# CAPÍTULO II
# Inferência a partir dos dados

As conclusões válidas para uma amostra, obtidas através dos métodos da Estatística Descritiva, não o são necessariamente para toda a população. Isso é notório, por exemplo, quando comparamos os resultados das sondagens feitas durante um determinado processo eleitoral com os resultados definitivos das eleições. A "ponte"para passar da amostra para a população é a Estatística Matemática que dispõe de instrumentos capazes de fazer inferências para a população a partir de amostras da mesma, medindo o grau de incerteza naturalmente associado a tais inferências.

## 1 Distribuições subjacentes aos dados

Para utilizar os procedimentos da Estatística Matemática, necessitamos de alguns conceitos mais teóricos (da Teoria das Probabilidades), em particular os de *variável aleatória* e *distribuição de probabilidade*, os quais tentaremos abordar de forma intuitiva.

Em muitos casos, os elementos da população não são números reais, podendo ser, por exemplo, uma molécula de um gás, uma planta ou um ser humano. No entanto, face à necessidade do tratamento matemático dos resultados, torna-se fundamental atribuir um valor real a cada elemento da população (ou mais do que um, se estiver em causa o estudo de mais do que uma característica da população), sempre que isso faz sentido. Este procedimento permite, por exemplo, avaliar a percentagem de indivíduos da população para os quais a característica em estudo assume determinados valores ou, de modo equivalente, calcular a probabilidade de que a característica em causa assuma valores num dado intervalo. Por exemplo, ao pretendermos estudar a obesidade dos portugueses, atribuímos a cada pessoa o seu IMC (índice de massa corporal - quociente entre o seu peso, em $kg$, e o quadrado da sua altura, em $m$). Temos assim uma correspondência $\omega \to X(\omega)$, onde $\omega$ representa uma determinada pessoa e $X(\omega)$ representa o seu IMC. Terá interesse conhecer, por exemplo, a probabilidade de se ter $X > 25$. Esta probabilidade multiplicada por 100 representa a percentagem de portugueses cujo IMC é superior a 25.

A correspondências deste tipo chamamos variáveis aleatórias. Mais precisamente, uma *variável aleatória* (v.a.) é uma função, $X$, que a cada elemento $\omega$ da população faz corresponder um número real $X(\omega)$, de modo que é sempre possível calcular a probabilidade de $X$ assumir valores em qualquer intervalo de números reais dado (no exemplo acima, tal intervalo é $]25, +\infty[$).

O facto dos elementos da população terem uma correspondência com números reais através da variável $X$ faz com que, por vezes, também se use a designação de "população"para $X$.

Na prática, temos essencialmente dois tipos de v.a.'s: discretas e contínuas. As primeiras só assumem uma quantidade finita ou infinita numerável de valores com probabilidade positiva; as segundas assumem valores em todo o conjunto dos números reais, $\mathbb{R}$, ou em intervalos de $\mathbb{R}$.

Uma v.a. discreta $X$ fica completamente caracterizada se conhecermos a probabilidade (positiva) de assumir cada um dos seus valores possíveis. O conjunto destes valores é designado por *suporte de* $X$.

A caracterização de uma v.a. contínua é feita, por exemplo, através de uma função real de variável real com determinadas características, chamada função densidade.

As *distribuições de probabilidade* (ou *leis de probabilidade*) são modelos utilizados para descrever populações reais ou, por outras palavras, para caracterizar o comportamento de v.a.'s aleatórias. Assim, tal como foi referido para as v.a.'s, temos dois tipos de distribuições mais usuais: as discretas e as contínuas.

## 1.1 Uma distribuição discreta: a binomial

Suponhamos que estamos interessados no número de vezes que ocorre determinado acontecimento quando repetimos um número finito de vezes um dado procedimento cujo resultado é, à partida, desconhecido (por esta razão, chamamos *experiências aleatórias* a tais procedimentos). Admita-se que a experiência em causa tem as seguintes características:

1. as repetições da experiência processam-se nas mesmas condições e os seus resultados são independentes;

2. a cada realização da experiência corresponde apenas um de dois resultados possíveis - "sucesso"ou "insucesso"(geralmente o sucesso corresponde àquilo que queremos contar);

3. a probabilidade de ocorrência de cada resultado mantém-se inalterada de experiência para experiência (designamos por $p$ a probabilidade de ocorrer um sucesso e, consequentemente, a probabilidade de ocorrer um insucesso será $1-p$).

As experiências que possuem estas características são designadas por *experiências de Bernoulli*.

Representando por $X$ o número de sucessos que ocorrem em $n$ repetições de uma experiência de Bernoulli, então $X$ é uma v.a. que pode tomar, com probabilidade positiva, os valores $0, 1, 2, \ldots, n$ (assim, o suporte de $X$ é o conjunto $S_X = \{0, 1, 2, \ldots, n\}$). A probabilidade de $X$ assumir cada um dos valores $k \in \{0, 1, 2, \ldots, n\}$ é dada por

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

onde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Nesta situação, dizemos que $X$ segue uma *distribuição binomial de parâmetros* $n$ *e* $p$ e denotamos este facto por $X \sim B(n, p)$.

Verifica-se que a média de uma população/v.a. com lei binomial é $np$ sendo a correspondente variância igual a $np(1-p)$.

**Exemplo** Sabe-se que com determinado tratamento se alcançam 90% de curas de uma doença quando o mesmo é aplicado a pacientes em condições bem definidas. Supondo que o tratamento é aplicado a 20 pacientes nessas condições, qual é a probabilidade de se obterem pelo menos 18 curas?

Neste caso, admitindo que os doentes reagem ao tratamento de forma independente, a v.a. $X$ que representa o número de doentes curados em 20 com o referido tratamento segue a lei $B(20, 0.9)$. Pretendemos $P(X \geq 18)$. Tem-se

$$P(X \geq 18) = P(X = 18) + P(X = 19) + P(X = 20) = 0.68.$$

Os valores das probabilidades $P(X = q)$, $q = 18, 19, 20$, podem ser calculados no SPSS: em primeiro lugar criamos uma variável com os valores $q$ pretendidos e depois usamos $Transform \rightarrow Compute\ Variable$ para fazer aparecer uma janela "calculadora". Inscrevemos em $Target\ Variable$ o nome de uma nova variável, à nossa escolha, onde serão colocados pelo programa os valores das probabilidades correspondentes aos valores $q$ da primeira variável. Em $Function\ group$ escolhemos $PDF\ \&\ Noncentral\ PDF$ e em $Functions\ and\ Special\ Variables$ seleccionamos $Pdf.Binom$. A seta lateral "envia"para $Numeric\ Expression$ a função que permite calcular os valores pretendidos. A soma final pode ser calculada em $Analyze \rightarrow Descriptive\ Statistics$.

Uma alternativa, conveniente para as situações em que o procedimento acima descrito envolve muitas parcelas, consiste em usar $CDF\ \&\ Noncentral\ CDF$ que dá a probabilidade da variável $X$ tomar valores inferiores ou iguais a determinado valor $q$, probabilidade esta que designamos por $F(q)$. Exemplificando: se $X$ tem suporte $\{0, 1, 2, \ldots, 40\}$, então a probabilidade $P(15 \leq X \leq 23)$ é igual a $F(23) - F(14)$.

No caso particular $n = 1$, que corresponde à realização da experiência apenas uma vez, a v.a. $X$ toma apenas os valores 0 e 1, tendo-se $P(X = 1) = p$ (que é a probabilidade da ocorrência de sucesso) e $P(X = 0) = 1 - p$ (que corresponde à probabilidade da ocorrência de insucesso). Nestas condições diz-se que $X$ segue a *distribuição de Bernoulli de parâmetro* $p$ e escrevemos $X \sim B(p)$.

Por exemplo, quando se faz um inquérito a $n$ pessoas com uma pergunta cuja resposta só pode ser uma de duas (por exemplo, "sim"ou "não", "branco"ou "preto", "zona rural"ou "zona urbana"), podemos associar o valor 0 a uma das respostas e o valor 1 à outra. No final obtemos uma sucessão de 0's e 1's que é uma amostra concreta de uma v.a. $X$ seguindo uma lei de Bernoulli.

## 1.2 Distribuições contínuas e curvas de densidade

Os gráficos adequados para amostras de variáveis contínuas são os histogramas. Quando se constrói um histograma (com todas as classes de igual amplitude), a sua forma não é alterada se alterarmos o tipo de unidades que usamos no eixo vertical. Podemos assim ter, entre outros, histogramas em que os valores do eixo vertical são frequências absolutas ou frequências relativas. Consideremos agora um histograma cujos rectângulos têm uma altura igual à frequência relativa da classe correspondente dividida pela amplitude desta. Neste caso, o histograma tem área total igual a 1. Nesta secção, vamos admitir que os histogramas são contruídos desta forma.

Na figura 1 estão representados 3 histogramas correspondentes a outras tantas amostras de uma mesma população, de dimensões 100, 500 e 1000.

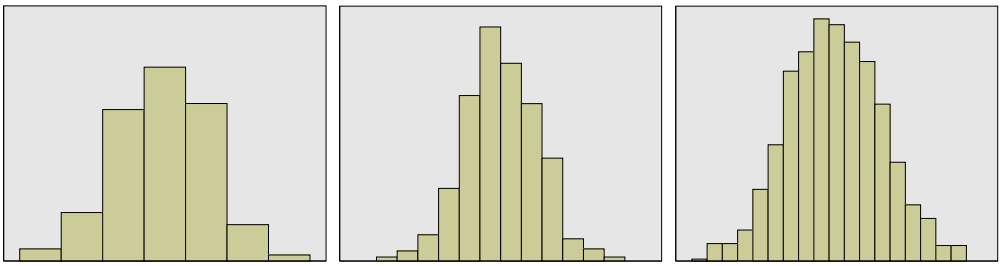

Figura 1: Histogramas correspondentes a três amostras da mesma população.

Podemos observar que os histogramas vão sendo modificados, mas mantêm uma certa forma que se acentua com o aumento da dimensão da amostra e a diminuição da amplitude das classes. Se pensarmos na amplitude das classes a tender para zero acompanhada do aumento da dimensão da amostra, facilmente podemos imaginar uma curva que se ajusta à parte superior do histograma e tal que a área delimitada superiormente por ela e inferiormente pelo eixo horizontal é igual a 1. Na figura 2 podemos observar o ajustamento acima referido.

Figura 2: Ajustamento da curva de densidade ao histograma.

Curvas do tipo acima descrito são chamadas curvas de densidade. Mais precisamente, damos o nome de *curva de densidade* a uma curva que está acima do eixo dos $xx$, podendo coincidir com este nalguns intervalos, e tal que que a medida da área entre este eixo e a curva é igual a 1. O(s) intervalo(s) de $\mathbb{R}$ onde a curva está estritamente acima do eixo dos $xx$ corresponde(m) ao *suporte* da distribuição correspondente.

Uma curva de densidade corresponde ao gráfico de uma função real de variável real chamada *função densidade* (ou apenas *densidade*).

A distribuição de uma v.a. contínua, $X$, é caracterizada por uma função densidade e as probabilidades de acontecimentos definidos à custa de $X$ são calculadas a partir da expressão matemática que define tal função. Por exemplo, sendo $a$ e $b$ dois números reais ($a \leq b$), a probabilidade de $X$ assumir valores no intervalo $[a, b]$, que denotamos por $P(a \leq X \leq b)$, é dada pela área delimitada pela curva de densidade de $X$, pelo eixo dos $xx$ e pelas rectas verticais $x = a$ e $x = b$, como é exemplificado na figura 3.

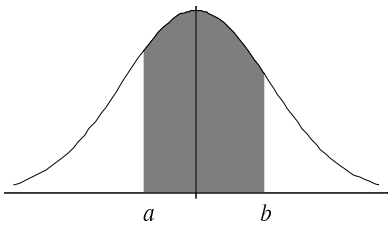


Figura 3: Área correspondente à probabilidade $P(a \leq X \leq b)$.

Note-se que a probabilidade de $X$ tomar um valor isolado é nula, i.e., $P(X = a) = 0$ (corresponde ao caso em que $X$ toma valores no intervalo $[a, a]$, obtendo-se uma área nula).

Nos gráficos apresentados nas figuras 2 e 3, as curvas de densidade são simétricas relativamente a um eixo vertical que passa pelo ponto mais elevado da curva. Além disso, as curvas estão sempre acima do eixo dos $xx$. Claro que não é sempre assim como mostram as curvas de densidade que se apresentam nas figuras 4 e 5.

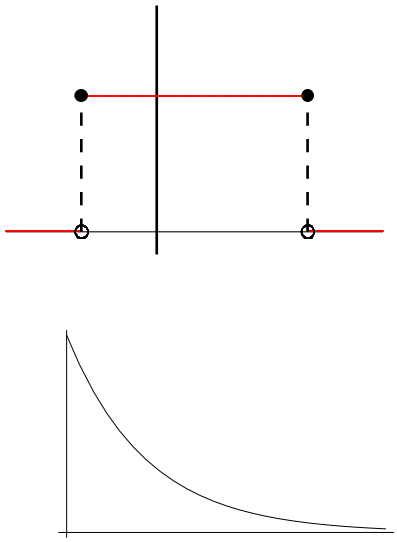

Figura 4: Curvas de densidade.

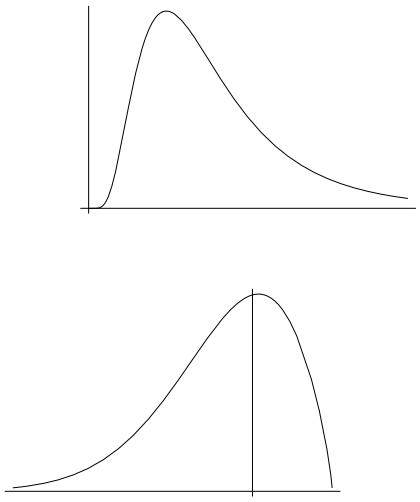

Figura 5: Curvas de densidade.

## 1.3 A distribuição normal

A distribuição normal, ou distribuição de Gauss, é a mais conhecida das distribuições contínuas. De facto, do ponto de vista das aplicações, tem-se observado que muitas características quantitativas de populações podem ser bem representados por variáveis com distribuição normal.

Os histogramas das figuras 1 e 2 correspondem a amostras de uma distribuição normal e a curva de densidade da figura 2 é dita *curva normal* ou *curva de Gauss*.

Na figura 6 apresentam-se duas curvas de Gauss.

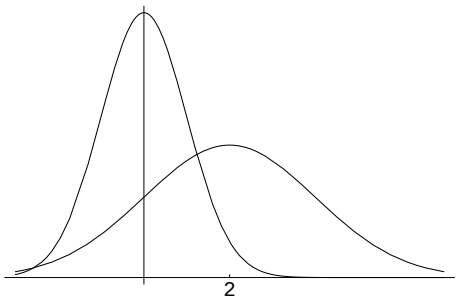


Figura 6: Curvas de Gauss.

A sua localização e a sua forma (mais ou menos "achatada") estão relacionadas, respectivamente, com a média, $m$, e o desvio padrão, $\sigma$, da variável $X$.

A expressão matemática da densidade correspondente a uma curva de Gauss de média $m$ ($m \in \mathbb{R}$) e desvio padrão $\sigma$ ($\sigma > 0$) é

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right], \quad x \in \mathbb{R}.$$

Se uma v.a. $X$ é caracterizada por uma densidade deste tipo diz-se que $X$ tem *distribuição normal (ou de Gauss) de parâmetros* $m$ *e* $\sigma$, escrevendo-se $X \sim N(m, \sigma)$.

Note-se que a função densidade acima descrita é sempre positiva; como tal, qualquer curva de Gauss está sempre acima do eixo dos $xx$. Assim, o suporte da distribuição normal é $\mathbb{R}$.

Qualquer curva de Gauss é simétrica relativamente à recta vertical $x = m$, i.e., relativamente à média da população correspondente.

Um caso particular importante ocorre quando $m = 0$ e $\sigma = 1$, correspondendo à chamada distribuição *normal centrada e reduzida* ou distribuição *normal standard*.

Uma população normal, $X$, tem a particularidade de verificar as seguintes propriedades:

- a proporção de indivíduos da população para a qual $X$ toma valores entre $m-\sigma$ e $m+\sigma$ é aproximadamente igual a 0.68, i.e., $P(m - \sigma \leq X \leq m + \sigma) \simeq 0.68$;
- a proporção de indivíduos da população para a qual $X$ toma valores entre $m - 2\sigma$ e $m + 2\sigma$ é aproximadamente igual a 0.95;
- a proporção de indivíduos da população para a qual $X$ toma valores entre $m - 3\sigma$ e $m + 3\sigma$ é aproximadamente igual a 0.997.

Na prática, sempre que dispomos de uma amostra que conduza a um histograma simétrico ou aproximadamente simétrico, devemos começar por verificar se uma distribuição normal é adequada para a variável em estudo. No entanto, há distribuições simétricas não normais cujas curvas de densidade são semelhantes à curva de Gauss (as mais conhecidas são a *distribuição t de Student* e a *distribuição de Cauchy*). Coloca-se então a questão: como saber, com alguma segurança, se a amostra pode ser efectivamente considerada como proveniente de uma distribuição normal?

Podemos começar por usar uma ferramenta gráfica, designada *papel de probabilidade* (Q-Q *plot*, no SPSS) que é construída da seguinte forma: para cada uma de determinadas percentagens (por exemplo $\frac{i}{n+1} \times 100\%$, $i = 1, \ldots, n$), consideram-se dois valores:

- o valor observado que tem à sua esquerda, incluindo-o, tal percentagem de observações,
- o valor da população que terá à sua esquerda, incluindo-o, tal percentagem de observações admitindo que a distribuição da população é de facto normal (a média $m$ e o desvio padrão $\sigma$ desta população são estimados, respectivamente, pela média, $\overline{x}$, e pelo desvio padrão corrigido, $s_c$, da amostra).

Os pontos definidos por cada um destes pares de valores são marcados num sistema de eixos dando origem a uma nuvem de pontos.

Se esta nuvem de pontos evidenciar uma relação linear entre abcissas e ordenadas, temos uma validação informal da normalidade da população de onde foi retirada a amostra, como podemos observar na figura 7.

No SPSS, a construção de um gráfico deste tipo é conseguida através de: *Analyze* $\rightarrow$ *Descriptive Statistics* $\rightarrow$ Q-Q *plots*, escolhendo a opção *Normal* em *Test Distribution*.

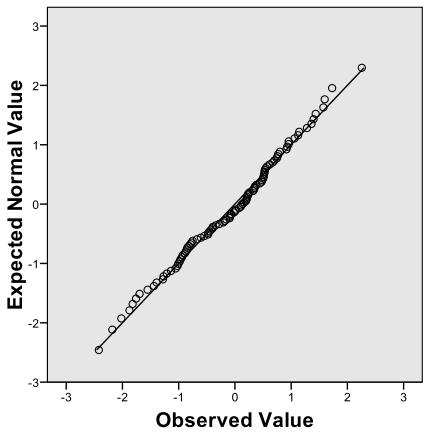


Figura 7: Papel de probabilidade para uma amostra de dimensão 100. .

## 1.4 Outras distribuições contínuas

Outras distribuições contínuas usuais são, por exemplo, a *distribuição uniforme*, a *distribuição exponencial*, a *distribuição lognormal* e a *distribuição de Weibull*. Nas figuras 8, 9, 10 e 11 apresentam-se histogramas de amostras de cada uma destas distribuições acompanhados de tipos de curvas de densidade que se adequam a tais distribuições .

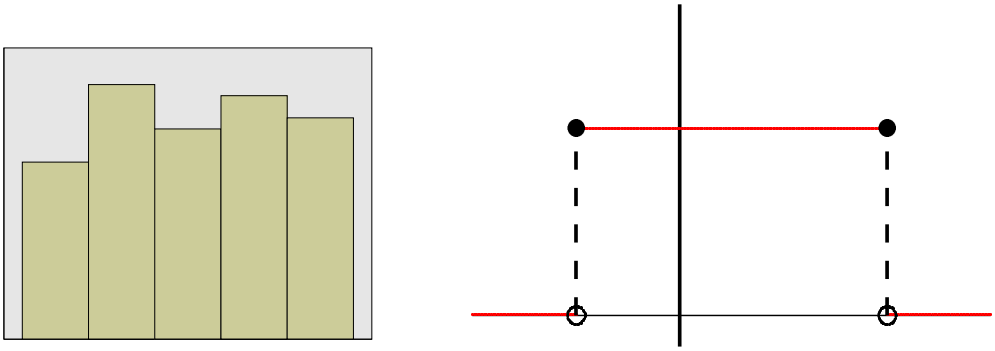

Figura 8: Histograma e curva de densidade do tipo uniforme.

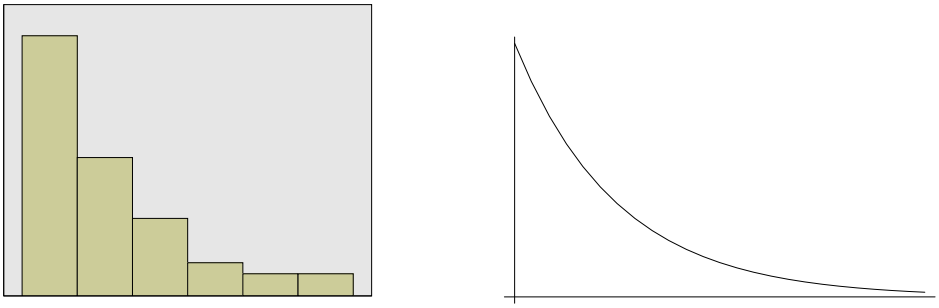

Figura 9: Histograma e curva de densidade do tipo exponencial.

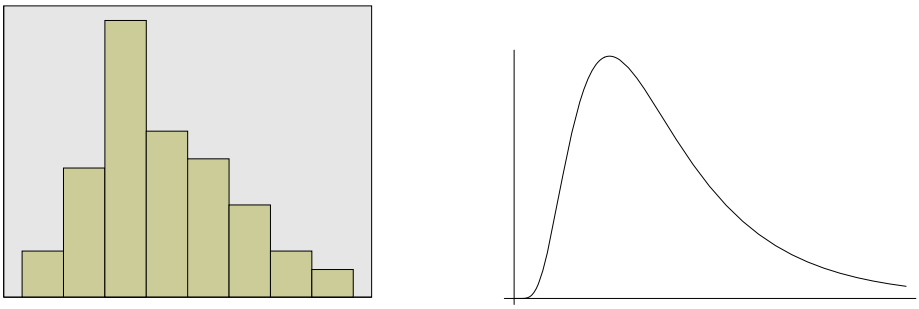

Figura 10: Histograma e curva de densidade do tipo lognormal.

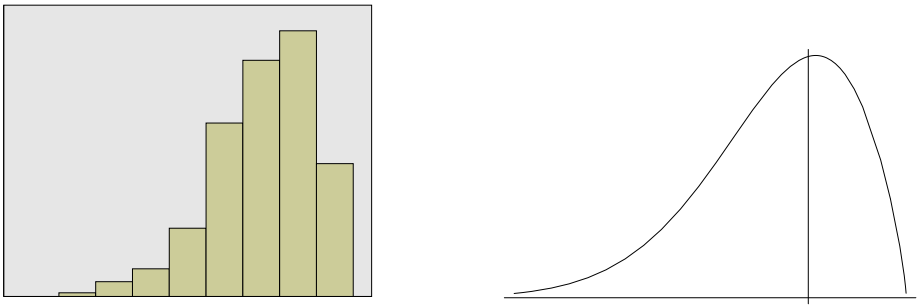

Figura 11: Histograma e curva de densidade do tipo Weibull.

Perante uma amostra, como escolher uma distribuição adequada para a v.a. subjacente?

Em primeiro lugar, é muito importante conhecer pelo menos as famílias de distribuições mais usuais e saber em que casos geralmente se aplicam para que seja mais fácil a escolha. Assim, para além da informação fornecida pela análise preliminar dos dados (gráficos, média, mediana, moda, assimetra, curtose,...), o conhecimento que por vezes temos do tipo de característica em estudo pode ajudar naquela escolha. Por exemplo, o tempo entre chegadas sucessivas de clientes a um determinado posto de serviço é geralmente bem modelado por uma distribuição exponencial.

Tendo sido seleccionada uma (ou mais) família(s) de distribuições, torna-se fundamental saber quais os parâmetros que a(s) especificam completamente e obter estimativas daqueles que são desconhecidos. Esses parâmetros não são necessariamente a média e o desvio padrão. Por exemplo, no caso da lei uniforme, os parâmetros são os extremos do intervalo onde a

característica em estudo assume os seus valores e são habitualmente estimados pelo mínimo e pelo máximo da amostra.

Seguidamente, verifica-se se o(s) modelo(s) escolhido(s) é (são) adequado(s) através, por exemplo, do papel de probabilidade. Claro que, no SPSS, devemos escolher a opção adequada em *Test Distribution*. Por exemplo, a opção adequada para o caso da figura 8 é $Uniform$.

Esta verificação também pode (e deve!) ser feita através de testes estatatísticos, ditos de ajustamento, que estudaremos adiante.

Se tivermos mais do que um modelo compatível com os dados, os papéis de probabilidade e a análise dos resultados dos testes estatísticos podem ajudar a seleccionar um deles.

## 1.5 Misturas de distribuições

Abordamos este tema com um exemplo. Foi feito um inquérito a 165 famílias duma certa região do país, tendo-se registado a zona de residência (rural ou urbana), o número de filhos e a despesa média mensal em electricidade.

Na figura seguinte apresenta-se o histograma relativo à variável "despesa média mensal em electricidade".

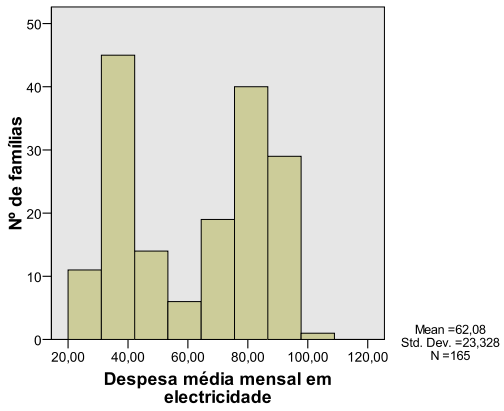


Figura 12: Histograma da despesa das famílias da região.

Trata-se de uma distribuição bimodal perante a qual devemos analisar a possibilidade da existência de dois grupos distintos na população de onde foi retirada a amostra aos quais poderão corresponder diferentes distribuições. Neste caso, a separação entre famílias rurais e urbanas conduziu aos dois histogramas da figura 13.

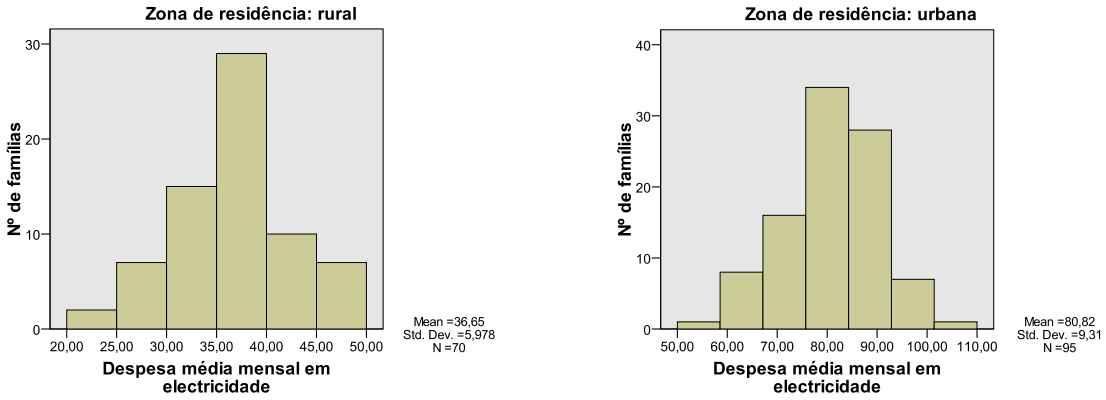


Figura 13: Histogramas das despesas das famílias rurais e urbanas.

A observação destes histogramas leva-nos a considerar a possibilidade de que tanto a despesa das famílias rurais como a despesa das famílias urbanas sejam normalmente distribuídas. Usamos o papel de probabilidade (QQ plot) para avaliar tal possibilidade.

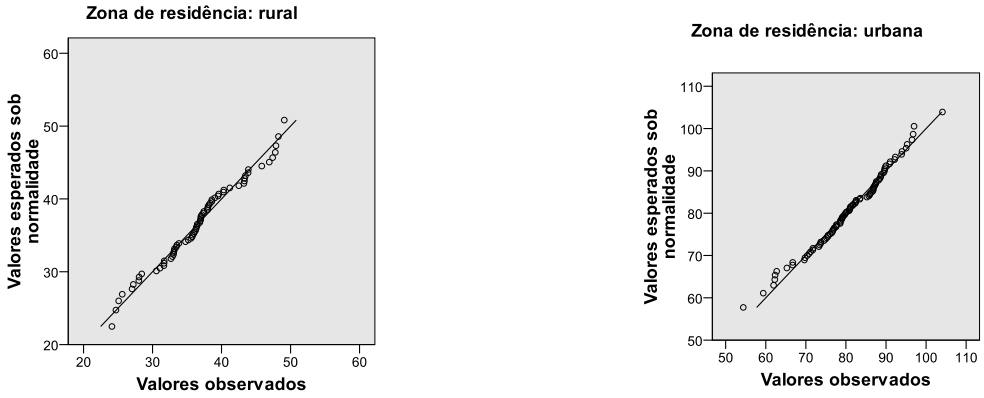


Figura 14: Papéis de probabilidade das despesas das famílias rurais e urbanas.

O papel de probabilidade sugere que, de facto, podemos admitir que ambas as amostras representadas graficamente na figura 14 são provenientes de populações normais (a população das famílias rurais da região em causa e a população das famílias urbanas da mesma região). Uma estimativa da média da despesa das famílias rurais da região é 36.7 euros, enquanto que a despesa média das famílias urbanas da região pode ser estimada por 80.8 euros. Os correspondentes desvios padrão são estimados por 6.0 euros e 9.3 euros. Assim, passaremos a considerar que a despesa das famílias rurais é uma variável aleatória seguindo a lei $N(36.7, 6)$ e que a despesa das famílias urbanas é uma variável aleatória seguindo a lei $N(80.8, 9.3)$.

Coloca-se agora a seguinte questão: qual é a distribuição da despesa das famílias da região em estudo?

Trata-se de uma *mistura* das duas leis acima identificadas. A proporção de famílias rurais na amostra inicial é aproximadamente 0.42 ($0.42 \simeq 70/165$), sendo 0.58 a proporção de famílias urbanas na mesma amostra. Assim, a distribuição da despesa das famílias daquela região pode

ser descrita por [1]

$$0.42\,N(36.7, 6) + 0.58\,N(80.8, 9.3). \tag{1}$$

A curva de densidade desta distribuição apresenta-se na figura 15.

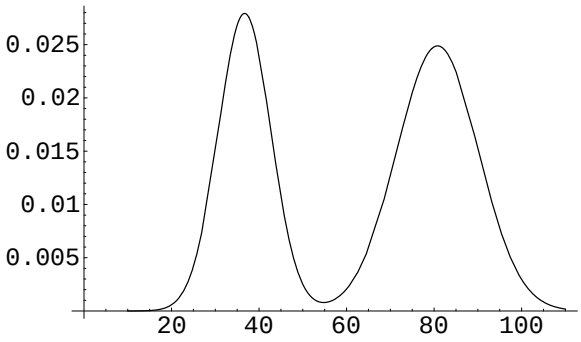


Figura 15: Curva de densidade da distribuição (1).

[1] Será que podemos "arredondar"os valores envolvidos e considerar para a população em causa a distribuição $0.4\,N(37, 6) + 0.6\,N(81, 9)$? A esta questão saberemos responder mais adiante.