e a curva $f_X(t)$, para $t \leq x$; ver Figura 3.6

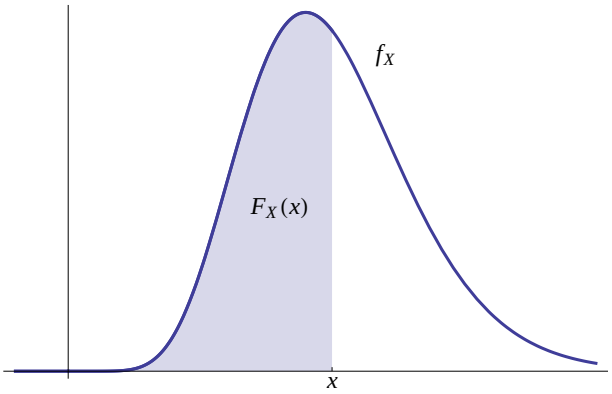

Figura 3.6: Interpretação geométrica de $F_X(x)$

A equação (3.11) mostra mostra que F_X é uma primitiva de f_X. Assim sendo, tem-se

$$P(a < x < b) = P(a < x \leq b) = P(a \leq x < b) = P(a \leq x \leq b) = F_X(b) - F_X(a).$$

A função F_X é contínua, não decrescente e tal que

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1.$$

Exemplo 3.2 (cont.) *No caso da v.a. considerada no Exemplo 3.2, tem-se:*

- *para* $x < 0$, $F_X(x) = \int_0^x f_X(t)dt = \int_0^x 0dt = 0$;
- *para* $0 \leq x \leq 2$,

$$\begin{aligned} F_X(x) &= \frac{3}{8}\int_0^x (4t - 2t^2)dt \\ &= \frac{3}{8}\left[2t^2 - \frac{2t^3}{3}\right]_{t=0}^{t=x} \\ &= \frac{3}{4}(x^2 - \frac{x^3}{3}) \end{aligned}$$

- *para* $x > 2$,

$$\begin{aligned} F_X(x) &= \frac{3}{8}\int_0^2 (4t - 2t^2)dt \\ &= 1 \end{aligned}$$

A função de distribuição desta v.a. é, assim, dada por

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{3}{4}(x^2 - \frac{x^3}{3}), & 0 \leq x \leq 2 \\ 1, & x > 2 \end{cases}$$

O gráfico de $F_X(x)$ *é apresentado na Figura 3.7.*

3.4 Características teóricas ou populacionais

Tal como em estatística descritiva, interessa agora o cálculo de medidas de localização central, dispersão e forma que caracterizam as distribuições de probabilidade de uma v.a. X. Estas carac-

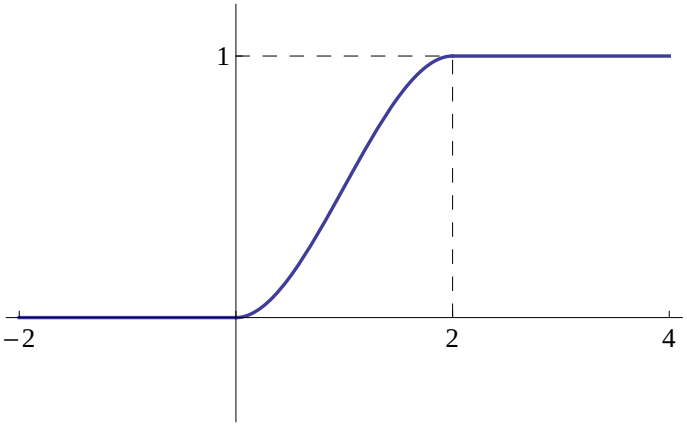

Figura 3.7: Função de distribuição correspondente ao Exemplo 3.2

terísticas são ditas **características teóricas** ou **populacionais**[4] e correspondem às características amostrais estudadas no capítulo anterior.

No que se segue, supomos que X é ou uma v.a. discreta com f.m.p. $\{(x_k, p_k) : k \in \mathbb{K}\}$ ou uma v.a. contínua com f.d.p. $f_X(x)$.

3.4.1 Valor médio ou valor esperado

O **valor médio de** X, ou **valor esperado de** X, ou **esperança de** X, designado por μ_X (por vezes, apenas por μ) ou $E(X)$, é definido do seguinte modo:

$$\mu_X = E(X) = \begin{cases} \displaystyle\sum_{k\in\mathbb{K}} x_k p_k, & \text{no caso discreto,} \\ \displaystyle\int_{-\infty}^{\infty} x f_X(x) dx, & \text{no caso contínuo.} \end{cases} \tag{3.12}$$

Vemos assim que o valor médio de X é a média pesada (de acordo com a f.m.p., no caso discreto, ou com a f.d.p., no caso contínuo) dos valores de X.

Nota: O valor médio é dado pelos expressões indicadas, apenas quando a série, no caso discreto,

[4]Quando estamos interessados no estudo de uma dada variável aleatória X associada a uma certa população, podemos pensar na população não como o conjunto de todos os indivíduos, mas como o conjunto de todos os valores que podem ser assumidos por essa variável. Nesse sentido, a f.m.p. (caso discreto), a f.d.p. (no caso contínuo) ou f.d. da v.a. X, ou seja a "lei de probabilidade" ou "distribuição" da v.a. X caracterizam a *população*, sendo razoável falar em características populacionais.

ou o integral, no caso contínuo, forem absolutamente convergentes, i.e. tivermos $\sum_{k\in\mathbb{K}} |x_k| p_k < \infty$ e $\int_{-\infty}^{\infty} |x| f_X(x)dx < \infty$. Se tal não acontecer, dizemos que o valor médio não existe. [5]

Exemplo 3.3. *Se X for a v.a. correspondente ao "número obtido no lançamento de um dado equilibrado, com as faces numeradas de 1 a 6", cuja f.m.p. é*

$$X : \begin{cases} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{cases}$$

ter-se-á

$$\begin{aligned} E(X) &= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} \\ &= \frac{1}{6}(1 + 2 \cdots + 6) = \frac{7}{2} \end{aligned}$$

Convém salientar que, neste caso, o valor médio ou valor esperado de X, $\frac{7}{2}$, não é um dos valores que X pode assumir (i.e., ao lançar o dado, o resultado nunca será a "saída do número $\frac{7}{2}$"). O que será de esperar é que, se lançarmos o dado um grande número de vezes e acharmos a média dos valores obtidos, esta deve ser aproximadamente $7/2$.

Exemplo 3.2 (cont.) *No caso da v.a. considerada no Exemplo 3.2, tem-se*

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \frac{3}{8}\int_0^2 (4x^2 - 2x^3)dx = \frac{3}{8}\left[\frac{4x^3}{3} - \frac{x^4}{2}\right]_{x=0}^{x=2} = 1.$$

Valor médio de uma função de X

Suponhamos que temos uma v.a. X com uma certa distribuição de probabilidade e que estamos interessados em calcular o valor médio, não da variável X, mas sim de uma certa função de X, digamos $g(X)$ (por exemplo, calcular o valor médio da v.a. $Y = X^2$, que corresponderia a tomar $g(t) = t^2$).[6] A seguinte proposição diz-nos como tal pode ser feito.

Proposição 3.1. *Sendo g uma função real, tem-se*

$$E[g(X)] = \begin{cases} \sum_{k\in\mathbb{K}} g(x_k)p_k, & \textit{no caso discreto}, \\ \int_{-\infty}^{\infty} g(x) f_X(x)dx, & \textit{no caso contínuo}, \end{cases} \tag{3.13}$$

[5] De notar que, se, por exemplo, a série $\sum_k x_k p_k$ fosse apenas convergente, sem ser absolutamente convergente, a sua soma dependeria da ordem dos termos, pelo que o valor médio não estaria bem definido.

[6] Estamos implicitamente a assumir que g é tal que $Y = g(X)$ define uma nova v.a.

desde que a série e o integral sejam absolutamente convergentes.

Da proposição anterior, segue-se facilmente o resultado seguinte.

$$E(aX + b) = aE(X) + b, \quad \text{para quaisquer } a, b \in \mathbb{R}. \tag{3.14}$$

Em particular, se $a = 0$, obtém-se $E(b) = b$, ou seja, tem-se que o valor médio de uma (v.a.) constante é igual ao valor dessa constante.

Exemplo 3.1 (cont.) *Retomando a v.a. discreta considerada no Exemplo 3.1, tem-se, por exemplo,*

$$E(X^2) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = \frac{3}{2}.$$

Exemplo 3.2 (cont.) *No caso da v.a. contínua considerada no Exemplo 3.2, tem-se*

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)dx \\ &= \frac{3}{8}\int_0^2 x^2(4x - 2x^2)dx \\ &= \frac{3}{8}\left[x^4 - \frac{2x^5}{5}\right]_0^2 = \frac{6}{5}. \end{aligned}$$

3.4.2 Variância populacional

Se X é uma v.a. com valor médio μ_X, a **variância de** X, denotada por $\text{var}(X)$ ou σ_X^2, ou por vezes, apenas por σ^2, é definida como

$$\text{var}(X) = \sigma_X^2 = E\Big((X - \mu_X)^2\Big). \tag{3.15}$$

Temos, assim

$$\text{var}(X) = \sigma_X^2 = \begin{cases} \displaystyle\sum_{k\in\mathbb{K}} (x_k - \mu_X)^2 p_k, & \text{no caso discreto,} \\ \displaystyle\int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx, & \text{no caso contínuo.} \end{cases} \tag{3.16}$$

Tem-se, também

$$\begin{aligned} \text{var}(X) &= E\Big((X - \mu_X)^2\Big) = E(X^2 - 2\mu_X X + \mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 = E(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= E(X^2) - \mu_X^2, \end{aligned}$$

ou seja, tem-se

$$\text{var}(X) = E(X^2) - \Big(E(X)\Big)^2,$$

sendo esta a fórmula mais usual para efectuar cálculos.

Exemplo 3.1 (cont.) *No caso do Exemplo 3.1, já vimos que* $E(X) = 1$ *e* $E(X^2) = \frac{3}{2}$*, pelo que será*

$$\text{var}(X) = \frac{3}{2} - 1 = \frac{1}{2}.$$

Exemplo 3.2 (cont.) *No caso da v.a. contínua considerada no Exemplo 3.2, já vimos que* $E(X) = 1$ *e* $E(X^2) = \frac{6}{5}$*, pelo que virá, para a variância de* X*:*

$$\text{var}(X) = E(X^2) - \Big(E(X)\Big)^2 = \frac{6}{5} - 1 = \frac{1}{5}.$$

Propriedades da variância

Temos:

$$\text{var}(aX + b) = a^2 \,\text{var}(X), \quad \text{para quaisquer } a, b \in \mathbb{R}. \tag{3.17}$$

De facto, tem-se

$$\begin{aligned}
\text{var}(aX + b) &= E\Big((aX + b - E(aX + b))^2\Big) \\
&= E\Big((aX + b - a\mu_X - b)^2\Big) \\
&= E\Big(a^2(X - \mu_X)^2\Big) \\
&= a^2 E\Big((X - \mu_X)^2\Big) \\
&= a^2 \,\text{var}(X).
\end{aligned}$$

Em particular, para $a = 0$, obtém-se

$$\text{var}(b) = 0$$

e, considerando $a = 1$, vem

$$\text{var}(X + b) = \text{var}(X).$$

Finalmente, considerando $b = 0$, tem-se

$$\text{var}(aX) = a^2 \,\text{var}(X).$$

A quantidade $\sqrt{\text{var}(X)} = \sigma$ é chamada **desvio padrão** teórico da v.a. X.

O resultado seguinte, cuja demonstração (muito simples) deixamos ao cuidado dos alunos, será usado com frequência.

Normalização (ou *standardização*) de uma v.a.

Teorema 3.1. *Se X é uma v.a. com valor médio $E(X) = \mu$ e variância $\text{var}(X) = \sigma^2$ (desvio padrão σ) (finitas), então a v.a.*

$$Y = \frac{X - \mu}{\sigma}$$

tem média nula e variância igual a 1 (logo, o seu desvio padrão também é igual a 1), i.e. tem-se

$$E(Y) = 0 \quad e \quad \text{var}(Y) = 1.$$

3.4.3 Mediana e quantis teóricos

Quanto aos quantis teóricos, denotados por χ_p, são definidos de forma análoga aos quantis empíricos. O **quantil (de probabilidade)** p, χ_p, é definido por

$$\chi_p = \inf\{x : F_X(x) \geq p\}. \tag{3.18}$$

De salientar que, acordo com esta definição, o quantil χ_p é sempre uma valor assumido pela variável, ou seja, o seu cálculo com o Mathematica deve ser feito usando a função `Quantile` com os parâmetros por defeito. Neste caso, contudo, para variáveis discretas, o valor dado pela função `Median[x]` não coincide necessariamente com o dado pela função `Quantile[x, 1/2]`. No caso de variáveis contínuas, o uso de qualquer das funções produz o mesmo valor.

No caso contínuo, se desenharmos o gráfico da f.d.p. f_X, $p \times 100\%$ da área compreendida entre a curva f_X e o eixo dos xx estará para a esquerda de χ_p e $(1 - p) \times 100\%$ para a sua direita; ver Figura 3.8.

Ainda no caso contínuo, se F_X for injectiva (ou equivalentemente, estritamente crescente), então $\chi_p = F^{-1}(p)$; ver Figura 3.9.

De notar também que, no caso contínuo, se a distribuição for simétrica (com valor médio μ), então os quantis χ_p e χ_{1-p} estão situados simetricamente em relação a μ; ver Figura 3.10.

Tal como no caso dos quantis empíricos, o quantil $\chi_{1/2}$ é chamado **mediana** (teórica) de X e, χ_p para $p = 1/4, 2/4, 3/4$ são chamados **quartis** (populacionais) (respectivamente primeiro, segundo e terceiro quartis).

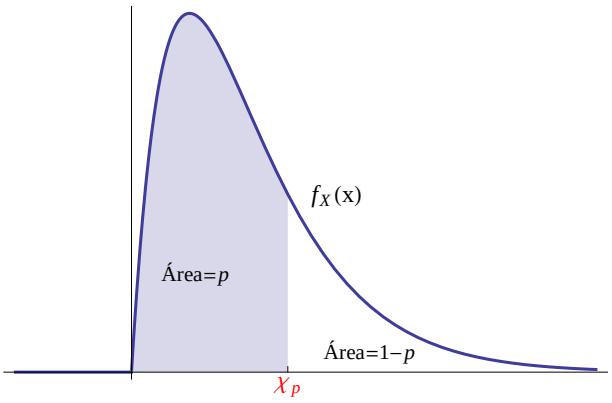

Figura 3.8: Quantil p, χ_p

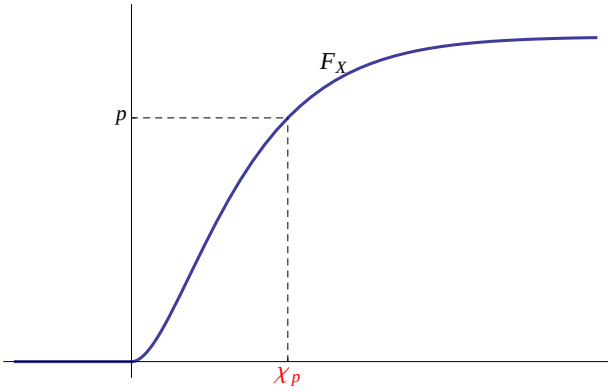

Figura 3.9: Quantil p, χ_p

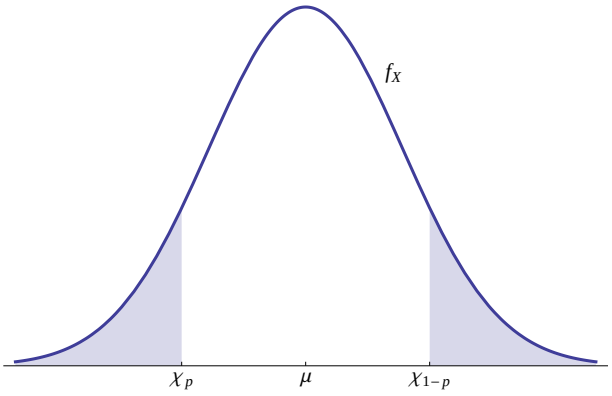

Figura 3.10: Quantis χ_p e χ_{1-p} de uma distribuição simétrica

3.4.4 Moda(s) teórica(s)

Um valor x_m é dito uma **moda** de X (ou diz-se que X tem uma moda em x_m), se verificar

- $P(X = x_m) = \max_{k\in\mathbb{K}} p_k$, no caso discreto
- $f_X(x_m) = \max_{x\in\mathbb{R}}\{f_X(x)\}$, no caso contínuo. [7]

Exemplo 3.4. *Por exemplo, no caso da v.a. discreta considerada no Exemplo 3.1 tem-se que $x_m = 1$ é (a única) moda.*

No caso da v.a. contínua do Exemplo 3.2, tem-se que a função $f_X(x)$ tem um máximo absoluto no ponto $x = 1$, pelo que 1 é uma (a única) moda de X.

3.4.5 Coeficiente de assimetria e coeficiente de achatamento

O **coeficiente de assimetria** populacional, β_1, é a característica populacional correspondente ao coeficiente de assimetria empírico b_1, dado por (2.9), e é definido por

$$\beta_1 = \begin{cases} \dfrac{1}{\sigma^3}\displaystyle\sum_{k\in\mathbb{K}}(x_k-\mu)^3 p_k, & \text{no caso discreto,} \\ \dfrac{1}{\sigma^3}\displaystyle\int_{-\infty}^{\infty}(x-\mu)^3 f(x)dx, & \text{no caso contínuo.} \end{cases} \tag{3.19}$$

[7] Por vezes consideram-se também *modas relativas*, i.e. consideram-se os pontos onde a função densidade tem máximos locais, chamando-se, por exemplo, bimodal a uma distribuição cuja f.d.p. tenha dois máximos locais, ainda que só tenha um máximo absoluto.

O coeficiente de assimetria dá-nos alguma indicação de quanto a distribuição se "afasta" de uma distribuição simétrica. Se a distribuição é simétrica, então $\beta_1 = 0$.[8]

O **coeficiente de achatamento** ou **de curtose** populacional, β_2, é definido por

$$\beta_2 = \begin{cases} \frac{1}{\sigma^4} \sum_{k \in \mathbb{K}} (x_k - \mu)^4 p_k, & \text{no caso discreto,} \\ \frac{1}{\sigma^4} \int_{-\infty}^{\infty} (x - \mu)^4 f(x) dx, & \text{no caso contínuo.} \end{cases} \tag{3.20}$$

3.5 Pares aleatórios; vectores aleatórios

Em muitas situações, tem interesse considerar duas ou mais variáveis em simultâneo e estudar o seu relacionamento mútuo. Deste modo, somos levados a generalizar o conceito de v.a., introduzindo o conceito de **vector aleatório**. Chamamos **vector aleatório** de dimensão n ($n \in \mathbb{N}$), o qual denotamos, em geral, por $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, a uma função de Ω em $\mathbb{R}^n$,

$$\begin{aligned} \mathbf{X} = (X_1, X_2, \ldots, X_n) : &\Omega \to \mathbb{R}^n \\ &\omega \mapsto (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)). \end{aligned}$$

Quando $n = 2$, temos os chamados **pares aleatórios**, vulgarmente representados por (X, Y).

É este o caso que estudaremos primeiramente.

No caso em que as v.a.'s X e Y que formam o par (X, Y) são ambas discretas, este par fica caracterizado pelo conhecimento da **função massa de probabilidade conjunta**, i.e., pela função que a cada ponto (x_k, y_ℓ) onde x_k pertence ao suporte de X e y_ℓ ao suporte de Y associa a probabilidade $p_{k\ell}$ dada por

$$p_{k\ell} = P(X = x_k, Y = y_\ell).$$

Tem-se, $p_{k\ell} \geq 0$ e $\sum_k \sum_\ell p_{k\ell} = 1$.

A distribuição deste par é, no caso em que X e Y têm suporte finito, dada geralmente dada na forma de uma tabela de dupla entrada que contém os valores de X, os valores de Y e os valores de

[8] Embora o recíproco não seja necessariamente verdadeiro.

pij.

$X \setminus Y$	y_1	y_2	$\cdots$	y_ℓ	$\cdots$	
x_1	p_{11}	p_{12}	$\cdots$	$p_{1\ell}$	$\cdots$	$p_{1\bullet}$
x_2	p_{21}	p_{22}	$\cdots$	$p_{2\ell}$	$\cdots$	$p_{2\bullet}$
$\vdots$						
x_k	p_{k1}	p_{k2}	$\cdots$	$p_{k\ell}$	$\cdots$	$p_{k\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		
	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet\ell}$	$\cdots$	

Exemplo 3.5. *Considere-se a experiência de lançamento de duas moedas equilibradas e seja X a variável já anteriormente considerada, correspondente ao "número de caras obtidas", e seja Y a variável correspondente à "diferença entre o número de caras e o número de coroas". A f.m.p. conjunta do par aleatório (X, Y) é dada na tabela seguinte:*

$X \setminus Y$	$y_1 = -2$	$y_2 = 0$	$y_3 = 2$	
$x_1 = 0$	$1/4$	0	0	$1/4$
$x_2 = 1$	0	$1/2$	0	$1/2$
$x_3 = 2$	0	0	$1/4$	$1/4$
	$1/4$	$1/2$	$1/4$	

No caso em que as v.a.'s X e Y são contínuas, o par aleatório fica caracterizado pela **função densidade de probabilidade conjunta**

$$\begin{aligned} f :&\mathbb{R}^2 \to \mathbb{R} \\ &(x, y) \mapsto f(x, y) \end{aligned}$$

tal que $f(x, y) \geq 0$ e $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$.

De modo análogo ao caso univariado, temos para qualquer conjunto $C \subseteq \mathbb{R}^2$,

$$P\big((X, Y) \in C\big) = \begin{cases} \displaystyle\sum_{k,\ell:(x_k,y_\ell)\in C} \sum p_{k\ell}, & \text{no caso discreto} \\ \displaystyle\iint_C f(x, y)dxdy, & \text{no caso contínuo.} \end{cases} \tag{3.21}$$

A **função de distribuição conjunta** para um par aleatório (X, Y) é definida, para cada $(x, y) \in \mathbb{R}^2$ por

$$F(x, y) = P(X \leq x, Y \leq y). \tag{3.22}$$

Tem-se, então

$$F(x,y) = \begin{cases} \displaystyle\sum_{k\,:\,x_k \le x} \sum_{\ell\,:\,y_\ell \le y} p_{k\ell}, & \text{no caso discreto,} \\ \displaystyle\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)du\,dv, & \text{no caso contínuo.} \end{cases}$$

As distribuições de X e Y, chamadas **distribuições marginais**, obtêm-se facilmente.

No caso discreto, comecemos por notar que o acontecimento $\{X = x_k\}$ pode ser escrito como a união, para todos os valores de ℓ, dos acontecimentos mutuamente exclusivos $\{X = x_k, y = y_\ell\}$, isto é,

$$X = x_k = \bigcup_{\ell} \{X = x_k, y = y_\ell\}.$$

Então, usando o axioma P3 da medida de probabilidade, tem-se, denotando por $p_{k\bullet}$ a probabilidade de $X = x_k$, i.e. $p_{k\bullet} = P(X = x_k)$,

$$\begin{aligned} p_{k\bullet} = P(X = x_k) &= P\left(\bigcup_{\ell} \{X = x_k, y = y_\ell\}\right) \\ &= \sum_{\ell} P(X = x_k, y = y_\ell) \\ &= \sum_{\ell} p_{k\ell} \end{aligned}$$

De modo análogo se obtém

$$p_{\bullet\ell} = P(Y = y_\ell) = \sum_{k} p_{k\ell},$$

onde $p_{\bullet\ell}$ denota a probabilidade de Y assumir o valor y_ℓ, i.e. $p_{\bullet\ell} = P(Y = y_\ell)$. Assim, sabida a f.m.p conjunta, consegue sempre determinar-se a f.m.p de cada uma das variáveis, somando as probabilidades conjuntas ao longo dos diferentes valores da outra variável (estas probabilidades estão indicadas na tabela anterior).

Note-se, contudo, que o recíproco não é verdadeiro, ou seja, o conhecimento das probabilidades $p_{k\bullet} = P(X = x_k)$ e $p_{\bullet\ell} = P(Y = y_\ell)$ não permite determinar a probabilidade conjunta $p_{k\ell} = P(X = x_k, Y = y_\ell)$.

No caso contínuo, o resultado é semelhante. Se A e B são dois conjuntos de $\mathbb{R}$, sendo $C = \{(x,y) : x \in A, Y \in B\}$, vemos, de acordo com (3.21), que

$$P(X \in A, Y \in B) = \int_B \int_A f(x,y)dxdy = \int_A \int_B f(x,y)dydx.$$

Então, tem-se

$$\begin{aligned} P(X \in A) &= P(X \in A, Y \in (-\infty, \infty)) \\ &= \int_A \int_{-\infty}^{\infty} f(x,y) dy dx \\ &= \int_A f_X(x) dx, \end{aligned}$$

onde

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy$$

é, portanto, a função densidade de probabilidade de X (**função densidade marginal de** X). De modo análogo, a **densidade marginal de** Y é dada por

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx.$$

O caso de vectores aleatórios de dimensão n é análogo ao caso bivariado sendo as f.m.p. e f.d.p. conjuntas funções de n variáveis.

3.5.1 Variáveis independentes

O conceito de acontecimentos independentes estende-se, de um modo natural, ao caso de variáveis.

Duas variáveis X e Y são ditas independentes se, para quaisquer conjuntos A e B tivermos

$$P(X \in A, Y \in B) = P(X \in A)\, P(X \in B).$$

Pode mostrar-se que isto é equivalente a ter-se

$$p_{k\ell} = p_{k\bullet} p_{\bullet \ell}, \quad \text{no caso discreto}$$

e a ter-se

$$f(x,y) = f_X(x) f_Y(y), \quad \text{no caso contínuo.}$$

Assim, as v.a. serão independentes se a f.m.p. conjunta, no caso discreto (respectivamente a f.d.p conjunta, no caso contínuo) for igual ao produto das f.m.p. marginais (respectivamente, igual ao produto das f.d.p. marginais).

Exemplo 3.5 (cont.) *Relativamente às variáveis aleatórias consideradas no Exemplo 3.5, tem-se, por exemplo:*

$$p_{12} = P(X = x_1, Y = y_2) = P(X = 0, Y = 0) = 0,$$

$$p_{1\bullet} = P(X = x_1) = P(X = 0) = \frac{1}{4}$$

e

$$p_{\bullet 2} = P(Y = y_2) = P(Y = 0) = \frac{1}{2}$$

ou seja,

$$p_{12} \neq p_{1\bullet} p_{\bullet 2},$$

pelo que X e Y não são independentes.

Dizemos que n variáveis $X_1, \ldots, X_n$ são independentes, se, para quaisquer conjuntos $B_1, \ldots, B_n$ tivermos

$$P(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = P(X_1 \in B_1)P(X_2 \in B_2) \ldots P(X_n \in B_n),$$

o que é equivalente a ter-se

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \ldots P(X_n = x_n),$$

no caso discreto, ou a ter-se

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n),$$

no caso contínuo.

Claro que se usarmos o conceito de função de distribuição conjunta, teremos que $X_1, X_2, \ldots, X_n$ são v.a.'s independentes se e só se a f.d. conjunta do vector $(X_1, X_2, \ldots, X_n)$ for igual ao produto das suas marginais, isto é, se para todo o $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ se tiver,

$$\begin{aligned} F(x_1, x_2, \ldots, x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &= P(X_1 \leq x_1)P(X_2 \leq x_2) \cdots P(X_n \leq x_n) \\ &= F_{X_1}(x_1)F_{X_2}(x_2) \cdots F_{X_n}(x_n). \end{aligned}$$

3.6 Operações com variáveis aleatórias

Sendo as v.a.'s funções, as operações entre elas (soma, diferença, produto,etc) são definidas do modo habitual. Por exemplo, sendo X e Y duas v.a.'s associadas ao mesmo espaço amostral Ω, define-se $X + Y$ como sendo a v.a. dada por

$$(X + Y)(\omega) = X(\omega) + Y(\omega), \quad \omega \in \Omega.$$

Não é difícil de mostrar que a soma $X + Y$, a diferença $X - Y$, o produto XY e o quociente X/Y (definido quando $\{\omega : Y(\omega) = 0\} = \emptyset$) são novas v.a.'s.

3.6.1 Valor médio da soma de variáveis aleatórias

Dadas duas v.a.'s X e Y, pode provar-se que

$$E(X+Y) = E(X) + E(Y)$$

(desde que $E(X)$ e $E(Y)$ existam). Mais geralmente, tem-se: tem-se

Teorema 3.2. *Dadas n v.a's $X_1, \ldots, X_n$, tem-se*

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n), \tag{3.23}$$

desde que $E(X_i); i = 1, \ldots, n$ existam.

3.6.2 Covariância de duas variáveis aleatória; variância da soma de variáveis aleatórias

Acabámos de referir que o valor médio da soma de v.a.'s é a igual à soma dos respectivos valores médios. Relativamente à variância, o mesmo tipo de resultado não é, em geral, verdadeiro. Por exemplo, tem-se

$$\text{var}(X+X) = \text{var}(2X) = 2^2\,\text{var}(X) = 4\,\text{var}(X) \neq \text{var}(X) + \text{var}(X).$$

Começamos por introduzir o seguinte conceito.

Dadas duas v.a.'s X e Y, a sua **covariância**, denotada por $\text{cov}(X,Y)$, é definida por

$$\text{cov}(X,Y) = E\Big((X-\mu_X)(Y-\mu_Y)\Big)$$

Segue-se, de imediato, da definição, que

$$\text{cov}(X,Y) = \text{cov}(Y,X)$$

e que

$$\text{cov}(X,X) = \text{var}(X).$$

Além disso, tem-se

$$\begin{aligned}
\text{cov}(X,Y) &= E\Big((X-\mu_X)(Y-\mu_Y)\Big) \\
&= E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y) \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\
&= E(XY) - E(X)E(Y) - E(Y)YE(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

A covariância goza também da propriedade seguinte:

$$\text{cov}(X \pm Y, Z) = \text{cov}(X, Z) \pm \text{cov}(Y, Z),$$

de onde se obtém facilmente o seguinte resultado

$$\text{cov}(X \pm Y, X \pm Y) = \text{cov}(X, X) \pm 2\,\text{cov}(X, Y) + \text{cov}(Y, Y).$$

Tem-se, então o seguinte resultado:

$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2\,\text{cov}(X, Y). \tag{3.24}$$

Pode provar-se também o seguinte resultado importante. Se X e Y são v.a.'s independentes, então

$$\text{cov}(X, Y) = 0$$

e, portanto, nesse caso, tem-se

$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y).$$

Mais geralmente, tem-se:

Teorema 3.3. *Se $X_1, \ldots, X_n$ são v.a.'s independentes, então*

$$\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$$

(desde que $\text{var}(X_i); i = 1, \ldots, n$ *existam).*

Chama-se **correlação populacional** de X e Y e designa-se por $\rho_{X,Y}$, ou simplesmente por ρ, a quantidade dada por

$$\rho = \rho_{X,Y} = E\Big(\frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}\Big). \tag{3.25}$$

É imediato concluir que

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

A correlação populacional tem propriedades semelhantes à correspondente versão empírica:

1. $-1 \leq \rho \leq 1$
2. $|\rho| = 1$ se e só se existir uma relação linear entre X e Y.[9]
3. Se X e Y são independentes, então $\rho = 0$ (a recíproca não é verdadeira).

[9] i.e., tivermos $P(Y = aX + b) = 1$, para algum a e algum b em $\mathbb{R}$.

4

Modelos Paramétricos

Existem certo tipo de distribuições que, quer do ponto de vista teórico, quer do ponto de vista prático, desempenham um papel especial na Estatística. Neste capítulo, estudaremos algumas dessas distribuições. Outros modelos serão estudados nas aulas práticas.

4.1 Modelos Discretos

4.1.1 Distribuição uniforme em n pontos

Se X é uma variável aleatória cujo contradomínio é o conjunto de n pontos, $\{x_1, \ldots, x_n\}$, e é tal que todos os pontos do seu contradomínio têm igual probabilidade, isto é, se a f.m.p. de X é dada por

$$P(X = x_k) = \frac{1}{n}; k = 1, 2, \ldots, n,$$

diz-se que X tem uma **distribuição uniforme** nos n pontos $x_1, \ldots, x_n$ e escreve-se, abreviadamente, $X \sim U\{x_1, \ldots, x_n\}$.

Exemplo 4.1. *Como exemplo de uma v.a. com distribuição uniforme, temos a v.a. que representa o número obtido no lançamento de um dado equilibrado, com as faces numeradas de 1 a 6, a qual tem distribuição uniforme no conjunto* $\{1, 2, \ldots, 6\}$.

No caso de ser $n = 1$, diz-se que a distribuição é degenerada no ponto x_1, tendo-se, então $P(X = x_1) = 1$, isto é, a probabilidade 1 está totalmente concentrada num ponto único ponto x_1.

Características teóricas do modelo uniforme

Sendo $X \sim U\{1, 2, \ldots, n\}$[1] pode mostrar-se que se tem:

- $\mu_X = \frac{1+n}{2}$
- $\sigma_X^2 = \frac{n^2-1}{12}$
- a distribuição é plurimodal
- $\chi_{1/2} = \lfloor \frac{N+1}{2} \rfloor$
- $\beta_1 = 0$
- $\beta_2 = \frac{3}{5}(3 - \frac{4}{N^2-1})$

No Mathematica, a função asociada à distribuição uniforme em n pontos é a função `DiscreteUniformDistribution`.

4.1.2 Distribução de Bernoulli e distribuição binomial

Distribuição de Bernoulli

Suponhamos que se efectua uma experiência cujos resultados são apenas dois: um, que designaremos por "sucesso"e outro, que designaremos por "insucesso". Uma experiência deste tipo é dita uma experiência de Bernoulli. Se for X a v.a. cujo valor é $X = 1$ quando o resultado é "sucesso"e $X = 0$, quando o resultado é "insucesso", então a f.m.p. desta variável aleatória é da forma

$$X = \begin{cases} 0 & 1 \\ 1-p & p \end{cases},$$

onde $p\,(0 < p < 1)$ é a probabilidade de "sucesso"(e, claro está, $1 - p$ é a probabilidade de "insucesso"). Uma v.a. cuja f.m.p. é deste tipo, é dita uma **variável aleatória de Bernoulli** (ou modelo de Bernoulli) com parâmetro p.[2] Se X tem uma distribuição de Bernoulli com parâmetro p, escrevemos $X \sim Ber(p)$.

[1] Por uma questão de simplicidade, consideramos a distribuição uniforme nos pontos $x_1 = 1, x_2 = 2, \ldots, x_n = n$.

[2] Esta distribuição tem o nome do matemático suiço Jacob Bernoulli, que viveu entre 1654 e 1705.

Características teóricas do modelo de Bernoulli

Se $X \sim Ber(p)$, com $0 < p < 1$, tem-se:

- $\mu_X = p$
- $\sigma_X^2 = p(1-p)$
- 0 é moda de X, se $p \leq \frac{1}{2}$; 1 é moda de X, se $p \geq \frac{1}{2}$; assim, X é bimodal se $p = \frac{1}{2}$ e unimodal se $p \neq \frac{1}{2}$.
- $\beta_1 = \frac{1-2p}{\sqrt{p(1-p)}}$
- $\beta_2 = 3 + \frac{1-6p(1-p)}{p(1-p)}$

A distribuição $Ber(p)$ é simétrica quando $p = \frac{1}{2}$, tem assimetria positiva quando $p < \frac{1}{2}$ e assimetria negativa quando $p > \frac{1}{2}$.

No Mathematica, a função asociada à distribuição de Bernoulli, é a função `BernoulliDistribution`.

Distribuição binomial

Suponhamos agora que efectuamos n repetições de uma experiência de Bernoulli (com probabilidade p de "sucesso"e $(1-p)$ de "insucesso"). Se X representa o número de sucessos que ocorrem nestas n repetições, então X diz-se uma v.a. **binomial com parâmetros n e p** e escreve-se $X \sim Bi(n,p)$. Note-se que X pode também ser vista como a soma de n v.a.'s independentes $X_1, \ldots, X_n$, em que cada $X_i \sim Ber(p)$. O caso em que $n = 1$ corresponde, naturalmente, ao de uma distribuição de Bernoulli. A f.m.p. de uma v.a. binomial com parâmetros n e p, em que $n \in \mathbb{N}$ e $0 < p < 1$, é dada por

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, 2, \ldots, n. \tag{4.1}$$

Características teóricas do modelo binomial

Se $X \sim Bi(n,p)$, então pode mostrar-se que:

- $\mu_X = np$
- $\sigma_X^2 = np(1-p)$

- $\lfloor p(n+1) \rfloor$ é moda de X
- $\beta_1 = \frac{1-2p}{\sqrt{np(1-p)}}$
- $\beta_2 = 3 + \frac{1-6p(1-p)}{np(1-p)}$

Nas figuras 4.1 – 4.3 apresentam-se gráficos da f.m.p. do modelo binomial para diferentes valores de p, para $n = 10$.

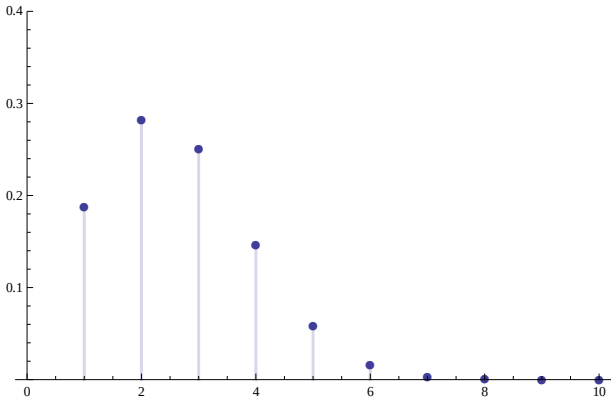

Figura 4.1: Função massa de probabilidade do modelo $Bi(10, 0.25)$

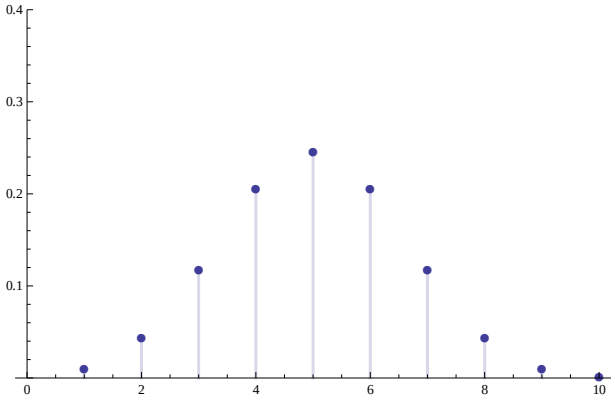

Figura 4.2: Função massa de probabilidade do modelo $Bi(10, 0.5)$

Tal como a distribuição de Bernoulli, a distribuição $Bi(n, p)$ é simétrica quando $p = \frac{1}{2}$, tem assimetria positiva quando $p < \frac{1}{2}$ e assimetria negativa quando $p > \frac{1}{2}$.

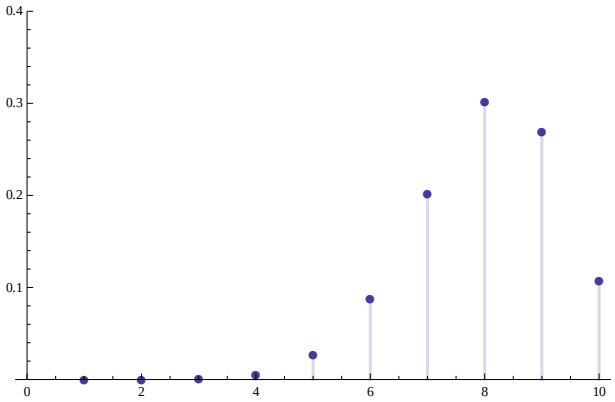

Figura 4.3: Função massa de probabilidade do modelo $Bi(10, 0.8)$

No Mathematica, para trabalhar com a distribuição binomial, deve usar a função `BinomialDistribution`.

4.1.3 Distribuição de Poisson

Uma v.a. X com contradomínio $0, 1, 2, \ldots$ diz-se uma **variável de Poisson** com parâmetro λ ($\lambda > 0$), se a sua f.m.p. for dada por

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, \ldots. \tag{4.2}$$

Se X tem uma distribuição de Poisson com parâmetro λ, escrevemos $X \sim Poi(\lambda)$.

Características teóricas do modelo de Poisson

Se $X \sim Poi(\lambda)$, tem-se:

- $\mu_X = \lambda$
- $\sigma^2_X = \lambda$
- $\lfloor \lambda \rfloor$ é moda de X (e, se λ é inteiro, então λ e $\lambda - 1$ são modas de X).
- $\beta_1 = \frac{1}{\sqrt{\lambda}}$ (a distribuição tem sempre assimetria positiva, sendo esta tanto mais acentuada quanto menor for o valor de λ)

- $\beta_2 = 3 + \frac{1}{\lambda}$ (a distribuição é leptocúrtica e a curtose é tanto maior quanto menor for o valor de λ).

Nas figuras 4.4 – 4.6 apresentam-se gráficos da f.m.p. da distribuição $Poi(\lambda)$ para diferentes valores de λ.[3]

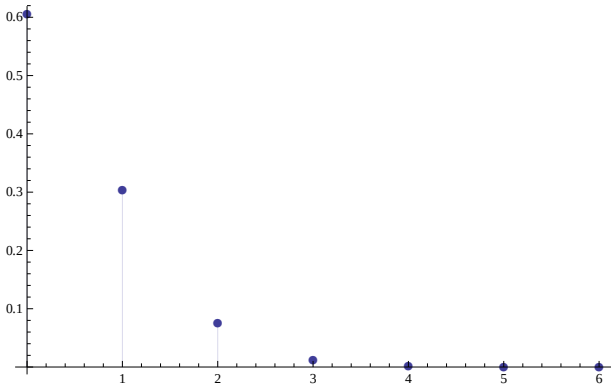

Figura 4.4: Função massa de probabilidade do modelo $Poi(0.5)$

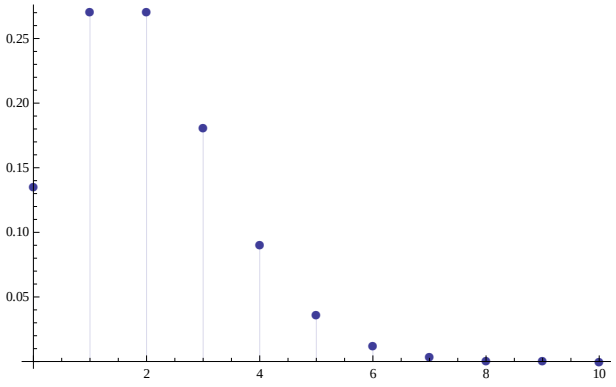

Figura 4.5: Função massa de probabilidade do modelo $Poi(2)$

O modelo de Poisson tem uma grande quantidade de aplicações em diversas áreas, porque pode ser usado como uma aproximação para o modelo binomial com parâmetros n e p, quando n é grande (n é maior que 150, digamos) e p é pequeno ($p < 0.01$, digamos), o que significa

[3]A distribuição de Poisson foi introduzida por Siméon Denis Poisson (1781-1840), no seu trabalho intitulado *Recherches sur la probabilité des jugements en matière criminelle et civile*, publicado em 1838.

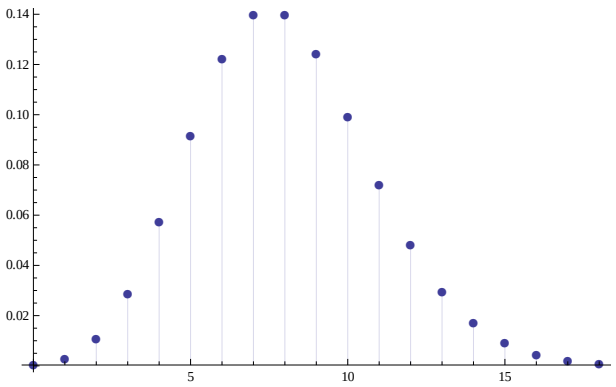

Figura 4.6: Função massa de probabilidade do modelo $Poi(8)$

que o "sucesso"é um acontecimento *raro*. De facto, pode mostrar-se que a distribuição $Bi(n,p)$, desde que n seja suficientemente grande e p suficientemente pequeno, coincide praticamente com a distribuição de Poisson com parâmetro $\lambda = np$. Este facto justifica o nome de "lei dos acontecimentos raros"atribuído geralmente ao modelo de Poisson. Para ver que assim é, seja $X \sim Bi(n,p)$ e seja $\lambda = np$. Então,

$$\begin{aligned} P(X=k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\ldots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} \end{aligned}$$

Para n grande e p pequeno, tem-se

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \quad \frac{n(n-1)\ldots(n-k+1)}{n^k} \approx 1 \quad \text{e} \quad \left(1-\frac{\lambda}{n}\right)^k \approx 1.$$

Vemos assim que, para n grande e p pequeno, se tem

$$P(X=k) \approx e^{-\lambda} \frac{\lambda^k}{k!},$$

onde $\lambda = np$.

Alguns exemplos de v.a.'s que obedecem (aproximadamente) ao modelo de Poisson são:

1. o número de erros tipográficos contidos numa página de um livro;
2. o número de pessoas de uma certa comunidade que têm mais de 100 anos de idade;

3. o número de lâmpadas que fundem na primeira hora de utilização;

4. o número de chamadas telefónicas recebidas num *call center*, por minuto.

No Mathematica, a função associada ao modelo de Poisson é a função `PoissonDistribution`.

4.2 Modelos Contínuos

4.2.1 Distribuição uniforme num intervalo

Diz-se que uma v.a. X tem **distribuição uniforme no intervalo** $[a, b] \subset \mathbb{R}$ e escreve-se, abreviadamente, $X \sim U[a, b]$, se a sua f.d.p. for dada por

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{se } a \leq x \leq b, \\ 0, & \text{outros valores de } x. \end{cases} \tag{4.3}$$

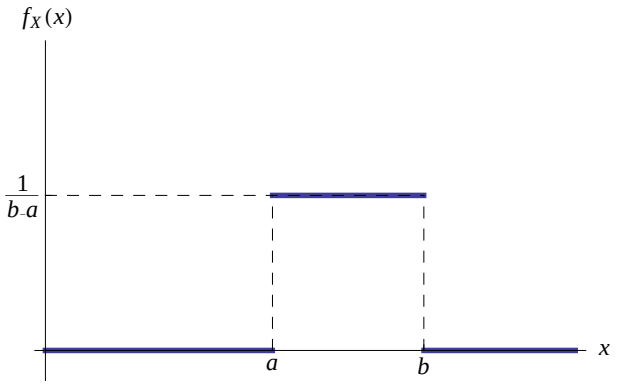

Figura 4.7: Função densidade de probabilidade do modelo uniforme

A função de distribuição correspondente é dada por

$$F_X(x) = \begin{cases} 0, & \text{se } x < a \\ \dfrac{x-a}{b-a}, & \text{se } a \leq x < b \\ 1, & \text{se } x \geq b \end{cases} \tag{4.4}$$

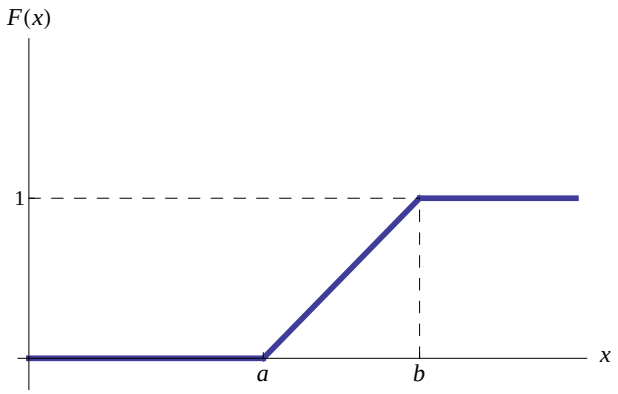

Figura 4.8: Função de distribuição do modelo uniforme

Características teóricas da distribuição uniforme

Se $X \sim U[a, b]$, tem-se:

- $\mu_X = \frac{a+b}{2}$
- $\sigma_X^2 = \frac{(b-a)^2}{12}$
- $\chi_{1/2} = \frac{a+b}{2}$
- $\beta_1 = 0$
- $\beta_2 = \frac{9}{5} = 1.8$ (a distribuição é platicúrtica).

No Mathematica, a função associada à distribuição uniforme é a função `UniformDistribution`.

4.2.2 Distribuição exponencial

Uma v.a. X cuja função densidade de probabilidade seja dada, para $\lambda > 0$, por

$$f_X(x) = \begin{cases} 0, & \text{se } x < 0, \\ \lambda e^{-\lambda x}, & \text{se } x \geq 0 \end{cases} \tag{4.5}$$

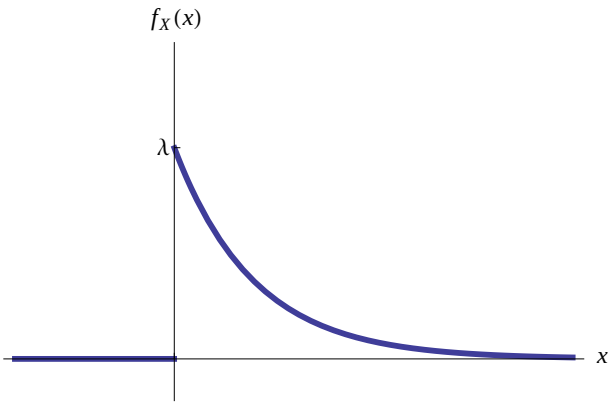

Figura 4.9: Função densidade de probabilidade do modelo exponencial

diz-se uma v.a. **exponencial com parâmetro** λ (ou ter distribuição exponencial com parâmetro λ). Se X tem uma distribuição exponencial com parâmetro λ, escreve-se $X \sim Exp(\lambda)$.

A função de distribuição correspondente é dada por

$$F_X(x) = \begin{cases} 0, & \text{se } x < 0, \\ 1 - e^{-\lambda x}, & \text{se } x \geq 0, \end{cases} \tag{4.6}$$

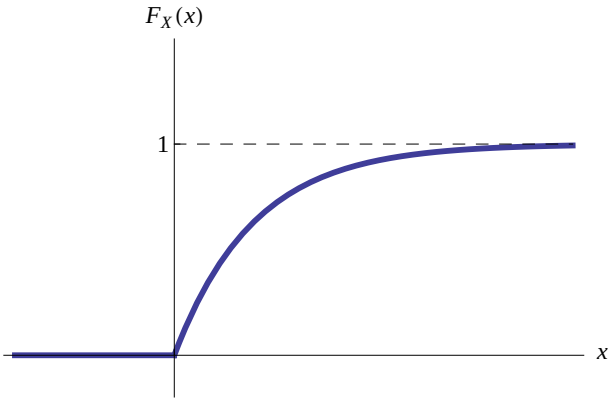

Figura 4.10: Função de distribuição do modelo exponencial

A distribuição exponencial está geralmente associada a variáveis que medem a quantidade de tempo que decorre até que um acontecimento específico ocorra: por exemplo, o tempo (contado a partir deste instante) que decorre até que ocorra um tremor de terra, ou até que recebamos

uma chamada telefónica por engano ou o intervalo de tempo entre falhas consecutivas de um dado aparelho etc.

Características teóricas da distribuição exponencial

Se $X \sim Exp(\lambda)$ $\lambda > 0$, tem-se:

- $\mu_X = \frac{1}{\lambda}$
- $\sigma_X^2 = \frac{1}{\lambda^2}$
- 0 é moda de X
- $\chi_{1/2} = \frac{\log 2}{\lambda}$
- $\beta_1 = 2$ (a distribuição tem sempre assimetria positiva, para qualquer λ)
- $\beta_2 = 9$ (a distribuição é sempre leptocúrtica, seja qual for o valor de λ)

No Mathematica, para trabalhar com a distribuição exponencial, deverá usar a função `ExponentialDistribution`.

4.2.3 Distribuição normal (ou Gaussiana)

Das distribuições contínuas, tem especial importância a distribuição normal ou Gaussiana[4], que os alunos já conhecem bem do Ensino Secundário, mas que vamos agora revisitar. Diz-se que uma v.a. X **distribuição normal** ou **Gaussiana** com parâmetros μ e σ ($\sigma > 0$), e escreve-se abreviadamente $X \sim N(\mu, \sigma)$, se tiver f.d.p. dada por

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{4.7}$$

É fácil de mostrar que, se $X \sim N(\mu, \sigma)$, então

- $E(X) = \mu$
- var$(X) = \sigma^2$

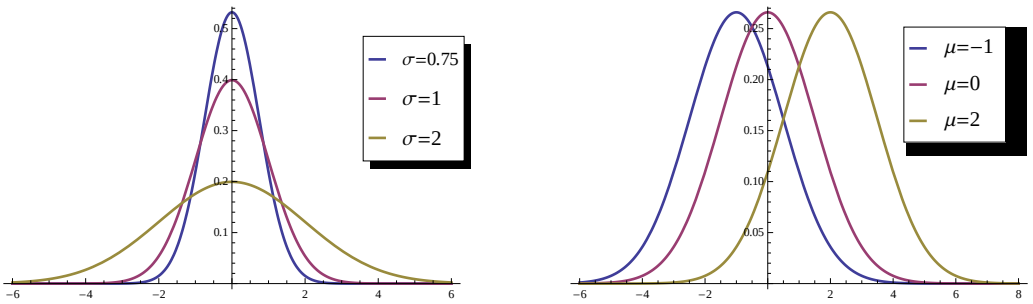

Figura 4.11: Função densidade de probabilidade do modelo $N(\mu, \sigma)$: Esquerda: $\mu = 0; \sigma = 0.75, 1, 2$; Direita: $\mu = -1, 0, 2; \sigma = 1.5$

Assim, os parâmetros μ e σ da v.a. $X \sim N(\mu, \sigma)$ são o seu valor médio e o seu desvio padrão, respectivamente (ou seja, se preferimos, σ^2 é a variância de X).

A distribuição normal é simétrica e a sua f.d.p.tem a bem conhecida "forma de sino".

Uma propriedade importante da distribuição normal é que, se $X \sim N(\mu, \sigma)$, então $Y = \alpha X + \beta$ também tem uma distribuição normal. Atendendo às propriedades do valor médio e da variância – veja as equações (3.14) e (3.17) – e ainda ao resultado relativo à standardização de v.a's – Teorema 3.1 – é imediato concluir que a variável Z definida por

$$Z = \frac{X - \mu}{\sigma} \tag{4.8}$$

é normal com valor médio 0 e variância 1, i.e. $Z \sim N(0, 1)$. Neste caso, diz-se que Z é **normal reduzida** ou **standard**. A f.d.p. da normal reduzida é usualmente denotada por ϕ e a sua expressão, obtida de (4.7) considerando $\mu = 0$ e $\sigma = 1$ é dada por

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{4.9}$$

A função de distribuição da normal reduzida Z é usualmente designada por Φ e é dada por

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{4.10}$$

Note-se que esta função não tem uma expressão analítica, uma vez que a função $\phi(x)$ não é integrável analiticamente.

[4] O termo *Gaussiana* vem do nome do matemático alemão Carl Friedrich Gauss (1777-1855), que desenvolveu e aplicou este modelo; o termo *normal* parece ter sido proposto por Francis Galton, matemático e estatístico inglês que viveu entre 1822 e 1911.

Características teóricas do modelo normal

Seja $X \sim N(\mu, \sigma)$. Tem-se, então:

- $\mu_X = \mu$;
- $\sigma_X^2 = \sigma^2$;
- μ é (a única) moda de X;
- $\chi_{1/2} = \mu$;
- $\beta_1 = 0$;
- $\beta_2 = 3$ (a distribuição tem sempre a mesma curtose, independentemente do valor dos parâmetros μ e σ).[5]

No Mathematica, a função associada à distribuição normal é a função `NormalDistribution`.

Outras propriedades do modelo normal

Para além das propriedades já referidas, o modelo normal goza ainda de outras propriedades importantes.

MN1 A soma de v.a.'s independentes com distribuição normal é ainda uma distribuição normal. Mais precisamente, se $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$,..., $X_n \sim N(\mu_n, \sigma_n)$ e $X_1, X_2, \ldots, X_n$ são independentes, então

$$S_n = (X_1 + X_2 + \cdots + X_n) \sim N(\mu, \sigma), \text{ onde } \mu = \sum_i \mu_i \text{ e } \sigma = \sqrt{\sum_i \sigma_i^2}.$$

Em particular, se as n variáveis $X_1, \ldots, X_n$ são independentes e todas delas têm uma distribuição normal com a mesma média e variância , i.e. se $X_i \sim N(\mu, \sigma)$, então

$$S_n = (X_1 + X_2 + \cdots + X_n) \sim N(n\mu, \sigma\sqrt{n}).$$

[5] Este resultado justifica, como referimos anteriormente, a classificação das distribuições (em platicúrticas, mesocúrticas e leptocúrticas) através da comparação do seu coeficiente de curtose com o valor 3; no fundo, estamos a considerar como "standard" o achatamento da distribuição normal.

MN2 Usando o resultado anterior em conjugação com a propriedade que já referimos de que, se $X \sim N(\mu, \sigma)$ e $Y = \alpha X + \beta$, então Y também é normal, tem-se que se $X_1 \sim N(\mu_1, \sigma_1)$,..., $X_n \sim N(\mu_n, \sigma_n)$ e $X_1, X_2, \ldots, X_n$ são independentes, então

$$(\alpha_1 X_1 + \cdots + \alpha_n X_n) \sim N(\mu, \sigma) \text{ onde } \mu = \sum_i \alpha_i \mu_i \text{ e } \sigma = \sqrt{\sum_i \alpha_i^2 \sigma_i^2}.$$

Em particular, se se $X_1, \ldots, X_n$ são independentes e todas as variáveis tiverem a mesma distribuição $X_i \sim N(\mu, \sigma)$, se designarmos por $\overline{X}$ a v.a. definida por

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}, \tag{4.11}$$

a que chamaremos **média** das v.a's $X_1, \ldots, X_n$, tem-se

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}). \tag{4.12}$$

Note-se que a variância da (v.a.) média $\overline{X}$ é menor do que a das v.a's individuais X_i (uma vez $\frac{\sigma^2}{n} < \sigma^2$), ou seja a média $\overline{X}$ tem tendência a estar mais próxima do valor médio μ do que a de cada uma das variáveis individuais X_i. Esta tendência aumenta à medida que o número de variáveis aumenta. Como consequência imediata do resultado 4.12, temos que

$$\frac{\overline{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1). \tag{4.13}$$

Como já referimos, algumas outras distribuições interessantes serão estudadas nas aulas práticas.

4.3 Teorema Limite Central e Lei dos Grandes Números

4.3.1 Amostragem

Consideremos uma dada experiência aleatória (por exemplo, lançamento de um dado com as faces numeradas de 1 a 6) e seja X uma v.a. associada a essa experiência, por exemplo, a v.a. que corresponde ao número da face que fica voltada para cima. Considerem-se n repetições nas mesmas condições dessa experiência e sejam $X_1, X_2, \ldots, X_n$ as v.a.'s associadas a cada uma das repetições. Então essas variáveis aleatórias são, naturalmente, independentes e, além disso, todas elas têm a mesma distribuição de probabilidade (distribuição igual à da v.a. X) – dizemos então que as variáveis são **independentes e identicamente distribuídas (i.i.d.)**.

Consideremos agora uma outra situação. Suponhamos que temos uma população com uma v.a. X em estudo (por exemplo, a variável *altura* na população dos estudantes da Universidade do Minho).

Consideremos uma v.a. X_1 associada à altura de um indivíduo *escolhido aleatoriamente dessa população* (i.e., o indivíduo é escolhido ao acaso, e supomos que todos os indivíduos da população têm igual probabilidade de serem escolhidos); seja X_2 a v.a. associada à altura de novo indivíduo escolhido aleatoriamente da população, e assim sucessivamente, até termos n ($n \in \mathbb{N}$) variáveis aleatórias correspondentes, respectivamente, às alturas de n indivíduos escolhidos aleatoriamente da população (indivíduos esses que constituem uma amostra casual da população). Isto corresponde a escolher ao acaso, com reposição, uma amostra da população (após "extrairmos" um indivíduo da população, ele é reposto na população) e a considerar as variáveis aleatórias associadas a cada um dos elementos da amostra.[6]

As variáveis $X_1, X_2, \ldots, X_n$ assim obtidas são também independentes e igualmente distribuídas, sendo a distribuição de cada X_i igual à distribuição de X.

- Quando, dada uma v.a. X com uma certa distribuição, tivermos n v.a.'s $X_1, X_2, \ldots, X_n$ i.i.d's com a mesma distribuição de X, dizemos que o vector aleatório $\mathbf{X} = (X_1, \ldots, X_n)$ é uma **amostra aleatória de dimensão** n de X.
- Se $\mathbf{x} = (x_1, \ldots, x_n)$ for um valor observado do vector aleatório $\mathbf{X} = (X_1, \ldots, X_n)$ [a] dizemos que $\mathbf{x} = (x_1, \ldots, x_n)$ **é uma amostra aleatória observada**.

[a] No caso de segundo exemplo acima considerado, x_1 seria a altura do primeiro indivíduo seleccionado, x_2 a altura do segundo indivíduo seleccionado, etc.

Nota: Neste momento, interessa-nos fazer a distinção entre a amostra aleatória, enquanto vector aleatório $\mathbf{X} = (X_1, \ldots, X_n)$, e a amostra observada (ou amostra de observações), constituída por dados, $\mathbf{x} = (x_1, \ldots, x_n)$. Como já referimos anteriormente, por vezes chamamos amostra ao conjunto dos dados, não sendo tão "rigorosos" nesta distinção.

4.3.2 Teorema Limite Central (TLC)

Vimos, na Secção 4.2.3, que a média $\overline{X}$ de uma amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$ da distribuição normal $X \sim N(\mu, \sigma)$ tem uma distribuição normal $N(\mu, \frac{\sigma}{\sqrt{n}})$ ou, equivalentemente, que $\frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0, 1)$.

Nesta secção apresentamos um teorema que contém um dos resultados mais interessantes da teoria da Probabilidade – o chamado Teorema Limite Central – o qual pode ser visto como uma ex-

[6] Na prática, é frequente fazer-se amostragem sem reposição (por exemplo, quando se faz uma sondagem, os indivíduos seleccionados são auscultados uma única vez) ; no entanto, desde que a dimensão N da população seja elevada e n seja pequeno, relativamente a N, podemos considerar que a amostragem com reposição é praticamente equivalente à amostragem sem reposição.

tensão dos resultados acima enunciados. Informalmente, o TLC estabelece que, sob certas condições, mesmo quando as variáveis X_i não são normais, mas apenas i.i.d., a distribuição da sua média $\overline{X}$, embora não normal, pode ser "bem aproximada"por uma distribuição normal. Mais precisamente, tem-se:

Teorema Limite Central

Sejam $X_1, X_2, X_3, \ldots$, v.a.'s i.i.d. com valor médio μ e variância finita σ^2 e sejam

$$S_n = X_1 + \cdots + X_n \tag{4.14}$$

e

$$\overline{X} = \overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{4.15}$$

Considerem-se, ainda, as v.a's Y_n e Z_n definidas por

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \quad \text{e} \quad Z_n = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}. \tag{4.16}$$

Então, a função de distribuição de Y_n e a função de distribuição de Z_n convergem pontualmente para a função de distribuição Φ da v.a. $Z \sim N(0,1)$, ou seja, tem-se

$$\lim_{n\to\infty} P\Big(Y_n \leq x\Big) = \Phi(x) \quad \text{e} \quad \lim_{n\to\infty} P\Big(Z_n \leq x\Big) = \Phi(x),$$

onde $\Phi(x)$ é a função definida por (4.10). [a]

[a] Dizemos, neste caso, que Y_n e Z_n **convergem em distribuição para** $Z \sim N(0,1)$ e escrevemos

$$Y_n \xrightarrow{d} Z \sim N(0,1) \quad \text{e} \quad Z_n \xrightarrow{d} Z \sim N(0,1)$$

.

Por outras palavras, S_n é aproximadamente $N(n\mu, \sigma\sqrt{n})$ e $\overline{X}$ é aproximadamente $N(\mu, \frac{\sigma}{\sqrt{n}})$, desde que n seja suficientemente grande. Por vezes, escrevemos

$$S_n \approx N(n\mu, \sigma\sqrt{n}) \quad \text{e} \quad \overline{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$$

com o significado acima indicado.

Na prática, isto significa que podemos aproximar probabilidades referentes à variável soma $S_n = X_1 + \cdots + X_n$, ou à variável média $\overline{X}$, por probabilidades calculadas a partir do modelo normal, seja qual for a distribuição subjacente às v.a.'s i.i.d X_i [7], desde que n seja grande. Se as v.a.'s não forem muito assimétricas, basta que seja $n > 30$ para obtermos uma aproximação razoável.

[7] Desde que estas tenham variância finita

Por exemplo, se $X_i \sim Bi(1,p)$, temos que $S_n \sim Bi(n,p)$ e o Teorema Limite Central diz-nos que, para n grande, S_n tem distribuição aproximadamente $N(np, \sqrt{np(1-p)})$.

Na Figura 4.12, apresentam-se, em sobreposição no mesmo gráfico, a f.m.p. de uma distribuição $Bi(100, 0.5)$ e a f.d.p. da $N(50, 5)$ (note-se que, para $n = 100$ e $p = 0.5$, vem $\sqrt{np(1-p)} = \sqrt{25} = 5$), onde é patente o "ajuste".

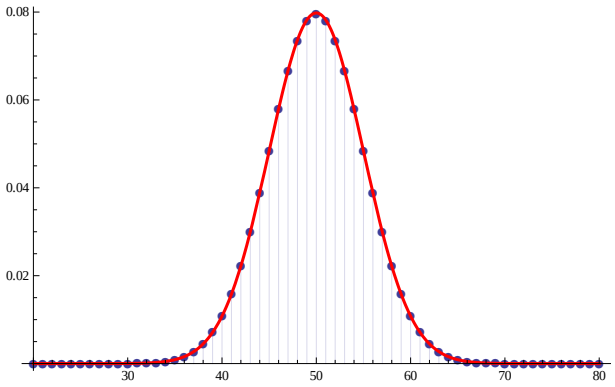

Figura 4.12: f.m.p $Bi \sim (100, 0.5)$ (azul) e f.d.p. $N(50, 5)$ (vermelho).

O TLC permite também explicar, de certo modo, por que razão se constata que muitos fenómenos naturais têm uma distribuição semelhante à normal, uma vez que muitos deles podem ser considerados como a "soma" de vários "acontecimentos" aleatórios.

4.3.3 Lei dos Grandes Números

Apresentamos agora outro resultado muito importante da teoria da probabilidade, conhecido por Lei dos Grandes Números (LGN).[8] Informalmente, podemos dizer que a LGN descreve o que acontece quando se repete uma mesma experiência um grande número de vezes. De acordo com a LGN, a média dos resultados obtidos com um elevado número de repetições deve estar próxima do valor esperado (ou valor médio) da distribuição correspondente à experiência em causa, tornando-se cada

[8] A LGN foi provada inicialmente por por Jacob Bernoulli, na quarta parte do seu trabalho *Ars Conjectandi*, publicado postumamente em 1713.

vez mais próxima quando o número de repetições aumenta. De um modo mais rigoroso, tem-se:

Lei dos Grandes Números (LGN)

Seja $X_1, X_2, \ldots,$ uma sequência de v.a.'s i.i.d., cada uma delas com valor médio $E(X_i) = \mu$. Para $n \in \mathbb{N}$, seja $\overline{X} = \overline{X}_n = \frac{X_1+\cdots+X_n}{n}$ a média de $X_1, \ldots, X_n$. Então, para qualquer $\epsilon > 0$, tem-se

$$\lim_{n\to\infty} P(|\overline{X} - \mu| < \epsilon) = 1.$$

Dizemos que $\overline{X}_n$ *converge em probabilidade* para μ e escrvemos $\overline{X}_n \xrightarrow{p} \mu$.

Como aplicação imediata da LGN, suponhamos que efectuamos uma sequência de repetições independentes de uma dada experiência aleatória. Seja A uma acontecimento fixo e seja $p = P(A)$ a probabilidade desse acontecimento ocorrer nessa experiência e $1 - p$ a probabilidade de não ocorrer. Seja

$$X_i = \begin{cases} 1, & \text{se } A \text{ ocorre na } i\text{-ésima repetição} \\ 0, & \text{se } A \text{ não ocorre na } i\text{-ésima repetição} \end{cases}$$

Então X_i são v.a.'s i.i.d. com distribuição $Bi(1, p)$ e $\mu = E(X_i) = p$. Neste caso, a v.a. $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ representa a frequência relativa da ocorrência de A nas n repetições da experiência, $f_n(A)$. A LGN estabelece, portanto, que $f_n(A)$ converge (em probabilidade) para $p = P(A)$. Assim sendo, faz sentido aproximar $P(A)$ pela frequência relativa da sua ocorrência, quando n for suficientemente grande.

4.4 Distribuições relacionadas com a distribuição normal

4.4.1 A distribuição qui-quadrado

Consideremos uma amostra aleatória $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ da v.a. $Z \sim N(0, 1)$, i.e. sejam Z_i v.a.s i.i.d. tais que $Z_i \sim N(0, 1)$. Chama-se **distribuição qui-quadrado com n graus de liberdade** à distribuição da variável

$$Q = Q_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2. \tag{4.17}$$

Se Q tem uma distribuição qui-quadrado com n graus de liberdade, escrevemos $Q \sim \chi^2_n$. Pode mostrar-se que a f.d.p. de $Q \sim \chi^2_n$ é dada por

$$f_Q(X) = f_n(x) = \begin{cases} 0, & \text{para } x \leq 0, \\ \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x^2/2} x^{n/2-1}, & \text{para } x > 0, \end{cases} \tag{4.18}$$

onde Γ designa a chamada *função gama*. [9]

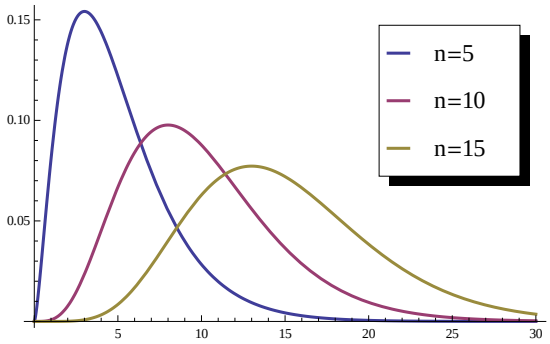

Figura 4.13: f.d.p. da distribuição $\chi^2_n; n = 5, 10, 15$.

Características teóricas da distribuição χ^2_n

Se $X \sim \chi^2_n$, tem-se

- $\mu_X = n$
- $\sigma^2_X = 2n$
- $\max\{n-2, 0\}$ é moda de X
- $\beta_1 = \sqrt{8/n}$ (quando n aumenta, o coeficiente de assimetria aproxima-se de zero).
- $\beta_2 = 3 + \frac{12}{n}$ (quando n aumenta, a curtose aproxima-se da curtose da normal).

É uma consequência imediata da definição que a soma de duas v.a.'s independentes com distribuição de qui-quadrado é ainda uma v.a. com distribuição de qui-quadrado. Mais especificamente, se $X \sim \chi^2_n$ e $Y \sim \chi^2_m$ são independentes, então $(X + Y) \sim \chi^2_{n+m}$.

Pelo TLC, uma vez que a distribuição de qui-quadrado é uma soma de n variáveis independentes[10] ela converge para uma distribuição normal. Mais especificamente, se $X \sim \chi^2_n$, então, quando

[9] A função gama pode ser vista como uma extensão da função factorial; tem-se $\Gamma(n) = (n-1)!$, se $n \in \mathbb{N}$, e $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$, para $\alpha > 0$.

[10] Sendo $Z_1, \ldots, Z_n$ independentes, pode provar-se que $Z^2_1, \ldots, Z^2_n$ também são independentes.

$n \to \infty$, $\frac{X-n}{\sqrt{2n}}$ converge (em distribuição) para a normal $Z \sim N(0,1)$, ou seja, para n grande, χ^2_n é aproximadamente $N(n, \sqrt{2n})$); veja a Figura4.18.

No Mathematica, a função associada à distribuição qui-quadrado é a função **ChiSquareDistribution**.

4.4.2 A distribuição t de Student

Se Z e Q_n são v.a.'s independentes tais que $Z \sim N(0,1)$ e $Q \sim \chi^2_n$, então dizemos que a v.a. $T = T_n$ definida por

$$T = T_n = \frac{Z}{\sqrt{Q_n}}\sqrt{n} \tag{4.19}$$

tem uma distribuição t **de Student** (por vezes apenas referida como distribuição t) **com** n **graus de liberdade**. Se T tem distribuição t de Student com n graus de liberdade, escrevemos simplesmente $T \sim t_n$. Pode provar-se que a f.d.p. de $X \sim t_n$ é dada por

$$f_T(x) = f_n(X) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \; x \in \mathbb{R}. \tag{4.20}$$

Na Figura 4.14 apresentam-se gráficos de f.d.p. da distribuição t_n para diversos valores de n.

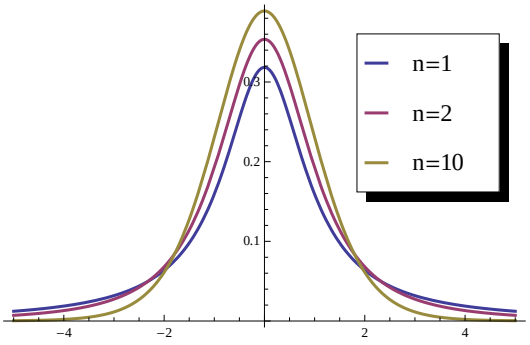

Figura 4.14: f.d.p. da distribuição $t_n; n = 1, 2, 10$.

Características teóricas da distribuição t de Student

Se $X \sim t_n$, então:

- Para $n > 1$, $\mu_X = 0$ (para $n = 1$, μ_X não existe)
- Para $n > 2$, $\sigma_X^2 = \frac{n}{n-2}$ (para $n \leq 2$, σ_X^2 não existe)
- Para $n > 3$, $\beta_1 = 0$ (para $n \leq 3$, β_1 não existe)
- Para $n > 4$, $\beta_2 = 3 + \frac{6}{n-4}$ (para $n \leq 4$, β_2 não existe)

Pode provar-se que esta distribuição é aproximadamente $N(0,1)$ para valores de n grandes.

No Mathematica, a função associada à distribuição t de Student é a função **StudentTDistribution**.

4.4.3 A distribuição F de Fisher-Snedecor

Se X e Y são duas v.a.'s independentes tais que $X \sim \chi_n^2$ e $Y \sim \chi_m^2$, então dizemos que a variável $V = V_{n,m}$ definida por

$$V = V_{n,m} = \frac{X/n}{Y/m} \tag{4.21}$$

tem uma **distribuição F de Fisher-Snedecor** (ou apenas uma **distribuição F**) **com n e m graus de liberdade**. Nesse caso, escrevemos $V \sim F_{n,m}$. Pode mostrar-se que, se $V \sim F_{n,m}$, então a sua f.d.p. é dada por

$$f_{n,m}(x) = \begin{cases} 0, & \text{para } x \leq 0, \\ c_{n,m} \dfrac{x^{(n-2)/2}}{(m+nx)^{(n+m)/2}}, & \text{para } x > 0, \end{cases} \tag{4.22}$$

onde

$$c_{n,m} = \frac{\Gamma(\frac{n+m}{2}) n^{n/2} m^{m/2}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}. \tag{4.23}$$

Na Figura 4.15 apresentam-se gráficos da f.d.p. da distribuição $F_{n,m}$ para $n = 2, 5, 20$ e $m = 10$.

Características teóricas da distribuição F

Se $X \sim F_{n,m}$, então:

- Para $m > 2$, $\mu_X = \frac{m}{m-2}$ (para $m \leq 2$, μ_X não existe)
- Para $m > 4$, $\sigma_X^2 = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$ (para $m \leq 4$, σ_X^2 não existe)

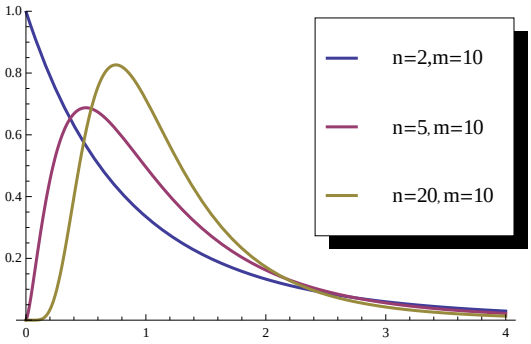

Figura 4.15: f.d.p. da distribuição $F_{n,m}; n = 2, 5, 20; m = 10$.

Não damos aqui as expressões do coeficiente de assimetria β_1 e do coeficiente de curtose β_2, por terem expressões complicadas.

No Mathematica, para trabalhar com a distribuição F, deve usar a função **FRatioDistribution**. Pode usá-la, por exemplo, para determinar as expressões dos coeficientes de assimetria e de curtose.

4.4.4 Distribuições por amostragem

Seja $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória da variável X e suponhamos que X tem valor médio μ e variância σ^2. Relembremos que a média amostral de $X_1, \ldots, X_n$ é a v.a. dada por

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n},$$

e que se tem

$$E(\overline{X}) = \mu \quad \text{e} \quad \text{var}(\overline{X}) = \frac{\sigma^2}{n}. \tag{4.24}$$

Definimos agora uma nova v.a., a que chamamos **variância amostral**, e que denotamos por S^2, do seguinte modo

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{4.25}$$

A v.a. $S = \sqrt{S^2}$ é chamada **desvio padrão amostral**.

Não é difícil de mostrar que

$$E(S^2) = \sigma^2 \tag{4.26}$$

isto é, que o valor médio da v.a. variância amostral é igual à variância populacional.

Distribuições por amostragem de uma v.a. normal

No caso particular em que $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ é uma amostra aleatória de uma v.a. $N(\mu, \sigma)$, já sabemos que a v.a. $\overline{X}$ também é normal. Mais precisamente, sabemos que

$$\overline{X} \sim N(\mu, \frac{\sigma}{n}),$$

e, portanto, que

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1).$$

O teorema seguinte (que não demonstraremos) diz-nos qual a distribuição da v.a. variância amostral, S^2, e garante, além disso, que $\overline{X}$ e S^2 são v.a.'s independentes.

Teorema 4.1. *Se $\mathbf{X} = (X_1, \ldots, X_n)$ é uma amostra aleatória da distribuição $N(\mu, \sigma)$, então:*

1. *$\overline{X}$ e S^2 são v.a.'s independentes*

2. *$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$*

3. *$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$.*

Nota: O resultado 2. já tinha sido referido anteriormente, sendo aqui apenas relembrado, por uma questão de completude.

Como corolário imediato do teorema anterior, tendo em atenção a definição da distribuição t de Student, tem-se o seguinte corolário.

Corolário 4.1. *Nas condiçoes do teorema anterior, tem-se*

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

O teorema seguinte caracteriza a distribuição do quociente das variâncias de duas amostras

aleatórias independentes provenientes de distribuições normais.

Teorema 4.2. *Sejam $\mathbf{X} = (X_1, \ldots, X_n)$ uma amostra aleatória de dimensão n de uma distribuição $N(\mu, \sigma)$ e $\mathbf{Y} = (Y_1, \ldots, Y_m)$ uma amostra aleatória de dimensão m de uma distribuição $N(\mu', \sigma')$ e suponhamos que as amostras são independentes uma da outra. Então, tem-se*

$$\frac{S^2/\sigma^2}{S'^2/\sigma'^2} \sim F_{n-1,m-1}$$

4.5 Distribuição normal bivariada

Quando temos simultaneamente duas variáveis em estudo, X e Y, eventualmente relacionadas, é necessário recorrer a pares aleatórios (X, Y), isto é, a distribuições bivariadas. De entre as distribuições bivaridas, tem especial importância a normal bivariada ou binormal, a qual pode ser vista como uma generalização da normal para duas variáveis.

Sejam X e Y duas v.a.'s com distribuição normal, $X \sim N(\mu_X, \sigma_x)$ e $Y \sim N(\mu_Y, \sigma_Y)$. Diz-se que o par aleatório (X, Y) tem uma **distribuição normal bivariada** ou **binormal** ou que é um **par gaussiano**, se a sua função densidade de probabilidade conjunta for dada por

$$f_{X,Y}(x, y) = f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{z}{2(1-\rho)^2}}, \tag{4.27}$$

onde

$$z = z(x, y) = \Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2 \tag{4.28}$$

e onde ρ designa a correlação de X e Y. Nestas condições, escrevemos $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$. Neste caso, é fácil de ver que, se $\rho = 0$, então

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2} \\ &= f_X(x) f_Y(y) \end{aligned}$$

ou seja a função densidade de probabilidade conjunta é o produto das densidades marginais, isto é, as v.a. X e Y são independentes.

Como sabemos, o recíproco é também verdadeiro, isto é, se X e Y são independentes, então $\rho = 0$. Assim, no caso de um par aleatório (X, Y) com distribuição normal bivariada, tem-se que X e Y são independentes se e só se $\rho = 0$.

É também imediato reconhecer que, se $X \sim N(\mu_X, \sigma_x)$ e $Y \sim N(\mu_Y, \sigma_Y)$ e X e Y forem independentes, então (X, Y) tem distribuição normal bivariada (mas o recíproco não é verdadeiro).

5

Inferência Estatística

A inferência estatística compreende um conjunto de métodos que têm por objectivo usar a informação constante nos dados (amostra) para responder a questões específicas sobre a população.

Tornamos a relembrar que, dada uma v.a. X com uma certa distribuição, dizemos que $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ é uma amostra aleatória de X (de dimensão n), se as v.a.'s X_i são i.i.d. com X, isto é, são réplicas independentes de X. A um valor observado $\mathbf{x} = (x_1, \ldots, x_n)$ desse vector aleatório, chamamos amostra aleatória observada.

No âmbito da inferência estatística chamamos **estatística** a qualquer função das v.a.'s que constituem a amostra aleatória, desde que tal função não inclua parâmetros desconhecidos. Note-se que, neste sentido, uma estatística é ainda uma v.a.. Também chamamos estatística a um valor observado dessa v.a.

Suponhamos que conhecemos a distribuição da qual os dados provêm, a menos de um ou mais parâmetros. Por exemplo, suponhamos que sabemos que os dados provêm de uma v.a. com distribuição normal $N(\mu, \sigma)$, mas que não conhecemos μ, ou que desconhecemos mesmo μ e σ, ou que sabemos que os dados vêm de uma distribuição $Poi(\lambda)$, cujo valor médio λ é desconhecido.

Problemas em que a forma da distribuição subjacente aos dados é especificada a menos de um conjunto de parâmetros desconhecidos e em que pretendemos *estimar* esses parâmetros são chamados problemas de estimação paramétrica ou de parâmetros.

5.1 Estimação de parâmetros

5.1.1 Estimação pontual

Suponhamos então que conhecemos o modelo X do qual provêm os dados, a menos de um ou mais parâmetros, i.e., por exemplo, que conhecemos a sua f.d. $F(x;\theta)$, a qual depende de um parâmetro θ (ou de um vector de parâmetros $\theta = (\theta_1, \ldots, \theta_r)$) não conhecido. Põe-se então o problema de estimar o valor do parâmetro desconhecido θ (ou de cada um dos parâmetros θ_i).

Conhecida uma amostra aleatória concreta $\mathbf{x} = (x_1, \ldots, x_n)$, podemos interpretar essa amostra como uma realização de uma amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$ de X.

Qualquer estatística (no sentido de v.a.) T_n baseada na amostra aleatória $\mathbf{X}$, isto é, qualquer função da amostra aleatória que não envolva o parâmetro desconhecido θ, usada para estimar θ, é chamada um **estimador** de θ.

O valor observado de T_n na amostra concreta $\mathbf{x} = (x_1, \ldots, x_n)$ é chamado uma **estimativa** de θ.

Assim, um estimador é uma v.a. (com uma certa distribuição) e uma estimativa é um número real (que, com uma "boa" escolha do estimador, deverá ser uma "aproximação razoável" para o valor do parâmetro).

Por exemplo, um estimador usualmente utilizado para estimar a média μ de uma distribuição normal $N(\mu, \sigma)$, baseado numa amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$ dessa população é a (v.a.) média amostral $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ e um valor $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ obtido a partir da amostra observada (i.e. concreta) $\mathbf{x} = (x_1, \ldots, x_n)$ é uma estimativa de μ (correspondente ao estimador $\overline{X}$).

Há vários métodos para encontrar estimadores pontuais e vários critérios que especificam o que se entende por um "bom estimador pontual" do parâmetro θ.

Seja $T = T_n$ um estimador de um dado parâmetro θ.[1] Dizemos que T_n é um estimador:

- **centrado** ou **não enviesado** de θ se verificar

$$E(T_n) = \theta.$$

  A $E(T_n) - \theta$ chamamos **viés** do estimador T_n para θ.

- **assimptoticamente centrado** se verificar

$$\lim_{n \to \infty} E(T_n) = \theta.$$

[1] Usamos o índice n para salientar qual o tamanho da amostra aleatória em que se baseia o estimador.

- **eficiente** se verificar
$$\lim_{n\to\infty} \text{var}(T_n) = 0$$
- **consistente** se T_n convergir (em probabilidade) para θ (na prática isto significa que, para n grande, a probabilidade de T_n estar próximo de θ é próxima de 1).

Dados dois estimadores centrados, diremos que é mais eficiente aquele que tiver menor variância (devendo esse ser o utilizado).

Nalguns modelos paramétricos, é possível encontrar um estimador do parâmetro, centrado e com variância mínima (i.e. com variância inferior à de qualquer outro estimador centrado desse parâmetro).

Pode mostrar-se que se T_n for um estimador centrado (ou assimptoticamente centrado) e eficiente, então T_n é consistente.

Exemplo 5.1. *Considere-se uma amostra aleatória (de tamanho n) $X_1, \ldots, X_n$ proveniente do modelo $N(\mu, 1)$. É natural considerar a média amostral $\overline{X}$ como estimador para μ. Como sabemos (veja 4.12), tem-se $E(\overline{X}) = \mu$ e $\text{var}(\overline{X}) = \frac{\sigma^2}{n} = \frac{1}{n}$. Então, podemos concluir que a média amostral é um estimador centrado e eficiente, pelo que é um estimador consistente. Pode provar-se que este é o "melhor"estimador centrado que existe para μ, no sentido de que não há outro estimador centrado para μ que tenha menor variância do que $\overline{X}$.*

Exemplo 5.2. *No modelo $N(\mu, \sigma)$, a variância amostral $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ é um estimador consistente de σ^2. De facto, se relembramos o Teorema 4.1, sabemos que*
$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$
Por outro lado, também sabemos que $E(\chi^2_{n-1}) = n - 1$ e que $\text{var}(\chi^2_{n-1}) = 2(n-1)$; veja a subsecção com as características teóricas da distribuição de qui-quadrado. Assim sendo, temos, devido às propriedades do valor médio e da variância,
$$E((n-1)\frac{S^2}{\sigma^2}) = n - 1 \Rightarrow \frac{n-1}{\sigma^2}E(S^2) = n - 1 \Rightarrow E(S^2) = \sigma^2$$
e
$$\text{var}((n-1)\frac{S^2}{\sigma^2}) = 2(n-1) \Rightarrow \frac{(n-1)^2}{\sigma^4}\text{var}(S^2) = 2(n-1) \Rightarrow \text{var}(S^2) = \frac{2\sigma^4}{n-1}.$$
Como $E(S^2) = \sigma^2$, concluímos que S^2 é um estimador centrado para σ^2. Por outro lado, como $\text{var}(S^2) \to 0$ quando $n \to \infty$, concluímos que o estimador é eficiente, pelo que será também consistente.

Mais geralmente, pode mostrar-se que, em qualquer modelo, S^2 é um estimador consistente da variância σ^2.

Estimativas (e estimadores) de máxima verosimilhança

Dos vários métodos de estimação pontual que existem, referimos, de forma muito breve, o chamado **método da máxima verosimilhança**, que permite, em geral, obter estimadores dos parâmetros com boas propriedades.

Novamente, suponhamos que temos uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$ de uma v.a. X, vista como um valor observado de uma amostra aleatória aleatória de X, $\mathbf{X} = (X_1, \ldots, X_n)$. Designemos por $f(x_1, x_2, \ldots, x_n; \theta)$ a f.m.p. conjunta, no caso de X_i serem discretas ou a f.d.p. conjunta, quando X_i são contínuas. Estamos a supor que $f(x_1, \ldots, x_n; \theta)$ é conhecida, a menos do parâmetro θ.

Podemos pensar nessa função como descrevendo a probabilidade de observar a amostra concreta $\mathbf{x} = (x_1, \ldots, x_n)$, para um determinado valor θ.

A **estimativa de máxima verosimilhança (m.v.)** de θ, denotada por $\hat{\theta}$, é obtida determinando o valor de θ que maximiza a função $f(x_1, \ldots, x_n; \theta)$ (considerando os valores de $x_1, \ldots, x_n$ fixos). Dito de outro modo, ao calcular a estimativa $\hat{\theta}$ de m.v., respondemos à questão: para que valor do parâmetro θ é que é "mais verosímil"que os dados observados tenham sido obtidos?

Segue-se um quadro com as estimativas de m.v. dos parâmetros de alguns modelos mais usuais

Modelo	Parâmetro	Estimativa m.v.
$Bi(1,p)$	p	$\hat{p} = \overline{x}$
$Poisson(\lambda)$	λ	$\hat{\lambda} = \overline{x}$
$N(\mu,\sigma)$	μ	$\hat{\mu} = \overline{x}$
$N(\mu,\sigma)$	σ	$\hat{\sigma} = \sqrt{m_2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
$Exp(\lambda)$	λ	$\hat{\lambda} = 1/\overline{x}$
$U[a,b]$	a	$\hat{a} = x_{(1)} = \min_i\{x_i\}$
$U[a,b]$	b	$\hat{b} = x_{(n)} = \max_i\{x_i\}$.

As estimativas de máxima verosimilhança têm a chamada propriedade de **invariância**, a qual significa que a estimativa de máxima verosimilhança de uma função de θ, $\tau(\theta)$, é dada por $\tau(\hat{\theta})$, i.e. que se tem

$$\widehat{\tau(\theta)} = \tau(\hat{\theta}).$$

Dada uma estimativa de m.v., a correspondente estatística (enquanto v.a.) designa-se por **estimador de máxima verosimilhança**.

Assim, por exemplo, sendo a estimativa de m.v. da média μ, no modelo normal $N(\mu, \sigma)$, dada pela média amostral (baseada na amostra observada $\mathbf{x} = \{x_1, \ldots, x_n\}$), $\overline{\mathbf{x}} = \frac{x_1+\cdots+x_n}{n}$, o correspondente estimador de m.v. será a (v.a.) média amostral $\overline{X} = \frac{X_1+\cdots+X_n}{n}$.

De notar que, de acordo com a tabela acima (e atendendo à propriedade de invariância das estimativas de m.v.), temos que o estimador de m.v. da variância populacional σ^2 é o momento centrado de ordem 2,

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Este estimador não é centrado, embora seja assimptoticamente centrado e consistente.

5.1.2 Estimação intervalar

Como acabámos de ver, no caso da estimação pontual para um certo parâmetro θ, é escolhido um estimador, i.e. uma estatística T (função da amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$), com "boas propriedades"e calcula-se uma estimativa para θ com base no valor observado dessa estatística.

Na chamada *estimação intervalar*, a ideia é encontrar duas estatísticas T_1 e T_2, tais que $T_1 \leq T_2$ e para as quais se tenha $T_1 < \theta < T_2$ com uma determinada probabilidade (grande e fixada *a priori*). Mais precisamente, escolhido um determinado valor de α (pequeno, tipicamente $\alpha = 0.05, 0.01, 0.005$), encontram-se duas estatísticas T_1 e T_2 para as quais seja

$$P(T_1 < \theta < T_2) = 1 - \alpha.$$

Nesse caso, dizemos que o intervalo (aleatório) cujos extremos são essas estatísticas T_1 e T_2, i.e. o intervalo (T_1, T_2), é um **intervalo de** $(1-\alpha)\times 100\%$ **de confiança** ou um **intervalo com uma margem de erro de** $\alpha \times 100\%$ para θ. Em vez de um estimador pontual para θ, temos agora um estimador intervalar para esse parâmetro.

Assim por exemplo, fixada uma probabilidade $1 - \alpha = 0.95$, falaremos num intervalo de 95% de confiança para o parâmetro que estamos a estimar ou diremos que esse intervalo tem uma margem de erro de 5%.

Quando consideramos valores observados das estatísticas que definem os extremos dos intervalos aleatórios, encontramos então verdadeiros intervalos da recta real, os quais são *estimativas intervalares* para o parâmetro em causa. A estes intervalos chamamos também intervalos de confiança (neste caso, deterministas, ou seja, não aleatórios) para esse parâmetro.

Intervalo de confiança para o valor médio em população normal $N(\mu, \sigma)$, com σ conhecido

Comecemos por analisar o caso correspondente a uma amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$ proveniente da distribuição normal $N(\mu, \sigma)$, em que σ é conhecido, mas μ é desconhecido. Vimos, no

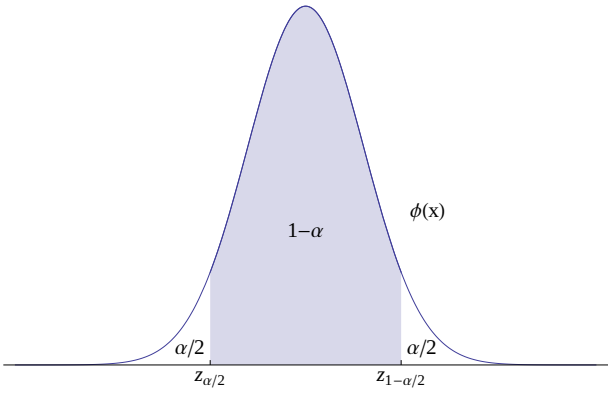

Figura 5.1: Quantis $-c = \chi_{\alpha/2}$ e $c = \chi_{1-\alpha/2}$ da distribuição $N(0, 1)$

capítulo anterior – relembre o Teorema 4.1 –, que a v.a. $\frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ tem uma distribuição $N(0, 1)$. Assim, se representarmos por $z_{1-\alpha/2}$ o quantil de probabilidade $1 - \alpha/2$ da distribuição $Z \sim N(0, 1)$ (donde, será $-z_{1-\alpha/2}$ o quantil de probabilidade $\alpha/2$) – veja a Figura 5.1 – temos que

$$P\Big(-z_{1-\alpha/2} < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < z_{1-\alpha/2}\Big) = 1 - \alpha$$

ou seja, podemos escrever, explicitando μ na dupla desigualdade

$$P\Big(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) = 1 - \alpha.$$

Podemos, pois, afirmar que o intervalo aleatório

$$\Big(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) \tag{5.1}$$

contém o valor μ desconhecido, com probabilidade $1 - \alpha$, ou seja, que o intervalo dado por (5.1) é um intervalo de $(1 - \alpha) \times 100\%$ de confiança para μ, no contexto de uma população $N(\mu, \sigma)$, supondo σ conhecido.

Neste caso, as duas estatísticas T_1 e T_2 de que falámos anteriormente são $T_1 = T_1(X_1, \ldots, X_n) = \overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ e $T_2 = T_2(X_1, \ldots, X_n) = \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Nota: Repare que a amplitude dos intervalos de confiança (5.1), que é dada por $2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$, diminui quando o tamanho da amostra, n, aumenta e aumenta quando a margem de erro diminui (ou seja, a confiança aumenta): quando α diminui, i.e quando $1-\alpha$ aumenta, o quantil $z_{1-\alpha/2}$ aumenta, já que a função $\Phi(x)$ é crescente.

Ao afirmarmos que o intervalo dado por (5.1) é um intervalo de $(1-\alpha)\times 100\%$ de confiança para μ queremos dizer que, se calculássemos intervalos concretos, com base em sucessivas amostras, a proporção de intervalos obtidos contendo o parâmetro μ convergiria (em probabilidade) para $(1-\alpha)$. Ou, de outro modo, se considerarmos um número elevado N de amostras, a frequência relativa

$$f_N = \frac{\text{número de intervalos que contêm } \mu}{N}$$

deverá ser próxima de $1-\alpha$.

Exemplo 5.3. *Admita que a variável aleatória que representa o tempo de vida (em horas) de uma determinada bactéria tem distribuição normal $N(\mu,\sigma)$ com $\sigma = 0.18$. Foi estudada uma amostra aleatória de 8 bactérias desse tipo e o tempo médio de vida dessas bactérias foi de 1.63. Determine um intervalo de 95% de confiança o parâmetro μ dessa distribuição.*

Resolução*: Temos $\overline{x} = 1.63$, $\sigma = 0.18$, $n = 8$ e $\alpha = 0.05$. Se determinarmos o quantil de probabilidade $p = 1-\alpha/2 = 0.975$ da distribuição $N(0,1)$, tem-se que $z_{0.975} = 1.96$. Então, vem*

$$z_{0.975}\frac{\sigma}{\sqrt{n}} = 1.96\frac{0.18}{\sqrt{8}} = 0.12,$$

pelo que um intervalo de 95% de confiança para μ é

$$(1.63-0.12, 1.63+0.12) = (1.51, 1.75).$$

É importante salientar que, ao dizermos que $(1.51, 1.75)$ é um intervalo de $95\times$ % de confiança para μ, não estamos a afirmar que a probabilidade de μ pertencer ao intervalo $(1.51, 1.75)$ é de 0.95. Este intervalo, obtido com a amostra particular, ou contém ou não contém μ – não resta nenhuma aleatoriedade quando usamos a amostra concreta, por isso falar em probabilidade não faz qualquer sentido. Quando dizemos que esse intervalo é um intervalo de 95% de confiança para μ, isto deve ser interpretado apenas no seguinte sentido: "esse intervalo foi obtido por um processo que, em aproximadamente 95% dos casos, fornece intervalos que contêm μ."

Nota: O intervalo de confiança (5.1) para o valor médio, supondo a variância conhecida (finita), pode também ser utilizado para um modelo não normal, no caso de termos uma amostra de grande dimensão, em consequência do Teorema Limite Central. Mais precisamente, neste caso apenas

poderemos afirmar que, para n grande, se tem

$$P\Big(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) \approx 1-\alpha.$$

Assim sendo, dizemos que o intervalo $\Big(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big)$ é um intervalo de aproximadamente $(1-\alpha)\times 100\%$ de confiança para μ, ou que é um intervalo de $(1-\alpha)\times 100\%$ de confiança *assimptótico* para o valor médio μ.

Intervalo de confiança para o valor médio em população normal $N(\mu,\sigma)$, com σ desconhecido

Vejamos agora o caso (mais realista) de termos uma amostra aleatória $\mathbf{X} = (X_1,\ldots,X_n)$ proveniente da distribuição normal $N(\mu,\sigma)$, em que μ e σ são ambos desconhecidos. Neste caso, o intervalo (5.1) não é um estimador intervalar para μ, uma vez que os seus extremos envolvem o parâmetro desconhecido σ, ou seja, não são estatísticas. Mas, sendo $X \sim N(\mu,\sigma)$, sabemos, pelo Teorema 4.1, que a v.a. $\frac{\overline{X}-\mu}{S}\sqrt{n}$, em que S é a v.a. desvio padrão amostral, tem distribuição t de Student com $n-1$ graus de liberdade. Assim sendo, tendo em conta a simetria da distribuição de t de Student, se designarmos por $t_{n-1,1-\alpha/2}$ o quantil de probabilidade $1-\alpha/2$ de t_{n-1}, teremos

$$P(-t_{n-1,1-\alpha/2} < \frac{\overline{X}-\mu}{S}\sqrt{n} < t_{n-1,1-\alpha/2}) = 1-\alpha$$

ou seja, teremos que

$$P\Big(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\Big) = 1-\alpha$$

ou seja, podemos dizer que

$$\Big(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\Big) \tag{5.2}$$

é um intervalo de $(1-\alpha)\times 100\%$ confiança para o valor médio μ, no contexto de uma distribuição $N(\mu,\sigma)$ com σ desconhecido.

Intervalo de confiança para a variância em população normal $N(\mu,\sigma)$

Suponhamos agora que temos uma amostra aleatória $\mathbf{X} = (X_1,\ldots,X_n)$ proveniente da distribuição normal $N(\mu,\sigma)$, em que σ é desconhecido, e que pretendemos obter um intervalo de confiança para σ (ou para a variância σ^2).

Sabemos que a v.a. $(n-1)\frac{S^2}{\sigma^2}$ tem uma distribuição χ^2_{n-1}. Sendo $u_{n-1,1-\alpha_2}$ o quantil de probabilidade $1-\alpha/2$ da distribuição χ^2_{n-1} e $u_{n-1,\alpha/2}$ o quantil de probabilidade $\alpha/2$ da mesma

distribuição (neste caso, como χ^2_{n-1} não é simétrica não temos $u_{n-1,\alpha/2} = -u_{n-1,1-\alpha_2}$), podemos escrever

$$P\Big(u_{n-1,\alpha/2} < (n-1)\frac{S^2}{\sigma^2} < u_{n-1,1-\alpha/2}\Big) = 1-\alpha$$

de onde se obtém

$$P\Big(\frac{n-1}{u_{n-1,1-\alpha/2}}S^2 < \sigma^2 < \frac{n-1}{u_{n-1,\alpha/2}}S^2\Big) = 1-\alpha.$$

Concluímos, assim, que

$$\Big(\frac{n-1}{u_{n-1,1-\alpha/2}}S^2, \frac{n-1}{u_{n-1,\alpha/2}}S^2\Big) \tag{5.3}$$

é um intervalo de $(1-\alpha)\times 100\%$ de confiança para σ^2. Consequentemente, ter-se-á o seguinte intervalo de $(1-\alpha)\times 100\%$ de confiança para o desvio padrão σ, no âmbito de uma distribuição normal $N(\mu,\sigma)$:

$$\Big(\sqrt{\frac{n-1}{u_{n-1,1-\alpha/2}}}S, \sqrt{\frac{n-1}{u_{n-1,\alpha/2}}}S\Big). \tag{5.4}$$

No quadro seguinte, apresenta-se uma tabela dos intervalos de confiança em população normal $N(\mu,\sigma)$.

Notações

Em tudo quanto se segue, os quantis p das distribuições $Z \sim N(0,1)$, $T \sim t_n$, $U \sim \chi^2_n$ e $V \sim F_{n,m}$ serão denotados por z_p, $t_{n,p}$, $u_{n,p}$ e $v_{n,m,p}$, respectivamente.

INTERVALOS DE DE CONFIANÇA EM POPULAÇÃO NORMAL $N(\mu,\sigma)$

- **Intervalo de $(1-\alpha)\times 100\%$ de confiança para μ (σ conhecido)**

$$\Big(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big)$$

- **Intervalo de $(1-\alpha)\times 100\%$ de confiança de para μ (σ desconhecido)**

$$\Big(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\ ,\ \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\Big)$$

- **Intervalo de $(1-\alpha)\times 100\%$ de confiança para σ^2**

$$\Big(\frac{n-1}{u_{n-1,1-\alpha/2}}S^2\ ,\ \frac{n-1}{u_{n-1,\alpha/2}}S^2\Big)$$

Intervalo de confiança (assimptótico) para p em população $Ber(p)$

Seja X uma v.a. de Bernoulli com parâmetro p (desconhecido) e $\mathbf{X} = (X_1, \ldots, X_n)$ uma amostra aleatória de X, i.e. $X_1, \ldots, X_n$ são independentes e $X_i \sim Ber(1, p)$. Como sabemos, o valor médio de X_i é $\mu = p$ e a sua variância é dada por $s^2 = p(1-p)$. Assim sendo, pelo Teorema do Limite Central, se n for grande, temos que a v.a. $\frac{\overline{X}-p}{\sqrt{p(1-p)}}\sqrt{n}$ pode ser aproximada pela $N(0,1)$. Então, se $c = z_{1-\alpha/2}$, podemos escrever, se n for grande

$$P\Big(-c < \frac{\overline{X}-p}{\sqrt{p(1-p)}}\sqrt{n} < c\Big) \approx 1-\alpha.$$

Com base na desigualdade anterior, e com algum esforço, pode então obter-se o seguinte intervalo de $100(1-\alpha)\%$ de confiança (assimptótico) para p:

$$\Big(\frac{2n\overline{X} + c^2 - z_{1-\alpha/2}\sqrt{4n\overline{X} + c^2 - 4n\overline{X}^2}}{2(n+c^2)}, \frac{2n\overline{X} + c^2 + c\sqrt{4n\overline{X} + c^2 - 4n\overline{X}^2}}{2(n+c^2)}\Big), \quad c = z_{1-\alpha/2}. \tag{5.5}$$

Como a expressão do intervalo anterior é bastante complicada, usa-se muitas vezes um intervalo de confiança aproximado mais simples, dado por

Intervalo de confiança em população $Ber(p)$
Intervalo de (aproximadamente) $(1-\alpha)\times 100\%$ de confiança para p
$$\Big(\overline{X} - z_{1-\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\,,\, \overline{X} + z_{1-\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\Big).$$

Nota: A justificação deste outro intervalo de confiança assimptótico baseia-se no facto de se poder mostrar que a v.a. $Y = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n}$, onde $\hat{p}$ é o estimador de m.v. de p (ou seja, $\hat{p} = \overline{X}$) ser aproximadamente $N(0,1)$, quando n é grande.

Intervalo de confiança para a diferença dos valores médios de duas populações normais (amostras independentes, σ e σ' conhecidos)

Passamos agora ao caso do estudo de duas amostras, começando com o caso de duas amostras independentes provenientes de modelos normais.

Consideremos, então, duas amostras aleatórias $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, independentes (uma da outra), provenientes de duas v.a.'s $X \sim N(\mu, \sigma)$ e $Y \sim N(\mu', \sigma')$, em que μ e μ' são desconhecidos, mas em que σ e σ' são conhecidos. Neste caso, devido aos resultados estabelecidos no Capítulo 4

(relativos a combinações lineares de v.a.'s normais e independentes) facilmente se conclui que a v.a. $\overline{X} - \overline{Y}$ tem distribuição $N(\mu - \mu', \sigma^*)$ onde $\sigma^* = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{m}}$. Isto significa que temos

$$\frac{\overline{X} - \overline{Y} - (\mu - \mu')}{\sigma^*} \sim N(0,1).$$

Então, de modo análogo ao que fizemos para encontrar um intervalo de confiança para o valor médio numa população normal (com σ conhecido), obtém-se o seguinte intervalo de $(1-\alpha)\times 100\%$ de confiança para a diferença dos valores médios $\mu - \mu'$:

$$\left(\overline{X} - \overline{Y} - z_{1-\alpha/2}\sigma^*, \overline{X} - \overline{Y} + z_{1-\alpha/2}\sigma^*\right),$$

onde, como habitualmente, reservamos a notação z_p para o quantil de probabilidade p da normal reduzida $Z \sim N(0,1)$.

Intervalo de confiança para a diferença dos valores médios de duas populações normais (amostras independentes, σ e σ' desconhecidos, $\sigma = \sigma'$)

No caso em que σ e σ' são desconhecidos, mas iguais entre si, isto é, no caso de sabermos que $\sigma = \sigma'$, pode provar-se que a v.a.

$$T = \frac{\overline{X} - \overline{Y} - (\mu - \mu')}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

onde

$$S_P^2 = \frac{(n-1)S^2 + (m-1)S'^2}{n+m-2},$$

tem uma distribuição t de Student com $n+m-2$ graus de liberdade. Assim, fazendo uso dessa variável, facilmente se determina o seguinte intervalo de $(1-\alpha)\times 100\%$ de confiança para a diferença dos valores médios, neste caso:

$$\left(\overline{X} - \overline{Y} - t_{n+m-2,1-\alpha/2}S_P\sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{X} - \overline{Y} + t_{n+m-2,1-\alpha/2}S_P\sqrt{\frac{1}{n} + \frac{1}{m}}\right).$$

Intervalo de confiança para a diferença dos valores médios de duas populações normais (amostras independentes, σ e σ' desconhecidos)

Se σ e σ' são desconhecidos e quaisquer, é ainda possível obter uma fórmula aproximada para o intervalo de confiança para a diferença dos valores médios (aproximação de Welch-Satterhwaite), o

qual envolve uma modificação nos graus de liberdade da distribuição de t de Student. A expressão do intervalo de (aproximadamente) $(1-\alpha)\times$ % confiança é a seguinte:

$$\Big(\overline{X}-\overline{Y}-t_{\nu,1-\alpha/2}\sqrt{\frac{S^2}{n}+\frac{S'^2}{m}}\,,\,\overline{X}-\overline{Y}+t_{\nu,1-\alpha/2}\sqrt{\frac{S^2}{n}+\frac{S'^2}{m}}\Big),$$

onde ν é natural mais próximo de

$$\tilde{\nu}=\frac{\left(\frac{S^2}{n}+\frac{S'^2}{m}\right)^2}{\frac{S^4}{n^2(n-1)}+\frac{S'^4}{m^2(m-1)}}.$$

Intervalo de confiança para o quociente das variâncias de duas populações normais (amostras independentes)

No caso que temos vindo a considerar, de duas amostras independentes provenientes de modelos normais $X\sim(\mu,\sigma)$ e $Y\sim(\mu',\sigma')$, sabemos, pelo Teorema 4.2, que a v.a.

$$V=\frac{S^2/\sigma^2}{S'^2/\sigma'^2}$$

tem uma distribuição $F_{n-1,m-1}$ (i.e. uma distribuição F de Fisher-Snedecor com $n-1$ e $m-1$ graus de liberdade). Com base nesta v.a., é fácil de obter o seguinte intervalo de confiança de $(1-\alpha)\times 100\%$ para o quociente de variâncias σ^2/σ'^2:

$$\Big(\frac{1}{v_{n-1,m-1,1-\alpha/2}}\frac{S^2}{S'^2},\frac{1}{v_{n-1,m-1,\alpha/2}}\frac{S^2}{S'^2}\Big),$$

onde $v_{n-1,m-1,p}$ designa o quantil de probabilidade p da distribuição $F_{n-1,m-1}$.

Nota: Se substituirmos os extremos deste intervalo de confiança, pelas suas raízes quadradas, obteremos um intervalo de confiança para o quociente dos desvios padrões σ/σ'.

No quadro seguinte, apresenta-se uma tabela dos intervalos de confiança para a diferença de valores médios e para o quociente de variâncias em duas populações normais, no caso de amostras independentes.

INTERVALOS DE DE CONFIANÇA EM DUAS POPULAÇÕES NORMAIS

Amostras Independentes

- **Intervalo de $(1-\alpha)\times 100\%$ de confiança para $\mu-\mu'$ (σ, σ' conhecidos)**

$$\left(\overline{X}-\overline{Y}-z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}+\frac{\sigma'^2}{m}}\,,\,\overline{X}-\overline{Y}+z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}+\frac{\sigma'^2}{m}}\right)$$

- **Intervalo de $(1-\alpha)\times 100\%$ de confiança para $\mu-\mu'$ ($\sigma=\sigma'$ desconhecido)**

$$\left(\overline{X}-\overline{Y}-t_{n+m-2,1-\alpha/2}S_P\sqrt{\frac{1}{n}+\frac{1}{m}}\,,\,\overline{X}-\overline{Y}+t_{n+m-2,1-\alpha/2}S_P\sqrt{\frac{1}{n}+\frac{1}{m}}\right),$$

$$S_P=\sqrt{\frac{(n-1)S^2+(m-1)S'^2}{n+m-2}}$$

- **Intervalo de $\approx(1-\alpha)\times 100\%$ de confiança para $\mu-\mu'$ (σ,σ' desconhecidos)**

$$\left(\overline{X}-\overline{Y}-t_{\nu,1-\alpha/2}\sqrt{\frac{S^2}{n}+\frac{S'^2}{m}}\,,\,\overline{X}-\overline{Y}+t_{\nu,1-\alpha/2}\sqrt{\frac{S^2}{n}+\frac{S'^2}{m}}\right),$$

$$\nu \text{ natural mais próximo de } \tilde{\nu}=\frac{\left(\frac{S^2}{n}+\frac{S'^2}{m}\right)^2}{\frac{S^4}{n^2(n-1)}+\frac{S'^4}{m^2(m-1)}}.$$

- **Intervalo de $(1-\alpha)\times 100\%$ de confiança para σ^2/σ'^2**

$$\left(\frac{1}{v_{n-1,m-1,1-\alpha/2}}\frac{S^2}{S'^2}\,,\,\frac{1}{v_{n-1,m-1,\alpha/2}}\frac{S^2}{S'^2}\right)$$

Intervalo de confiança assimptótico para a diferença de proporções em modelos $Ber(p)$ e $Ber(p')$ (amostras independentes)

Consideremos agora o caso de duas amostras aleatórias independentes, de dimensões n e m, provenientes de distribuições $Ber(p)$ e $Ber(p')$, respectivamente, em que p e p' são desconhecidos. Então,

como consequência do Teorema Limite Central, podemos concluir que, desde que n e m sejam grandes, se tem

$$\overline{X} - \overline{Y} \approx N\left(p - p', \sqrt{\frac{p(1-p)}{n} + \frac{p'(1-p')}{m}}\right),$$

ou seja, que

$$Z = \frac{\overline{X} - \overline{Y} - (p - p')}{\sqrt{\frac{p(1-p)}{n} + \frac{p'(1-p')}{m}}} \approx N(0,1).$$

Com base neste resultado, é possível encontrar um intervalo assimptótico de aproximadamente $(1-\alpha)100\%$ de confiança para a diferença dos parâmetros p e p'. Tal como no caso do intervalo de confiança para p, também aqui a expressão obtida directamente usando a v.a. Z acima é complicada. Um intervalo de confiança (menos preciso), mas mais simples, é o seguinte.

$$\Big(\overline{X} - \overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n} + \frac{\overline{Y}(1-\overline{Y})}{m}}\,,\, \overline{X} - \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n} + \frac{\overline{Y}(1-\overline{Y})}{m}}\Big).$$

Intervalo de confiança para a diferença de valores médios (amostras emparelhadas, modelo binormal)

No caso de duas amostras aleatórias que não são independentes entre si, mas sim emparelhadas, $(X_1, Y_1), \ldots, (X_n, Y_n)$, podemos encará-las como uma amostra aleatória de dimensão n de um par aleatório (X, Y) (em que as variáveis X e Y não são necessariamente independentes).

Por exemplo na amostra emparelhada anterior, se as v.a.'s X e Y representam, respectivamente, o peso de um indivíduo antes e depois de um certo tratamento de emagrecimento, cada par (X_i, Y_i) diz respeito aos pesos antes e depois do tratamento do indivíduo i. Se a amostra é aleatória, este par é independente de qualquer outro par (X_j, Y_j). No entanto, para cada i, X_i e Y_i estão relacionados (não são independentes).

No que se segue, vamos considerar o caso em que amostra provém de um par aleatório (X, Y) com uma distribuição binormal $N(\mu, \mu', \sigma, \sigma', \rho)$. Neste caso, pode provar-se que a v.a. diferença, $D = X - Y$, tem uma distribuição normal, com valor médio $\mu - \mu'$ (e com variância dada por $\sigma^2 + \sigma'^2 - 2\,\mathrm{cov}(X, Y) = \sigma^2 + \sigma'^2 - 2\rho\sigma\sigma'$, onde ρ é a correlação de X e Y). Se usarmos o intervalo de confiança usual para uma população normal com variância desconhecida para a amostra $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ proveniente desta v.a. D, obtemos de imediato o intervalo de $(1-\alpha) \times 100\%$ de confiança para a diferença dos valores médios $\mu - \mu'$ apresentado no quadro seguinte.

INTERVALO DE DE CONFIANÇA EM POPULAÇÃO BINORMAL

Amostras Emparelhadas

Intervalo de $(1-\alpha)\times 100\%$ de confiança para $\mu-\mu'$ (σ, σ' desconhecidos)

$$\Big(\overline{D}-t_{n-1,1-\alpha/2}\frac{S_D}{\sqrt{n}}\,,\,\overline{D}+t_{n-1,1-\alpha/2}\frac{S_D}{\sqrt{n}}\Big),\qquad S_D=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i-\overline{D})^2}$$

Intervalos unilaterais

Todos os intervalos de confiança referidos até agora são intervalos bilaterias, baseados no uso de duas estatísticas T_1 e T_2, para as quais se tem $P(T_1<\theta<T_2)=1-\alpha$. Podemos também considerar outro tipo de intervalos, ditos unilaterais, com um dos extremos infinito. Mais precisamente, se encontramos uma estatística T_1 tal que $P(T_1<\theta)=1-\alpha$, então diremos que o intervalo $(T_1,+\infty)$ é um intervalo (unilateral) de $(1-\alpha)$% de confiança para θ. De modo análogo, se tivermos uma estatística T_2 tal que $P(\theta<T_2)=1-\alpha$, diremos que o intervalo $(-\infty,T_2)$ é um intervalo (unilateral) de $(1-\alpha)\times 100\%$ para θ.

Consideremos novamente o caso de uma amostra $\mathbf{X}=(X_1,\ldots,X_n)$ proveniente de uma distribuição normal $N(\mu,\sigma)$ com σ conhecido. Como $Z=\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\sim N(0,1)$, se $z_{1-\alpha}$ for o quantil de probabilidade $1-\alpha$ da normal reduzida, podemos afirmar que

$$P\Big(\frac{\overline{X}-\mu}{\sigma}\sqrt{n}<z_{1-\alpha}\Big)=1-\alpha$$

ou seja, que

$$P\Big(\overline{X}-z_{1-\alpha}\frac{\sigma}{\sqrt{n}}<\mu\Big)=1-\alpha.$$

Temos, então, o seguinte intervalo de $(1-\alpha)\times 100\%$ de confiança unilateral para μ:

$$\Big(\overline{X}-z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,,\,+\infty\Big). \tag{5.6}$$

De modo análogo, temos também (tendo em conta que $z_\alpha=-z_{1-\alpha}$)

$$P\Big(\frac{\overline{X}-\mu}{\sigma}\sqrt{n}>-z_{1-\alpha}\Big)=1-\alpha$$

ou seja, temos

$$P\Big(\frac{\overline{X}-\mu}{\sigma}\sqrt{n}>-z_{1-\alpha}\Big)=1-\alpha,$$

de onde se obtém

$$P\Big(\mu < \overline{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\Big) = 1 - \alpha.$$

Tem-se, assim, outro intervalo de $(1-\alpha)\times 100\%$ de confiança unilateral para μ:

$$\Big(-\infty\,,\, \overline{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\Big). \tag{5.7}$$

Outros intervalos de confiança unilaterais correspondentes as casos bilaterias anteriormente estudados seriam obtidos de modo análogo. Por exemplo, deixamos ao cuidado dos alunos verificar que um intervalo de $(1-\alpha)\times 100\%$ de confiança para a diferença de valores médios $\mu - \mu'$, em populações normais, com variâncias conhecidas, será

$$\Big(\overline{X} - \overline{Y} - z_{1-\alpha}\sigma*, +\infty\Big), \quad \text{onde} \quad \sigma* = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{m}}.$$

5.2 Testes de Hipóteses

5.2.1 Generalidades

Os chamados testes estatísticos constituem uma parte muito importante da inferência estatística. Neste curso introdutório, limitamo-nos a estudar alguns exemplos de testes paramétricos, i.e. testes sobre *valores de parâmetros de uma distribuição*. Tal como no caso da estimação de um parâmetro, partimos de uma amostra, $(x_1, \ldots, x_n)$, que admitimos ser uma realização de uma amostra aleatória $(X_1, \ldots, X_n)$ de uma determinada população, cuja distribuição $F(x;\theta)$ é conhecida a menos de um parâmetro θ (ou de um vector de parâmetros θ).[2] Contudo, em vez estimar explicitamente um parâmetro desconhecido, o que pretendemos agora é usar essa amostra para *testar* uma *hipótese* (i.e. uma conjectura) particular sobre esse parâmetro desconhecido.

Uma hipótese estatística é, em geral, uma conjectura ou afirmação sobre um (ou mais) parâmetro(s) de uma distribuição. É chamada uma hipótese, porque não sabemos se é verdadeira ou falsa.

Por exemplo, considere-se uma distribuição normal $N(\mu, 1)$, com valor médio desconhecido e cuja variância é conhecida (igual a 1). A afirmação "μ é igual a 1"é uma hipótese estatística, bem como a afirmação "μ é menor do que 1".

[2]Consideramos o caso de uma única amostra, por uma questão de simplicidade na descrição; no entanto, iremos também estudar testes envolvendo duas populações, por exemplo, caso estejamos interessados em fazer testes para a diferença de médias, etc.

Chamamos **hipótese nula** à hipótese que pretendemos testar e designamo-la por H_0. Em geral, é formulada uma hipótese alternativa, que designamos por H_1.[3] Nesse caso, dizemos que estamos perante um *teste de hipóteses* (sobre o parâmetro que estamos a testar) e que vamos testar a hipótese H_0 *versus* a hipótese H_1 (escrevemos, abreviadamente H_0 *vs* H_1). Quando referimos que H_1 é uma hipótese alternativa, queremos dizer que H_0 e H_1 devem ser incompatíveis, i.e., se se verificar H_0 não pode verificar-se H_1.

Exemplo 5.4. *Suponhamos que temos uma moeda e que pretendemos saber se ela é ou não equilibrada. Estamos, assim, perante um modelo de Bernoulli $Ber(p)$, em que p é a probabilidade de sair cara. Neste caso, portanto, queremos testar a hipótese $H_0 : p = \frac{1}{2}$ vs $H_1 : p \neq \frac{1}{2}$.*

Chama-se **teste** de uma hipótese estatística a uma regra usada para decidir se rejeitamos ou não a hipótese H_0.

Essa regra vai ser baseada na utilização de uma **estatística de teste**, que é uma função $T = T(X_1, \ldots, X_n)$ da amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$. Nos casos que vamos considerar, essa estatística envolve sempre um estimador do parâmetro que estamos a testar e é tal que a sua distribuição, supondo a hipótese nula verdadeira, é totalmente conhecida.

Por exemplo, suponhamos que estamos perante uma amostra proveniente de uma distribuição $N(\mu, \sigma)$, com μ desconhecido e σ conhecido, e que pretendemos testar $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$. Neste caso, será natural considerar como estatística de teste a v.a. $T = T(X_1, \ldots, X_n) = \overline{X}$ (uma vez que $\overline{X}$ é um "bom"estimador para μ) ou, se for mais conveniente, $T = T(X_1, \ldots, X_n) = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$, a qual envolve o estimador $\overline{X}$ para μ e é tal que, se $\mu = \mu_0$, tem uma distribuição conhecida, $N(0, 1)$.

Com a ajuda desta v.a. T e dependendo de um valor α de probabilidade, por nós fixado, vamos definir uma região do plano, $\mathscr{C} = \mathscr{C}(\alpha)$, a chamada **região crítica** ou **região de rejeição**, sendo o teste da da forma: "rejeite-se H_0 se e só se $T \in \mathscr{C}$".

Na prática, naturalmente, a hipótese H_0 será rejeitada se e só se o valor observado da estatística, $T(x_1, \ldots, x_n)$, estiver na região crítica $\mathscr{C}$.

Note-se que ao tomar a decisão de rejeitar ou não a hipótese nula, podemos cometer dois tipos de erro: o primeiro consiste em rejeitar H_0, sendo H_0 verdadeira e o segundo, consiste em não rejeitar H_0, sendo H_1 verdadeira (e, portanto, sendo H_0 falsa). No primeiro caso dizemos que cometemos um **erro de tipo I** ou **erro de primeira espécie** e no segunda caso dizemos que cometemos um **erro de tipo II** ou **erro de segunda espécie**.

[3] Alguns autores preferem a notação H_A.

Em geral, consideramos mais grave cometer um erro de tipo I do que um erro do tipo II. Assim, especificamos um certo valor α pequeno (tipicamente $\alpha = 0.1, \alpha = 0.05$ ou $\alpha = 0.005$) e escolhemos um teste que garanta que a probabilidade de cometer um erro do tipo I seja igual a α. A este este valor de α, chamamos **nível de significância do teste**.

Denota-se por β a probabilidade de cometer um erro de segunda espécie, i.e.

$$\beta = P(\text{não rejeitar} H_0 | H_1 \text{verdadeira}).$$

O valor $1 - \beta$ (i.e. a probabilidade de rejeitar H_0 sendo H_1 verdadeira) é a chamada **potência do teste**.

5.2.2 Testes paramétricos em modelo normal

Testes sobre o valor médio em população normal (σ conhecido)

Suponhamos que dispomos de uma amostra observada $(x_1, \ldots, x_n)$ correspondente a uma amostra aleatória $(X_1, \ldots, X_n)$ proveniente de uma distribuição $N(\mu, \sigma)$, com μ desconhecido e σ conhecido e que pretendemos testar a hipótese nula

$$H_0 : \mu = \mu_0$$

versus a hipótese alternativa

$$H_1 : \mu > \mu_0,$$

onde μ_0 é uma certa constante dada. Este tipo de hipótese alternativa é dita uma hipótese unilateral.

Como $\overline{X}$ é um estimador natural de μ, faz sentido, atendendo à forma da hipótese alternativa H_1, procurar uma região de rejeição definida por uma expressão da forma

$$\overline{X} \geq k. \tag{5.8}$$

Assim, se desejarmos um teste com nível de significância α, devemos determinar o valor k para o qual se tenha um erro de tipo I igual α, isto é, k deverá ser tal que se tenha

$$P\Big(\overline{X} \geq k \,|\, H_0 \text{ verdadeira}\Big) = P\Big(\overline{X} \geq k \,|\, \mu = \mu_0\Big) = \alpha. \tag{5.9}$$

Neste caso, como σ é conhecido, se a hipótese nula se verificar, isto é, se $\mu = \mu_0$, sabemos que a v.a.

$$Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

tem distribuição $N(0,1)$. Então, a equação (5.9) é equivalente a ter-se

$$P\Big(Z \geq \frac{k-\mu_0}{\sigma}\sqrt{n}\Big) = \alpha. \tag{5.10}$$

Conclui-se então que $\frac{k-\mu_0}{\sigma}\sqrt{n}$ deverá ser o quantil de probabilidade $1-\alpha$ da distribuição normal standard, i.e. deve ter-se

$$\frac{k-\mu_0}{\sigma}\sqrt{n} = z_{1-\alpha},$$

ou seja, deverá ter-se

$$k = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \tag{5.11}$$

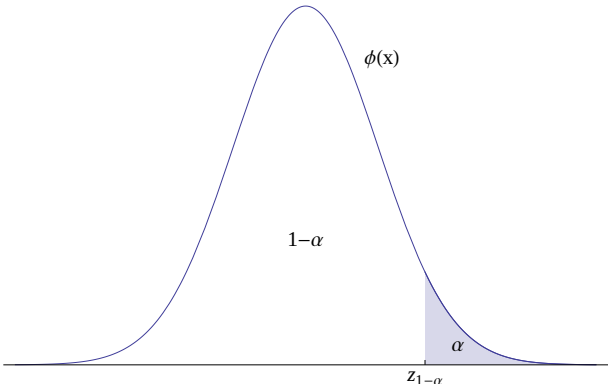

Figura 5.2: Quantil de probabilidade $1-\alpha$ da distribuição $N(0,1)$, $z_{1-\alpha}$

Substituindo o valor de k dado por (5.11) na equação (5.8), vemos que o teste (ao nível de significância α) é descrito do seguinte modo:

Rejeite-se H_0 se e só se

$$\overline{X} \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

ou, de modo equivalente,

Rejeite-se H_0 se e só se

$$Z = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \geq z_{1-\alpha}.$$

Claro está, que, perante a amostra observada concreta $\mathbf{x} = (x_1, \ldots, x_n)$, a decisão será de rejeitar H_0 se e só se valor observado da estatística de teste Z estiver na região crítica, i.e. se e só se for

$$\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n} \geq z_{1-\alpha}.$$

Note-se que

$$\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n} \geq z_{1-\alpha} \iff \mu_0 \leq \overline{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

$$\iff \mu_0 \notin \left(\overline{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,,\, +\infty\right)$$

Mas, $\left(\overline{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\,,\, +\infty\right)$ é o intervalo de confiança unilateral para μ, no caso σ conhecido, deduzido anteriormente; veja a Equação (5.6). Em conclusão: este teste equivale a tomar a decisão com base num intervalo de confiança unilateral para μ – rejeita-se H_0 se e só se μ_0 **não pertencer** a esse intervalo de confiança.

Seguindo uma metodologia totalmente análoga se obtêm os testes para o caso de outra hipótese unilateral

$$H_0 : \mu = \mu_0 \;\; \textit{vs} \;\; H_1 : \mu < \mu_0$$

ou para o caso de uma hipótese bilateral

$$H_0 : \mu = \mu_0 \;\; \textit{vs} \;\; H_1 : \mu \neq \mu_0.$$

No primeiro caso, o teste será equivalente a tomar a decisão com base no outro intervalo de confiança unilateral para μ, dado por (5.7); no segundo, a decisão deve ser tomada com base no intervalo de confiança bilateral correspondente a (5.1).

Valor de probabilidade p (valor-p)

Em alternativa à abordagem de fixar *a priori* o nível de significância do teste (e, consequentemente, rejeitar ou não H_0 em função do valor observado da estatística do teste), pode calcular-se o chamado **valor-p** (ou **nível de significância do resultado**), o qual é dado pela probabilidade de observar um valor da estatística de teste tão ou mais "extremo"do que aquele que foi observado, supondo H_0 verdadeiro. Por exemplo, no caso do teste para a média de uma população normal $N(\mu, \sigma)$ com σ conhecido, em que queiramos testar $\mu = \mu_0$ *vs* $\mu \neq \mu'$, o valor p é dado por

$$p = P\Big(|Z| \geq |z|\Big),$$

onde z é o valor observado da estatística de teste, $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$, e em que que $Z \sim N(0,1)$.

A decisão de rejeição é, então, baseada nessa probabilidade. Assim, se ao valor observado da estatística corresponder um determinado valor-p, não há razões para rejeitarmos H_0 para níveis de significância inferiores a esse valor, mas rejeitaremos H_0 para níveis de significância iguais ou superiores a esse valor. Por exemplo, se o valor-p calculado for inferior a 0.05, considera-se o

resultado significativo ao nível 5%, havendo evidência para rejeitar H_0 (a esse nível). Se o valor-p for inferior a 0.01, considera-se o resultado significativo ao nível 1%, havendo grande evidência para rejeitar H_0.

Até aqui considerámos apenas o caso em que a hipótese nula é da forma $H_0 : \mu = \mu_0$. No entanto, pode mostrar-se que, se, por exemplo, pretendermos testar a hipótese $H_0 : \mu \leq \mu_0$ *vs* $\mu > \mu_0$, podemos considerar a mesma região crítica que no caso $\mu = \mu_0$ *vs* $H_1 : \mu > \mu_0$. De modo análogo, a região crítica para $H_0 : \mu \geq \mu_0$ *vs* $H_1 : \mu < \mu_0$ é a mesma do que a correspondente às hipóteses $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu < \mu_0$.

Testes sobre o valor médio em população normal (σ desconhecido)

Consideremos agora o caso em que a amostra provém de uma distribuição normal $N(\mu, \sigma)$, mas em que σ é desconhecido. Suponhamos que pretendemos testar a hipótese $H_0 : \mu = \mu_0$ *vs* a hipótese $H_1 : \mu > \mu_0$. Neste caso, em analogia com o que fizemos para o caso dos intervalos de confiança para o valor médio com σ desconhecido, a v.a. apropriada para usar como estatística de teste será

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n},$$

a qual, como sabemos, tem uma distribuição t_{n-1} (t de Student com $n-1$ graus de liberdade). Seguindo o processo descrito na secção anterior passo a passo, com as devidas adaptações, facilmente se conclui que o teste, ao nível de significância α, consistirá no seguinte:

Rejeite-se H_0 se e só se

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n} \geq t_{n-1,1-\alpha},$$

onde $t_{n-1,1-\alpha}$ designa o quantil de probabilidade $1-\alpha$ da distribuição t_{n-1}.

De modo análogo se deduzem os testes adequados a outras hipóteses, neste caso em que σ é desconhecido. No quadro seguinte indicam-se as regiões de rejeição dos testes (nível de significância α) para o valor médio μ em modelo normal, em cinco casos diferentes de hipóteses, quer para σ conhecido quer para σ desconhecido.

Nota Uma vez mais, recordamos que denotamos por z_p, $t_{n,p}$, $u_{n,p}$ e $v_{n,m,p}$, os quantis de probabilidade p das distribuições $Z \sim N(0,1)$, $T \sim t_n$, $U \sim \chi^2_n$ e $V \sim F_{n,m}$.

Testes sobre o valor médio em população normal $N(\mu, \sigma)$

- **Caso σ conhecido**

Hipóteses	Região de rejeição
$H_0 : \mu = \mu_0$ (ou $H_0 : \mu \leq \mu_0$) *vs* $H_1 : \mu > \mu_0$	$Z \geq z_{1-\alpha}$
$H_0 : \mu = \mu_0$ (ou $H_0 : \mu \geq \mu_0$) *vs* $H_1 : \mu < \mu_0$	$Z \leq -z_{1-\alpha}$
$H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$	$\lvert Z\rvert \geq z_{1-\alpha/2}$

$$Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}.$$

- **Caso σ desconhecido**

Hipóteses	Região rejeição
$H_0 : \mu = \mu_0$ (ou $H_0 : \mu \leq \mu_0$) *vs* $H_1 : \mu > \mu_0$	$T \geq t_{n-1,1-\alpha}$
$H_0 : \mu = \mu_0$ (ou $H_0 : \mu \geq \mu_0$) *vs* $H_1 : \mu < \mu_0$	$T \leq -t_{n-1,1-\alpha}$
$H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$	$\lvert T\rvert \geq t_{n-1,1-\alpha/2}$

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}.$$

Exemplo 5.5. *[Ros87, p.208]*

Sabe-se que, quando um sinal com valor μ é enviado de uma certa localidade A para uma localidade B, ele é recebido em B com uma distribuição $N(\mu, 2)$ (por outras palavras, o ruído adicionado ao sinal na transmissão tem uma distribuição $N(0, 2)$). Suponha que um determinado sinal foi enviado, independentemente e nas mesmas condições, 5 vezes, e que a média dos valores recebidos foi $\overline{x} = 9.5$. Diga, se, ao nível de significância de 5%, podemos aceitar que o sinal enviado foi $\mu = 8$.

Resolução: *Neste caso, estamos perante um teste bilateral $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$, com $\mu_0 = 8$, no âmbito de um modelo normal com desvio padrão conhecido, $\sigma = 2$. A amostra seleccionada tem dimensão $n = 5$ e o nível de significância é $\alpha = 0.5$. O valor observado da*

estatística de teste é

$$z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{9.5 - 8}{2}\sqrt{5} = 1.68.$$

Por outro lado, temos $z_{1-\alpha/2} = z_{0.975} = 1.96$. *Como* $|z| = 1.68 < 1.96$, *o valor observado da estatística* **não está** *na região de rejeição, pelo que não rejeitamos a hipótese* $\mu = 8$.

Vejamos como seria a abordagem usando o valor-p.

Neste caso,

$$\begin{aligned} p &= P\Big(|Z| \geq |z|\Big) = P\Big(|Z| \geq 1.68\Big) = 2P\Big(Z \geq 1.68\Big) \\ &= 2\Big(1 - P(Z < 1.68)\Big) = 2\Big(1 - \Phi(1.68)\Big) = 0.093. \end{aligned}$$

(Na expressão acima, Φ *denota a função de distribuição da normal reduzida; o valor* $\Phi(1.68)$ *pode calcular-se facilmente usando o Mathematica.) Como* $p = 0.093 > 0.05$, *não rejeitamos* H_0 *ao nível de significância* 5%.

5.2.3 Alguns outros testes

Para todos os casos de intervalos de confiança estudados na Secção 5.1.2, vamos encontrar testes de hipóteses "correspondentes", seguindo uma metodologia análoga à descrita acima. Limitamo-nos, aqui, a apresentar quadros-resumo desses testes.

No quadro seguinte, apresentamos os testes para a igualdade de valores médios de duas amostras aleatórias independentes $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_m)$ provenientes de populações normais $N(\mu, \sigma)$ e $N(\mu', \sigma')$, respectivamente. Note-se que testar $H_0 : \mu = \mu'$ equivale a testar $H_0 : \mu - \mu' = 0$.

Testes sobre a igualdade dos valores médios em duas populações normais $N(\mu,\sigma)$ e $N(\mu',\sigma')$

Amostras Independentes

- **Caso σ e σ' conhecidos**

Hipóteses	Região de rejeição
$H_0 : \mu = \mu'$ (ou $H_0 : \mu \leq \mu'$) *vs* $H_1 : \mu > \mu'$	$Z \geq z_{1-\alpha}$
$H_0 : \mu = \mu'$ (ou $H_0 : \mu \geq \mu'$) *vs* $H_1 : \mu < \mu'$	$Z \leq -z_{1-\alpha}$
$H_0 : \mu = \mu'$ *vs* $H_1 : \mu \neq \mu'$	$\lvert Z\rvert \geq z_{1-\alpha/2}$

$$Z = \frac{\overline{X} - \overline{Y} - (\mu - \mu')}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma'^2}{m}}}.$$

- **Caso $\sigma = \sigma'$ desconhecido**

Hipóteses	Região de rejeição
$H_0 : \mu = \mu'$ (ou $H_0 : \mu \leq \mu'$) *vs* $H_1 : \mu > \mu'$	$T \geq t_{m+n-2,1-\alpha}$
$H_0 : \mu = \mu'$ (ou $H_0 : \mu \geq \mu'$) *vs* $H_1 : \mu < \mu'$	$T \leq -t_{n+m-2,1-\alpha}$
$H_0 : \mu = \mu'$ *vs* $H_1 : \mu \neq \mu'$	$\lvert T\rvert \geq t_{n+m-2,1-\alpha/2}$

$$T = \frac{\overline{X} - \overline{Y} - (\mu - \mu')}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

$$S_P^2 = \frac{(n-1)S^2 + (m-1)S'^2}{n+m-2}.$$

TESTES SOBRE A IGUALDADE DOS VALORES MÉDIOS EM DUAS POPULAÇÕES NORMAIS $N(\mu,\sigma)$ E $N(\mu',\sigma')$

Amostras Independentes

- **Caso σ,σ' desconhecidos**

Hipóteses	Região de rejeição
$H_0:\mu=\mu'$ (ou $H_0:\mu\leq\mu'$) *vs* $H_1:\mu>\mu'$	$T\geq t_{\nu,1-\alpha}$
$H_0:\mu=\mu'$ (ou $H_0:\mu\geq\mu'$) *vs* $H_1:\mu<\mu'$	$T\leq -t_{\nu,1-\alpha}$
$H_0:\mu=\mu'$ *vs* $H_1:\mu\neq\mu'$	$\lvert T\rvert\geq t_{\nu,1-\alpha/2}$

$$T=\frac{\overline{X}-\overline{Y}-(\mu-\mu')}{\sqrt{\frac{S^2}{n}+\frac{S'^2}{m}}},$$

$$\nu \text{ natural mais próximo de } \tilde{\nu}=\frac{\left(\frac{S^2}{n}+\frac{S'^2}{m}\right)^2}{\frac{S^4}{n^2(n-1)}+\frac{S'^4}{m^2(m-1)}}.$$

Testes para a igualdade de médias relativos a amostras emparelhadas (modelo binormal)

O quadro seguinte diz respeito ao caso de termos uma amostra aleatória bivariada $((x_1,y_1),\ldots,(x_n,y_n))$ proveniente de uma distribuição binormal $(X,Y)\sim N(\mu,\mu',\sigma,\sigma',\rho)$.

Testes sobre a igualdade dos valores médios em população binormal $N(\mu, \mu', \sigma, \sigma', \rho)$

Hipóteses	Região de rejeição
$H_0 : \mu = \mu'$ (ou $H_0 : \mu \leq \mu'$) *vs* $H_1 : \mu > \mu'$	$T \geq t_{n-1,1-\alpha}$
$H_0 : \mu = \mu'$ (ou $H_0 : \mu \geq \mu'$) *vs* $H_1 : \mu < \mu'$	$T \leq -t_{n-1,1-\alpha}$
$H_0 : \mu = \mu'$ *vs* $H_1 : \mu \neq \mu'$	$\lvert T \rvert \geq t_{n-1,1-\alpha/2}$

$$T = \frac{\overline{D}}{S_D}\sqrt{n}, \quad D = X - Y, \quad S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}.$$

Testes sobre variâncias em modelos normais

No quadro seguinte apresentam-se os testes para a variância σ^2 em população normal $N(\mu, \sigma)$.

Testes sobre a variância σ^2 em população normal $N(\mu, \sigma)$

Hipóteses	Região de rejeição
$H_0 : \sigma^2 = \sigma_0^2$ (ou $H_0 : \sigma^2 \leq \sigma_0^2$) *vs* $H_1 : \sigma^2 > \sigma_0^2$	$U \geq u_{n-1,1-\alpha}$
$H_0 : \sigma^2 = \sigma_0^2$ (ou $H_0 : \sigma^2 \geq \sigma_0^2$) *vs* $H_1 : \sigma^2 < \sigma_0^2$	$U \leq u_{n-1,\alpha}$
$H_0 : \sigma^2 = \sigma_0^2$ *vs* $H_1 : \sigma^2 \neq \sigma_0^2$	$U \leq u_{n-1,\alpha/2}$ ou $U \geq u_{n-1,1-\alpha/2}$

$$U = \frac{(n-1)S^2}{\sigma_0^2}.$$

Têm-se também os seguintes testes de igualdade de variâncias em duas populações normais $N(\mu, \sigma)$ e $N(\mu', \sigma')$, baseados em duas amostras independentes extraídas dessas populações.

Testes sobre a igualdade de variâncias em duas populações normais $N(\mu, \sigma)$ e $N(\mu', \sigma')$
Amostras Independentes

Hipóteses	Região de rejeição
$H_0 : \sigma^2 = \sigma'^2$ (ou $H_0 : \sigma^2 \leq \sigma'^2$) *vs* $H_1 : \sigma^2 > \sigma'^2$	$V \geq v_{n-1,m-1,1-\alpha}$
$H_0 : \sigma^2 = \sigma'^2$ (ou $H_0 : \sigma^2 \geq \sigma'^2$) *vs* $H_1 : \sigma^2 < \sigma'^2$	$V \leq v_{n-1,m-1,\alpha}$
$H_0 : \sigma^2 = \sigma'^2$ *vs* $H_1 : \sigma^2 \neq \sigma'^2$	$V \leq v_{n-1,m-1,\alpha/2}$ ou $V \geq v_{n-1,m-1,1-\alpha/2}$

$$V = \frac{S^2}{S'^2}.$$

Testes relativos a proporções (modelos de Bernoulli)

Segue-se um quadro contendo a descrição de testes sobre o valor de uma proporção p relativa a uma amostra aleatória retirada de uma população $Ber(p)$

Testes sobre uma proporção p de um modelo $Ber(p)$

Hipóteses	Região de rejeição
$H_0 : p = p_0$ (ou $H_0 : p \leq p_0$) *vs* $H_1 : p > p_0$	$Z \geq z_{1-\alpha}$
$H_0 : p = p_0$ (ou $H_0 : p \geq p_0$) *vs* $H_1 : p < p_0$	$Z \leq z_{\alpha}$
$H_0 : p = p_0$ *vs* $H_1 : p \neq p_0$	$\lvert Z \rvert \geq z_{1-\alpha/2}$

$$Z = \frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

O quadro seguinte contém a descrição de testes sobre a igualdade de proporções a partir de duas amostras aleatórias independentes extraídas de duas populações $Ber(p)$ e $Ber(p')$.

Testes de igualdade de duas proporções p e p' em duas populações $Ber(p)$ e $Ber(p')$
Amostras Independentes

Hipóteses	Região de rejeição
$H_0 : p = p'$ (ou $H_0 : p \leq p'$) *vs* $H_1 : p > p'$	$Z \geq z_{1-\alpha}$
$H_0 : p = p'$ (ou $H_0 : p \geq p'$) *vs* $H_1 : p < p'$	$Z \leq z_{\alpha}$
$H_0 : p = p'$ *vs* $H_1 : p \neq p'$	$\lvert Z \rvert \geq z_{1-\alpha/2}$

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}, \quad \hat{p} = \frac{n\overline{X} + m\overline{Y}}{n+m}.$$

Bibliografia

[Ath07] M. E. Athayde. *estatistica.R*. Departamento de Matemática, Universidade do Minho, 2007.

[Hay02] Anthony J. Hayter. *Probability and Satistics for Engineers and Scientists*. Duxbury, 2002.

[Kit98] Larry J. Kitchens. *Exploring Statistics: A Modern Introduction to Data Analysis and Inference*. Brooks/Cole Publishing Comp., 1998.

[PV08] D. D. Pestana and S. F. Velosa. *Introdução à Probabilidade e à Estatística*. Fundação Calouste Gulbenkian, 3 edition, 2008.

[Ros87] Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. John Wiley, 1987.

[Zar84] Jerrold H. Zar. *Biostatistical Analysis*. Pentice Hall, 1984.