Análise de dados para negócios

Cesaltina Pires

Janeiro de 2003

Conteúdo

Capítulo 1

Representação gráfica de dados

1.1 Variáveis discretas e contínuas

- *População* designa um conjunto de todos os elementos com alguma característica comum e com interesse para o estudo concreto que se está a fazer. Por vezes, distingue-se entre *população objectivo* e *população inquirida*. A

  população objectivo inclui a totalidade dos elementos que estão sobre estudo. Se não for possível construir uma amostra da população objectivo e esta for selecionada a partir de uma outra população essa é a população inquirida. $\Rightarrow$ designamos por N o número de elementos da população.

  A amostra é um subconjunto finito da população $\Rightarrow$ designamos por n o número de elementos da amostra.

- É importante notar que mesmo que uma dada característica seja qualitativa, há formas de representá-la quantitativamente. Por exemplo, se estivermos interessados no sexo de um indivíduo podemos decidir representar por 0 se o sexo for masculino e por 1 se for feminino.

- Vamos supor que temos uma colecção de elementos, ou *amostra*. E vamos supor que conhecemos o valor de um conjunto de características para cada um dos elementos da colecção. Cada característica pode ser representada por uma variável. Exemplo:

Observ.	Rendimento	Idade	Sexo	Anos de escol.
1	500	45	0	20
2	300	30	1	15
3	450	35	0	20
4	150	25	0	15
5	150	32	1	10

$\Rightarrow$ cada linha representa um caso (uma observação)

$\Rightarrow$ cada coluna representa uma variável

- As variáveis podem ser de dois tipos: *discretas* ou *contínuas*.

  - Uma variável é discreta se só puder tomar um n^{o}finito de valores ou uma infinidade numerável de valores.

    $\Rightarrow$ exemplos: n^{o}de divisões por unidade de alojamento,

  - Uma váriável é contínua se puder tomar qualquer valor dentro dum intervalo de números reais

    $\Rightarrow$ exemplos: tempo de vida de uma máquina, despesa do agregado familiar.

1.2 Distribuições de frequência ou empíricas

1.2.1 Variáveis discretas

- Um exemplo: Num inquérito aos orçamentos familiares (1989-90), numa amostra de 9640 unidades de observação obtiveram-se os seguintes dados sobre o número de indíviduos por agregado doméstico:

N^{o}de indivíduos	Frequência Absoluta	Frequência Relativa
1	1138	0,118
2	2748	0,285
3	2304	0,239
4	2082	0,216
5	848	0,088
≥ 6	520	0,054
Total	9640	1

- O número de vezes que um acontecimento ou fenómeno é observado na amostra desina-se por *frequência absoluta.*

- Consideremos uma variável discreta que pode tomar um de k valores diferentes (α_1, $\alpha_2, \cdots, \alpha_k$). Seja n o número todas de observações na amostra e designemos por n_1 o número de observações que registaram o valor α_1, por n_2 o número de observações

com o valor α_2, e por aí adiante. As frequências absolutas são exactamente n_1, n_2,$\cdots$. Note-se que

$$n_1 + n_2 + \cdots + n_k = n$$

- O número de vezes em que um acontecimento é observado em relação ao número total de dados desina-se por *frequência relativa.*

$$f_i = \frac{n_i}{n}$$

Repare-se que:

$$n_1 + n_2 + \cdots + n_k = n \Rightarrow f_1 + f_2 + \cdots + f_k = 1$$

- Uma outra noção importante é a de *função cumulativa das frequências relativas.* $F(x)$ indica-nos a frequência relativa de observações com um valor igual ou inferior a x. Deste modo, se considerarmos uma variável discreta que pode tomar um de k valores diferentes (α_1, α_2,$\cdots$, α_k), obtemos:

$$F(x) = \begin{cases} 0 & \text{se } x < \alpha_1 \\ f_1 & \text{se } \alpha_1 \leq x < \alpha_2 \\ f_1 + f_2 & \text{se } \alpha_2 \leq x < \alpha_3 \\ \vdots & \\ f_1 + f_2 + \cdots + f_i & \text{se } \alpha_i \leq x < \alpha_{i+1} \\ \vdots & \\ 1 & \text{se } x \geq \alpha_k \end{cases}$$

1.2.2 Variáveis contínuas

- Neste caso o processo para construir o quadro de frequências é um pouco mais moroso. Há dois passos essenciais:

  – definição das classes de valores – intervalos de classe. Os intervalos devem ser disjuntos, $I_j \cap I_k = \emptyset$. E a sua união deve conter todos os valores possíveis que a variável pode tomar.

  Para definir os intervalos, basta definir os limites inferior e superior do intervalo. Uma possibilidade é considerar intervalos abertos à esquerda e fechados à direita:

$$x_i \in I_j \Leftrightarrow l_{j-1} < x_i \leq l_j$$

$\Rightarrow$ a diferença $l_j - l_{j-1}$ é a *amplitude* da classe j.

$\Rightarrow$ é normal considerar classes com amplitude constante. Mas, pode haver casos em que esse procedimento não é adequado.

– contagem dos valores pertencentes a cada classe

$\Rightarrow$ note-se que ao agruparmos em classes há sempre perda de informação, porque deixamos de observar a variabilidade dentro de cada classe.

- Exemplo de uma situação em que não é prático considerar classes de igual amplitude. O quadro a seguir apresenta dados das exploraçãoes agrícolas de Trás os Montes, relativamente à superfície agrícola utilizável:

Classes (ha)	Frequências Absolutas	Frequências Relativas
$0 < s < 0,5$	4391	$0,0546$
$0,5 \leq s < 1,5$	8557	$0,1064$
$1,5 \leq s < 2,5$	17104	$0,2126$
$2,5 \leq s < 5$	22900	$0,2846$
$5 \leq s < 10$	14684	$0,1825$
$10 \leq s < 20$	8694	$0,1080$
$20 \leq s < 50$	3467	$0,0431$
$50 \leq s < 100$	497	$0,0062$
$s \geq 100$	163	$0,0020$
Total	80457	1

$\Rightarrow$ Neste exemplo, por causa da frequência de explorações agrícolas de pequena dimensão é conveniente definir várias classes entre 0 e 10 hectares, para se ver com maior detalhe a distribuição. Mas, já não teria interesse estar a distinguir, por exemplo, entre explorações com 100 ou com 102 hectares.

$\Rightarrow$ Deve evitar-se o uso de *classes abertas*, como a última classe neste exemplo, onde não é claramente definido o limite superior. Isto pode levantar problemas em certos cálculos (como médias).

- Um outro exemplo com dados classificados – peso de 500 cigarros SG filtro (em miligramas)

Peso	Freq. Absol.	Freq. Relativa
760-780	4	0,008
780-800	43	0,086
800-820	118	0,236
820-840	168	0,336
840-860	117	0,234
860-880	39	0,078
880-890	11	0,022

- Quando se passa dos dados originais para uma tabela de frequências há sempre perda de informação, uma vez que deixamos de considerar a variabilidade dentro de cada classe.

- Tal como fizemos para as variáveis discretas, também podemos definir a *função cumulativa* das frequências relativas. $F(x)$ indica-nos qual é a frequência relativa de valores iguais ou inferiores a x. A função $F(x)$ tem as seguintes propriedades:

  - $0 \leq F(x) \leq 1$ com $-\infty < x < +\infty$
  - $F(x)$ é uma função não decrescente
  - $F(-\infty) = 0$, $F(+\infty) = 1$

1.3 Representação gráfica

1.3.1 Variáveis discretas

- A distribuição de frequências pode ser representada graficamente usando o *diagrama de barras*. No eixo das abscissas representam-se os vários valores que a variável pode tomar. Depois traçam-se barras cuja altura é igual à frequência.

  $\Rightarrow$ fazer o diagrama de barras no exemplo do número de indivíduos do agregado familiar.

- Função cumulativa

Nºde indivíduos	Freq.Absoluta	Freq. Relativa	Freq. Acumulada
1	1138	0, 118	0, 118
2	2748	0, 285	0, 403
3	2304	0, 239	0, 642
4	2082	0, 216	0, 858
5	848	0, 088	0, 946
≥ 6	520	0, 054	1
Total	9640	1	

$\Rightarrow$ reparar que temos uma função em escada, que varia entre 0 e 1, e que é não decrescente.

1.3.2 Variáveis contínuas

- A representação gráfica de distribuições de frequência de variáveis contínuas é feita pelo *histograma*. Um histograma é uma colecção de rectângulos adjacentes, cuja base é um intervalo de classe e a altura é a frequência relativa ou absoluta dividida pela amplitude da classe. Desta forma a área do rectângulo é igual à frequência relativa ou absoluta.

$$A_j = h_j \times \frac{f_j}{h_j} = f_j \text{ ou } A_j = h_j \times \frac{n_j}{h_j} = n_j$$

  $\Rightarrow$ Quando as classes tem todas igual amplitude é normal fazer a altura do rectângulo igual à frequência relativa ou absoluta.

  $\Rightarrow$ Se aumentarmos indefinidamente o número de classes, tendendo a amplitude das classes para zero, o histograma tende para uma curva contínua. A essa curva chama-se *curva de frequências* e é representação gráfica da *função de frequências*.

- Uma representação alternativa é feita pelo *polígono de frequências* que resulta de se unirem por segmentos de recta os pontos médios dos lados superiores dos rectângulos

- A função cumulativa de frequências relativas também pode ser representada graficamente – é o *polígono integral.*

Capítulo 2

Medidas de localização e dispersão

2.1 Medidas de localização

2.1.1 Média

- A média é muitas vezes usada como valor representativo de uma amostra. A média é o «centro» da distribuição. É comum falar-se em rendimento médio, média das idades, nota média,...

- A média de uma amostra constituída pelos n valores $x_1, x_2, \cdots, x_n$ define-se pela expressão:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Se só dispusermos dos dados classificados, podemos calcular a média usando a hipótese de que os valores de cada classe são todos iguais ao ponto médio da classe. Designemos por x'_j o ponto médio da classe j (é igual ao limite inferior da classe + metade da amplitude da classe). A média é dada por

$$\overline{x} = \frac{n_1 x'_1 + n_2 x'_2 + \cdots + n_k x'_k}{n}$$

  $\Rightarrow$ isto é uma média ponderada, cada valor é ponderado pela frequência com que ocorre.

- Propriedades da média – designemos por $m(x_1, x_2, \cdots, x_n)$ a função média

  - Se adicionarmos um constante k a cada um dos valores da amostra, a média também aumenta k.

$$m(x_1 + k, x_2 + k, \cdots, x_n + k) = m(x_1, x_2, \cdots, x_n) + k$$

  - Se multiplicarmos cada um dos valores por k, a média também será multiplicada por k

$$m(kx_1, kx_2, \cdots, kx_n) = km(x_1, x_2, \cdots, x_n)$$

  - A média da soma de duas variáveis é igual à soma das médias

$$m(x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n) = m(x_1, x_2, \cdots, x_n) + m(y_1, y_2, \cdots, y_n)$$

    $\Rightarrow$ pode generalizar-se para mais variáveis

  - Se as n observações de uma amostra estiverem repartidas por k subamostras podemos calcular a média como função das médias da subamostras:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_k\overline{x}_k}{n}$$

  - A média dos desvios em relação à média é zero:

$$m(x_1 - \overline{x}, x_2 - \overline{x}, \cdots, x_n - \overline{x})$$

- A média é o *centro de gravidade* da distribuição. Se considerarmos os desvios positivos em relação à média, e os desvios negativos em relação à média, eles compensam-se exactamente. Mas, é importante notar que duas distribuições que tenham a mesma média podem ser muito diferentes.

  $\Rightarrow$ É importante estudar também a dispersão dos valores em torno da média

  $\Rightarrow$ «Se eu comer um frango e tú comeres zero, em média comemos meio frango»

- A média tem a vantagem de incluir no seu cálculo todos os valores da amostra. Mas por essa razão é sensível à existência de valores extremos na amostra.

  $\Rightarrow$ os valores aberrantes costumam designar-se por *outliers*.

Média geométrica

- A média geométrica é adequada quando estamos interessados em calcular médias de variáveis que têm um efeito multiplicativo, como taxas de crescimento ou taxas de juro quando se admite capitalização.

  Suponha-se que se conhecem as taxas de crescimento anuais do PIB entre 1990 e 1995

Taxa	1991	1992	1993	1994	1995
%	2	3	1	4	5

$\Rightarrow$ o PIB de 1995 pode ser calculado uma vez conhecido o PIB de 1990 fazendo:

$$\begin{aligned} PIB_{1995} &= PIB_{1990}(1+0,02)(1+0,03)(1+0,01)(1+0,04)(1+0,05) \\ &= 1,1587 PIB_{1990} \end{aligned}$$

$\Rightarrow$ A questão é: qual é a taxa de crescimento anual média? Qual é a taxa tal que se o PIB crescer todos os anos aquela taxa obtemos o mesmo crescimento que o verificado?

$$PIB_{1990}(1+g)^5 = 1,1587 PIB_{1990} \Leftrightarrow (1+g)^5 = 1,1587 \Leftrightarrow g = \sqrt[5]{1.1587} - 1$$

$\Rightarrow$ a taxa de crescimento média foi de 0.02989.

- No caso geral, se designarmos por g_i a taxa de crescimento do ano i, temos:

$$g = \left[\prod_{i=1}^{n}(1+g_i)\right]^{\frac{1}{n}} - 1 = \sqrt[n]{\prod_{i=1}^{n}(1+g_i)} - 1$$

2.1.2 Mediana

- A mediana é o *centro posicional*. Em termos aproximados, a mediana é o valor que tem 50% de observações com valores mais baixos e 50% das observações com valores mais elevados.

- Para se calcular a mediana tem que se ordenar as observações, da mais pequena para a maior:

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

Designando a mediana por M, temos que:

$$\begin{aligned} M &= x_{k+1} \quad \text{se } n = 2k+1 \\ M &= \frac{x_k + x_{k+1}}{2} \quad \text{se } n = 2k \end{aligned}$$

- No caso de dados classificados pode calcular-se a mediana usando a função cumulativa. De facto, o valor da função cumulativa é igual a $\frac{1}{2}$ se o argumento for a mediana.

$$F(M) = \text{freq. relativa de valores inferiores ou iguais a } M = \frac{1}{2}.$$

- Exemplo do cálculo da mediana com valores classificados

Peso	Freq. abs. acum.	Freq. Relativa	Freq. rel. acumul.
760-780	4	0,008	0,008
780-800	47	0,086	0,094
800-820	165	0,236	0,330
820-840	333	0,336	0,666
840-860	450	0,234	0,900
860-880	489	0,078	0,978
880-890	500	0,022	1

$\Rightarrow$ verificamos imediatamente que a mediana se situa na classe 820-840. Fazendo interpolação linear (a ideia é que os valores se distribuem uniformemente na classe) obtemos:

$$\frac{M - 820}{840 - 820} = \frac{0.5 - 0.333}{0.666 - 0.333} \quad \Leftrightarrow \quad M = 830.03$$

- A mediana é menos sensível do que a média a valores aberrantes.
- A mediana é uma *estatística de ordem*. Existem muitas outras estatísticas de ordem:
  - O valor máximo e o valor mínimo.
  - Os quartis – O primeiro quartil é valor tal que tal que 25% dos observações têm um valor inferior aquele valor, o segundo quartil é a mediana, o terceiro quartil é o valor com 75% das observações com valores mais baixos.
  - Os decis,...

2.1.3 Moda

- A *moda* é o valor mais frequente.

  $\Rightarrow$ em amostras pequenas a moda não faz muito sentido porque é natural não haver repetições.

$\Rightarrow$ em amostras grandes pode ser uma medida com algum interesse.

$\Rightarrow$ podemos ter mais que uma moda

- No caso de dados classificados é fácil identificar a *classe modal*, ou seja a classe com maior frequência.

2.2 Medidas de dispersão

- Como já referimos, para se caracterizar uma dada distribuição é importante não só conhecer a média, mas também a dispersão dos valores em torno da média. A Figura 2.1 ilustra a ideia da dispersão em relação à média de um conjunto de observações. Nesta secção vamos falar de medidas de dispersão.

2.2.1 Desvio padrão e variância

- Qual é o comportamento do conjunto de desvios em relação à média? Quando há pouca dispersão os desvios são globalmente pequenos, se houver muita dispersão os desvios são globalmente grandes. Como medir a dispersão?

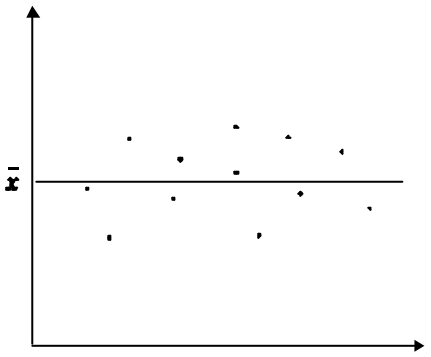

Figura 2.1: Média e dispersão em torno da média.

Não podemos limitar-nos a somar os desvios em relação à média, porquê?

$\Rightarrow$ porque a soma dos desvios em relação à média é zero (os desvios positivos e negativos compensam-se).

Por isso temos que considerar uma medida que não leve em conta o sinal dos desvios, só leve em conta a sua magnitude.

$\Rightarrow$ uma forma de fazer isto é considerar o quadrado dos desvios (ao elevar ao quadrado obtemos sempre número positivo)

$\Rightarrow$ outra forma de fazer isto é considerar o valor absoluto dos desvios.

- A variância é a média dos desvios quadrados em relação à média:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n}$$

Se os dados forem classificados:

$$s^2 = \frac{\sum n_j (x'_j - \overline{x})^2}{n}$$

- O desvio padrão é a raiz quadrada da variância:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

- No caso de amostras pequenas devem calcular-se a variância e desvio-padrão corrigidos. A fórmula é idêntica excepto que se divide por $n - 1$:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n - 1}$$

- *Propriedades da variância*
  - $s^2 = \frac{\sum x_i^2}{n} - \overline{x}^2$

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \overline{x})^2}{n} = \frac{\sum \left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right)}{n} = \\ &= \frac{\sum x_i^2}{n} - \frac{2\overline{x}\sum x_i}{n} + \frac{n\overline{x}^2}{n} = \frac{\sum x_i^2}{n} - \overline{x}^2 \end{aligned}$$

  - Se as n observações de uma amostra estiverem repartidas em k subamostras, as variâncias das subamostras estão relacionadas pela expressão seguinte:

$$ns^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \overline{x})^2 = \sum_{j=1}^{k} n_j s_j^2 + \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2$$

    onde x_{ji} designa o i-ésimo elemento da subamostra j.

    $\Rightarrow$ A interpretação desta expressão é que a variação total no conjunto de todas as subamostras é igual à soma das variações dentro de cada subamostra mais a variação entre subamostras.

- Exemplo de cálculo da variância

2.2.2 Desvio–médio

- O desvio médio é a média dos valores absolutos dos desvios em relação à média:

$$d = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

- Tanto o desvio padrão como o desvio médio são medidas sensíveis à existência de *outliers*.

2.2.3 Extremos-quartos e mediana

- As estatísticas de ordem podem também ser utilizadas para avaliar a dispersão. Uma medida possível é a diferença entre os extremos (valor máximo−valor mínimo). Mas a diferença entre os extremos não é uma medida *resistente*. É preferível usar a *dispersão quartal* que é a diferença entre o terceiro e o primeiro quartil. Representa a amplitude do intervalo onde se situam as observações centrais (50%).

2.2.4 Medidas de dispersão relativas

- O desvio padrão, o desvio médio e dispersão quartal são todas medidas de dispersão que são expressas na mesma unidade que a variável está a ser expressa. Se a unidade de medida for alterada, o valor da medida de dispersão também virá alterado (exemplo dos cigarros - mudar de miligramas para quilos). Estas medidas de dispersão são designadas por medidas de *dispersão absoluta*.

- Por vezes, é conveniente dispor de medidas de dispersão que sejam independentes das unidades de medida. Um caso em que isso acontece é quando se pretendem fazer comparações entre distribuições. Vamos por isso estudar medidas de *dispersão relativa*.

- A medida de dispersão mais usada é o *coeficiente de dispersão*:

$$\frac{s}{\overline{x}}$$

  por vezes é apresentado em %, multiplicando por 100 o coeficiente de dispersão. Nessa forma é designado por *coeficiente de variação*.

- Uma outra medida de dispersão relativa é dada pelo rácio da dispersão quartal e da mediana.

2.2.5 Índice de concentração e curvas de Lorenz

- Nalguns fenómenos económicos há interesse em estudar o *grau de concentração* de uma dada variável pelos vários elementos. Por exemplo, se conhecermos a riqueza total de um país podemos estar interessados na forma como essa riqueza está distribuída pelos cidadãos desse país. Pode acontecer que a riqueza esteja igualmente distribuída por todos os indivíduos, mas também pode acontecer que uma fracção substancial da riqueza esteja nas mãos de uma pequena fracção de indíviduos.

- Consideremos o exemplo das explorações agrícolas. Uma primeira ideia da concentração pode ser obtida comparando o andamento das frequências acumuladas com a área acumulada em percentagem da área total.

Classes (ha)	Freq. Abs.	Área Total	Fre. rel acum.	Área acum.
$0 < s < 0,5$	4391	2646	$0,054575$	$0,0040904$
$0,5 \leq s < 1,5$	8557	10295	$0,160930$	$0,020023$
$1,5 \leq s < 2,5$	17104	38366	$0,373516$	$0,079388$
$2,5 \leq s < 5$	22900	108352	$0,658140$	$0,247043$
$5 \leq s < 10$	14684	136584	$0,840647$	$0,458383$
$10 \leq s < 20$	8694	150401	$0,840647$	$0,691101$
$20 \leq s < 50$	3467	124220	$0,9487505$	$0,883310$
$50 \leq s < 100$	497	41484	$0,991796$	$0,947499$
$s \geq 100$	163	33930	1	1
Total	80457	6462781	1	1

$\Rightarrow$ interpretação: 5,5% das explorações agrícolas possuem 0,4% da superfície agrícola, 16,1% das explorações agrícolas possuem 2% da superfície agrícola,...

- Consideremos uma distribuição de frequências com k classes. Seja t_j o total do atributo correspondente aos n_j elementos da classe j. Se definirmos:

$$p_i = \frac{\sum_{j=1}^{i} n_j}{\sum_{j=1}^{k} n_j} \quad \text{e} \quad q_i = \frac{\sum_{j=1}^{i} t_j}{\sum_{j=1}^{k} t_j}$$

ou seja, p_i representa a proporção de elementos com um valor do atributo inferior ou igual ao limite superior da classe i. A variável q_i representa a proporção da totalidade do atributo que é possuída pelos mesmos elementos.

$\Rightarrow$ Note-se que $p_i \geq q_i$, variando ambos entre 0 e 1.

$\Rightarrow$ fazendo $(p_0, q_0) = (0, 0)$ e unindo por segmentos de recta os pontos (p_i, q_i) obtemos uma linha poligonal, que quando se consideram um número infinito de pontos tende para *curva de Lorenz*.

$\Rightarrow$ se a distribuição for equitativa, temos que $p_i = q_i$ e a curva de Lorenz é a diagonal do quadrado unitário.

$\Rightarrow$ quanto mais afastada estiver a curva de Lorenz da diagonal, maior é a concentração, maior é a desigualdade na distribuição do atributo.

- O *índice de concentração de Gini* é baseado na ideia de que quanto maior for a área entre a curva de Lorenz e a diagonal, maior é a concentração. O índice é dado por:

$$G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

$\Rightarrow G = 0$ se houver igual repartição

$\Rightarrow G = 1$ quando existe concentração máxima, isto é quando $q_i = 0$ para todas as classes excepto a última.

$\Rightarrow 0 \leq G \leq 1$ e cresce com a concentração.

2.3 Assimetria

- A ideia da simetria tem a ver com a forma como os valores se distribuem em torno do «centro», se se distribuem de forma simétrica ou não.

- Nas distribuições simétricas a média, a mediana e a moda coincidem. Nas distribuições assimétricas a média é «puxada» para o lado mais longo da distribuição.

  $\Rightarrow$ se a distribuição é assimétrica positiva temos média $>$ mediana $>$ moda.

  $\Rightarrow$ se a distribuição é assimétrica negativa temos média $<$ mediana $<$moda.

  $\Rightarrow$ o grau de assimetria de Pearson é baseado nesta ideia

$$g = \frac{\overline{x} - \text{mod}}{s}.$$

- Uma outra medida de assimetria, proposta por Bowley é baseada na ideia que em distribuições simétricas os quartis estão a igual distância da mediana, ou seja:

$$(F_u - M) - (M - F_l) = 0$$

O grau de assimetria de Bowley é definido por:

$$g' = \frac{(F_u - M) - (M - F_l)}{(F_u - M) + (M - F_l)}$$

$\Rightarrow$ se a distribuição é assimétrica positiva $(F_u - M) > (M - F_l)$ logo $g' > 0$.

Capítulo 3

Algumas distribuições

3.1 Distribuição normal

- A distribuição normal é extremamente importante em estatística por várias razões:
  - Vários fenómenos parecem seguir uma distribuição normal, ou podem ser aproximadamente descritos por uma distribuição normal.
  - A distribuição normal pode ser usada para aproximar várias distribuições discretas.
  - É a distribuição base em *inferência estatística*.
- A função densidade de probabilidades normal tem várias características interessantes:

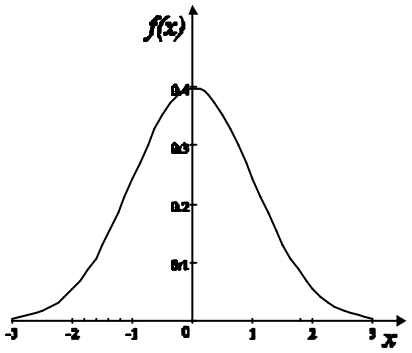

Figura 3.1: Função densidade da normal estandartizada.

  - Tem uma forma de sino e é simétrica.
  - As medidas de localização central (média, moda e mediana) são todas iguais.

  - O intervalo inter-quantis está contido em $\left[\mu - \frac{2}{3}\sigma, \mu + \frac{2}{3}\sigma\right]$, onde μ é a média e σ é o desvio padrão.
  - A variável aleatória (v.a. contínua) pode tomar qualquer valor entre $-\infty$ e $+\infty$.

    $\Rightarrow$ notar que a maior parte da probabilidade está concentrada em torno da média.

- Recordar que como a v.a. é contínua a probabilidade de um valor particular de x ocorrer é zero. Contudo, é possível calcular a probabilidade de x estar entre a e b.

$$P(a < x < b) = \int_a^b f(x)dx$$

  $\Rightarrow$ em termos geométricos isto é igual á área abaixo de $f(x)$ entre a e b.

  $\Rightarrow$ como $P(-\infty < x < +\infty) = 1$ a área abaixo da função densidade tem de ser igual a 1. Ou seja, $\int_{-\infty}^{\infty} f(x)dx = 1$

- Em aplicações práticas é natural que as propriedades da normal não sejam exactamente verificadas: é possível que não haja perfeita simetria, é possível que a variável aleatória não varie num intervalo infinito. Na prática é natural que a v.a. tome valores que se situam no intervalo $[\mu - 3\sigma, \mu + 3\sigma]$

- A função densidade de probabilidade da distribuição normal é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{com } -\infty < x < \infty$$

  onde $e \simeq 2.71828$, $\pi \simeq 3.14159$, μ é o valor esperado da variável aleatória x e σ é o desvio-padrão.

  $\Rightarrow$ se uma v.a. segue uma distribuição normal, ela é completamente caracterizada por μ e σ, são os únicos dois parâmetros da distribuição.

  $\Rightarrow$ em termos de notação diz-se que $x \sim N(\mu, \sigma)$ (lê-se: x segue uma distribuição normal com média μ e desvio-padrão σ).

  $\Rightarrow$ mostrar duas distribuições normais com o mesmo desvio-padrão mas média diferentes

  $\Rightarrow$ mostrar duas distribuições normais com a mesma média mas desvios-padrões diferentes.

$\Rightarrow$ note-se que seria bastante trabalhoso ter que usar a expressão anterior para calcular a probabilidade de x tomar um valor num certo intervalo. Felizmente não é nessário fazermos essas contas, como veremos já de seguida.

3.2 A normal estandartizada

- A distribuição normal estandartizada é a normal no caso em que $\mu = 0$ e $\sigma = 1$. Se $z \sim N(0,1)$ a função densidade é:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- Há várias tabelas para a normal estandartizada. Na maioria dos casos, essas tabelas indicam-nos a $P(z \leq a)$.

  $\Rightarrow$ mostrar os diferentes tipos de tabelas

  $\Rightarrow$ Qual é $P(0.5 \leq z \leq 1.2)$?

  $\Rightarrow$ Qual é $P(0.2 \leq z \leq 0.5)$?

  $\Rightarrow$ Qual é $P(-0.5 \leq z \leq 0.5)$?

  $\Rightarrow$ Qual é $P(-0.5 \leq z \leq 0.5)$?

- Se tivermos uma variável aleatória x com uma distribuição normal com média μ e desvio padrão σ é possível «estandartizar» essa variável. Para tal, basta definirmos uma nova variável z que resulta de «transformarmos» a variável x de acordo com:

$$z = \frac{x - \mu}{\sigma}$$

  Intuição: ao retirarmos μ a cada um dos valores da variável aleatória x, vamos obter uma v.a. com média zero (o que estamos a fazer é a «deslocar» a distribuição de forma a ficar centrada em 0. De forma semelhante, ao dividirmos pelo desvio padrão estamos a alterar a dispersão em torno da média de forma a que $\sigma_z = 1$.

  Formalmente:

$$E[z] = E\left[\frac{x - \mu}{\sigma}\right] = \frac{1}{\sigma}\left[\underbrace{E[x] - \mu}_{0}\right] = 0$$

e

$$\begin{aligned}\sigma_z^2 &= E\left[(z-E(z))^2\right] = E(z^2) = E\left[\frac{(x-\mu)^2}{\sigma^2}\right] \\ &= \frac{1}{\sigma^2}E\left[(x-\mu)^2\right] = \frac{\sigma^2}{\sigma^2} = 1\end{aligned}$$

$\Rightarrow$ A estandartização pode ser vista como um «reescalar» da variável original, sendo a nova unidade de medida o desvio padrão.

$\Rightarrow$ o valor da variável z indica-nos quantos desvios padrões é que estamos afastados da média. Por exemplo, se a variável $x \sim N(15,2)$ o valor $x = 19$ corresponde:

$$z = \frac{19-15}{2} = 2$$

Ou seja, $x = 19$ está dois desvios padrões acima da média, $19 = \mu + 2\sigma = 15 + 2 \times 2$.

- Qual é a vantagem de estandartizarmos a váriavel com distribuição $N(\mu, \sigma)$? a vantagem é que depois de estandartizada podemos usar as tabelas da normal estandartizada para calcular probabilidades.

  $\Rightarrow$ Para qualquer distribuição normal, a probabilidade da variável aleatória distar da média menos que um desvio-padrão é 68.26%.

  $\Rightarrow$ A probabilidade da v.a. distar da média menos de dois desvios padrões é 95.44%

  $\Rightarrow$ A probabilidade da v.a. distar da média menos de três desvios padrões é 99.73%.

- Exemplo: as notas numa cadeira de Estatística seguem uma distribuição aproximadamente normal com média 13 e desvio padão 2.

  - Qual é percentagem de alunos que passa na cadeira (isto é, nota é superior ou igual a 9.5)?
  - Qual é a probabilidade de um aluno escolhido aleatoriamente ter mais de 17?
  - Qual é a percentagem de alunos com notas entre 11 e 15?

- Também se pode usar a tabela da normal para encontrar os valores da v.a. que correspondem a uma dada probabilidade. No exemplo anterior, podiamos querer calcular qual é o intervalo inter-quartis da variável aleatória. Entre que notas é que se situam 50% de notas «centrais».

$\Rightarrow$ na normal estandartizada o valor k tal que:

$$P(z \geq k) = 0.25 \quad \Leftrightarrow \quad k = 0.675$$

Mas como:

$$z = \frac{x - \mu}{\sigma} \quad \Leftrightarrow \quad x = \mu + z\sigma$$

temos que o valor do terceiro quartil da v.a. x é:

$$x = 13 + 2 \times 0.675 = 14.35$$

e o valor do primeiro quartil é:

$$x = 13 - 2 \times 0.675 = 11.65.$$

3.2.1 Como testar a normalidade

- Uma boa ideia é comparar a distribuição empírica e as suas propriedades com a distribuição normal:

  - Construir histograma e polígono de frequências da variável que se está a analisar e comparar com a função densidade de uma v.a. normal.
  - Calcular medidas descritivas e comparar as suas propriedades com a de uma distribuição normal

    $\Rightarrow$ calcular média, mediana, moda, midrange e verificar se estas medidas tem valores próximos uns dos outros.

    $\Rightarrow$ Calcular o intervalo de variação da variável aleatória e ver se ele é aproximadamente igual a 6 vezes o desvio-padrão da variável.

    $\Rightarrow$ verificar se o intervalo inter-quartis é aproximadamente igual a 1.33 vezes o desvio padrão.
  - Testar como é que as observaçõe se distribuem:

    $\Rightarrow$ ver se aproximadamente $\frac{2}{3}$ das observações distam da média menos de 1 desvio-padrão.

    $\Rightarrow$ ver se aproximadamente 95% das observações distam da média menos de 2 desvios-padrões.

- Há testes formais da normalidade, que eventualmente referiremos quando falarmos de testes de hipóteses. Estes testes são baseados em medidas de simetria e de achatamento. A simetria é medida usando:

$$\frac{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^3}{n}}{s^3}$$

Esta medida designa-se por *skewness* em inglês.Se a distribuição da variável em estudo for simétrica esta medida será igual a zero. Se a variável apresentar assimetria positiva o que acontece é que teremos desvios positivos com valores elevados, e desvios negativos com valores menos elevados. Como ao elevarmos ao cubo o sinal dos desvios se vai manter, o que acontece ao elevarmos ao cubo, é que a soma dos desvios positivos vai «dominar» a soma dos desvios negativos ao cubo, obtendo-se um valor positivo para a medida de assimetria.

Como a normal não é a única distribuição simétrica, para verificarmos o ajustamento à normal temos de analisar também o *achatamento* da distribuição. O achatamento está relacionado com o «peso» das abas. O achatamento, ou *kurtosis*, da distribuição é dado por:

$$\frac{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^4}{n}}{s^4}$$

Para uma variável normal o achatamento é igual a 3.

O teste de Bowman-Shelton é baseado na proximidade da *skewness* a 0 e na proximidade da *kurtosis* a 3.

3.3 Distribuição do $\mathcal{X}^2$

Se Z é uma v.a. com distribuição $N(0,1)$, então Z^2 é uma v.a. com distribuição qui-quadrado com 1 grau de liberdade:

$$Z^2 \sim \mathcal{X}^2(1).$$

Um resultado importante é que a soma de variáveis aleatórias *independentes* com distribuição qui-quadrado também segue uma distribuição qui-quadrado, em que os graus de liberdade são iguais à soma dos graus de liberdade.

Em termos mais formais: Sejam $X_1, X_2, \cdots, X_k$ variáveis aleatórias $\mathcal{X}^2(r_1), \mathcal{X}^2(r_2), \cdots, \mathcal{X}^2(r_k)$, respectivamente. Se $X_1, X_2, \cdots, X_k$ forem independentes e $Y = X_1 + X_2 + \cdots + X_k$ então Y é $\mathcal{X}^2(r_1 + r_2 + \cdots + r_k)$.

Uma consequência do resultado anterior é que se $Z_1, Z_2, \cdots, Z_r$ forem variáveis aleatórias $N(0,1)$ e mutuamente independentes, então $W = Z_1^2 + Z_2^2 + \cdots + Z_r^2$ tem distribuição $\mathcal{X}^2(r)$.

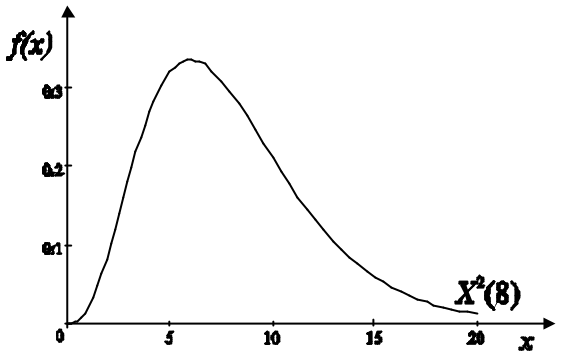

Figura 3.2: Função densidade da qui-quadrado com 8 graus de liberdade.

É de realçar que uma variável aleatória com distribuição qui-quadrado, só pode tomar valores maiores ou iguais a zero. Para além disso, a distribuição qui-quadrado depende apenas dos graus de liberdade. Quanto mais elevado for o número de graus de liberdade, menos assimétrica é a distribuição.

3.4 A distribuição t

A distribuição t é muito importante em Estatística e Econometria porque é a distribuição da média na amostra quando a variância da população não é conhecida.

Se Z é uma v.a. com distribuição $N(0,1)$, U é uma v.a. $\mathcal{X}^2(r)$ e Z e U são independentes, então

$$T = \frac{Z}{\sqrt{U/r}}$$

tem uma distribuição $t-$*student* com r graus de liberdade.

Observações:

- Tal como a normal estandartizada, a distribuição t é simétrica em torno do zero e tem forma de sino.
- Tem mais área nas abas e menos área no centro que a normal. A intuição é que na t não se conhece o desvio-padrão da população, usando-se o desvio padrão da amostra

para o estimar. Essa incerteza sobre o valor de σ faz com a t seja mais variável do que z.

- Converge para $N(0,1)$ quando o número de graus de liberdade aumenta. Mostrar gráfico comparando para diferentes valores de n.
- f.d.p. é função só dos graus de liberdade. Mostrar tabela.

Exemplo 3.1 Seja T uma variável com distribuição t com 7 graus de liberdade então:

$$P(t \leq 1.415) = 0.9\blacklozenge$$

3.5 A distribuição F

Se U e V são variáveis aleatórias independentes com distribuição qui-quadrado com r_1 e r_2 graus de liberdade, respectivamente, então

$$F = \frac{\frac{U}{r_1}}{\frac{V}{r_2}}$$

tem uma distribuição F com r_1 e r_2 graus de liberdade. A Figura 3.3 apresenta a função distribuição F com $(10,6)$ graus de liberadade.

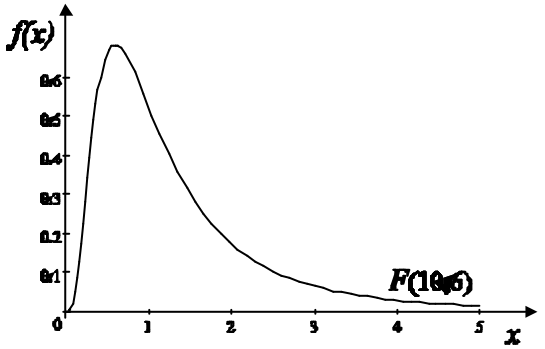

Figura 3.3: A função densidade F com $(10,6)$ graus de liberdade.

Observações:

- f.d.p. é função só de r_1 e r_2. Mostrar tabela.
- Mostrar gráficos da F

Capítulo 4

Amostragem e estimação

4.1 População e amostra

Uma parte importante da estatística relaciona-se com o problema da fazer *inferências* acerca da *população* relevante com base na informação de um subconjunto dessa população, com base numa *amostra*.

Exemplo 4.1 Queremos conhecer a distribuição etária em Portugal mas não temos dinheiro para fazer um census à população toda. Colhemos informação sobre uma amostra da população e tentamos inferir o que se passa na população.♦

Exemplo 4.2 Sondagens eleitorais.♦

- *Porquê amostras*?
  - Obter informação sobre toda a população ⇒ custos + elevados (custos)
  - É mais rápido obter informação sobre amostra (tempo).
  - Amostra permite aumentar a qualidade da informação obtida (precisão)
- *Como escolher a amostra*?
  - A amostra tem que ser *representativa* da população ⇒ princípio da aleatoriedade
  - *Amostra aleatória simples*: amostra de dimensão n de uma população de N objectos, todas as amostras possíveis de n objectos têm igual probabilidade de serem escolhidas.
    - exemplo de por N objectos num chapéu e tirar grupo de n objectos

    - tabelas de números aleatórios.
  - Outros procedimentos: *amostra estratificada*.

- *Passos de uma sondagem*:
  - Qual a informação que se quer obter
  - Qual é a população relevante
  - Como escolher a amostra
  - Como é que a informação é obtida
  - Como é que a informação na amostra pode ser usada para fazer inferências
  - Que conclusões podem ser retiradas sobre a população

4.2 Distribuição por amostragem

Normalmente estamos interessados em fazer inferências sobre certas características da população como a média e a variância. A ideia é fazer essas inferências usando a informação na amostra (calculando, por exemplo, a média e a variância na amostra, $\overline{X}$ e S^2). Mas devemos estar conscientes que, se amostra fosse diferente a média e variância na amostra também seriam diferentes. Por outras palavras, devemos olhar para $\overline{X}$ e S^2 como variáveis aleatórias (o seu valor depende da amostra recolhida).

A questão seguinte é: qual é a função densidade de probabilidades destas *estatísticas*? Qual é a sua *distribuição por amostragem*? É importante salientar que é o conhecimento da distribuição por amostragem que nos permite fazer inferências sobre a população.

Comecemos por definir o conceito de *estatística*. O que é uma *estatística*? Uma *estatística* é uma função da informação da amostra, isto é, posso calcular o valor da estatística uma vez conhecidas as observações da amostra. Por exemplo:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{e} \quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})}{n-1}$$

são estatísticas.

Em termos de notação usaremos sempre letras maiscúlas para designar as estatísticas e letras minúsculas para nos referirmos a valores particulares dessas estatísticas. Por exemplo, $\overline{X}$ e S^2 designam as variáveis aleatórias média na amostra e desvio-padrão na amostra, enquanto que $\overline{x}$ e s^2 se referem a valores que essas variáveis aleatórias tomam para uma amostra em particular.

O que é *distribuição por amostragem de uma estatística*? É a distribuição de probabilidades dos valores que essa estatística poderia tomar para todas as amostras de tamanho n que é possível escolher da população.

Exemplo 4.3 Seis empregados, variável de interesse é anos de experiência

$$2 \quad 4 \quad 6 \quad 6 \quad 7 \quad 8 \quad \Rightarrow \mu = 5.5$$

Suponha-se que escolhemos aleatoriamente um grupo de 5 trabalhadores. Qual é a distribuição de $\overline{X}$? Admitindo que a amostragem é feita sem reposição, há seis amostras possíveis com 5 elementos

Amostra	Média na Amostra
$4, 6, 6, 7, 8$	6.2
$2, 6, 6, 7, 8$	5.8
$2, 4, 6, 7, 8$	5.4
$2, 4, 6, 7, 8$	5.4
$2, 4, 6, 6, 8$	5.2
$2, 4, 6, 6, 7$	5.0

Qual e a função densidade de probabilidades de $\overline{X}$?♦

Exemplo 4.4 Suponhamos que a população são 4 amigos. A variável é a idade deles: $x_1 = 18, x_2 = 20, x_3 = 22, x_3 = 24$. Qual é a distribuição por amostragem da média, se a dimensão da amostra for $n = 2$?

Estas contas assumem que não há reposição (uma vez escolhido um elemento da população, ele não pode pode voltar a sair naquela amostra)

Samples	x_1, x_2	x_1, x_3	x_1, x_4	x_2, x_3	x_2, x_4	x_3, x_4
$\bar{x}$	19	20	21	21	22	23

Mais uma vez vemos que $\overline{X}$ é uma variável aleatória.♦

Porque é que estamos interessados na distribuição da estatística? Quando fazemos inferência temos só uma amostra com n elementos. Podemos olhar para esta amostra como uma das amostras possíveis no conjunto de todas as amostras de dimensão n retiradas da população em questão. Para a amostra que dispomos podemos calcular a estatística em que estamos interessados. Depois, levando em conta o nosso conhecimento sobre a distribuição por amostragem da estatística, podemos fazer inferências sobre a população. Isto são as ideias básicas de inferência estatística.

4.2.1 Distribuição da média da amostra

Suponhamos que a população tem média μ e variância σ^2. O que podemos dizer sobre a distribuição amostral de $\overline{X}$?

- A média da distribuição amostral de $\overline{X}$ é igual à média da população:

$$E(\overline{X}) = \mu$$

- A variância da distribuição amostral de $\overline{X}$ é igual a:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Estas propriedades resultam das propriedades do valor esperado e da variância e da definição de uma amostra aleatória. Um aspecto curioso é que a variância da média na amostra decresce com a dimensão da amostra. Isto significa que, à medida que dimensão na amostra aumenta a média na amostra é um estimador cada vez mais preciso da média na população.

As propriedades anteriores são interessantes, mas não nos indicam qual é a distribução por amostragem de $\overline{X}$. Será que podemos dizer alguma coisa sobre a função densidade de $\overline{X}$?

Se a variável que estamos a estudar tiver uma distribuição normal na população, então a distribuição da média da amostra segue também uma distribuição normal (veja a Figura 4.1):

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

E se não conhecermos a distribuição na população da variável, ou se a distribuição na população não for normal, que podemos dizer sobre a função densidade de $\overline{X}$?

Um resultado muito importante em estatística diz-nos que, qualquer variável aleatória X, com média μ e variância σ^2, *seja qual for a sua distribuição*, se a dimensão da amostra for elevada, então $\overline{X}$ tem aproximadamente uma distribuição normal com média μ e desvio padrão $\frac{\sigma^2}{n}$. Este resultado é o famoso *Teorema do Limite Central*.

Os resultados anteriores são baseados na hipótese de que a amostragem é feita com reposição. Contudo, na prática, a maioria dos estudos são feitos *sem reposição*. Nestas condições, se a população for finita e a dimensão da amostra não for pequena em relação

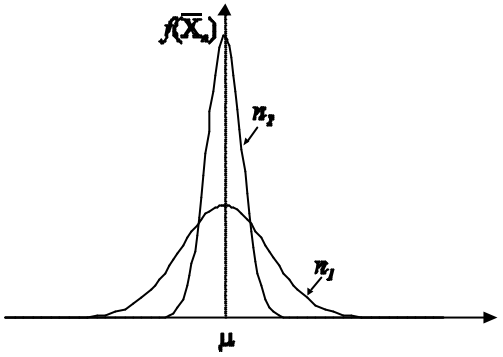

Figura 4.1: Distribuição de $\overline{X}$ para duas amostras de dimensão diferente ($n_2 > n_1$).

à dimensão da população, no cálculo do desvio padrão da distribuição da média amostral deve usar-se uma *factor de correção para populações finitas*:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\frac{N-n}{N-1} \quad \Leftrightarrow \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

onde N é a dimensão da população e n é a dimensão da amostra.

Normalmente, se $n < 0.05N$ (a dimensão da amostra é inferior a 5% da dimensão da população) não se usa o factor de correção.

Repare-se que o factor de correção é sempre inferior a 1. Logo, o desvio padrão corrigido da média na amostra é inferior. Isto está de acordo com a intuição porque estamos a considerar casos em que a amostra é uma fracção relativamente elevada da população.

4.2.2 Distribuição da diferença entre duas médias

Vamos supor que estamos interessados em estimar a diferença na média de uma determinada variável para duas populações diferentes (por exemplo: homens versus mulheres, portugueses versus americanos,...).

Seja n_1 a dimensão da amostra retirada da primeira população e n_2 a dimensão da amostra retirada da segunda população. Sejam μ_1 e μ_2 as médias em cada uma das populações e σ_1^2, σ_2^2 as respectivas variâncias. A distribuição amostral da diferença das médias tem as seguintes propriedades:

- $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$
- $\text{var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Tal como no caso da média, se as populações tiverem distribuição normal a distribuição por amostragem da diferença de médias também é normal. Para além disso, independentemente da distribuição das populações, se n_1 e n_2 forem elevados então a distribuição por amostragem da diferença de médias é aproximadamente normal.

4.2.3 Distribuição da proporção

Suponhamos que estamos interessados em estimar a fracção da população que possui uma determinada característica (fuma ou não, tem olhos azuis ou não, usa a internet ou não, vota num dado candidato ou não...). Note-se que este tipo de variável pode ser representado por 0 ou 1. A população tem uma distribuição Bernoulli.

Vamos admitir que a proporção da população com a característica em causa é p. Se recolhermos uma amostra com dimensão n e calcularmos a proporção da amostra com a característica obtemos a estatística $\widehat{P}$, com as seguintes propriedades:

- $E\left[\widehat{P}\right] = p$
- $\text{var}\left[\widehat{P}\right] = \frac{p(1-p)}{n}$

4.2.4 Distribuição de $\frac{(n-1)S^2}{\sigma^2}$

Se $X_1, X_2, \cdots, X_n$ forem as observações de uma amostra aleatória de dimensão n retirada de uma população normal $N(\mu, \sigma^2)$ então $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$ é $\mathcal{X}^2(n-1)$

Notar que o número de graus de liberdade é $n-1$. Isto é bastante intuitivo, ao estimarmos $\overline{X}$ perdemos um grau de liberdade.

4.3 Estimação

Nesta secção vamos abordar a questão de fazer inferências sobre a população quando temos informação para uma amostra dessa população. Muitas das vezes a distribuição da população depende só de alguns parâmetros (por exemplo: se soubermos que a distribuição é normal basta saber média e variância) ou então nós estamos interessados só em certos parâmetros. A questão é: «será que podemos inferir algo sobre o valor desses parâmetros de interesse com base na informação da amostra?» O nosso objectivo é arranjar uma forma de *estimar* o valor do parâmetro.

Um *estimador* de um parâmetro θ da população é uma variável aleatória que depende da informação da amostra, e que é usada para estimar o valor de θ. O valor do estimador para uma amostra específica chama-se *estimativa*.

Exemplo: $\overline{X} = \frac{\sum x_i}{n}$ é um estimador, $\bar{x} = 5$ é uma estimativa.

4.3.1 Propriedades desejáveis dos estimadores

Mas qual é o critério para escolher estimadores? Se $\widehat{\theta}$ é um estimador de θ, que propriedades é que $\widehat{\theta}$ deve ter para ser um bom estimador?

- Não enviesamento
- Consistência
- Eficiência
- Erro quadrado médio minímo

Não enviesamento

Um estimador $\widehat{\theta}$ diz-se não enviesado se a média desse estimador for igual ao valor do parâmetro θ que queremos estimar, ou seja

$$E(\widehat{\theta}) = \theta$$

O que é que esta propriedade nos diz? É preciso não esquecer que $\widehat{\theta}$ é uma variável aleatória. O valor de $\widehat{\theta}$ depende de qual é a amostra que é recolhida. O que a propriedade nos diz é que se nós repetíssemos o processo de amostragem muitas vezes a média das estimativas obtidas é igual ao valor do parâmetro na população. A Figura 4.2 ilustra graficamente o que significa um estimador ser não enviesado.

Vejamos alguns exemplos:

- $\overline{X}$ é um estimador não enviesado de μ

$$E(\overline{X}) = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} [n\mu] = \mu$$

- S^2 é um estimador não enviesado de σ^2

$$E(S^2) = E\left[\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \overline{X})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} \left(X_i - \mu + \mu - \overline{X}\right)^2\right]$$

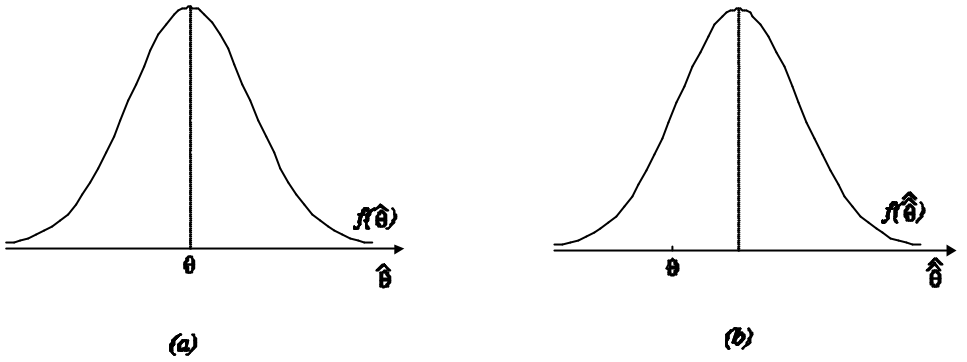

Figura 4.2: *(a)* $\widehat{\theta}$ é um estimador não enviesado de θ. *(b)* $\widehat{\widehat{\theta}}$ é um estimador enviesado de θ.

o que é equivalente a

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\left((X_i-\mu)^2 + (\mu-\overline{X})^2 - 2(X_i-\mu)(\overline{X}-\mu)\right)\right]$$

mas isto é

$$\frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i-\mu)^2 + n(\mu-\overline{X})^2 - 2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)\right]$$

ou seja

$$\frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\mu-\overline{X})^2\right] = \frac{1}{n-1}\left[n\sigma^2 - n\sigma^2/n\right] = \sigma^2$$

Neste exemplo dividimos a soma dos desvios ao quadrado por $n-1$. $n-1$ são os *graus de liberdade* na estimação da variância (perdemos um grau de liberdade ao estimar $\overline{x}$).

Consistência

Esta propriedade refere-se ao comportamento do estimador à medida que a dimensão da amostra se aproxima de infinito. Em termos intuitivos é desejável que à medida que a amostra se torna maior o estimador esteja cada vez mais *próximo* do parâmetro. Consistência significa que quando o tamanho da amostra é muito elevado a distribuição da estatística fica muito muito concentrada em torno do parâmetro da população.

A Figura 4.3 ilustra graficamente a ideia da consistência. A figura apresenta a função densidade para três amostras de dimensão diferente ($n_3 > n_2 > n_1$). Quanto maior a

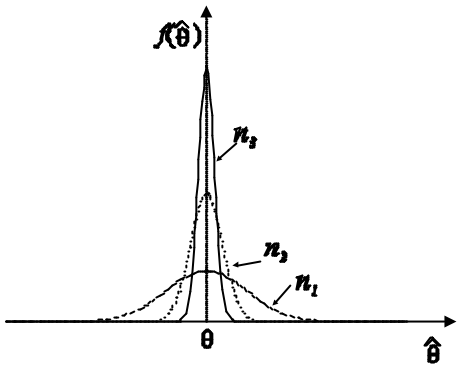

Figura 4.3: Função densidade do estimador $\widehat{\theta}$ para amostras com dimensão $n_3 > n_2 > n_1$.

amostra mais «concentrada» é a função densidade em torno do valor do parâmetro.Em termos um pouco mais formais: seja $\widehat{\theta}_n$ o estimador quando a amostra é de dimensão n e seja ε um qualquer número positivo (tão pequeno quanto nós quizermos), o estimador diz-se *consistente* se

$$\lim_{n\to\infty} P\left[(\widehat{\theta}_n - \theta) < \varepsilon\right] \to 1.$$

Ou seja, se o estimador for consistente, quando n tende para infinito o estimador converge (em probabilidade) para o verdadeiro valor do parâmetro. Ou ainda, é possível aproximar, tanto quanto desejarmos, o estimador do verdadeiro valor do parâmetro desde que a amostra seja suficientemente grande.

Um estimador pode ser enviesado mas ser consistente. Este facto é ilustrado na Figura 4.4, onde estão representadas as funções densidade do estimador com amostras de dimensão diferentes ($n_3 > n_2 > n_1$). O estimador $\widehat{\theta}$ é um estimador enviesado de θ (isto é particularmente visível para amostras de pequena dimensão). Contudo, à medida que a dimensão da amostra aumenta a função densidade concentra-se cada vez mais em torno do valor do parâmetro. Repare-se que à medida que n se torna maior o enviesamento do estimador fica cada vez mais pequeno e tende para zero quando n tende para infinito.

Exemplo 4.5 o estimador

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

é um estimador enviesado de σ^2 mas, no entanto, é consistente.♦

Em contrapartida, um estimador pode ser não enviesado e não ser consistente. Isto acontece se a a variância do estimador não tender para zero quando n tender para infinito.

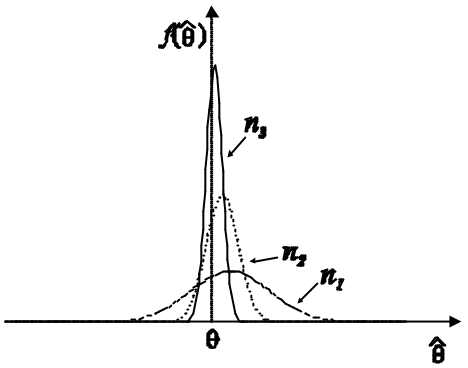

Figura 4.4: $\widehat{\theta}$ é um estimador enviesado mas consistente.

Eficiência

Podem existir muitos estimadores não enviesados. Como escolher entre eles? É natural escolher o estimador cuja f.d.p. está mais concentrada em relação ao valor do parâmetro da população. Ou seja aquele que tem menor dispersão em torno da média. Se nós tivermos dois estimadores não enviesados de θ, $\widehat{\theta}_1$ e $\widehat{\theta}_2$, baseados em amostras de igual dimensão, então dizemos que o estimador $\widehat{\theta}_1$ é *mais eficiente* se

$$\text{Var}(\widehat{\theta}_1) < \text{Var}(\widehat{\theta}_2)$$

e a eficência relativa de um estimador em relação ao outro é

$$\text{eficiência relativa} = \frac{\text{Var}(\widehat{\theta}_2)}{\text{Var}(\widehat{\theta}_1)}$$

Na Figura 4.5 estão representadas as funções densidade de dois estimadores não enviesados do parâmetro θ. O estimador $\widehat{\theta}$ é mais eficiente que o estimador $\widehat{\widehat{\theta}}$. É importante sublinhar que a eficiência é uma propriedade relativa (estamos a comparar estimadores). No entanto a comparação é feita só entre estimadores que são *não enviesados*.

Exemplo 4.6 A média e a mediana são ambos estimadores não enviesados de μ quando a distribuição é normal. No entanto a média tem menor variância

$$\text{eficiência relativa} = \frac{1.57\sigma^2/n}{\sigma^2/n} = 1.57.\blacklozenge$$

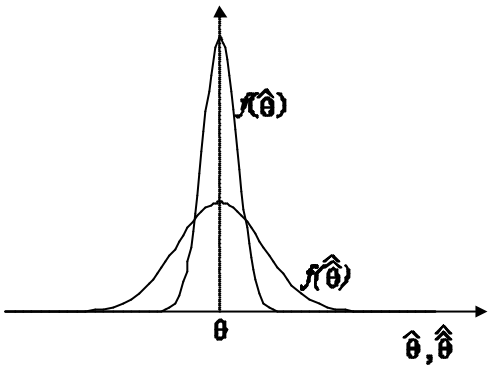

Figura 4.5: O estimador $\widehat{\theta}$ é mais eficiente que o estimador $\widehat{\widehat{\theta}}$.

Erro quadrado médio mínimo

Embora a propriedade de não enviesamento seja desejável pode acontecer que nenhum dos estimadores não enviesados seja muito preciso, pode acontecer que todos eles tenham uma variância elevada em torno de θ. É possível que haja estimadores desse parâmetro que tenham algum enviesamento mas que tenham menor variância. Nestes casos não é óbvio que o estimador não enviesado seja o mais apropriado. Esta ideia é apresentada na Figura 4.6 onde estão representadas as funções densidade de dois estimadores de θ. O estimador $\widehat{\theta}$ é um estimador enviesado de θ, mas tem um variância relativamente pequena. Em contrapartida, o estimador $\widehat{\widehat{\theta}}$ é um estimador não enviesado de θ, mas tem um variância relativamente elevada. Qual dos dois estimadores é melhor?

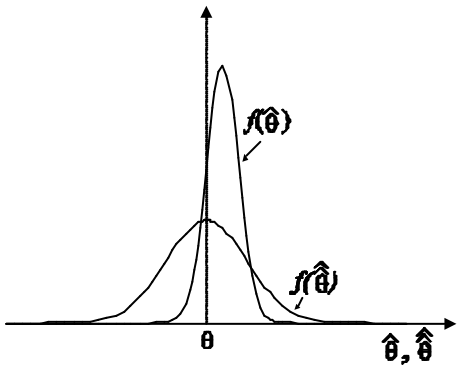

Figura 4.6: O estimador $\widehat{\theta}$ tem um erro quadrado médio inferior ao estimador $\widehat{\widehat{\theta}}$.

Um critério que à partida parece bastante lógico para decidir nestes casos é escolher o estimador que em média tem um menor erro quadrado (porquê quadrado?). O erro

quadrado médio de um estimador $\widehat{\theta}$ é dado por

$$EQM(\widehat{\theta}) = E\left[(\widehat{\theta} - \theta)^2\right]$$

Pode mostrar-se que:

$$EQM(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + \mathrm{Env}^2$$

Isto resulta de

$$E\left[\left(\widehat{\theta} - E(\widehat{\theta}) + E(\widehat{\theta}) - \theta\right)^2\right] = E\left[\left(\widehat{\theta} - E(\widehat{\theta})\right)^2 + \left(E(\widehat{\theta}) - \theta\right)^2 + 2\left(\widehat{\theta} - E(\widehat{\theta})\right)\left(E(\widehat{\theta}) - \theta\right)\right]$$

Isto sugere que se queremos minimizar EQM pode ser preferível um estimador enviesado, desde que a variância desse estimador seja mais pequena e que mais que compense pelo enviesamento (veja a Figura 4.6).

É interessante notar que, se estivermos a considerar só estimadores não enviesados a minimização de EQM corresponde à minimização da variância. Logo obteremos o estimador mais eficiente.

4.3.2 Como encontrar estimadores?

Até aqui enunciamos algumas propriedades desejáveis de um estimador. Mas, há tantas funções que é possível construir com base em $X_1, X_2, \cdots, X_n$! Como encontrar possíveis estimadores? Há muitos métodos: método dos momentos, método da máxima verosimilhança, método dos mínimos quadrados.

Método dos momentos

Se existirem k parâmetros que têm que ser estimados, o método dos momentos, consiste em igualar os primeiros k momentos da amostra aos primeiros k momentos da população. Os k momentos da população dependem dos k parâmetros a estimar. Obtemos assim um sistema com k equações e k incógnitas e resolvendo o sistema encontramos os estimadores dos k parâmetros.

Exemplo 4.7 Consideremos uma v.a. X com distribuição $N(\mu, \sigma^2)$. Neste caso

$$E(X) = \mu \quad \text{e} \quad E(X^2) = \sigma^2 + \mu^2$$

Dada uma amostra de dimensão n os dois primeiros momentos na amostra são dados por:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{e} \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

igualando os momentos da amostra aos momentos na população obtemos

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} x_i = \mu \\ \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma^2 + \mu^2 \end{cases}$$

e a solução deste sistema em relação a μ e σ^2 dá-nos os estimadores do método dos momentos:

$$\begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \blacklozenge \end{cases}$$

Método da máxima verosimilhança

Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória retirada de uma distribuição com função densidade de probabilidade $f(x;\theta_1,\theta_2,\cdots,\theta_k)$ em que $\theta_1,\theta_2,\cdots,\theta_k$ são parâmetros desconhecidos, com $(\theta_1,\theta_2,\cdots,\theta_k) \in$ (espaço dos parâmetros - conjunto de valares que os parâmetros podem tomar).

A função densidade de probabilidade da amostra aleatória é

$$L(\theta_1,\theta_2,\cdots,\theta_k) = f(x_1;\theta_1,\theta_2,\cdots,\theta_k) f(x_2;\theta_1,\theta_2,\cdots,\theta_k)\cdots f(x_k;\theta_1,\theta_2,\cdots,\theta_n)$$

quando interpretada como função dos parâmetros é chamada a *função de verosimilhança*. Repare-se que a f.d.p. da amostra aleatória depende dos valores dos parâmetros. Dada uma amostra em particular aquilo que se pergunta é: quais são os valores de $\theta_1,\theta_2,\cdots,\theta_k$ que com maior probabilidade geraram esta amostra. Ou seja, queremos encontrar os valores dos parâmetros que maximizam o valor da f.d.p. da amostra.

Suponhamos que as funções $u_1(x_1,x_2,\cdots,x_n),\cdots,u_k(x_1,x_2,\cdots,x_n)$ maximizam o valor da função de verosimilhança. Então os estimadores de máxima verosimilhança são

$$\begin{cases} \widehat{\theta}_1 = u_1(x_1,x_2,\cdots,x_n) \\ \widehat{\theta}_2 = u_2(x_1,x_2,\cdots,x_n) \\ \qquad\qquad \vdots \\ \widehat{\theta}_k = u_k(x_1,x_2,\cdots,x_n) \end{cases}$$

Exemplo 4.8 Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória retirada de uma distribuição exponencial com f.d.p.

$$f(x,\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \quad 0 < x < \infty, \ \theta \in \quad = \{\theta : 0 < \theta < \infty\}$$

Recorde-se que o valor esperado desta variável é θ e a variância é θ^2. A função de verosimilhança é

$$L(\theta) = \left(\frac{1}{\theta} e^{-\frac{x_1}{\theta}}\right)\left(\frac{1}{\theta} e^{-\frac{x_2}{\theta}}\right)\cdots\left(\frac{1}{\theta} e^{-\frac{x_n}{\theta}}\right) = \frac{1}{\theta^n} e^{-\frac{\sum_i x_i}{\theta}}$$

Se tomarmos o logaritmo desta função, como o logaritmo é uma função crescente a solução do problema de maximização será a mesma (e como isto envolve produtos, logaritmo ajuda porque ficamos com somas)

$$\ln L(\theta) = -n \ln \theta - \frac{\sum_{i=1}^n x_i}{\theta}$$

Para encontrarmos o máximo derivamos e igualamos a zero

$$\frac{d \ln L(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2} = 0 \Leftrightarrow -n\theta + \sum_{i=1}^n x_i = 0 \Leftrightarrow \theta = \frac{\sum_{i=1}^n x_i}{n}$$

logo o estimador de máxima verosimilhança de θ é a média na amostra.♦

4.3.3 Estimação pontual versus estimação por intervalos

Quando escolhemos uma amostra e calculamos o valor do estimador para essa amostra obtemos uma estimativa. Uma estimativa é simplesmente um dos muitos valores que o estimador poderia tomar. Contudo, as nossas inferências sobre o parâmetro da população são baseadas nessa estimativa. Por exemplo, às observações $x_1, x_2, \cdots, x_n$ corresponde a estimativa $\overline{x}$. Se usarmos $\overline{x}$ como estimativa de μ estamos a fazer *estimação pontual* (obtemos um certo valor que supostamente é um bom palpite do valor de μ). Mas, qual é o nosso grau de confiança nessa estimativa? Por exemplo: se na sondagem sobre as eleições presidenciais se obteve que 54% dos individuos da amostra pretende votar no partido A, qual é o grau de confiança de que a verdadeira percentagem de votos esteja entre 51% e 57%? Este tipo de perguntas leva-nos a pensar em *estimação por intervalos*.

Um estimador por intervalos de um parâmetro da população é uma regra para determinar um intervalo que com certa probabilidade contém o parâmetro da população em que estamos interessados. Naturalmente, há um trade-off entre o grau de confiança e o

tamanho do intervalo. Quanto maior for o grau de confiança (quanto mais certos nós quisermos estar de que o verdadeiro valor do parâmetro está no intervalo) maior terá de ser o intervalo (menos precisa será a nossa estimativa).

Para construirmos intervalos de confiança devemos relembrar-nos (mais uma vez) que o estimador é uma variável aleatória. A precisão do estimador depende da sua distribuição ⇒ para construir intervalos de confiança precisamos de conhecer a distribuição ⇒ importância da distribuição por amostragem.

Vamos ver isto usando um exemplo. Consideremos uma população normal em que a média da população é desconhecida, mas σ^2 é conhecido e que queremos estimar μ. Consideremos o estimador $\overline{X}$, sabemos que $\overline{X}$ tem distribuição $N(\mu, \sigma^2/n)$, ou ainda que

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

usando a tabela para a normal estandartizada, dada a probabilidade $1 - \alpha$ é possível encontrar o *valor critíco* $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

É claro que o *valor critíco* $z_{\alpha/2}$ depende de α. Por exemplo: se $1 - \alpha = 0.95$, então $z_{0.025} = 1.96$, se $1 - \alpha = 0.90$, então $z_{0.05} = 1.645$. Na Figura 4.7 está representada a função densidade da normal estandartizada e indicados os valores critícos necessários para os três níveis de confiança mais usados na prática: 90%, 95% e 99%. Note-se que quanto mais elevado for o nível de confiança desejado maior é o valor critíco.

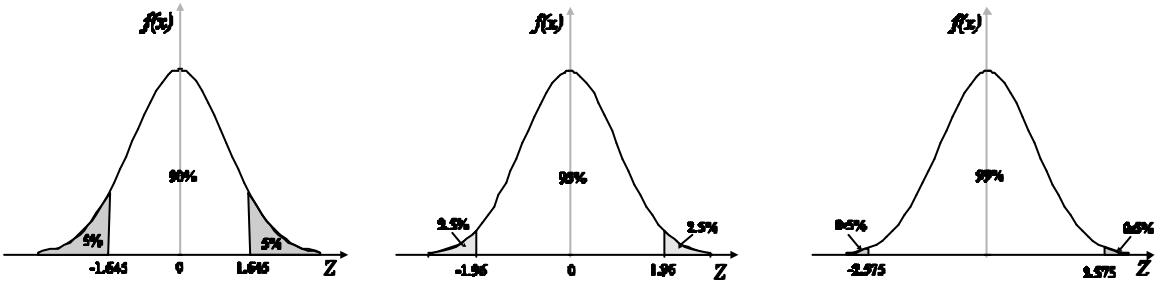

Figura 4.7: Distribuição normal para determinar valor de $z_{\alpha/2}$ necessário para um nível de confiança de *(a)* 90%, *(b)* 95% e *(c)* 99%.

Mas, isto é equivalente a:

$$P\left(\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}\right) = 1 - \alpha$$

ou por palavras, a probabilidade que o *intervalo aleatório*

$$\left[\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

contenha μ é $1 - \alpha$. Outra forma de dizer, é que se repetirmos a amostragem muitas vezes e construirmos o intervalo de confiança para cada amostra em $100(1-\alpha)\%$ dos casos o intervalo contém o verdadeiro valor do parâmetro. Este intervalo é um *intervalo de confiança* de $100(1-\alpha)\%$ de μ.

A interpretação do conceito de intervalo de confiança é ilustrada na Figura 4.8. Nesta figura estão representados os intervalos de confiança da média na população, μ, para dez amostras diferentes mas com igual dimensão. Como a média obtida em cada uma das amostras é diferente, os intervalos de confiança vão também ser diferentes para as várias amostras. No exemplo da Figura 4.8, o intervalo de confiança para uma das amostras não contém μ. Como o nível de confiança é 90%, se o processo de amostragem fosse repetido um número muito elevado de vezes, 90% dos intervalos conteriam o verdadeiro valor do parâmetro da população e 10% dos intervalos não incluiriam μ.

Alguns comentários sobre o intervalo de confiança $\left[\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$:

- $1 - \alpha$ chama-se o *coeficiente de confiância*.
- O intervalo de confiança é centrado em $\overline{x}$ e obtém-se subtraindo e somando $z_{\alpha/2}\sigma/\sqrt{n}$.
- A amplitude do intervalo depende de n, de σ e de α:
  - Quanto maior for a variabilidade na população, σ, maior é a amplitude do intervalo $\Rightarrow$ menos precisa é a estimativa.
  - Quanto maior for n, menor é $\sigma/\sqrt{n}$ e logo menor é a amplitude do intervalo correspondente a um dado nível de confiança $\Rightarrow$ mais precisa é a estimativa.
  - Quanto maior for o nível de confiança, $1 - \alpha$, maior é o valor de $z_{\alpha/2}$ e logo maior é a amplitude do intervalo. Isto é lógico, se quizermos estar mais certos de que o intervalo contém μ teremos que, para uma mesma dimensão da amostra, aumentar a amplitude do intervalo. Maior grau de confiança $\Rightarrow$ menos precisão na estimativa.
  - Ideia do trade-off.

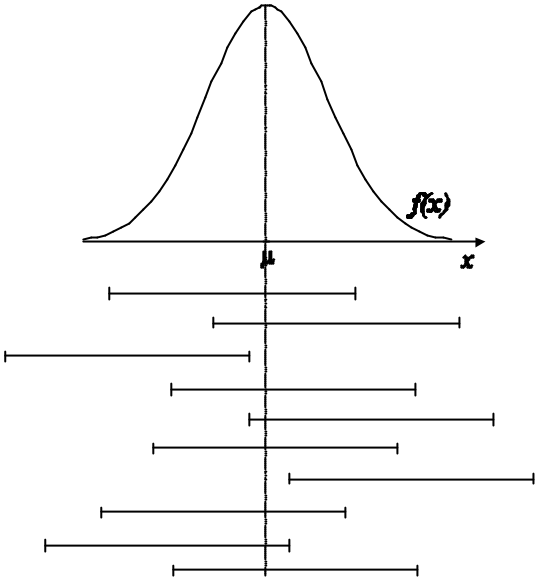

Figura 4.8: Intervalos de confiança de 90% para a média na população, considerando dez amostras diferentes.

4.4 Intervalos de confiança para a média

4.4.1 Variância conhecida

Quando estudamos distribuições por amostragem vimos que há dois casos em que utilizar a normal como distribuição de $\overline{X}$ é apropriado:

- Se a população for normal $N(\mu, \sigma^2) \Rightarrow \overline{x} \sim N(\mu, \sigma^2/n)$
- Se a população tem média μ e variância σ^2, independentemente da sua distribuição, quando a amostra é grande a distribuição de $\overline{X}$ é aproximadamente normal pelo teorema do limite central.

Nestes casos se conhecermos σ^2 é possível construír intervalos de confiança para μ, com base na estimativa da média na amostra. Isto é precisamente aquilo que fizemos anteriormente quando introduzimos o conceito de intervalo de confiança.

Resumindo: se tivermos uma amostra de dimensão n de uma população com média μ e variância σ^2 , se σ^2 for conhecido e $\overline{x}$ for a média observada na amostra, então o intervalo

de $100(1-\alpha)\%$ de confiança é dado por

$$\left[\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

Exemplo 4.9 Seja x a duração de uma lâmpada de 60-watts comercializada por um certo produtor. A experiência passada permite concluir que a distribuição da duração de lâmpadas é normal com variância 1269. Numa amostra aleatória de 27 lâmpadas de 60-watts verificou-se que a duração média foi de 1478 horas. Construa um intervalo com um nível de confiança de 95% para a duração média das lâmpadas de 60-watts daquele produtor.

O intervalo é:

$$\left[1478 - 1.96\left(\frac{36}{\sqrt{27}}\right),\ 1478 + 1.96\left(\frac{36}{\sqrt{27}}\right)\right] = [1464.42, 1491, 58]\ \blacklozenge$$

4.4.2 Variância desconhecida

Na maior parte dos casos a variância da população é tão desconhecida quanto a média. Se a variância for desconhecida teremos que estimar a variância da população se quizermos construir intervalos de confiança para a média.

Se a distribuição da população for normal, pode mostrar-se que

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

ou seja a distribuição $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ é uma t com $n-1$ graus de liberdade. Se n for grande a t e a normal são praticamente idênticas e pode construir-se o intervalo usando a normal. Mas, se t for pequeno ($n < 30$) deve usar-se a distribuição t.

Sabendo isto é fácil construir o intervalo de confiança.

$$P\left(-t_{\alpha/2}(n-1) \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right) = 1 - \alpha$$

depois de manipulações semelhantes às feitas acima obtemos

$$P\left(\overline{X} - t_{\alpha/2} \cdot S/\sqrt{n} \leq \mu \leq \overline{X} + t_{\alpha/2} \cdot S/\sqrt{n}\right) = 1 - \alpha$$

Logo *intervalo aleatório*:

$$\left[\overline{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right),\ \overline{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]$$

contém o verdadeiro valor do parâmetro com probabilidade $1-\alpha$.

O intervalo de confiança para uma amostra concreta de dimensão n, com média $\overline{x}$ e desvio-padrão s é:

$$\left[\overline{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right),\ \overline{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right].$$

Exemplo 4.10 Uma amostra aleatória de seis automóveis com o mesmo modelo e ano foi recolhida e o seu consumo médio de gasolina registado: 6.0, 6.2, 5.9, 6.1, 6.2, 6.3. Construa um intervalo de confiança a um nível de confiança de 90% para o consumo médio deste modelo de automóveis.

A média na amostra é 6.12, o desvio padrão na amostra é 0.177, o valor de $t_{5,0.5} = 2.015$, logo o intervalo de confiança é

$$\left[6.12 - 2.015\left(\frac{0.177}{\sqrt{6}}\right), 6.12 + 2.015\left(\frac{0.177}{\sqrt{6}}\right)\right] = [5.9744,\ 6.2656]\ \blacklozenge$$

4.5 Intervalos de confiança para diferença de médias

Muitas vezes estamos interessados em comparar as médias de duas populações. Por exemplo, um produtor tem dois fornecedores diferentes e quer testar se há ou não diferença na qualidade do produto fornecido por ambos os produtores.

4.5.1 Variâncias conhecidas

Se as populações de onde as amostras são retiradas forem independentes e tiverem distribuição normal, a diferença entre as médias tem também distribuição normal, com média $\mu_x - \mu_y$ e variância $\sigma_x^2/n_x + \sigma_y^2/n_y$. Isto implica que:

$$P\left((\overline{X} - \overline{Y}) - z_{\alpha/2} \cdot \sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y} \leq \mu_x - \mu_y \leq (\overline{X} - \overline{Y}) + z_{\alpha/2} \cdot \sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}\right) = 1 - \alpha.$$

4.5.2 Variâncias desconhecidas – amostra grande

Se as variâncias não forem conhecidas temos que estimá-las com base nas amostras. Se as amostras forem grandes a distribuição normal será uma boa aproximação para a distribuição de $(\overline{x} - \overline{y})$. Isto significa que

$$\left[(\overline{x} - \overline{y}) \pm z_{\alpha/2}\sqrt{s_x^2/n_x + s_x^2/n_y}\right]$$

é um intervalo de $100(1-\alpha)\%$ de confiança de $\mu_x - \mu_y$.

Exemplo 4.11 Num estudo sobre as consequências do tabaco no trabalho recolheram-se duas amostras aleatórias independentes de fumadores e não fumadores. Nos 96 fumadores o absentismo médio por mês foi de 2.15 horas, com o desvio padrão na amostra igual a 2.09. Nos 206 empregados não fumadores o absentismo médio foi de 1.69 horas por mês com um desvio padrão de 1.91 horas por mês. Construa um intervalo de confiança de 99% para a diferença das médias das duas populações.

Solução: Os resultados nas amostras são os seguintes:

$$\begin{array}{lll} \overline{x} = 2.15 & n_x = 96 & s_x = 2.09 \\ \overline{y} = 1.69 & n_y = 206 & s_y = 1.91 \end{array}$$

Como as amostras são grandes podemos utilizar a distribuição normal. O valor de $z_{\alpha/2} = z_{0.005} = 2.575$. O intervalo de confiança é

$$(2.15 - 1.69) \pm 2.575\sqrt{\frac{(2.09)^2}{96} + \frac{(1.91)^2}{206}}$$

ou seja

$$-.19 \leq \mu_x - \mu_y \leq 1.11$$

como o valor zero está incluído neste intervalo a evidência na amostra contra a hipótese de que as duas médias são iguais não é muito forte.♦

4.6 Intervalos de confiança para proporções

Queremos estimar qual é a proporção da população que tem um certo atributo. Um estimador que pode ser utilizado para este efeito é a proporção na amostra com aquele atributo. A questão que se coloca é: «qual é a distribuição desse estimador?»

Se a amostra for grande a distribuição do estimador é aproximadamente normal. Seja n a dimensão da amostra e seja $\widehat{P}$ a fracção de elementos da amostra com o atributo em causa. Para n grande

$$\frac{\widehat{P} - p}{\sqrt{\widehat{P}(1-\widehat{P})/n}}$$

tem uma distribuição aproximadamente $N(0,1)$.

- Logo, o intervalo

$$\left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

é um intervalo de $100(1-\alpha)\%$ nível de confiança de p.

Exemplo 4.12 Numa certa campanha eleitoral um dos candidatos manda realizar uma sondagem (aleatória) entre a população com capacidade de voto. Os resultados foram que em 351 eleitores 194 dizem favorecer o candidato. O candidato deve ou não sentir-se confiante que vai ganhar?

Solução: A proporção de eleitores na amostra favorecendo o candidato é $\widehat{p} = \frac{194}{351} = 0.553$. Se construirmos um intervalo de confiança de 95% obtemos

$$0.553 \pm 1.96\sqrt{\frac{0.553 \times 0.447}{351}} \Leftrightarrow [.501, 0.605]$$

como este intervalo está todo acima de 50% o candidato pode sentir-se relativamente confiante de que ganha. Mas, repare-se que se o nível de confiança for maior a amplitude do intervalo será maior e logo a possibilidade de ter menos de 50% dos votos existe.♦

4.7 Intervalos de confiança para variância

4.7.1 Intervalo para variância de população normal

Nesta secção vamos estudar intervalos de confiância para a variância de uma *população normal*. Naturalmente este intervalo é baseado na variância da amostra

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

onde usamos o facto de $(n-1)S^2/\sigma^2$ ser $\mathcal{X}^2(n-1)$ para definirmos os intervalos de confiança. Se designarmos por $\mathcal{X}^2_{n-1,\alpha/2}$ o valor b tal que a probabilidade de a v.a. com distribuição qui-quadrado com $n-1$ graus de liberdade ser maior ou igual que b é igual a $\alpha/2$, então temos

$$\begin{aligned} P\left[\mathcal{X}^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \mathcal{X}^2_{n-1,\alpha/2}\right] &= 1-\alpha \Leftrightarrow \\ P\left[\frac{(n-1)S^2}{\mathcal{X}^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\mathcal{X}^2_{n-1,\alpha/2}}\right] &= 1-\alpha \end{aligned}$$

Exemplo 4.13 Um produtor está preocupado com a variabilidade nos níveis de impureza contidos na matéria-prima recebida de um fornecedor. Uma amostra aleatória de 15 encomendas mostrou um desvio padrão de 2.36% no nível de concentração de impurezas. Assuma que a população é normal. Encontre um intervalo de confiança de 95% para a variância na população.

O valor de $\mathcal{X}^2_{14,0.975} = 5.629$ e $\mathcal{X}^2_{14,0.025} = 26.12$ e $14(2.36)^2 = 77.974$ logo

$$\frac{77.974}{26.12} \leq \sigma^2 \leq \frac{77.974}{5.629} \Leftrightarrow 2.99 \leq \sigma^2 \leq 13.85 \blacklozenge$$

4.7.2 Intervalo para rácio de variâncias de populações normais independentes

Se estivermos interessados em comparar a variância de duas populações normais independentes podemos fazê-lo construindo um intervalo de de confiança para $\frac{\sigma^2_X}{\sigma^2_Y}$.

Como $(n_x - 1)S^2_X/\sigma^2_X$ e $(n_y - 1)S^2_Y/\sigma^2_Y$ têm ambas distribuição qui-quadrado, com $n_x - 1$ e $n_y - 1$ graus de liberdade, respectivamente, se tomarmos o rácio delas divido pelos respectivos graus de liberdade obtemos uma variável aleatória com distribuição F_{n_x-1,n_y-1}. Ou seja

$$\frac{\frac{(n_x-1)S^2_X}{(n_x-1)\sigma^2_X}}{\frac{(n_y-1)S^2_Y}{(n_y-1)\sigma^2_Y}} = \frac{\frac{S^2_X}{\sigma^2_X}}{\frac{S^2_Y}{\sigma^2_Y}} = F$$

tem distribuição F_{n_x-1,n_y-1}.

Exemplo 4.14 Sejam X e Y a quantidade (em miligramas) de nicotina em cigarros com filtro e sem filtro, respectivamente. Assuma que as distribuições de X e Y são normais $N(\mu_X, \sigma^2_X)$ e $N(\mu_Y, \sigma^2_Y)$. Considere as duas amostras aleatórias independentes: uma amostra de 9 elementos de X

$$0.9 \;\; 1.1 \;\; 0.1 \;\; 0.7 \;\; 0.3 \;\; 0.9 \;\; 0.8 \;\; 1.0 \;\; 0.4$$

e uma amostra de 11 elementos de Y

$$1.5 \;\; 0.9 \;\; 1.6 \;\; 0.5 \;\; 1.4 \;\; 1.9 \;\; 1.0 \;\; 1.2 \;\; 1.3 \;\; 1.6 \;\; 2.1$$

encontre um intervalo de confiança de 95% para $\frac{\sigma^2_X}{\sigma^2_Y}$. $\blacklozenge$

4.8 Escolha da dimensão da amostra

Quão grande deve ser a amostra para estimar um parâmetro com um certo nível de precisão? A resposta a esta pergunta depende da variabilidade na população. Por exemplo, se quizermos estimar a média da população e soubermos que a variância na população é nula basta $n = 1$ para estimar com exactidão a média da população. Mas, se a variância na população for elevada e desejarmos estimar com bastante precisão μ a amostra necessária será elevada.

A dimensão da amostra depende também do nível de precisão com que queremos estimar o parâmetro. Se quizermos obter uma estimativa mais precisa (isto é, com menor amplitude do intervalo de confiança) teremos que ter uma amostra mais elevada.

Se a variância da amostra for conhecida é fácil calcular qual é a dimensão da amostra que nos garante uma dada amplitude do intervalo de confiança. De facto, nós sabemos que

$$\left[\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

é o intervalo de $100\,(1-\alpha)\,\%$ nível confiança da média da população. Este intervalo está centrado na média observada na amostra e expande-se $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ para cada um dos lados. Suponha-se que queremos garantir que

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq L$$

Isso implica que

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{L}\right)^2$$

Como seria de esperar, quanto menor for a amplitude do intervalo que desejamos maior terá que ser n (maior precisão $\Rightarrow$ maior n). Para além disso, quanto maior for a variância na população, maior terá que ser n.

Um outro caso com interesse é o da proporções. Vimos atrás que o intervalo de confiança para a proporção é centrado na proporção na amostra e extende-se

$$z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

para cada lado. O problema aqui é que não é possível saber a amplitude do intervalo sem primeiro estimar a proporção na amostra. Contudo, é possível escolher n de forma

a garantir que a amplitude não ultrapassa um certo valor. Basta notar que o valor mais elevado que $\widehat{p}(1-\widehat{p})$ pode tomar é 0.25. Logo, se escolhermos

$$n \geq \frac{0,25 \times z^2_{\alpha/2}}{L^2}$$

temos a certeza que o intervalo se estende para cada lado num valor não superior a L.

Capítulo 5

Teste de hipóteses

5.1 Conceitos básicos

No capítulo anterior vimos como a informação na amostra pode ser usada para estimar parâmetros da distribuição da população. Neste capítulo vamos estudar como é que a informação na amostra pode ser utilizada para testar a validade de conjecturas, ou *hipóteses*, que tenhamos formado sobre a população.

Por exemplo, sou um produtor de um certo produto e gostaria de assegurar que menos de 2% dos produtos produzidos são defeituosos. Podemos testar se neste momento a quantidade de produtos defeituosos é inferior ou igual a 2% fazendo uma verificação a uma amostra aleatória de produtos e decidir depois com base nos resultados obtidos nessa amostra. Outro exemplo, testar se o salário é o mesmo para mulheres e homens com mesma qualificação e experiência.

Resumindo, temos uma certa hipótese sobre a população, conclui-se sobre o mérito ou não da hipótese usando informação na amostra.

Seja θ o parâmetro de interesse (as ideias podem ser generalizadas para um vector de parâmetros) e suponhamos que temos uma certa hipótese formada sobre o valor do parâmetro, hipótese essa que continuaremos a admitir a não ser que haja forte evidência de que a hipótese é falsa. A esta hipótese, que designaremos por H_0, chama-se em estatística a *hipótese nula*.

Se a hipótese nula não for verdadeira então alguma hipótese alternativa terá de o ser. Ao efectuarmos um teste de hipótese formulamos sempre qual é a *hipótese alternativa* em relação à qual a hipótese nula está a ser testada. A hipótese alternativa é designada por H_1.

Tanto a hipótese nula como a alternativa podem ser *simples* ou *compósitas*. Uma hipótese é simples se especificar um valor único para o parâmetro, é compósita se especificar um conjunto de valores.

Uma outra distinção com interesse é entre testes unilaterais e bilaterais. Por exemplo, o teste de $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$ é um teste bilateral porque a hipótese alternativa considera valores do parâmetro inferiores e superiores aos valores do parâmetro se a hipótese nula for verdadeira. Isto é, a hipótese alternativa considera valores à esquerda e à direita de μ_0.

Depois de especificadas as hipóteses nula e alternativa e de termos recolhida uma amostra temos que decidir se devemos ou não rejeitar a hipótese nula com base na informação da amostra. Temos que ter algum critério para decidir. Consideremos o exemplo de testar se há ou não discriminação nos salários. O parâmetro de interesse é a diferença entre média de salários das mulheres e homens. Podemos formular $H_0 : \mu_H - \mu_M = 0$ e $H_1 : \mu_H - \mu_M \neq 0$. Em termos intuitivos se obtivermos uma diferença das médias na amostra muito elevada é natural que rejeitemos a hipótese nula, se obtivermos uma diferença pequena talvez não haja evidência para a hipótese de *não discriminação* ser rejeitada. Mais à frente, veremos que o critério de decisão tem uma base estatística: não é só a diferença das médias na amostra que é importante, também temos de levar em consideração a variabilidade do estimador $\overline{X}_H - \overline{X}_M$.

Antes de avançarmos, há um pormenor de linguagem que gostaria de discutir. Qual é a expressão mais correcta – «aceitar a hipótese nula» ou «não rejeitar hipótese nula»? Não rejeitar a hipótese nula está mais de acordo com o estatuto de H_0 como hipótese mantida.

Como a nossa decisão é baseada só numa amostra, não é possível conhecermos o valor do parâmetro na população, logo não é possível ter a certeza se H_0 é falsa ou verdadeira. Nestas circunstâncias, pode acontecer que a decisão tomada sobre a rejeição ou não da hipótese nula seja errada. Há dois tipos de erros que podem ocorrer: um é rejeitarmos a hipótese nula quando ela é verdadeira, este é chamado *erro do tipo I*. O outro erro que podemos cometer é não rejeitar a hipótese nula quando ela é falsa, este é chamado *erro do tipo II*. Resumindo em tabela:

	H_0 Verdadeira	H_0 Falsa
Não Rejeitar	Decisão correcta $1-\alpha$	Erro tipo II β
Rejeitar	Erro do tipo I α nível de significância	Decisão Correcta $1-\beta$ potência do teste

A Figura 5.1 ilustra graficamente a região de não rejeição e de rejeição para um teste bilateral. A zona em que H_0 é rejeitada é também chamada *região crítica*. Nesta figura está representada a função densidade de probabilidade do estimador se a hipótese nula for verdadeira. A hipótese nula é rejeitada se a estimativa na amostra divergir muito do valor do parâmetro sendo H_0 verdadeira. O erro do tipo I, é a probabilidade de a hipótese nula ser rejeitada quando ela é verdadeira. Por outras palavras, o erro do tipo I é a probabilidade do valor do estimador cair na região critíca, quando H_0 é verdadeiro. Na Figura 5.1 o erro do tipo I é dado pela área a cinzento.

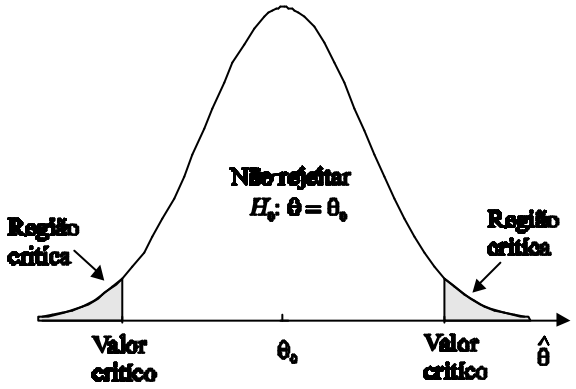

Figura 5.1: Região critíca e região de não rejeição num teste bilateral.

A Figura 5.2 ilustra os conceitos de erro do tipo II e da potência do teste. Na parte superior da Figura é determinada a região crítica para um teste unilateral com um nível de significância α. Se a hipótese alternativa for verdadeira, a distribuição por amostragem da estatística é a apresentada na parte inferior da figura. Logo, a probabilidade de não rejeitar H_0 quando esta hipótese é falsa é dada pela área a cinzento. Ou seja, o erro do tipo II é dado por β. A *potência do teste*, ou seja, a probabilidade de rejeitar a hipótese nula quando ela é falsa, é a área em branco abaixo da função densidade.

A Figura 5.2 pode ser usada para mostrar que existe um tradeoff entre α e β. Se queremos baixar α isso implica que β aumenta. De facto, um α menor implica uma região critíca mais pequena (o valor crítico dimimui no exemplo do gráfico). Mas isso faz aumentar a probabilidade de se cometer um erro do tipo II, faz aumentar β. Tendo em conta este trade-off, uma questão interessante é: «como escolher o valor de α?» O valor óptimo de α depende dos custos associados aos dois tipos de erros. Se o custo de cometermos um erro do tipo I for muito elevado relativamente ao custo do erro do tipo II, é preferível optar por um valor de α muito pequeno.

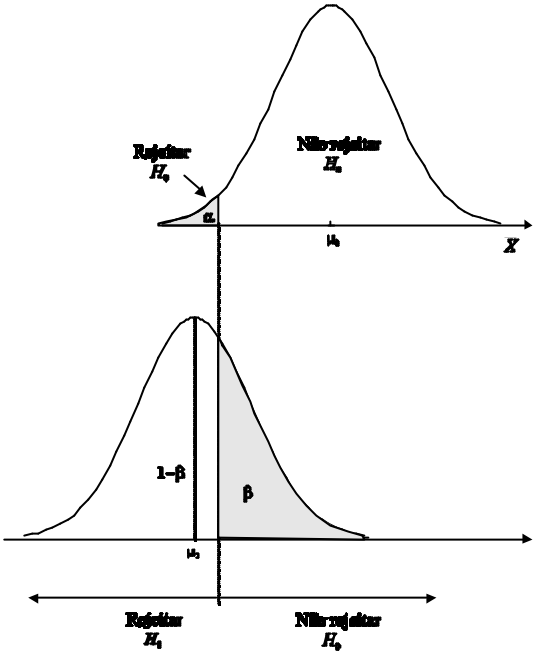

Figura 5.2: Erro do tipo II e potência do teste.

Na discussão anterior sobre o trade-off entre α e β admitimos que a dimensão da amostra é fixa. Contudo, é importante realçar que, se aumentarmos a dimensão da amostra, é possível diminuir simultaneamente α e β.

A *potência do teste* é a probabilidade de rejeitar a hipótese nula quando a hipótese alternativa é verdadeira. A potência do teste depende da valor do parâmetro na hipótese alternativa. A função que relaciona o valor do parâmetro com a potência do teste chama-se *função potência*. Uma função potência bem comportada assume valores mais baixos para valores do parâmetro próximos da H_0 e aumenta à medida de que verdadeiro valor do parâmetro se afasta mais do valor de H_0.

Exemplo 5.1 Eu tenho uma hipótese que é a de que a proporção de pessoas que prefere o Sporting ao Benfica é maior ou igual a 1/2. $H_0 \geq \frac{1}{2}$ e $H_1 < \frac{1}{2}$. Vamos imaginar que eu pergunto a 20 pessoas, escolhidas aleatoriamente qual dos clubes preferem. O critério de decisão é o seguinte: se o número de pessoas que dizem preferir o Sporting for inferior ou igual a 6 rejeito a hipótese nula.

Solução: Assumindo que p é a proporção de pessoas que prefere o Sporting, então o número de pessoas que prefere o Sportingn, Y, numa amostra de 20 pessoas segue a

distribuição binomial $b(20, p)$ Calculemos:

1. *Probabilidade do erro do tipo I – nível de significância:*

$$\alpha = P\left(Y \leq 6; p = \frac{1}{2}\right) = \sum_{y=0}^{6} \binom{20}{y} (1/2)^{20-y}(1/2)^{y} = 0.0577$$

2. *Probabilidade do Erro do tipo II* - depende qual dos valores da alternativa é que consideramos. Se escolhermos $p = \frac{1}{4}$ o valor de β é :

$$\beta = P\left(7 \leq Y \leq 20; p = \frac{1}{4}\right) = \sum_{y=7}^{20} \binom{20}{y} (1/4)^{y}(3/4)^{20-y} = 0.2142$$

enquanto que se $p = \frac{1}{10}$ o valor de β é:

$$\beta = P\left(7 \leq Y \leq 20; p = \frac{1}{10}\right) = \sum_{y=7}^{20} \binom{20}{y} (1/10)^{y}(9/10)^{20-y} = 0.0024$$

3. *Função Potência* - o que queremos aqui é $1 - \beta$ para os diferentes valores de p que constituem a hipótese alternativa

$$K(p) = 1 - \beta(p) = \sum_{y=0}^{6} \binom{20}{y} p^{y}(1-p)^{20-y}$$

$K(1/2) = \alpha = 0.0577$, $K(1/4) = 0.7858$, $K(1/10) = 0.9976$. Verifica-se que quanto mais baixo for o valor de p relativamente à hipótese nula $p = \frac{1}{2}$, maior é a potência do teste.♦

Um outro conceito muito importante é o de *valor de probabilidade do teste* ou *valor p*. O valor p é a probabilidade de obter um valor da estatística tão ou mais extremo do que o resultado obtido, se H_0 for verdadeiro. Na determinação do valor de p é importante ter em conta se o teste é unilateral ou bilateral. A Figura 5.3 ilustra o conceito de valor p num teste bilateral da hipótese $H_0 : \mu = \mu_0$ contra a alternativa $H_1 : \mu \neq \mu_0$. O valor da média na amostra foi de $\overline{x}$. A probabilidade da média na amostra ser superior ou igual a $\overline{x}$ ou inferior ou igual a $-\overline{x}$, quando a média na população é μ_0 é igual a p. O valor p é a área a cinzento na figura. A Figura 5.4 ilustra o conceito de valor no teste unilateral de $H_0 : \mu = \mu_0$ contra a alternativa $H_1 : \mu > \mu_0$.

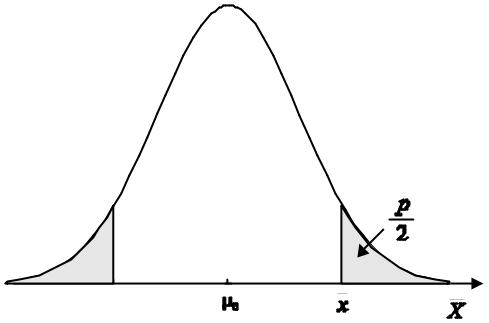

Figura 5.3: O valor p num teste bilateral.

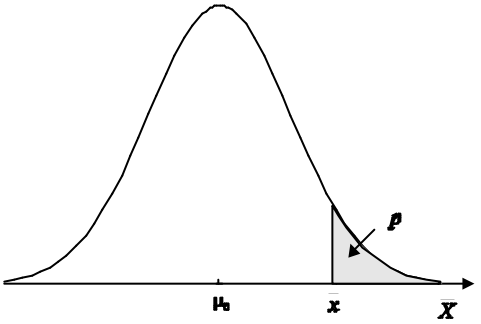

Figura 5.4: O valor p num teste unilateral.

O valor p pode ser usado no teste de hipóteses. De facto, se o valor p for inferior ao nível de significância então devemos rejeitar a hipótese nula. Caso contrário, se o valor p for superior ao nível de significância pretendido, não se deve rejeitar a hipótese nula. Aliás, é frequente definir o valor p como o valor mínimo do nível de significância para o qual H_0 é rejeitado tendo em conta o valor da estatística. Por exemplo, se o valor p é 0.005 isso significa que a hipótese nula é rejeitada mesmo para que o nível de significância seja 0.5%.

5.2 Ensaio de hipóteses sobre a média

5.2.1 População normal, variância conhecida

Na maior parte dos casos a hipótese nula é simples. Queremos testar:

$$H_0 : \mu = \mu_0$$

contra uma das três alternativas

$$\begin{aligned} &i)\ H_1 : \mu \neq \mu_0 \\ &ii)\ H_1 : \mu < \mu_0 \\ &iii)\ H_1 : \mu > \mu_0 \end{aligned}$$

Se retirarmos uma amostra aleatória da população e calcularmos a média na amostra, podemos usar essa média para testar a hipótese. Se a média na amostra divergir *pouco* de μ_0 podemos considerar a evidência em favor de H_0, se a média na amostra divergir *muito* de μ_0 podemos considerar isso evidência contra a H_0. O divergir muito ou pouco deve ser avaliado em termos do desvio padrão da média na amostra (da variabilidade de $\overline{X}$).

Estamos fartos de saber que, se a população for normal ou se n for grande:

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

Se a hipótese nula for verdadeira, isso implica que

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Esta informação é suficiente para podermos determinar a região crítica, para um dado nível de significância, α. A região critíca, em cada um dos casos, é:

1. $H_1 : \mu \neq \mu_0$

   Queremos escolher a região critíca de forma a que a probabilidade de rejeitar a hipótese nula quando ela é verdadeira é α. Como na alternativa o valor do parâmetro pode estar acima ou abaixo de μ_0 isto equivale a escolher o valor $z_{\alpha/2}$ tal que $P\left(Z \leq -z_{\alpha/2} \text{ ou } Z \geq z_{\alpha/2}\right) = \alpha$. Ou seja, a região critíca é dada pelos valores de Z abaixo de $-z_{\alpha/2}$ e pelos valores de Z acima de $z_{\alpha/2}$.

   Dada uma amostra em particular a regra de decisão é: *rejeitar* H_0 *se* $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$ ou se $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > -z_{\alpha/2}$. Isto é equivalente a *rejeitar* H_0 *se* $\overline{x} < \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$ ou se $\overline{x} > \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$. Na Figura5.5 está representada a região critíca para um nível de significância de 5%.

2. $H_1 : \mu < \mu_0$

   Queremos escolher a região critíca de forma a que a probabilidade de rejeitar a hipótese nula quando ela é verdadeira é α. Como na alternativa o valor do parâmetro está abaixo de μ_0 isto equivale a escolher o valor z_α tal que $P\left(Z \leq -z_\alpha\right) = \alpha$. Neste caso H_0 é rejeitado se $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha$, ou equivalentemente, se $\overline{x} < \mu_0 - z_\alpha\sigma/\sqrt{n}$.

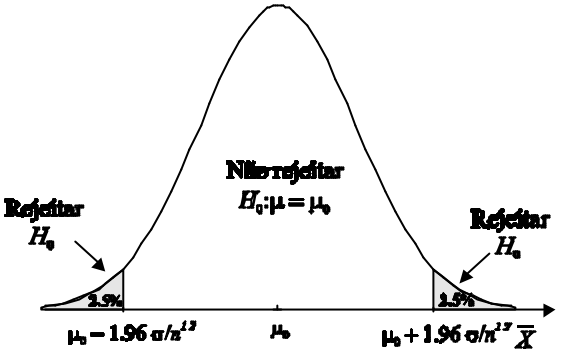

Figura 5.5: Região critíca num teste bilateral de $H_0 : \mu = \mu_0$, com $\alpha = 5\%$.

3. $H_1 : \mu < \mu_0$

   Queremos escolher a região critíca de forma a que a probabilidade de rejeitar a hipótese nula quando ela é verdadeira é α. Como na alternativa o valor do parâmetro está abaixo de μ_0 isto equivale a escolher o valor z_α tal que $P\,(\,Z \geq z_\alpha) = \alpha$. Neste teste H_0 é rejeitado se $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha$, ou equivalentemente, se $\overline{x} > \mu_0 + z_\alpha \sigma/\sqrt{n}$.

Exemplo 5.2 Um produtor de detergentes argumenta que a média do peso das caixas do seu detergente é 500 gramas. Sabe-se que a distribuição do peso é normal, com desvio padrão igual 12.5 gramas. Numa amostra aleatória de 20 caixas o peso médio foi de 485 gramas. Teste o argumento do produtor contra a alternativa que o peso é inferior a 500 gramas, para um nível de significância de 5%.

Solução: O valor critíco z_α tal que $P\,(\,Z \leq -z_\alpha) = 0.05$ é -1.645. Por conseguinte a hipótese nula deve ser rejeitada se $Z < -1.645$. Para a amostra recolhida o valor de z é:

$$z = \frac{485 - 500}{12.5/\sqrt{20}} = -5.37.$$

Logo, a hipótese nula é rejeitada.♦

5.2.2 População normal, variância desconhecida

Se a variância não for conhecida teremos que estimá-la usando a amostra. Neste caso sabemos que se a hipótese nula for verdadeira,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

isto sugere que T seja uma estatística apropriada para usar no teste de $H_0 : \mu = \mu_0$ contra a alternativa $H_1 : \mu \neq \mu_0$. Com $\mu = \mu_0$ sabemos que

$$P\left(T \leq -t_{\alpha/2,n-1} \text{ ou } T \geq t_{\alpha/2,n-1}\right) = \alpha$$

Dada uma amostra específica de dimensão n com média $\overline{x}$ e desvio padrão s a regra de decisão é: rejeitar $H_0 : \mu = \mu_0$ se e só se

$$\frac{\overline{x} - \mu_0}{\sqrt{s^2/n}} \leq -t_{\alpha/2,n-1} \text{ ou } \frac{\overline{x} - \mu_0}{\sqrt{s^2/n}} \geq t_{\alpha/2,n-1}$$

Se a alternativa fosse $H_1 : \mu < \mu_0$ ou $H_1 : \mu > \mu_0$ é fácil por paralelo com o que fizemos atrás construir o teste com nível de significância α.

Exemplo 5.3 Uma empresa produtora de papel tomou várias medidas para reduzir a descarga de poluentes num rio vizinho. Os responsáveis da empresa acreditam ter reduzido o conteúdo de poluentes nas descargas de uma média anterior $\mu = 500$ (avaliando a poluição na àgua em ppm). Para testar se a média baixou a empresa pensa fazer leituras do nível de poluição da àgua em 25 dias consequtivos. Tratando estes 25 valores como uma amostra aleatória, construir um teste com $\alpha = 0.01$. Suponha que depois de feitas as leituras se obteve $\overline{x} = 308.8$ e $s = 115.15$.

O teste mais apropriado é $H_0 : \mu = 500$ contra a alternativa $H_0 : \mu < 500$. Neste caso o teste é só de um dos lados da hipótese nula. A hipótese nula deve ser rejeitada se e só se

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{25}} \leq -t_{0.01,24} = -2.492$$

Para a amostra recolhida o valor da estatística t é

$$t = \frac{308.8 - 500}{115.15/\sqrt{25}} = -8.3$$

logo a hipótese nula é rejeitada e aceitamos a hipótese de que $\mu < 500$.

A resposta anterior não nos diz se o decréscimo na poluição é tanto quanto o desejável. Talvez seja interessante construir o intervalo de 99% nível confiança para μ:

$$308.8 \pm 2.797 \times \frac{115.5}{5} \Rightarrow [244.2, 373.4] \blacklozenge$$

Observações:

- Se a amostra for grande e a variância desconhecida a distribuição normal é apropriada quer a população tenha distribuição normal ou não.

- Se a hipótese nula fosse do tipo $H_0 : \mu \leq \mu_0$ contra a alternativa $H_0 : \mu > \mu_0$ o teste é o mesmo que se $H_0 : \mu = \mu_0$ contra a alternativa $H_0 : \mu > \mu_0$. *Qual é a intuição*? Se a média da população for exactamente μ_0 a probabilidade de rejeitar a hipótese nula com o teste proposto é α. Mas, se a média da população for inferior a μ_0 a probabilidade de a estatística cair na região crítica é ainda menor. Ou seja, a probabilidade do erro do tipo 1 é no máximo α.

5.3 Ensaio sobre a variância de uma população normal

Tal como seria de esperar estes ensaios são baseados na variância da amostra s^2. A base para o teste é o facto da variável aleatória

$$\mathcal{X}^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$$

ter distribuição qui-quadrado com $n-1$ graus de liberdade.

Imaginemos que queremos testar a hipótese de que a variância na população é igual a um certo valor, $H_0 : \sigma^2 = \sigma^2_0$. Se a variância da população for de facto σ^2_0 então a estatística $(n-1)S^2/\sigma^2_0$ tem uma distribuição qui-quadrado. Dada uma amostra em particular com variância s^2, se o valor de s^2 for *muito diferente* de σ^2_0 rejeitamos a hipótese nula. Por exemplo, no caso do teste bilateral com nível de significância α a regra de decisão é: rejeitar a hipótese nula se e só se:

$$\frac{(n-1)s^2}{\sigma^2_0} > \mathcal{X}^2_{n-1,\alpha/2} \quad \text{ou} \quad \frac{(n-1)s^2}{\sigma^2_0} < \mathcal{X}^2_{n-1,1-\alpha/2}$$

$\Rightarrow$ *Questão*: Como seria se a alternativa fosse $H_1 : \sigma^2 > \sigma^2_0$? E se $H_0 : \sigma^2 < \sigma^2_0$?

Exemplo 5.4 Um professor de psicologia argumenta que a variância nos testes de inteligência (I.Q.) para estudantes universitários é de 100. Para testar este argumento resolveu construir-se uma amostra com 30 estudantes que foram submitidos ao teste de inteligência. Nesta amostra $s^2 = 147.82$. Faça o ensaio da hipótese $H_0 : \sigma^2 = 100$ contra a alternativa $H_1 : \sigma^2 \neq 100$, para $\alpha = 0.05$.

Na tabela da Qui-quadrado podemos verificar que $\mathcal{X}^2_{29,0.025} = 45.72$ e que $\mathcal{X}^2_{29,0.975} = 16.05$. O valor da estatística na amostra é

$$\frac{(30-1)\times 147.82}{100} = 42.86$$

Logo com base nesta amostra não é possível rejeitar a hipótese nula de que $\sigma^2 = 100$.

Se construirmos o intervalo de confiança (95%) para a variância obtinhamos

$$\frac{(30-1)\times 147.82}{45.72} \leq \sigma^2 \leq \frac{(30-1)\times 147.82}{16.05} \Rightarrow [93.76, 267]$$

ou seja o intervalo de confiança contém 100, o que é consistente com o resultado do teste.♦

5.4 Ensaio sobre proporções

Muitas vezes estamos interessados em testar hipóteses sobre a proporção de elementos da população que possuem uma certa característica. O teste é baseado na proporção de elementos na amostra que possui a característica e no facto de sabermos que, para n elevado, a variável aleatória

$$Z = \frac{\widehat{P} - p}{\sqrt{p(1-p)/n}}$$

ter uma distribuição que se aproxima da $N(0,1)$.

Seja $H_0 : p = p_0$ a hipótese nula. Se a proporção na população for de facto p_0 sabemos que

$$Z = \frac{\widehat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

tem distribuição normal. Se a hipótese nula estiver a ser testada contra a alternativa $H_0 : p \neq p_0$ e o nível de significância desejado for α, a regra de decisão é: rejeitar a hipótese nula se

$$\frac{|\widehat{p} - p_0|}{\sqrt{p_0(1-p_0)/n}} > z_{\alpha/2}$$

Exemplo 5.5 Numa amostra de 802 compradores, 378 foram capazes de dizer qual era o preço do produto que tinham acabado de colocar no carrinho de compras. Faça um ensaio da hipótese de que pelo menos 50% dos compradores são capazes de dizer correctamente

o preço contra a alternativa de que aquela proporção na população é inferior a 50% com um nível de significância de 10%. Encontre também o valor-p deste teste.

Solução: Queremos testar $H_0 : p \geq 0.5$ contra a alternativa $H_1 : p < 0.5$. A regra de decisão é rejeitar a hipótese nula se

$$\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha = -1.28$$

Mas o valor da estatística na amostra é

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{.471 - .5}{\sqrt{.5 \times .5/802}} = -1.64$$

logo a hipótese nula é rejeitada.

O valor p do teste é a probabilidade de Z ter um valor inferior ou igual ao valor de $z = -1.64$ obtido na amostra, ou seja, $P(Z < -1.64) = P(Z > 1.64) = 0.0505$. Ou seja, a hipótese nula é rejeitada desde que o nível de significância do teste seja superior a 5.05%.♦

5.5 Ensaio sobre igualdade de médias

5.5.1 Variância conhecida com populações normais ou amostra grande

Se tivermos uma amostra de dimensão n_x de uma população com distribuição normal $N(\mu_X, \sigma_X^2)$ e uma amostra de dimensão n_y de uma população com distribuição normal $N(\mu_Y, \sigma_Y^2)$ sabemos que a variável aleatória

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}}$$

tem distribuição $N(0,1)$. Se as variâncias das duas populações forem conhecidas podem fazer-se ensaios de hipóteses baseados neste resultado. Mesmo que as variâncias não sejam conhecidas desde que as amostras sejam grandes é possível substituir a variância na população pela variância na amostra e continuar a usar a distribuição normal (pelo teorema do limite central) e isto é verdade mesmo que a população não seja normal.

Seja $H_0 : \mu_X - \mu_Y = d_0$ a hipótese nula que queremos testar contra $H_1 : \mu_X - \mu_Y \neq d_0$ então a regra de decisão é: rejeitar H_0 se

$$\frac{|(\overline{x} - \overline{y}) - d_0|}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}} > z_{\alpha/2}$$

Nos testes para os casos das alternativas $H_1 : \mu_X - \mu_Y < d_0$ e $H_1 : \mu_X - \mu_Y > d_0$ basta retirar o módulo do numerador e a estatística tem que ser menor que $-z_\alpha$ no primeiro caso e maior que z_α no segundo caso.

Exemplo 5.6 Num inquérito à administração pública pediu-se aos funcionários inquiridos para classificarem numa escala de 1 (discorda completamente) a 5 (concorda plenamente) a afirmação "As mulheres na administração pública são afectadas ao mesmo tipo de tarefas que os homens." Numa amostra de 186 funcionários masculinos a resposta média foi 4.059 e o desvio padrão 0.839. E numa amostra independente de 172 funcionárias públicas a resposta média foi 3.680 e o desvio padrão 0.966. Teste a hipótese de que a percepção média sobre o tratamento das mulheres na função pública é a mesma para funcionários e funcionárias públicas contra a alternativa de que os funcionários têm uma média mais elevada.

Solução: Designando por μ_X a média para os funcionários e μ_Y a média para as funcionárias, queremos testar $H_0 : \mu_X - \mu_Y = 0$ contra $H_0 : \mu_X - \mu_Y > 0$. A regra de decisão é: rejeitar a hipótese nula se

$$\frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} > z_\alpha$$

para as amostras recolhidas o valor da estatística é

$$\frac{4.095 - 3.680}{\sqrt{\frac{(.839)^2}{186} + \frac{(.966)^2}{172}}} = 3.95$$

Mesmo escolhendo um nível de significância muito baixo a hipótese nula é rejeitada. Por exemplo, para $\alpha = 0.0001$ (ou seja .01%) o valor de $z_\alpha = 3.75$ o que significa que a hipótese nula deve ser rejeitada mesmo a este nível de significância.♦

5.5.2 Amostras pequenas

Se as amostras forem pequenas e se for razoável admitir que a variância das duas populações é a mesma podemos usar o facto de a variável aleatória

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S\sqrt{\frac{n_x + n_y}{n_x n_y}}}$$

ter distribuição t com $(n_x + n_y - 2)$ graus de liberdade, e onde S^2 é

$$S^2 = \frac{(n_x - 1)S_X^2 + (n_y - 1)S_Y^2}{n_x + n_y - 2}$$

5.6 Ensaio sobre a igualdade da variância de duas populações normais

Tomemos duas variáveis aleatórias independentes com distribuição normal $N(\mu_X, \sigma_X^2)$ e $N(\mu_Y, \sigma_Y^2)$. Queremos testar a hipótese $H_0 : \sigma_X^2 = \sigma_Y^2$ (o que é equivalente a $H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$). Para testar esta hipótese podemos construir amostras aleatórias independentes de X e Y, calcular a variância em cada uma das amostras. Acontece que, se a hipótese nula for verdadeira, o rácio das variâncias nas amostras tem uma distribuição F com $(n_x - 1), (n_y - 1)$ graus de liberdade.

$$F = \frac{\frac{(n_x-1)S_X^2}{(n_x-1)\sigma_X^2}}{\frac{(n_y-1)S_Y^2}{(n_y-1)\sigma_Y^2}} = \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \Rightarrow \text{ para } H_0 \Rightarrow F = \frac{S_X^2}{S_Y^2}$$

Os valores extremos da região critíca dependem do tipo de teste que desejamos efectuar (se hipótese alternativa tem que estar só para um dos lados da nula ou se o teste é bilateral).

Exemplo 5.7 Um biologista que estuda aranhas está convencido que, numa certa espécie de aranhas, as fêmeas são mais compridas que o macho e que o comprimento nas fêmeas varia mais do que o comprimento nos machos. Assumindo que o comprimento é uma variável aleatória normal e que o comprimentos das fêmeas, X, e machos, Y, são indepentes teste a hipótese de que a variância no comprimento das fêmeas é igual à variância no comprimento dos machos contra a alternativa de que é a variância no comprimento das fêmeas é maior com base em amostras de 30 fêmeas e 30 machos para um nível de significância $\alpha = 0.01$. Os resultados nas amostras foram os seguintes: $\overline{x} = 8.153$, $s_x^2 = 1.410$, $\overline{y} = 5.917$, $s_y^2 = 0.4399$.

Solução: Queremos testar $H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$ contra a alternativa $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$. A estatistica F

$$\frac{s_x^2}{s_y^2} = \frac{1.410}{0.4399} = 3.2053 > F_{0.01}(29, 29) = 2.42$$

Logo a hipotese nula e rejeitada. A evidência suporta o biologista.♦

Capítulo 6

Regressão e correlação simples

QTOmarkbothmyheadings

Neste e nos próximos capítulos estamos interessados em problemas involvendo duas ou mais variáveis. Vamos discutir duas técnicas de análise: correlação e regressão.

A análise de correlação é usada para medir o grau de associação entre variáveis quantitativas. Em contrapartida, a análise de regressão é usada para prever o valor da variável *dependente* ou *explicada*, tendo em conta o valor de uma ou várias *variáveis independentes* ou *explicativas*. Neste capítulo concentramos a nossa atenção em modelos de *regressão linear simples*, onde só há uma variável explicativa e onde se admite a existência de uma relação linear entre a variável explicativa e a variável explicada. Mais tarde exploraremos o caso em que há várias variáveis explicativas - *regressão múltipla*.

6.1 Diagrama de dispersão e correlação

Até aqui fizemos análise de dados de uma variável. Mas pode acontecer estarmos interessados em analisar duas ou mais variáveis numa determinada amostra. Nestas circunstâncias, para além do estudo individual de cada uma das variáveis, podemos ter interesse em estudar eventuais relações entre as variáveis.

A relação a que nos estamos a referir é uma *relação estatística*. Por exemplo, consideremos a relação entre a idade do marido e a idade da mulher. Embora não exista uma relação exacta entre a idade do marido e da mulher, em *termos médios* quanto mais velho é o marido, mais velha é a mulher. As variáveis «idade do marido» e «idade da mulher» são *positivamente correlacionadas*.

O ponto de partida para se estudar a relação entre duas variáveis é termos uma colecção

de observações das duas variáveis:

$$\underbrace{(x_1, y_1)}_{1^a \text{observação}} \ , \ \underbrace{(x_2, y_2)}_{2^a \text{observação}} \ , \cdots , \ \underbrace{(x_n, y_n)}_{\text{n}^a \text{observação}} \ .$$

Se representarmos graficamente os n pontos no plano (num dos eixos temos a variável x, no outro a variável y podemos ficar com uma primeira ideia sobre a forma coma as duas variáveis se relacionam. Essa representação é chamada *diagrama de dispersão*.

Exemplo 6.1 Considere a seguinte amostra de 10 casais:

	Idade marido	Idade mulher
Casal 1	32	30
Casal 2	25	27
Casal 3	50	30
Casal 4	45	40
Casal 5	20	20
Casal 6	35	32
Casal 7	60	55
Casal 8	42	34
Casal 9	27	28
Casal 10	30	28

Construa o respectivo diagrama de dispersão.♦

Se, no diagrama de dispersão, o conjunto de pontos da amostra estiverem mais ou menos agrupados ao longo de uma linha recta, isso sugere que as duas variáveis aleatórias estão *linearmente relacionadas*.

Se conhecermos a distribuição conjunta das duas variáveis e quisermos medir a associação entre as duas variáveis de uma forma númerica podemos calcular a covariância entre x e y

$$\text{Cov}(x, y) = E\left[(x - \mu_x)(y - \mu_y)\right]$$

o problema da covariância é que o seu valor é sensível às unidades de medida de X e Y e, nesse sentido, não mede bem o *grau de associação linear* das duas variáveis. Mas, se dividirmos a covariância pelo desvio padrão de X e desvio padrão de Y obtemos uma medida que não depende das unidades - é o coeficiente de correlação

$$\rho = \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

O coeficiente de correlação toma valores entre -1 e +1. Se $\rho = 1$ significa que há uma associação linear perfeita entre as variáveis x e y e que essas v.a. são positivamente relacionadas. Se $\rho = -1$ as variáveis são negativamente relacionadas sendo a relação linear entre elas perfeita. Se $\rho = 0$ não há relação linear entre as variáveis (elas podem, contudo, ser relacionadas de outras formas). As Figuras 6.1, 6.2 e 6.3 ilustram vários casos.

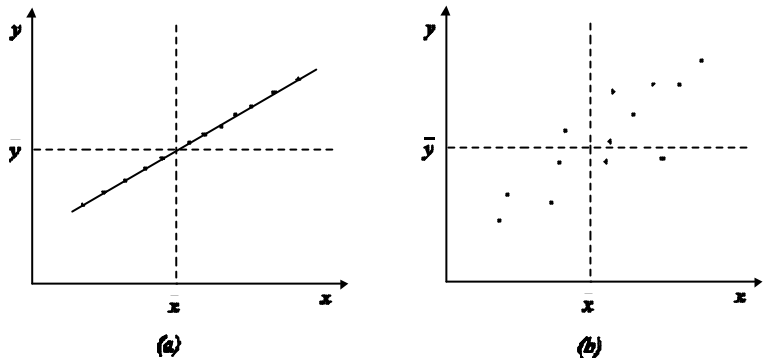

Figura 6.1: Correlação linear positiva: *(a)* $\rho = 1$ e *(b)* $\rho < 1$.

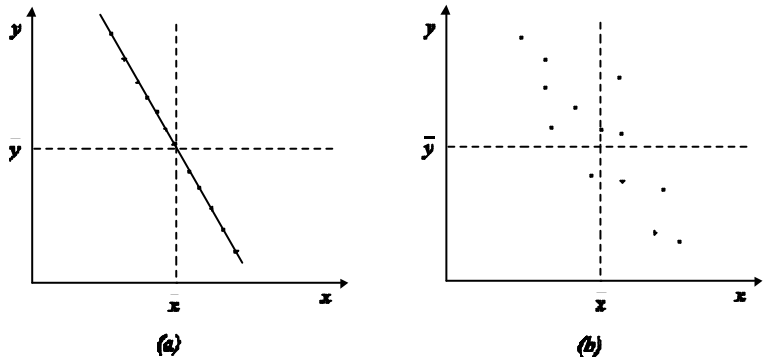

Figura 6.2: Correlação linear negativa: *(a)* $\rho = -1$ e *(b)* $\rho > -1$.

Na prática aquilo de que dispomos é uma amostra. O coeficiente de correlação na amostra pode ser estimado usando

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

que é uma estimador pontual. É possível mostrar que, se a distribuição conjunta das

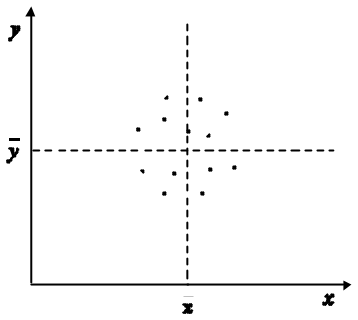

Figura 6.3: Correlação linear nula.

variáveis x e y for normal bivariada, o estimador

$$\frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t_{n-2}.$$

Podemos usar este facto para testar a hipótese nula $H_0 : \rho = 0$.

É importante sublinhar que a existência de correlação entre duas variáveis por si só nada nos diz sobre a existência de uma relação causal entre as variáveis. Ou seja, um valor elevado de r não significa que x seja causa de y ou que y seja causa de x. As variáveis x e y podem ser linearmente correlacionadas por muitas razões:

- Pode existir uma *relação causal unilateral*: a produção de trigo é afectada pela pluviosidade.
- Pode existir *interdependência*: é o o que se passa no exemplo da idade do marido e da idade da mulher.
- Pode existir uma *dependência indirecta* quando as variáveis estão associadas pelo facto de estarem sujeitas à influência de uma causa comum. Exemplo: a forte correlação entre o número anual de casos de insolação e a produção de trigo é explicada pelo facto de verões quentes originarem simultaneamente muitos casos de insolações e boas produções de trigo.

Um outro aspecto importante é que o facto de duas variáveis não serem linearmente correlacionadas nada nos diz sobre a existência de outro tipo de relação. As duas variáveis podem estar relacionadas de forma não-linear. Normalmente, a observação do diagrama de dispersão é muito útil para identificar o tipo de relação que existe entre as variáveis (se existir).

6.1.1 Teste de correlação de Spearman

O coeficiente de correlação da secção anterior é muito sensível à existência de *outliers*. Para além disso, a validade dos teste baseados naquele estimador depende da hipótese da normalidade. É possível obter medidas de correlação menos sensíveis à presença de outliers e que são válidos seja qual for a função distribuição da população.

O teste de correlação de ordem de Spearman é um teste não paramétrico. A ideia base é muito simples: começam por ordenar-se de forma ascente as observações de x e as observações de y. Para cada observação (x_i, y_i) ficamos assim a conhecer a ordem de x_i e a ordem de y_i. A partir daqui podemos calcular o coeficiente de correlação entre as ordem dos x_i e a ordem dos y_i.

6.2 Regressão linear simples

A ideia essencial nesta secção é a de estudar a dependência entre duas variáveis aleatórias, X e Y. Se a v.a X toma um certo valor, qual é o valor que esperamos que Y tome (o valor de X influencia o valor de Y).

Podemos interpretar isto no contexto da distribuição conjunta das variáveis X e Y. Aquilo em que estamos interessados é na distribuição condicionada de Y dado X, $E[Y|X = x]$. Em particular, a pergunta feita anteriormente refere-se ao valor esperado de Y dado X (o valor esperado da distribuição condicionada). Exemplo, X = tempo de estudo, Y = nota.

O objectivo da regressão é modelar a relação referida. Á partida o valor esperado de Y dado X pode assumir qualquer forma funcional (linear, exponencial, log-linear,...). Mas, muitas vezes é razoável admitir que esta relação é linear no intervalo relevante

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

onde o parâmetro β_0 é a intersecção na origem e β_0 é o declive da recta.

Se a dependência linear entre X e Y não for perfeita o valor de Y divergirá do seu valor esperado condicionado. Por outras palavras o modelo da população que estamos a admitir é:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

onde ε_i tem média zero. Uma interpretação do termo ε_i é que ele incorpora a influência de uma variedade de pequenos factores independentes que influenciam Y, para além de X.

Um aspecto muito importante na relação anterior é a interpretação de β_1. O parâmetro β_1 mede a sensibilidade da variável Y a variações em X. Se X aumentar de 1 unidade o valor de Y aumenta β_1 unidades.

Por sua vez, o parâmetro β_0 indica-nos qual é o valor esperado da variável Y quando $X = 0$. Contudo, é de realçar que embora esta interpretação esteja correcta do ponto de vista matemático ela pode não fazer sentido em termos económicos. Em termos económicos pode não fazer sentido o caso em que $X = 0$. Para além disso, a hipótese de que a relação entre Y e X é linear pode verificar-se para um certo intervalo de valores de X, mas não se verificar para valores de X muito afastados daquele intervalo, e em particular não ser válida na vizinhança do ponto $X = 0$.

A Figura 6.4 ilustra o modelo da população que estamos a admitir. O valor da variável dependente, y_i, pode divergir do seu valor esperado tendo em conta x_i. Essa diferença é o termo residual ε_i.

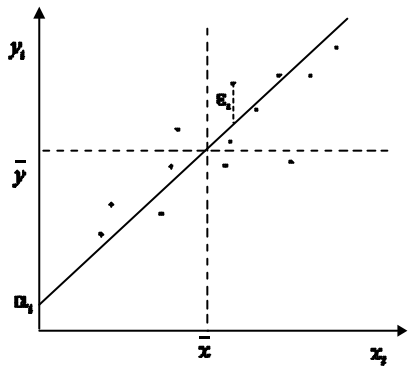

Figura 6.4: Recta de regressão de y sobre x.

O modelo de regressão da população é muito interessante. Contudo, na prática, nunca o poderemos determinar de forma completamente precisa. Na prática, aquilo que fazemos é usar uma amostra para estimar o modelo anterior. A questão que se coloca a seguir é: «como estimar este modelo com base na informação de uma amostra $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$»?. Teremos que estimar os parâmetros β_0 e β_1, para isso podemos utilizar o método dos mínimos quadrados. Se soubermos qual a distribuição conjunta de ε_i (normalmente assume-se normal multivariada) podemos também utilizar o método da máxima verosimilhança.

Exemplo 6.2 Função consumo keynesiana

Na seu livro General theory (1936) Keynes defende que o Consumo depende do rendi-

mento. Ou seja, se designarmos por C o consumo e por Y o rendimento, temos que $C = f(Y)$. Para além disso, Keynes sugere que quando o rendimento aumenta o consumo também aumenta, mas menos que o rendimento. Por outras palavras, a derivada $\frac{dC}{dY}$ é positiva mas inferior a 1 ($\frac{dC}{dY}$ é a propensão marginal ao consumo). A formulação mais usada da função consumo keynesiana é:

$$C = \alpha + \beta Y,$$

onde, de acordo com a teoria, $0 < \beta < 1$.

É claro que o modelo económico $C = \alpha + \beta Y$ é uma abstracção da realidade. Seria irrealista pensar que existe uma relação exacta entre consumo e rendimento. O modelo estatístico leva isto em consideração ao introduzirmos um termo residual não observável. Admitindo que esse termo entra de forma aditiva na relação anterior o modelo estatístico será:

$$C = \alpha + \beta Y + \varepsilon$$

O termo ε é uma variável aleatória não observável que combina o efeito de todos os outros factores que influenciam o consumo e que leva em conta o facto de a relação na realidade não ser exacta.♦

6.2.1 Método dos mínimos quadrados

Designemos por b_0 e b_1 os estimadores de β_0 e β_1. A diferença entre o valor observado da variável explicada e o valor previsto pela recta de regressão para a observação i, ou seja, o erro cometido na observação i é:

$$e_i = \underbrace{y_i}_{\text{valor observado}} - \underbrace{(b_0 + b_1 x_i)}_{\text{valor previsto}}.$$

A soma dos quadrados dos resíduos é

$$SS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

Os estimadores de β_0 e β_1 são os valores de b_0 e b_1 que minimizam a soma dos quadrados dos erros. Ou seja:

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

As condições de primeira ordem deste problema são:

$$\begin{cases} \frac{\partial SS}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i)) = 0 \\ \frac{\partial SS}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))(-x_i) = 0. \end{cases}$$

Estas equações são frequentemente designadas por *equações normais*. Resolvendo o sistema obtemos:

$$\begin{cases} b_0 = \overline{y} - b_1\overline{x} \\ b_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} \end{cases}$$

A primeira equação permite-nos concluir que a recta de regressão passa no ponto médio, $(\overline{x},\overline{y})$. Isto é um facto muito útil, porque facilita imenso o cálculo de b_0 uma vez conhecido o valor de b_1. Para além disso, a segunda equação diz-nos que valor de b_1 é dado pela covariância na amostra entre x e y dividida pela variância de x, o que se pode também exprimir usando o coeficiente de correlação, ou seja:

$$b_1 = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \rho\frac{s_y}{s_x}.$$

Exemplo 6.3 Uma empresa de fast-food está interessada em estudar a influência das despesas de publicidade nas vendas. Na tabela seguinte estão indicadas as variações percentuais, relativamente ao ano anterior, nas despesas de publicidade e nas vendas nas 8 regiões do país onde a empresa opera:

Variação % nas despesas publicidade (x_i)	0	4	14	10	9	8	6	1
Variação % nas vendas (y_i)	24	7.2	10.3	9.1	10.2	4.1	7.6	3.5

Estime a recta de regressão $y_i = \beta_0 + \beta_1 x_i + \varepsilon$. Talvez seja interessante efectuarmos os cálculos para este exemplo.

	x_i	y_i	$x_i y_i$	x_i^2
	0	24	0	0
	4	7.2	28.8	16
	14	10.3	144.2	196
	10	9.1	91	100
	9	10.2	91.8	81
	8	4.1	32.8	64
	6	7.6	45.6	36
	1	3.5	3.5	1
Soma	52	54.4	437.7	494

Logo

$$\begin{aligned} b_1 &= \frac{437.7 - 8 \times \frac{52}{8} \times \frac{54.4}{8}}{494 - 52} = 0.19027 \\ b_0 &= 6.8 - 0.19027 \times 6.5 = 5.5632. \blacklozenge \end{aligned}$$

Exemplo 6.4 Estimação da função consumo keynesiana com dados dos Estados-Unidos para período (1950-1985). Os resultados são:

$$\widehat{C} = \underset{(9.629)}{11.374} + \underset{(0.006)}{0.898} Y$$

onde os valores entre parenteses são os desvios-padrões dos estimadores. Repare-se que a propensão marginal a consumir é 0.898 e, logo, é positiva mas inferior a 1, como a teoria prevê. Se o rendimento aumentar de 1 unidade monetária as despesas de consumo aumentam 0.898 unidades monetárias.$\blacklozenge$

6.2.2 Poder explicativo da regressão

A regressão pode ser vista como uma tentativa de explicar o comportamento da v.a. Y usando informação sobre a v.a. X. Qual é a capacidade do modelo para explicar as variações ocorridas na amostra na variável Y? Se Y tem uma certa variabilidade na amostra que proporção dessa variabilidade pode ser explicada através da dependência linear de Y sobre X?

Podemos decompor a variabilidade total de Y em duas componentes: a variabilidade explicada pela regressão e a variabilidade residual (veja a Figura 6.5). Designemos por $\widehat{y}_i$ o valor previsto da variável y de acordo com a regressão, ou seja, $\widehat{y}_i = b_0 + b_1 x_i$. Tendo

em conta os valores da amostra a regressão estimada pode escrever-se:

$$y_i = b_0 + b_1 x_i + e_i \quad \Leftrightarrow \quad y_i = \widehat{y}_i + e_i.$$

Mas então podemos exprimir o desvio de y_i em relação à média $\overline{y}$ da seguinte forma:

$$y_i - \overline{y} = \widehat{y}_i + e_i - \overline{y} = (\widehat{y}_i - \overline{y}) + e_i.$$

Por palavras, a distância de y_i à média $\overline{y}$ tem duas componentes: a componente explicada e a componente residual.Mas então

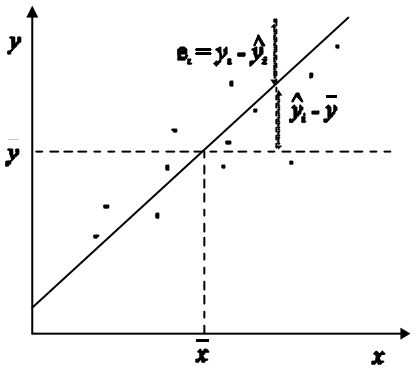

Figura 6.5: Componente explicada, $\widehat{y}_i - \overline{y}$, e componente não explicada, $y_i - \widehat{y}_i$, da diferença de y_i em relação a $\overline{y}$.

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n} 2(\widehat{y}_i - \overline{y})e_i$$

mas o último termo é igual a zero (usando equações normais), logo

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{variação total}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\text{variação explicada}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{variação residual}}$$

Ao rácio

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

dá-se o nome de *coeficiente de determinação*. R^2 diz-nos qual é a proporção da variância total da variável dependente que é explicada pelo modelo linear. É claro que $0 \leq R^2 \leq 1$, e quanto maior for R^2 maior é o poder explicativo da regressão.

Exemplo 6.5 No exemplo da publicidade calcular o valor previsto das vendas, os resíduos em cada um das observações, a varição explicada e não explicada.♦

6.2.3 Hipóteses do OLS e teorema de Gauss-Markov

Se certas condições forem satisfeitas, os estimadores obtidos usando o método dos mínimos quadrados (*ordinary least squares* – OLS) possuem propriedades bastante desejáveis. Nestas secção vamos enunciar as hipóteses tradicionais do modelo de regressão linear simples e enunciar uma consequência dessas hipóteses: o teorema de Gauss-Markov.

Consideremos o modelo da população:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

As hipótese seguintes são normalmente feitas:

1. As observações x_i ou são números fixos (fixados, por exemplo, por um experimentador), ou são realizações de variáveis aleatórias X_i, que são independentes do termo residual ε_i :

$$\operatorname{cov}[X_i, \varepsilon_i] = E[X_i, \varepsilon_i] = 0.$$

2. Os termos residuais ε_i são variáveis aleatórias com média 0:

$$E[\varepsilon_i] = 0, \quad i = 1, 2, \cdots, n$$

3. As variáveis aleatórias ε_i têm todas a mesma variância:

$$\operatorname{var}[\varepsilon_i] = E\left[\varepsilon_i^2\right] = \sigma_\varepsilon^2 \quad i = 1, 2, \cdots, n$$

4. As variáveis aleatórias ε_i não estão correlacionadas umas com as outras:

$$\operatorname{cov}[\varepsilon_i, \varepsilon_j] = E[\varepsilon_i, \varepsilon_j] = 0, \text{ para todo } i \neq j.$$

Se estas condições forem verificadas, e dispusermos de uma amostra com n observações, (x_1, y_1), (x_2, y_2), $\cdots$,(x_n, y_n) os estimadores dos mínimos quadrados, b_0 e b_1, são os estimadores que têm variância mínima na classe de *estimadores lineares e não enviesados*. Este é o teorema de Gauss-Markov. Por esta razão, diz-se que os estimadores dos mínimos quadrados são BLUE (best linear unbiased estimators).

Por outras palavras, os estimadores dos mínimos quadrados são os mais eficientes na classe de estimadores lineares, assumindo que se verificam as hipóteses acima mencionadas.

6.3 Testes de hipóteses e intervalos de confiança

Os estimadores dos mínimos quadrados são estimadores pontuais que são não enviesados e têm variância mínima nas hipóteses do modelo. Contudo, muitas vezes estamos interessados em construir intervalos de confiança para β_0 e β, ou testar hipóteses sobre estes parâmetros da população. Nestes casos, é preciso conhecer a distribuição dos estimadores.

É fácil mostrar que b_0 e b_1 são estimadores não enviesados. Por exemplo,

$$\begin{aligned} E(b_1) &= E\left[\frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^n (x_i-\overline{x})^2}\right] = E\left[\frac{\sum_{i=1}^n (x_i-\overline{x})(\beta_1(x_i-\overline{x})+\varepsilon_i)}{\sum_{i=1}^n (x_i-\overline{x})^2}\right] \\ &= \beta_1 + \frac{\sum_{i=1}^n E\left[(x_i-\overline{x})\varepsilon_i\right]}{\sum_{i=1}^n (x_i-\overline{x})^2} = \beta_1 \end{aligned}$$

Usando as propriedades sobre a variância é também possível mostrar que

$$\text{var}(b) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i-\overline{x})^2}$$

Isto não resolve ainda o problema porque σ_ε^2 é desconhecido. Mas, σ_ε^2 pode ser estimado usando como estimador a variância dos resíduos na amostra

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

onde a divisão por $n-2$ resulta do facto de dois parâmetros terem sido estimados e logo haver a perda de dois graus de liberdade. A s chama-se o *desvio-padrão da estimativa*.

A distribuição de b_0 e a distribuição de b_1 dependem da dimensão da amostra e da função de distribuição dos erros. Se a amostra for grande, a distribuição de b_j é aproxima-se da normal seja qual for a distribuição dos erros (isto é uma consequência do teorema do limite central). Se a amostra for pequena e os erros forem normais a distribuição de

$$\frac{b_j-\beta_j}{s_{b_j}} \quad j=0,1$$

é uma t com $(n-2)$ graus de liberdade.

Conhecendo a distribuição do estimador b_j podemos construir intervalos de confiança para o parâmetro β_j, ou fazer testes de hipóteses.

Se os resíduos tiverem uma distribuição normal e as hipóteses do OLS forem satisfeitas, um intervalo de confiança de $100(1-\alpha)\%$ para β_j é dado por:

$$b_j - s_{b_j} t_{n-2,\alpha/2} < \beta < b_j + s_{b_j} t_{n-2,\alpha/2}$$

onde $t_{n-2,\alpha/2}$ é o valor critíco tal que a probabilidade de uma variável aleatória t_{n-2} seja superior a esse valor é $\frac{\alpha}{2}$.

De forma semelhante, se os resíduos tiverem distribuição normal, podemos fazer testes de hipóteses usando o facto de $\frac{b_j-\beta_j}{s_{b_j}}$ ser uma t com $(n-2)$ graus de liberdade. Para um nível de significância α, para testar a hipótese nula $H_0 : \beta_j = \beta_j^0$ contra a alternativa $H_1 : \beta_j \neq \beta_j^0$, a regra de decisão é rejeitar a hipótese nula se

$$\frac{b_j - \beta_j^0}{s_{b_j}} < -t_{n-2,\alpha/2} \quad \text{ou se} \quad \frac{b_j - \beta_j^0}{s_{b_j}} > t_{n-2,\alpha/2}.$$

Podemos também estar interessados em testes unilaterais. por exemplo, se quisermos testar $H_0 : \beta_j = \beta_j^0$ contra a alternativa $H_1 : \beta > \beta_j^0$, a regra de decisão é rejeitar a hipótese nula se

$$\frac{b_j - \beta_j^0}{s_{b_j}} > t_{n-2,\alpha}.$$

Um caso de interesse particular é quando o valor de $\beta_1^0 = 0$. Neste caso, se a hipótese nula for verdadeira o modelo de regressão da população é:

$$Y_i = \beta_0 + \varepsilon_i$$

Isto significa que, seja qual for o valor da variável independente, a variável dependente é uma variável aleatória de média α e variância σ_ε^2. Por outras palavras, a variável explicada não depende (linearmente) da variável explicativa.

Se a hipótese nula $H_0 : \beta_1 = 0$ for rejeitada dizemos que a variável X é estatisticamente significativa. Caso contrário, se não for possível rejeitar a hipótese nula dizemos que X não é estatisticamente significativa.

Muitos softwares de estatística indicam o valor da estatística t para o teste da hipótese nula $H_0 : \beta_1 = 0$ contra a alternativa $H_1 : \beta_1 \neq 0$, e é normal na apresentação dos resultados de estudos empíricos indicar aquele valor.

6.4 Previsão

Podemos utilizar o modelo de regressão para *prever* o valor da variável explicada, tendo em conta um determinado valor da variável explicativa. Suponhamos que a variável independente é igual a x_{n+1} e que a relação linear estimada continua a ser verificada,

então:

$$Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$$

e

$$E\left[Y_{n+1}|x_{n+1}\right] = \beta_0 + \beta_1 x_{n+1}.$$

É claro que β_0 e β_1 não são conhecidos e também não sabemos qual vai ser o valor de ε_{n+1}. É natural substituir os parâmetros β_0 e β_1 pelas estimativas b_0 e b_1. Por conseguinte, uma *estimativa pontual* de Y_{n+1} é:

$$\widehat{Y}_{n+1} = b_0 + b_1 x_{n+1}.$$

Embora a estimativa pontual seja interessante, em muitos casos estamos interessados em saber qual é o grau de incerteza associado à previsão. Nessas condições devemos construir intervalos de confiança para a variável a prever. Como sempre isso requere o conhecimento da distribuição da variável aleatória. Em particular, o intervalo de confiança dependerá da variância de Y_{n+1}. Em termos intuitivos, há várias fontes de variabilidade. Por um lado, a variável aleatória ε_{n+1} tem uma certa variância, que pode ser estimada usando o desvio-padrão da estimativa. Por outro lado, os estimadores dos mínimos quadrados também têm uma determinada variância.

Se estivermos interessados em construir um intervalo de confiança com um nível de confiança de $100(1-\alpha)\%$ para Y_{n+1} ele é dados por:

$$\widehat{Y}_{n+1} \pm t_{n-2,\alpha/2}\sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] s_e^2}$$

Também se podem construir intervalos de confiança para $E\left[Y_{n+1}|x_{n+1}\right]$, a ideia é estimar o «valor médio» de Y_{n+1} tendo em conta que o valor da variável independente é x_{n+1}. A variância deste valor esperado condicionado é menor que a variância de Y_{n+1} porque aqui a variância de ε_{n+1} não é incluída. Neste caso, o intervalo de confiança é dado por:

$$\widehat{Y}_{n+1} \pm t_{n-2,\alpha/2}\sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] s_e^2}$$

É interessante analisar como é que os diferentes factores afectam o intervalo de confiança. Por um lado, quanto maior for n menor é a variância dos estimadores b_0 e b_1 e, logo, menor é a amplitude do intervalo de confiança.

Para além disso, quanto menor for s_e^2, menor é a amplitude do intervalo de confiança. Isto é bastante intuitivo porque s_e^2 é o estimador de σ_ε^2, e é claro que quanto menor a variabilidade dos resíduos, menor será a variabilidade do valor observado de Y em relação ao seu valor esperado.

Um aspecto interessante é a influência do termo $\sum_{i=1}^{n}(x_i - \overline{x})^2$. Repare-se que isto é um múltiplo da variância da variável explicativa. Quanto maior for a variabilidade na variàvel explicativa, maior é a precisão dos estimadores dos mínimos quadrados (ou seja, menor é a sua variância). Mas isso reduz a amplitude do intervalo de confiança.

Por último, quanto mais x_{n+1} estiver afastado da média $\overline{x}$, maior é a amplitude do intervalo de confiança. Ou seja, a precisão com que conseguimos estimar Y_{n+1} decresce à medida que x_{n+1} toma valores mais afastados da média.

6.5 Outras formas funcionais

Até aqui assumimos que a relação entre a variável explicativa e a variável explicada era linear. Mas, é possível que o modelo teórico de que partimos, ou dados usados, ou ambos, sugiram que a relação é não linear. É curioso que o modelo de regressão linear que acabamos de estudar se pode aplicar a muitas outras formas funcionais. De facto, em muitos casos é possível, usando transformações das variáveis originais, continuar a ter um modelo que é *linear nos parâmetros*. Nestes casos, podemos usar o modelo de regressão linear simples. Vejamos exemplos destas ideias

Exemplo 6.6 Consideremos a seguinte relação entre y e x:

$$y_i = \beta_0 + \beta_1(\frac{1}{x_i}) + \varepsilon_i$$

Esta forma funcional é não linear na variável explicativa. Contudo, o modelo é linear nos parâmetros β_0 e β_1 e, por conseguinte, podemos usar o OLS para o estimar. A única coisa que temos que fazer é começar por calcular $\frac{1}{x_i}$ para todas as observações e, depois, basta regredir y_i sobre a «nova» variável $\frac{1}{x_i}$.

É claro que, se a forma funcional for a descrita e estivermos interessados em calcular quanto é que varia a variável explicada quando a variável explicativa aumenta de 1 unidade, a resposta não é tão imediata como no modelo linear nas variáveis. Mas para responder basta calcular a derivada de y relativamente a x:

$$\frac{dy_i}{dx_i} = -\frac{\beta_1}{x_i^2}.\blacklozenge$$

Exemplo 6.7 Suponhamos que a relação entre y e x é descrita por:

$$y_i = \alpha x_i^{\beta_1} \exp(\varepsilon_i)$$

Apesar deste modelo ser não linear, podemos transformá-lo num modelo linear. Para isso basta calcular o logaritmo de ambos os membros :

$$\ln y_i = \ln \alpha + \beta_1 \ln x_i + \varepsilon_i$$

Este modelo é frequentemente designado por log-linear (existe uma relação linear entre o logaritmo das variáveis). Para estimar o modelo começamos por calcular os logaritmos das variáveis explicada e explicativa para todas as observações e depois fazemos uma regressão linear entre $\ln y$ e $\ln x$.

O parâmetro β_1 neste modelo tem uma interpretação muito curiosa: é a elastecidade de y relativamente a x. Ou seja, se x aumentar de 1% a variável explicada aumenta β_1%. É fácil mostrar este resultado derivando ambos os lados em ordem a x_i :

$$\frac{d(\ln y_i)}{dy_i}\frac{dy_i}{dx_i} = \beta_1 \frac{d(\ln x_i)}{dx_i} \quad \Leftrightarrow \quad \frac{1}{y_i}\frac{dy_i}{dx_i} = \beta_1 \frac{1}{x_i} \quad \Leftrightarrow \quad \frac{dy_i}{dx_i}\frac{x_i}{y_i} = \beta_1 . \blacklozenge$$

Capítulo 7

Regressão múltipla

7.1 Modelo de regressão múltipla

No modelo de regressão simples o comportamento de uma única variável independente foi usado para explicar o comportamento da variável dependente. Contudo, em muitos modelos económicos a variável dependente é influenciada por várias variáveis independentes. Por exemplo, a quantidade produzida é normalmente uma função da quantidade utilizada de vários *inputs*. Outro exemplo, os custos de produção dependem da quantidade produzida, mas dependem também dos preços dos factores produtivos.

Quando passamos de um modelo económico com várias variáveis explicativas para um modelo estatístico linear, obtemos o modelo de regressão múltipla. Tal como no capítulo anterior a ideia é calcular o valor esperado da variável independente condicionado no valor das variáveis explicativas. Se admitirmos que há k variáveis explicativas o modelo de regressão da múltipla na população é:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

onde o índice i diz respeito à observação i.

A interpretação dos parâmetros β_0, $\beta_1, \beta_2, \cdots$, e β_k é semelhante à dos parâmetros no modelo de regressão linear simples.

O parâmetro β_0 indica-nos o valor esperado da variável explicada quando as variáveis explicativas são todas iguais a zero ($x_1 = 0$, $x_2 = 0$, $\cdots$, $x_k = 0$). Embora esta interpretação seja teoricamente correcta, em certos contextos pode não fazer sentido a situação em que todas as variáveis são iguais a zero. Para além disso, quando os valores na amostra das variáveis explicativas são bastante diferentes de zero, pode ser irrealista assumir que o modelo linear é válido na vizinhança do ponto nulo. Ou seja, a regressão linear pode ajustar-se bem na vizinhança dos pontos da amostra, mas pode ser inadequado admitir que o mesmo tipo de relação se observa em regiões afastadas.

O parâmetro β_1 indica-nos a variação esperada na variável explicada quando x_1 aumenta de uma unidade, assumindo que todas as outras variáveis se mantêm constantes. Por outras palavras, β_1 mede a sensibilidade da variável explicada relativamente a variações em x_1.

De forma semelhante, o parâmetro β_i indica-nos a variação esperada na variável explicada quando x_i aumenta de 1 unidade, assumindo que todas as outras variáveis se mantêm constantes. Os parâmetros β_i são frequentemente designados por *coeficientes de regressão parciais*, porque fornecem uma medida da influência de cada uma das variáveis independentes na variável explicativa.

7.1.1 Modelo em notação matricial

Para trabalhar com o modelo de regressão linear múltipla facilita bastante utilizar notação matricial. Tendo em conta o conjunto de n observações, o modelo de regressão é descrito por:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_1 \end{cases}$$

Em termos matriciais estas equações podem escrever-se da seguinte forma:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

Ou seja:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

7.2 Método dos mínimos quadrados

O princípio dos mínimos quadrados aplicado na estimação do modelo de regressão múltipla é em tudo idêntico ao que vimos na regressão simples: tendo em conta a amostra quer

escolher-se os valores dos estimadores de forma a minimizar a soma dos quadrados da diferenças entre os valores observados e os valores previstos da variável explicada.

Admitamos que temos uma amostra de n observações com os valores das k variáveis explicativas e da variável explicada. Ou seja:

$$\begin{gathered}(x_{11}, x_{21}, \cdots, x_{k1}, y_1)\\(x_{12}, x_{22}, \cdots, x_{k2}, y_2)\\\vdots\\(x_{1n}, x_{2n}, \cdots, x_{kn}, y_n)\end{gathered}$$

Dadas estas n observações o problema é encontrar estimadores dos parâmetros β_0, β_1, β_2, $\cdots$, β_k. O método dos mínimos quadrados considera os estimadores b_0, b_1, b_2,$\cdots$, b_k que minimizam a soma dos quadrados dos resíduos:

$$\min_{b_0,b_1,\cdots,b_k} SS = \min_{b_0,b_1,\cdots,b_k} \sum_{i=1}^{n} \left[y_i - \underbrace{(b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki})}_{\widehat{y}_i} \right]^2 .$$

No ponto óptimo deste problema de optimização livre as derivadas parciais de SS em relação a b_0, b_1, b_2,$\cdots$, b_k têm de ser todas iguais a zero. Para encontrar a solução tem de resolver-se um sistema de $k+1$ equações. As contas não são lá muito simpáticas, mas felizmente há softwares que as fazem rapidamente. Contudo, usando notação matricial, a fórmula dos estimadores OLS é muito idêntica à obtida no modelo de regressão simples. De facto, designando por $\mathbf{b}$ o vector de estimadores dos mínimos quadrados, ou seja, $\mathbf{b} = (b_0, b_1, \cdots, b_k)$, pode mostrar-se que:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Exemplo 7.1 Considere o seguinte modelo explicativo do aumento de peso durante o primeiro ano dos «caloiros»:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

onde y – aumento de peso durante o primeiro ano na Universidade, x_1 – número médio de refeições por semana, x_2 – número médio de horas de exercício físico por semana e x_3 – número médio de cervejas consumidas por semana. Este modelo foi estimado usando uma amostra de 30 alunos da Universidade de Évora. As estimativas obtidas na amostra foram:

$$b_0 = 7.35, \quad b_1 = 0.653, \quad b_2 = -1.345 \quad \text{e} \quad b_3 = .613.$$

Será que neste modelo é possível dar uma interpretação adequada à estimativa b_0? Interprete as estimativas dos restantes coeficientes e verifique se o sinal desses coeficientes é aquele que esperaria obter apriori tendo em conta o modelo teórico considerado.♦

7.3 Hipóteses do modelo e teorema de Gauss-Markov

Tal como no modelo de regressão simples, se certas condições forem satisfeitas, os estimadores dos mínimos quadrados tem propriedades muito desejáveis.

Consideremos o modelo da população:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

e admitamos que temos conjunto de dados com n observações. As hipóteses seguintes são normalmente feitas:

1. As observações x_{1i}, x_{2i}, $\cdots$, x_{ki} ou são números fixos (fixados, por exemplo, por um experimentador), ou são realizações de variáveis aleatórias X_{1i}, $X_{2,i}$, $\cdots$, $X_{k,i}$ que são independentes do termo residual.

2. Os termos residuais ε_i são variáveis aleatórias com média 0:

$$E[\varepsilon_i] = 0, \quad i = 1, 2, \cdots, n$$

3. As variáveis aleatórias ε_i têm todas a mesma variância:

$$\text{var}[\varepsilon_i] = E\left[\varepsilon_i^2\right] = \sigma_\varepsilon^2 \quad i = 1, 2, \cdots, n$$

4. As variáveis aleatórias ε_i não estão correlacionadas umas com as outras:

$$\text{cov}[\varepsilon_i \varepsilon_j] = E[\varepsilon_i \varepsilon_j] = 0, \text{ para todo } i \neq j.$$

5. Não é possível encontrar um conjunto de números, c_0, c_1,$\cdots$, c_k tais que

$$c_0 + c_1 x_{1i} + c_2 x_{2i} + \cdots + c_k x_{ki} = 0, \text{ para todo } i = 1, 2, \cdots, n.$$

   Outra forma de dizer isto é que nenhuma das variáveis explicativas se pode exprimir como combinação linear das outras variáveis explicativas.

Se estas condições forem verificadas, e dispusermos de uma amostra com n observações, $(x_{11}, x_{21}, \cdots, x_{k1}, y_1)$, $(x_{12}, x_{22}, \cdots, x_{k2}, y_2)$, $\cdots$, $(x_{1n}, x_{2n}, \cdots, x_{kn}, y_n)$ os estimador dos mínimos quadrados, b_0, b_1, b_2,$\cdots$,b_k são os estimadores que têm variância mínima na classe de *estimadores lineares e não enviesados*. Este é o teorema de Gauss-Markov. Por esta razão, diz-se que os estimadores dos mínimos quadrados são BLUE (best linear unbiased estimators).

Para além das hipóteses mencionadas, é comum admitir que os resíduos, ε_i, seguem uma distribução normal. Esta hipótese é particularmente importante se quisermos fazer teste de hipóteses ou construir intervalos de confiança e a amostra não for muito grande relativamente ao número de parâmetros a estimar. Para amostras de dimensão elevada, a hipótese da normalidade é menos importante por causa do teorema do limite central.

7.4 O poder explicativo da regressão

Qual é a capacidade do modelo de regressão múltipla para explicar as variações ocorridas na amostra na variável Y? Se Y tem uma certa variabilidade na amostra que proporção dessa variabilidade pode ser explicada através da dependência linear entre a variável dependente e as variáveis explicativas?

Tal como no caso da regressão simples, podemos decompor a variabilidade total de Y em duas componentes: a variabilidade explicada pela regressão e a variabilidade residual. Designemos por $\widehat{y}_i$ o valor previsto da variável y de acordo com a regressão, ou seja, $\widehat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$. Tendo em conta os valores da amostra a regressão estimada pode escrever-se:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i \quad \Leftrightarrow \quad y_i = \widehat{y}_i + e_i.$$

Mas então podemos exprimir o desvio de y_i em relação à média $\overline{y}$ da seguinte forma:

$$y_i - \overline{y} = \widehat{y}_i + e_i - \overline{y} = (\widehat{y}_i - \overline{y}) + e_i.$$

Por palavras, a distância de y_i à média $\overline{y}$ tem duas componentes: a componente explicada e a componente residual.Mas então

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{variação total} - SST} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\text{variação explicada} - SSR} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{variação residual} - SSE} + \underbrace{2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})e_i}_{=0 \text{ pelas equações normais}}$$

Ao rácio

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

dá-se o nome de *coeficiente de determinação*. R^2 diz-nos qual é a proporção da variância total da variável dependente que é explicada pelo modelo de regressão múltipla. É claro que $0 \leq R^2 \leq 1$, e quanto maior for R^2 maior é o poder explicativo da regressão.

Se as hipóteses do modelo enunciadas na secção 7.3 forem verificadas, um estimador não enviesado de σ_ε^2 é dado por:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}.$$

Intituitivamente, o facto de se dividir por $n-k-1 = n-(k+1)$ tem a ver com o facto de termos estimado $k+1$ parâmetros da população $(\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$ e, por conseguinte, quando estimamos σ_ε^2 com base nas n observações da amostra já «perdemos» $k+1$ graus de liberdade. Note-se que, para podermos estimar s_e^2 o número de observações tem necessariamente que ser superior ao número de parâmetros a estimar, $n > k+1$. Um exemplo trivial disto era imaginar estimarmos uma regressão simples só com duas observações. Como dois pontos definem de forma única a recta, nesse caso $e_i = 0$ para ambas as observações, ou seja, não há qualquer grau de liberdade nos erros na amostra e, logo, os erros na amostra não podem ser usados para estimar σ_ε^2.

Embora o coeficiente de determinação seja um indicador da capacidade das variáveis explicativas explicarem o comportamento da variável explicada, é importante mencionar que ele tem algumas limitações. Se o número de observações não for grande relativamente ao número de parâmetros a estimar, pode obter-se um R^2 elevado pelo simples facto de haver poucos graus de liberdade na estimação, mesmo que na realidade a relação entre Y e as variáveis explicativas seja fraca. O problema é que o coeficiente de determinação não leva em conta os graus de liberdade!

Uma outra limitação, também relacionada com a questão dos graus de liberdade tem a ver com o que acontece quando aumentamos o número de variáveis explicativas. Se aumentar o número de variáveis explicativas o R^2 aumenta. Contudo, ao aumentarmos o número de variáveis explicativas o número de graus de liberdade diminui. Uma medida que leva em consideração esta perda de graus de liberdade é o *coeficiente de determinação ajustado*, $\overline{R}^2$, definido da seguinte forma:

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

Por último, é importante mencionar que a decomposição da variação total em variação explicada e variação residual *não é válida* se o modelo não incluir o termo constante, β_0. Consequentemente, neste caso é preferível não usar o R^2.

Exemplo 7.2 Considere a regressão estimada do exemplo 7.1. Suponha que a soma dos quadrados dos resíduos e a soma dos quadrados explicada foi:

$$SSE = 45.9 \quad \text{e} \quad SSR = 79.2$$

Determine e interprete o coeficiente de determinação. Encontre o coeficiente de determinação ajustado. Encontre um estimador não enviesado da variância dos resíduos.♦

7.5 Intervalos de confiança e teste de hipóteses de parâmetros individuais

Consideremos o modelo da população:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad \Leftrightarrow \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

e admita-se que se verificam as hipóteses do OLS apresentadas na secção 7.3. Admita-se ainda que os erros têm uma distribuição normal multivariada, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, ou em termos matriciais:

$$\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

Se designarmos por $\mathbf{b} = (b_0, b_1, \cdots, b_k)$, o vector de estimadores dos mínimos quadrados, sabemos que:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Estes estimadores são não enviesados e, nas hipóteses enunciadas seguem uma distribuição normal multivariada:

$$\mathbf{b} \sim \mathbf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

onde $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ deve ser interpretada como a matrix das variâncias e covariâncias dos estimadores. Na diagonal principal da matriz temos a variância de cada um dos estimadores, fora da diagonal principal teremos a covariância entre os vários estimadores.

Na prática, como não conhecemos o valor da variância dos resíduos na população, σ^2, teremos que usar um estimador daquele parâmetro. Já vimos atràs que

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} \quad \Leftrightarrow \quad s_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

é um estimador não enviesado de σ^2.

Por conseguinte, podemos usar o seguinte estimador da matriz das variâncias e covariâncias de $\mathbf{b}$:

$$\hat{\text{cov}}(\mathbf{b}) = s_e^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Designemos por b_0, b_1,$\cdots$, b_k os estimadores dos mínimos quadrados e por s_{b_0}, s_{b_1}, s_{b_2}, $\cdots$, s_{b_k} os respectivos desvios padrões. Nestas circunstâncias, a variável aleatória:

$$\frac{b_j - \beta_j}{s_{b_j}}$$

segue uma distribuição t-student com $n-k-1$ graus de liberdade.

Usando o facto de $\frac{b_j-\beta_j}{s_{b_j}} \sim T_{(n-k-1)}$ podemos construir intervalos de confiança para o parâmetro β_j, ou fazer testes de hipóteses sobre esse parâmetro da forma habitual.

Por exemplo, para testar a hipótese nula $H_0 : \beta_j = \beta_j^0$ contra a alternativa $\beta_j \neq \beta_j^0$ para um nível de significância α, a regra de decisão é:

$$\text{Rejeitar } H_0 \text{ se } \frac{b_j - \beta_j}{s_{b_j}} < -t_{(n-k-1),\alpha/2} \quad \text{ou se } \quad \frac{b_j - \beta_j}{s_{b_j}} > t_{(n-k-1),\alpha/2},$$

onde $t_{(n-k-1),\alpha/2}$ é o valor crítico tal que $P(T_{(n-k-1)} > t_{(n-k-1),\alpha/2}) = \frac{\alpha}{2}$.

Um teste que é muito utilizado é $H_0 : \beta_j = 0$. Repare-se que se a hipótese nula fosse verdadeira isso significaria que a variável x_j não influencia a variável dependente. Se o valor da estatística $\frac{b_j}{s_{b_j}}$ for muito diferente de zero, a hipótese nula será rejeitada. Por outras palavras, os dados da amostra parecem sugerir que x_j é importante para explicar o comportamento da variável dependente. Quando isto acontece, também se diz que a variável x_j é *estatisticamente significativa.*

É importante salientar que o valor da estatística $\frac{b_j}{s_{b_j}}$ depende do valor da estimativa b_j, mas depende também do desvio padrão do estimador dos mínimos quadrados, s_{b_j}. Se o estimador for muito preciso (isto é, se s_{b_j} for muito pequeno) é natural que se rejeite a hipótese nula $H_0 : \beta_j = 0$ mesmo que b_j tenha um valor próximo de zero.

Exemplo 7.3 Uma cadeia de hamburguers está a decidir quanto dinheiro deve gastar em publicidade e se deve ou não dar descontos especiais durante a próxima semana. Para estudar o efeito destas variáveis nas receitas da empresa partiu-se do seguinte modelo económico:

$$R = \beta_0 + \beta_1 p + \beta_2 d$$

onde R representa as receitas durante a semana, p o preço praticado durante a semana e d as despesas de publicidade durante a semana (as receitas e as despesas são medidas em milhares de euros e o preço é medido em euros). O modelo estatístico associado é:

$$R_i = \beta_0 + \beta_1 p_i + \beta_2 d_i + \varepsilon_i,$$

sendo satisfeitas todas as hipóteses do teorema de Gauss-Markov e ainda a hipótese de que os resíduos seguem uma distribuição normal multivariada. Este modelo foi estimado usando as observações das 52 semanas do ano anterior, tendo-se obtido os seguintes resultados:

$$\widehat{R} = \underset{(6.482)}{104.785} - \underset{(3.191)}{6.6419}p + \underset{(0.167)}{2.9843}d \qquad R^2 = 0.862$$

onde os termos entre parênteses são os desvios padrões dos estimadores.

Interprete os resultados obtidos. Teste as hipóteses (i) $H_0 : \beta_1 = 0$ ($H_1 : \beta_1 < 0$) e (ii) $H_0 : \beta_2 = 0$ ($H_1 : \beta_2 > 0$) para $\alpha = 5\%$.♦

7.6 Teste de hipóteses sobre conjuntos de parâmetros

Na secção anterior vimos como é que podemos realizar teste de hipóteses sobre parâmetros individuais. Contudo, pode acontecer que estejamos interessados em testar a hipótese de que, em simultâneo, os parâmetros tomam determinados valores.

7.6.1 Teste de aderência global do modelo

Um caso particular de teste simultâneo é o teste da hipótese nula de que os coeficientes de regressão associados a cada uma das variáveis explicativas são todos iguais a zero, ou seja:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

Se a hipótese nula fosse verdadeira o modelo da população seria:

$$Y_i = \beta_0 + \varepsilon_i,$$

o que significaria que, tomadas como um grupo, as variáveis explicativas não ajudam a explicar o comportamento da variável explicada. Por conseguinte, este teste pode ser visto como um teste à aderência global do modelo que estamos a estimar.

A regra de decisão neste teste é baseada na relação entre a variação explicada pela regressão e a variação residual. Quanto maior for a variação explicada pela regressão relativamente à variação residual, maior é a evidência contra a hipótese nula. Mais concretamente a regra de decisão é baseada na estatística:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

ou seja, são levados em conta os graus de liberdade associados a cada uma das somas dos desvios ao quadrado.

Um resultado importante para podermos efectuar o teste é o facto de F seguir uma distribuição F com k graus de liberdade no numerador e $n-k-1$ graus de liberdade no denominador.

Usando a tabela da $F_{(k,n-k-1)}$ é possível calcular o valor critíco para um nível de significância α. Se o valor da estatística F for superior a esse valor critíco a hipótese nula é rejeitada.

É interessante notar que a estatística F pode ser calculada a partir do coeficiente de determinação:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{SSR}{SSE}\frac{n-k-1}{k} = \frac{SSR}{SST-SSR}\frac{n-k-1}{k} = \frac{R^2}{1-R^2}\frac{n-k-1}{k}.$$

Tal com R^2 a estatística F é um indicador da aderência global do modelo. Mas a estatística F tem a vantagem de nos possibilitar testar a aderência global em termos estatísticos.

Exemplo 7.4 Considere o modelo estimado no exemplo 7.3 das receitas da cadeia de hamburguers. Suponha que nesse modelo se obteve $SSR = 11776.18$, $SSE = 1805.17$ e $SST = 13581.35$. Note-se que esta informação é fornecida pela maioria dos softwares de estatística, sendo apresentada numa tabela designada por Análise de Variância. Calcule o valor da estatística F e teste a hipótese nula de que $\beta_1 = \beta_2 = 0$, para $\alpha = 5\%$.

7.6.2 Teste de um subconjunto de coeficientes de regressão

Suponhamos que o modelo que estamos a estimar tem k variáveis explicativas e que estamos interessados em testar se k_1 ($k_1 < k$) daquelas variáveis são ou não, em conjunto, significativas.

A hipótese nula que queremos testar é:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k_1} = 0.$$

Se a hipótese nula for verdadeira o modelo da regressão é:

$$Y_i = \beta_0 + \beta_{k_1+1} x_{k_1+1,i} + \beta_{k_1+2} x_{k_1+2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i,$$

ou seja só incluí as restante $k-k_1$ variáveis explicativas e o termo constante. É importante notar que, se estimarmos este modelo, os estimadores obtidos para os $k+1-k_1$ coeficientes serão diferentes dos estimadores obtidos quando se incluem na regressão todas as variáveis explicativas. Designemos por SSE^* a soma dos quadrados dos resíduos da regressão que incluí só as últimas $k - k_1$ variáveis explicativas e por SSE a soma dos quadrados dos resíduos da regressão que incluí todas as variáveis explicativas.

A ideia do teste, é que se a hipótese nula é verdadeira SSE^* e SSE devem divergir pouco (mas SSE será sempre inferior ou igual a SSE^*). Em concreto, a regra de decisão é baseada na estatística:

$$F = \frac{\frac{(SSE^* - SSE)}{k_1}}{\frac{SSE}{n-k-1}} \sim F_{k_1, n-k-1}.$$

Se designarmos por $F_{(k_1,n-k-1),\alpha}$ o valor critíco para um nível de significância α, a hipótese nula é rejeitada se:

$$\frac{\frac{(SSE^* - SSE)}{k_1}}{\frac{SSE}{n-k-1}} > F_{(k_1,n-k-1),\alpha}$$

7.6.3 Teste de uma combinação linear de parâmetros

Por vezes é útil testar se os coeficientes de regressão satisfazem uma determinada restrição linear. Suponhamos que a hipótese nula é a seguinte:

$$H_0 : c_0\beta_0 + c_1\beta_1 + \cdots + c_k\beta_k = r \quad \Leftrightarrow \quad H_0 : \mathbf{c}'\boldsymbol{\beta} = r$$

Em geral, alguns dos coeficientes c_i serão iguais a zero. Tendo em conta a amostra, a estimativa de $\mathbf{c}'\boldsymbol{\beta}$ é $\mathbf{c}'\mathbf{b}$:

$$\mathbf{c}'\mathbf{b} = c_0 b_0 + c_1 b_1 + \cdots + c_k b_k = \widehat{r}$$

Se $\mathbf{c}'\mathbf{b}$ estiver próximo de r, a evidência é consistente com a hipótese nula. O teste estatístico é baseado na estatística

$$\frac{\mathbf{c}'\mathbf{b} - r}{\sqrt{\text{var}(\mathbf{c}'\mathbf{b})}} = \frac{\mathbf{c}'\mathbf{b} - r}{\sqrt{\mathbf{c}' \left[s_e^2 (\mathbf{X}'\mathbf{X})^{-1} \right] \mathbf{c}}}$$

que segue uma distribuição T de student com $n - k - 1$ graus de liberdade.

Exemplo 7.5 Consideremos a seguinte função de produção:

$$Y = AK^{\alpha} L^{\beta}$$

onde Y é o output produzido, K é a quantidade de capital utilizada e L é a quantidade de trabalho utilizada. A soma $\alpha + \beta$ indica-nos se a função de produção apresenta rendimentos constantes à escala ($\alpha + \beta = 1$), crescentes à escala ($\alpha + \beta > 1$) ou decrescentes à escala ($\alpha + \beta < 1$).

Suponhamos que o modelo estatístico associado é:

$$Y_i = Y = AK_i^{\alpha} L_i^{\beta} e^{\varepsilon_i}$$

Este modelo pode ser transformado num modelo linear nos parâmetros tomando o logaritmo de ambos os termos:

$$\ln Y_i = \underbrace{\ln A}_{\gamma} + \alpha \ln K + \beta \ln L + \varepsilon_i.$$

Para testar se a tecnologia apresenta rendimentos constantes à escala podemos fazer o teste da hipótese nula:

$$H_0 : \alpha + \beta = 1 \;\Leftrightarrow\; \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \gamma \\ \alpha \\ \beta \end{bmatrix} = 1. \blacklozenge$$

7.6.4 Teste de várias combinações lineares de parâmetros

Podemos generalizar os resultados da secção anterior para o caso em que estamos interessados em testar simultaneamente j restrições lineares sobre os parâmetros. Ou seja a hipótese nula é:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{r}$$

onde $\mathbf{C}$ é uma matriz de dimensão $j \times (k+1)$, em que cada linha se refere a uma restrição. O teste vai ser baseado na diferença $\mathbf{Cb} - \mathbf{r}$ (repare-se que isto corresponde a um vector de variáveis aleatórias). O teste é baseado na estatística F :

$$F = \frac{(\mathbf{Cb} - \mathbf{r})' \left[s_e^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C} \right]^{-1} (\mathbf{Cb} - \mathbf{r})}{j}$$

que segue uma F com j graus de liberdade no numerador e $n - k - 1$ graus de liberdade no denominador.

7.7 Previsão

Uma aplicação importante do modelo de regressão múltipla é a *previsão* do valor da variável dependente, tendo em conta que as variáveis explicativas tomam determinados. Suponhamos que os valores das k variáveis explicativas são iguais a $x_{1,n+1}$, $x_{2,n+1}$, $\cdots$, $x_{k,n+1}$ e que o modelo de regressão múltipla continua a verificar-se, ou seja:

$$Y_{n+1} = \beta_0 + \beta_1 x_{1,n+1} + \beta_2 x_{2,n+1} + \cdots + \beta_k x_{2,n+1} + \varepsilon_{n+1}, \quad \text{onde } E(\varepsilon_{n+1}) = 0.$$

Se usarmos os estimadores dos mínimos quadrados dos coeficientes da regressão, obtemos a seguinte estimativa pontual de Y_{n+1}:

$$\widehat{Y}_{n+1} = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_k x_{k,n+1} = \mathbf{x}'_{n+1}\mathbf{b}$$

Tendo em conta o teorema de Gauss-Markov, sabemos que este é o previsor mais eficiente de Y_{n+1} na classe de estimadores lineares e não enviesados.

Se, em vez de um estimador pontual, estivermos interessados em obter intervalos de confiança para a variável dependente, necessitamos de estimar a variância do erro de previsão:

$$\text{var}\left[\widehat{Y}_{n+1} - Y_{n+1}\right] = \sigma^2 \left[1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1}\right].$$

Substituindo σ^2 pelo estimador s_e^2 ficamos com estimador da variância do erro de previsão, e a partir daqui podemos construir intervalos de confiança para $\widehat{Y}_{n+1}$.

Capítulo 8

Tópicos de econometria

O primeiro passo num estudo econométrico é especificar um modelo algébrico que descreva de forma relativamente correcta o sistema que estamos interessados em estudar.

A realidade é extremamente complexa. É claro que é impossível descrever com absoluta precisão o comportamento de variáveis económicas. Quando se constroí um modelo económico a ideia é captar os factores mais importantes para explicar as variáveis de interesse, reconhecendo que é impossível levar em conta todos os factores que influenciam essas variáveis.

É importante salientar que quando se passa do modelo económico para o modelo estatístico se devem especificar as hipóteses feitas sobre os resíduos, uma vez que as propriedades dos estimadores obtidos e qualquer inferência estatística feita com base no modelo dependem daquelas hipóteses serem verificadas ou não.

Uma vez especificado o modelo estatístico, o passo seguinte é usar os dados para estimar os parâmetros do modelo. O método de estimação apropriado para estimar os coeficientes do modelo depende das propriedades estatísticas dos resíduos.

Depois da estimação do modelo deve verificar-se se os resultados obtidos estão de acordo ou não com a teoria económica. Por exemplo, na estimação de uma função procura espera-se que a quantidade procurada dependa negativamente do preço. Se isso não acontecer pode ser porque a especificação do modelo não fosse a mais correcta. Talvez se tenha omitido alguma variável explicativa importante, talvez a forma funcional assumida não seja a mais correcta,...

Para além da verificação dos sinais dos coeficientes, devem também verificar-se as hipóteses feitas sobre as propriedades dos resíduos. Será que a evidência empírica não contradiz a hipótese de que os resíduos têm todos igual variância? Será que a hipótese de que os resíduos são não correlacionados não é «suportada» pelos dados?

Neste capítulo estudamos como testar a validade das hipóteses do modelo de regressão clássica, as consequências dessas violações e como proceder nessas circunstâncias. Para

além disso, estudamos ainda outros tópicos importante de econometria: as variáveis dummy, problemas de especificação e minímos quadrados não lineares.

8.1 Multicolinearidade

Suponhamos que estamos interessados em estimar um regressão múltipla com duas variáveis explicativas:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$

Por outras palavras, queremos «isolar» o impacto de que cada uma das variáveis explicativas na variável dependente. Queremos estimar β_1 e β_2.

Para que seja possível estimar o contributo de cada uma das variáveis explicativas, é necessário que, nos dados de que dispomos, as duas variáveis explicativas não estejam perfeitamente correlacionadas. Imaginemos, por exemplo, que nos dados se verificava a seguinte relação entre x_1 e x_2: $x_{2i} = 2x_{1i}$. Ou seja, sempre que x_1 aumenta de uma unidade, x_2 aumenta de duas unidades. Nestas circunstâncias, é impossível «isolar» o efeito de cada uma das variáveis explicativas. Será possível dizer o que acontece ao valor esperado de Y se x_1 aumentar de 1 unidade e x_2 aumentar de 2 unidades, mas é impossível saber o que acontece a Y se x_1 aumentar, mantendo x_2 constante. Os dados não possuem variabilidade suficiente para responder a esta pergunta.

O problema que acabamos de descrever é o problema de existência de multicolinearidade perfeita: existe uma relação linear perfeita entre as variáveis explicativas. Se este problema existir, não conseguiremos obter os estimadores dos mínimos quadrados. A intuição para isto é o que vimos no exemplo anterior. Em termos formais o problema é que, se existir dependência linear entre as variáveis explicativas, a matriz $X'X$ é uma matriz singular e logo não é invertível. Por isso, é impossível calcular $(X'X)^{-1}X'Y$.

É importante salientar que no modelo geral de regressão múltipla, o problema da multicolinearidade não se coloca necessariamente por haver dependência linear entre duas variáveis explicativas. No caso geral, o problema surge sempre que uma das variáveis explicativas se possa escrever como combinação linear das restantes variáveis explicativas.

A multicolinearidade perfeita é muito fácil de identificar (quando mandarmos estimar os parâmetros, seja qual for o software usado, não é possível obter os estimadores).

Como resolver um problema de multicolinearidade perfeita? Retomemos o exemplo anterior. Já sabemos que é pura e simplesmente impossível isolar o efeito de x_1 e x_2. Mas

sabemos que $x_2 = 2x_1$, logo o nosso modelo pode ser reescrito da seguinte forma:

$$Y_i = \beta_0 + \beta_1 x_{1i} + 2\beta_2 x_{1i} + \varepsilon_i = \beta_0 + \underbrace{(\beta_1 + 2\beta_2)}_{\beta_1'} x_{1i} + \varepsilon_i$$

E agora é possível estimar os coeficientes β_0 e β_1'. Repare-se que β_1' capta o efeito de uma variação conjunta em x_1 e em x_2.

O caso de multicolinearidade perfeita é bastante extremo. Um caso menos extremo é quando existe uma relação linear muito forte, mas não perfeita, entre as variáveis explicativas. Neste caso, é possível obter os estimadores dos mínimos quadrados. Os dados fornecem *alguma (mas pouca)* informação sobre a influência de cada uma das variáveis explicativas. O problema é que os estimadores de β_1 e β_2 vão ser muito pouco precisos. Por outras palavras, o desvio padrão dos estimadores, s_{b_1} e s_{b_2} é muito elevado. Isto leva a que, frequentemente, as variáveis sejam classificadas como estatisticamente não significativas, quando na realidade elas são importantes para explicar o comportamento da variável explicada.

Uma indicação clara da presença de multicolinearidade ocorre quando, no seu conjunto, um grupo de variáveis explicativas é importante para explicar o comportamento da variável explicada, mas depois quando testamos cada uma das variáveis separadamente elas parecem não ser estatisticamente significativas. Nestas circunstâncias, há que ter cuidado em não se concluir que as variáveis não são significativas. É melhor concluir que, no seu conjunto aquele grupo de variáveis é importante, mas os dados não suficientemente informativos para nos permitirem «isolar» com precisão o efeito separado de cada uma das variáveis.

8.2 Variáveis dummy

Muitas vezes acontece que a variável explicada é influenciada por factores que não são quantitativos. Por exemplo, a venda de gelados depende da estação do ano, a procura de turismo pode ser influenciada por uma greve dos pilotos, o consumo agregado pode ser diferente num período de guerra dos restantes períodos. Como incorporar estes factores qualitativos no modelo de regressão?

Uma forma de incorporar os factores qualitativos é usar *variáveis dummy* (também chamadas variáveis dicotómicas e variáveis binárias). Uma variável dummy é uma variável que só pode tomar dois valores: 0 ou 1, para indicar a presença ou ausência da característica

relevante. Ou seja:

$$D = \begin{cases} 1 & \text{se a característica está presente.} \\ 0 & \text{se a característica não está presente.} \end{cases}$$

Uma questão importante é como é que a variável dummy afecta a relação que estamos a estudar. Será que afecta só a intersecção na origem (ou seja, o coeficiente β_0 é diferente para observações em que a característica está presente). Ou será que a dummy afecta a forma como alguma das variáveis explicativas influencia a variável dependente?

8.2.1 Alteração na intersecção na origem

Vamos começar por ver um exemplo em que a dummy só afecta a intersecção na origem. Consideremos a função consumo keynesiana:

$$C_i = \beta_0 + \beta_1 Y_i + \varepsilon_i, \quad \text{com } i = 1930, \cdots, 1980.$$

É natural admitir que durante o período da segunda guerra mundial o nível de consumo autonómo não foi igual ao dos restantes anos. Ou seja, o termo β_0 varia consoante o ano em causa seja um ano de guerra ou não. Podemos definir a variável dummy:

$$D_i = \begin{cases} 1 & \text{se } i = 1939, \cdots, 1945 \\ 0 & \text{caso contrário.} \end{cases}$$

O modelo de regressão que queremos estimar é:

$$C_i = \beta_0 + \delta D_i + \beta_1 Y_i + \varepsilon_i$$

Repare-se que isto é equivalente a dizer que o termo constante é igual a β_0 se $D_t = 0$ (ou seja, em anos que não são de guerra o consumo autonomo é dado por β_0). Em contrapartida, se $D = 1$ o termo constante é igual a $\beta_0 + \delta$ (ou seja, em anos de guerra o consumo autónomo é dado por $\beta_0 + \delta$). Por conseguinte, o coeficiente δ associado à variável dummy, indica-nos o deslocamento na intersecção na origem em anos de guerra.

O modelo anterior é estimado da forma habitual e podemos construir intervalos de confiança para o parâmetro δ, ou fazer teste de hipóteses sobre aquele parâmetro. Um teste com particular interesse é $H_0 : \delta = 0$. Se a hipótese nula for rejeitada a evidência empírica parece sugerir que a presença da característica afecta de facto a intersecção na origem.

Exemplo 8.1 No modelo anterior, estimado com dados para os Estados Unidos entre 1929 e 1970, obtiveram-se os seguintes resultados:

$$\widehat{C}_i = \underset{(3.98)}{101.36} - \underset{(-10.91)}{204.95}\, D_i + \underset{(58.73)}{0.86}\, Y_i$$

onde os valores entre parenteses são as estatísticas t.

Repare-se que o coeficiente associado à dummy é negativo e estatisticamente significativo (se testarmos $H_0 : \delta = 0$ contra a alternativa $H_1 : \delta \neq 0$ a hipótese nula é rejeitada mesmo a um nível de signifiucância de 1%. Isto sugere que durante os anos de guerra o consumo diminuiu consideravelmente.♦

8.2.2 Alteração do declive

Até aqui assumimos que o factor qualitativo só influência a intersecção na origem, mas pode acontecer que a presença ou ausência da característica afecte a forma como as outras variáveis explicativas influenciam a variável dependente. Por outras palavras, o coeficiente β_i pode ser diferente consoante a característica esteja presente ou não.

Consideremos o exemplo anterior da função consumo. É possível que durante os anos de guerra a propensão marginal a consumir seja diferente da propensão marginal a consumir durante os restantes anos. O modelo a estimar seria então:

$$C_i = \beta_0 + \beta_1 Y_i + \gamma D_i Y_i + \varepsilon_i \quad \Leftrightarrow \quad C_i = \beta_0 + (\beta_1 + \gamma D_i)\, Y_i + \varepsilon_i.$$

Isto significa que em anos «normais» o declive da recta de regressão é igual a β_1, enquanto que em anos de guerra o declive da recta de regressão é igual a $\beta_1 + \gamma$. Por conseguinte, o parâmetro γ mede a diferença na propensão marginal a consumir entre anos de guerra e anos normais. À partida esperamos que o coeficiente γ seja negativo, ou seja, espera-se que a propensão marginal a consumir seja mais baixa durante os anos de guerra.

É claro que podemos assumir que a guerra afecta simultaneamente o consumo autónomo e a propensão marginal a consumir. Neste caso, o modelo a estimar é:

$$C_i = \beta_0 + \delta D_i + \beta_1 Y_i + \gamma D_i Y_i + \varepsilon_i$$

8.2.3 Variáveis qualitativas com mais de duas classes

Muitas variáveis qualitativas têm mais do que duas classes. Por exemplo, se considerarmos a variável estado civil podemos ter quatro estados: solteiro, casado, divorciado, viúvo. É

possível utilizar variáveis dummy mesmo nestes casos. A ideia é definir variáveis dummy para cada um dos estados. Por exemplo:

$$D_1 = \begin{cases} 1 \text{ se o indíviduo é solteiro} \\ 0 \quad \text{se o indivíduo não é solteiro} \end{cases}, \quad D_2 = \begin{cases} 1 \text{ se o indíviduo é casado} \\ 0 \quad \text{se o indivíduo não é casado} \end{cases}$$

$$D_3 = \begin{cases} 1 \text{ se o indíviduo é divorciado} \\ 0 \quad \text{se o indivíduo não é divorciado} \end{cases}, \quad D_4 = \begin{cases} 1 \text{ se o indíviduo é viúvo} \\ 0 \quad \text{se o indivíduo não é viúvo} \end{cases}$$

Um cuidado a ter quando pretendemos incluir variáveis dummy referentes a um variável qualitativa com mais de dois estados é que, se houver j estados só devemos incluir $j-1$ variáveis dummy na regressão. A intuição para isto é muito simples, a soma da j dummies é necessariamente igual a 1 (cada indíviduo pertence a uma e só a uma classe). Isto implica que se conhecermos $j-1$ das variáveis, sabemos automaticamente o valor da $j-$ésima variável (esta variável depende linearmente das outras $j-1$). Se incluíssemos as j variáveis teríamos um problema de multicolinearidade perfeita.

8.3 Heterocedasticidade

No modelo clássico de regressão linear admitimos que a variância do termo residual é a mesma para todas as observações e que o termo residual das diferentes observações não eram correlacionados. Nesta e na próxima secção vamos discutir o que acontece se estas hipóteses não forem verificadas. Nesta secção veremos o que acontece se a variância do termo residual não for a mesma para todas as observações.

Vamos supor, por exemplo, que estamos interessados em estimar como é que a produção num dado sector se relaciona com as quantidades de factores produtivos utilizadas e vamos admitir que possuímos dados para várias empresas no sector (neste exemplo cada empresa é uma observação). É natural que a variabilidade do termo residual seja diferente para empresas de diferente dimensão. Se isso acontecer, a hipótese da homoscedasticidade dos resíduos (igual variância) não é verificada. Neste caso dizemos que o modelo apresenta *heterocedasticidade*.

Como é que se pode detectar a presença de heterocedasticidade? Se tivermos uma ideia da forma como a variância dos resíduos varia podemos graficamente tentar detectar a presença de heterocedasticidade. Começamos por estimar o modelo e calcular os resíduos observados:

$$e_i = y_i - \widehat{y}_i.$$

Depois podemos fazer um gráfico relacionando a variável que nós pensamos que influencia a variância com os resíduos estimados. No exemplo anteriormente descrito, a dimensão da empresa pode ser medida, por exemplo, pela quantidade de output produzida, y_i, (que neste caso é a variável dependente). Fazendo um gráfico relacionando y_i com e_i podemos conseguir detectar a presença de heterocedasticidade. Se a distribuição dos resíduos em torno do zero (valor esperado dos resíduos) for muito mais dispersa para valores elevados de y_i do que para valores baixos de y_i, isso sugere a presença de heterocedasticidade.

8.3.1 Teste de heterocedasticidade de Breusch-Pagan

Há muitos procedimentos para testar a presença de heterocedasticidade, tais como o teste de Goldfeld-Quandt, o teste de White e o teste de Breusch-Pagan. Por falta de tempo, aqui apresentamos apenas o teste de Breusch-Pagan, porque é um teste que sugere a forma de corrigir o problema. Suponhamos que a variância dos resíduos está relacionada com as variáveis $z_1, z_2, \cdots, z_m$ da seguinte forma:

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_m z_{mi}, \quad \text{com } i = 1, \cdots, n.$$

As variáveis z são quaisquer variáveis observáveis dos quais pensamos que a variância possa depender. Elas podem incluir variáveis explicativas do modelo de regressão que estamos a estimar, podem incluir a variável dependente desse modelo, ou quaisquer outras variáveis que pensemos serem relevantes.

O teste de Breusch–Pagan é um teste da hipótese nula $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$ contra a alternativa de que pelo menos um destes coeficientes é diferente de zero. Note-se que se H_0 for verdadeira os resíduos são homocedásticos, uma vez que teríamos $\sigma_i^2 = \alpha_0$ para todo o i.

Para efectuar o teste de Breusch–Pagan temos que fazer o seguinte:

1. Estimar o modelo original e obter os resíduos dos mínimos quadrados

$$e_i = y_i - (b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + \beta_k x_{k,i})$$

2. Calcular o quadrado dos resíduos e estimar a seguinte *regressão auxiliar*:

$$e_i^2 = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_m z_{mi} + v_i$$

   o termo residual v_i resulta do facto de os resíduos estimados divergirem dos verdadeiros resíduos.

3. Calcular a SSR da regressão auxiliar e $\widetilde{s}^2 = \sum e_i^2/n$.

4. Calcular a estatística de Breusch-Pagan que é dada por:

$$BP = \frac{SSR}{2\widetilde{s}^4}$$

Quando a hipótese nula é verdadeira e a amostra é grande, BP tem uma distribuição aproximadamente $\mathcal{X}^2$ com m graus de liberdade, onde m é o número de variáveis explicativas da regressão auxiliar.

8.3.2 Implicações da presença de heterocedasticidade

O que acontece se usarmos os mínimos quadrados ordinários para estimar um modelo em que os resíduos não têm variância constante? Há duas consequências importantes. A primeira é que na presença de heterocedasticidade os estimadores dos mínimos quadrados continuam a ser não enviesados, mas deixam de ter variância mínima na classe de estimadores lineares e não enviesados. Por outras palavras, os estimadores deixam de ser eficientes naquela classe. A precisão com que o OLS estima os coeficiente não é a maior possível.

A segunda consequência, e talvez a mais importante, é que o desvio-padrão dos estimadores calculado pelo OLS não é correcto porque assume que a matriz de variâncias e covariâncias dos resíduos é $\sigma^2 I$, quando isso não é verdade. Por conseguinte, se fizermos inferência estatística usando os resultados do OLS podemos retirar conclusões erradas.

Se conhecermos a estrutura da variância dos resíduos podemos transformar o modelo original de tal forma que o modelo transformado é um modelo homocedástico e estimar, usando o OLS, o modelo transformado. Uma interpretação alternativa deste procedimento é a de que na soma dos quadrados dos resíduos, não damos igual ponderação a todas as observações. As observações cujos resíduos tem maior variância são «menos ponderadas». A ideia é encontrar os estimadores dos *mínimos quadrados ponderados* (weighted least squares).

Vamos ilustrar a técnica de transformar o modelo original com um exemplo simples. Suponhamos que o modelo original é dado por:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

onde $\varepsilon_i \sim (0, x_i\sigma^2)$ e $E[\varepsilon_i \varepsilon_j] = 0$ para $i \neq j$ e a variável explicativa é não estocástica.

Podemos verificar de imediato que se dividirmos o termo residual por $\sqrt{x_i}$ a variável residual resultante tem variância constante. Mas então se dividirmos ambos os lados da

equação original por $\sqrt{x_i}$ obteremos um *modelo transformado* em que os resíduos são homocedásticos:

$$\frac{y_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1 \frac{x_i}{\sqrt{x_i}} + \frac{\varepsilon_i}{\sqrt{x_i}}$$

De facto:

$$\text{var}\left[\frac{\varepsilon_i}{\sqrt{x_i}}\right] = \left(\frac{1}{\sqrt{x_i}}\right)^2 \text{var}\left[\varepsilon_i\right] = \frac{x_i \sigma^2}{x_i} = \sigma^2.$$

Definindo as variáveis transformadas $y_i^* = \frac{y_i}{\sqrt{x_i}}$, $x_{1i}^* = \frac{1}{\sqrt{x_i}}$, $x_{2i}^* = \frac{x_i}{\sqrt{x_i}}$ e $\varepsilon_i^* = \frac{\varepsilon_i}{\sqrt{x_i}}$ o modelo transformado pode escrever-se:

$$y_i^* = \beta_0 x_{1i}^* + \beta_1 x_{2i}^* + \varepsilon_i^*.$$

Note-se que o modelo transformado satisfaz as hipóteses do OLS. A estimação deste modelo não levanta quaisquer problemas porque é muito fácil calcular as variáveis transformadas. Um aspecto importante é que os parâmetros a estimar são exactamente os mesmos do modelo original, β_0 e β_1. A única diferença é que estimando o modelo transformado vamos conseguir obter estimadores mais precisos daqueles parâmetros. A partir daqui, a interpretação dos coeficientes e a inferência estatística é feita da forma tradicional.

8.4 Autocorrelação

Nesta secção estudamos o que acontece se a hipótese de que os resíduos são não correlacionados não for satisfeita. O problema da correlação surge normalmente quando os dados são séries temporais (*time series*). Neste caso, é frequente os resíduos de um período estarem correlacionados com os resíduos do(s) período(s) anterior(es).

As perguntas que vamos tentar responder são: como é que o facto de os resíduos serem correlacionados afecta as propriedades dos estimadores dos mínimos quadrados ordinários? Como é que podemos testar a existência de autocorrelação? Como estimar os parâmetros se os resíduos forem correlacionados?

A resposta à primeira pergunta é semelhante aquela que demos no caso de heterocedasticidade. Por um lado, os estimadores dos mínimos quadrados não são os mais eficientes (embora continuem a ser não enviesados). Por outro lado, os desvios padrões dos estimadores calculados pelo OLS são enviesados, o que implica que intervalos de confiança ou teste de hipóteses neles baseados não são válidos.

Para responder à segunda e terceira pergunta temos que especificar a estrutura de autocorrelação. Há varias «formas» de os resíduos estarem correlacionados. Para exemplicar vejamos uma estrutura de correlação extremamente usada: processo autoregressivo de primeira ordem (também desingnado por $AR(1)$):

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

onde usamos o subscrito t para designar o período de tempo (estamos a considerar séries temporais, cada observação refere-se a um período de tempo). A variável u_t tem média 0, variância constante, não existe correlação entre u_t e $u_{t'}$. O parâmetro $-1 < \rho < 1$.

A intuição para este processo autoregressivo é bastante simples. O termo residual no período t tem duas componentes. A primeira $\rho\varepsilon_{t-1}$ está relacionada com o resíduo do período anterior e está associada à inercia existente nos sistemas económicos. O parâmetro ρ reflecte a intensidade desta inércia. A segunda componente u_t é o novo «choque» na variável económica.

Qual é a implicação deste processo autoregressivo nas propriedades dos resíduos? Se $-1 < \rho < 1$ o processo autoregressivo é estacionário, o que significa que os resíduos tem as mesmas propriedades ao longo do tempo. É relativamente fácil mostrar que o valor esperado é zero:

$$\begin{aligned} E[\varepsilon_t] &= \rho E[\varepsilon_{t-1}] + E[u_t] = \rho E[\varepsilon_t] + 0 \\ (1-\rho)E[\varepsilon_t] &= 0 \Leftrightarrow E[\varepsilon_t] = 0 \end{aligned}$$

A variância é:

$$\begin{aligned} \operatorname{var}[\varepsilon_t] &= \operatorname{var}[\rho\varepsilon_{t-1} + u_t] \\ &= \rho^2 \operatorname{var}[\varepsilon_{t-1}] + \operatorname{var}[u_t] + 2\operatorname{cov}[\varepsilon_{t-1}, u_t] \end{aligned}$$

Como u_t não está correlacionado com u_{t-i} o último termo é zero. Usando a estacionaridade obtemos:

$$(1-\rho^2)\operatorname{var}[\varepsilon_t] = \operatorname{var}[u_t] \Leftrightarrow \operatorname{var}[\varepsilon_t] = \frac{\operatorname{var}[u_t]}{1-\rho^2}$$

Qual é a covariância entre ε_t e ε_{t-1}?

$$\begin{aligned} \operatorname{cov}[\varepsilon_t, \varepsilon_{t-1}] &= E[\varepsilon_t\varepsilon_{t-1}] = E[(\rho\varepsilon_{t-1} + u_t)\varepsilon_{t-1}] \\ &= \rho \operatorname{var}[\varepsilon_t] + E[u_t\varepsilon_{t-1}] \\ &= \rho \operatorname{var}[\varepsilon_t] \end{aligned}$$

Repare-se que isto implica que o coeficiente de correlação entre ε_t e ε_{t-1} é ρ.

É também fácil mostrar que:

$$\operatorname{cov}\left[\varepsilon_t, \varepsilon_{t-i}\right] = \rho^i \operatorname{var}\left[\varepsilon_t\right].$$

8.4.1 Modelo transformado

Suponhamos que o modelo original é:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

onde $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ e $E\left[u_t\right] = 0$, $\operatorname{var}\left[u_t\right] = \sigma^2$ e $E\left[u_t u_{t-i}\right] = 0$.

Como a equação anterior se verifica para todas as observações, em $t-1$ temos:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}$$

Multiplicando por ρ esta equação e subtraindo à primeira obtemos:

$$\underbrace{y_t - \rho y_{t-1}}_{y_t^*} = \beta_0 \underbrace{(1-\rho)}_{x_{1,t}^*} + \beta_1 \underbrace{(x_t - \rho x_t)}_{x_{2,t}^*} + \underbrace{\varepsilon_t - \rho\varepsilon_{t-1}}_{u_t}$$

Obtemos assim um modelo transformado em que os resíduos têm variância constante e não são correlacionados e que, por conseguinte, pode ser estimado usando o método dos mínimos quadrados ordinário.

Note-se que no processo de transformação perdemos uma observação, porque a primeira observação não pode ser transformada. Para além disso, este procedimento presupõe o conhecimento de ρ, mas normalmente esse parâmetro não é conhecido.

Na prática o que se é começar por estimar usando o OLS o modelo original. Depois, usam-se os resíduos estimados para calcular $\widehat{\rho}$:

$$\widehat{\rho} = \frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=2}^{n} \widehat{e}_{t-1}}$$

A partir daí transformam-se as variáveis e usa-se o OLS para estimar os parâmetros β_0 e β_1. Este procedimento é adequado se a amostra for grande.

8.4.2 Teste de autocorrelação

Como é que podemos detectar a presença de autocorrelação nos resíduos? Uma primeira ideia é a de representar graficamente os resíduos estimados como função do tempo. Se

observamos que, quando o resíduo de um dado período é elevado, os resíduos dos períodos seguintes também tendem a ser elevados, observando-se sequências relativamente longas de resíduos com o mesmo sinal, isso é indicativo da presença de autocorrelação.

Apesar da representação gráfica ajudar a detectar problemas de autocorrelação, dada a «gravidade» do problema é conveniente efectuar testes mais formais. O teste mais usado para testar a presença de $AR(1)$ é o teste de *Durbin-Watson*. O teste é baseado nos resíduos obtidos do OLS, a partir dos quais se calcula a seguinte estatística:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

Vejamos como é que esta estatística se relaciona com o modelo $AR(1)$.

$$\begin{aligned} d &= \frac{\sum_{t=2}^{n} e_t^2}{\sum_{t=1}^{n} e_t^2} + \frac{\sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2} - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \\ &\simeq 1 + 1 - 2\widehat{\rho} \end{aligned}$$

Ou seja $d \simeq 2(1-\widehat{\rho})$. Repare-se que se $\widehat{\rho} = 0$ o valor da Durbin-Watson será próximo de 2, e isso é indicativo de que os erros não estão correlacionados. Em contrapartida se $\widehat{\rho} = 1$ o valor da estatística d seria próximo de zero, e isso indiciaria a presença de correlação positiva dos resíduos. De forma similar, se $\widehat{\rho} = -1$ o valor de d será próximo de 4, indicando que os resíduos são negativamente correlacionados.

É claro que para podermos efectuar um teste teremos que conhecer a distribuição da estatística d. Para isso, consideremos o modelo de regressão:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

onde $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ e $E[u_t] = 0$, $\text{var}[u_t] = \sigma^2$ e $E[u_t u_{t-i}] = 0$, com a hipótese adicional de que u_t seguem uma distribuição normal.

Consideremos o seguinte teste de hipóteses:

$$H_0 : \rho = 0, \quad H_1 : \rho > 0.$$

É claro um valor de d próximo de 2 sugere que a hipótese nula é falsa. A questão é, para um dado nível de significância, qual é o valor critíco, d_c, abaixo do qual a hipótese nula é rejeitada. Outra forma de colocar a questão é se calcularmos a estatística d, qual é o valor p dessa estatística, se o valor p for inferior ao nível de significância então a hipótese nula é rejeitada.

Até aqui tudo bem, só que a estatística d tem um problema. A distribuição de d depende da matriz X (depende das variáveis explicativas e da amostra concreta usada). Isto é uma chatice, porque significa que o valor critíco depende do problema concreto. Contudo, há programas que fornecem o valor p da estatística d, o que possibilita verificar se a hipótese nula é rejeitada ou não. Para além disso, embora d_c dependa de X, é possível definir limites inferior e superior para d_c, limites esses que não dependem de X. Por outras palavras, seja qual for o problema $d_{Lc} < d_c < d_{Uc}$ e existem tabelas para estes limites, para os diferentes níveis de significância.

Usando os limites d_{Lc} e d_{Uc} a regra de decisão é a seguinte: (i) se $d < d_c$ rejeitar $H_0 : \rho = 0$, (ii) se $d > d_{Uc}$ não rejeitar $H_0 : \rho = 0$ e (iii) se $d_{Lc} < d < d_{Uc}$ o teste é inconclusivo.

Para testar a presença de correlação negativa dos resíduos a ideia é semelhante. O teste é $H_0 : \rho = 0, \quad H_1 : \rho < 0$. A regra de decisão é: (i) se $d > 4 - d_{Lc}$ rejeitar H_0, (ii) se $d < 4 - d_{Uc}$ não rejeitar H_0, (iii) se $4 - d_{Uc} < d < 4 - d_{Lc}$ o teste é inconclusivo.

Exemplo 8.2 O gestor de uma empresa está convencido que os custos médios de produção (y) dependem do salário (x_1), do custo de outros inputs (x_2), das despesas gerais (x_3) e das despesas de publicidade (x_4). Usando uma série de 24 observações mensais o gestor estimou um modelo de regressão múltipla, obtendo os seguintes resultados:

$$\begin{aligned} y_t &= 0.75 + \underset{(.07)}{.24}\, x_{1t} + \underset{(.12)}{0.56} x_{2t} - \underset{(.23)}{0.32} x_{3t} + \underset{(.05)}{0.23} x_{4,t} \\ R^2 &= .79 \ \text{ e } \ d = 0.85. \end{aligned}$$

onde os valores entre parenteses são os desvios-padrões dos estimadores. Que é que se pode concluir destes resultados?

Neste exemplo, $n = 24$ e $k = 4$. Se considerarmos um nível de significância $\alpha = 5\%$ e analisarmos a tabela para os limites critícos da estatística Durbin-Watson verificamos que $d_{Lc} = 1.01$ e $d_{Uc} = 1.78$. Como $d = 0.85 < d_{Lc}$ rejeita-se a hipótese nula de não correlação contra a alternativa de que os resíduos são positivamente correlacionados. Tendo em conta este resultado devemos estimar o modelo transformado de forma a obtermos estimadores mais consistentes e a podermos efectuar testes de hipóteses sobre os coeficientes. Não é conveniente fazer inferência estatística usando os resultados obtidos com os mínimos quadrados ordinários porque os desvios-padrões indicados entre parenteses são enviesados e podem levar-nos a tirar conclusões erradas.♦

8.5 Problemas de especificação

Nesta secção estudamos o seguinte problema: o que é que acontece se não incluirmos na regressão alguma variável explicativa relevante? Aquilo que vamos ver é que, exceptuando o caso em que a variável omitida não está correlacionada com as variáveis explicativas incluídas, as consequências deste erro de especificação são extremamente graves. Por um lado, os estimadores dos mínimos quadrados deixam de ser não enviesados. Pelo outro, a inferência estatística feita com base naqueles estimadores pode levar a conclusões erradas.

Exemplo 8.3 Usando dados para 63 países estimou-se o seguinte modelo:

$$\begin{aligned} y &= .058 - \underset{(.019)}{.052} x_1 - \underset{(.042)}{.005} x_2 \\ R^2 &= 0.17 \end{aligned}$$

onde y – taxa de crescimento do PIB, x_1 – Rendimento real per capita e x_2 – Taxa de tributação média. Estimando o modelo sem incluir a variável x_1 obtiveram-se os seguintes resultados:

$$y = .06 - \underset{(.034)}{0.74} x_2 \quad \text{com } R^2 = .072$$

Comente estes resultados.♦

Vamos supor que o verdadeiro modelo é:

$$y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \varepsilon$$

onde ε satisfaz as hipóteses clássicas do modelo de regressão linear. E vamos admitir que no modelo estimado só se incluiram com variáveis explicativas as variáveis X_1:

$$y = X_1\beta_1 + \widetilde{\varepsilon}$$

Qual é a consequência de se omitirem as variáveis explicativas X_2? Uma primeira consequência é que, se o verdadeiro modelo for o apresentado, no modelo estimado o termo residual não tem valor esperado zero, porque incorpora a influência das variáveis X_2 :

$$\widetilde{\varepsilon} = X_2\beta_2 + \varepsilon$$

Isto, obviamente, viola uma das hipóteses do modelo clássico. A outra consequência é que os estimadores dos mínimos quadrados de β_1 serão enviesados:

$$\begin{aligned} \mathbf{b}_1 &= (X_1'X_1)^{-1}X_1'Y = (X_1'X_1)^{-1}X_1'(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2 + (X_1'X_1)^{-1}X_1'\varepsilon \end{aligned}$$

Por conseguinte, o valor esperado dos estimadores é:

$$E[\mathbf{b}_1] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2$$

O que esta expressão nos diz é que, ao estimamor qual é o efeito das variáveis X_1 em y sem incluir as variáveis X_2, os estimadores obtidos incluem não só o efeito das variáveis X_1 em y, mas incluem também o efeito das váriáveis omitidas se estas estiverem correlacionadas com as variáveis incluídas. Os estimadores obtidos só não são enviesados se as variáveis X_1 e X_2 forem «ortogonais». Nesse caso, $X_1'X_2 = 0$ e o enviesamento é nulo.

Até aqui vimos o que acontece se omitirmos variáveis relevantes. Mas também podemos pensar no caso contrário. O que acontece se incluirmos no modelo variáveis que não são relevantes. Neste caso, o problema é menos grave que no caso anterior. Os estimadores são não enviesados e o único problema é alguma perda de eficiência pelo facto de se incluirem variáveis a mais.

8.6 Minímos quadrados não lineares

Nos modelos considerados até aqui a forma funcional relacionando a variável explicada com as variáveis explicativas foi sempre linear nos parâmetros β_0, β_1,$\cdots$,β_k. Estudamos alguns casos em que existia uma relação não linear entre a variável explicada e as variáveis explicativas, mas em que o modelo era linear nos parâmetros e, logo, podia ser estimado usando regressão linear múltipla.

Contudo, na prática, há muitos modelos económicos e modelos estatísticos correspondentes em que não temos funções lineares nos parâmetros desconhecidos.

Exemplo 8.4 Consideremos o seguinte modelo estatístico de uma função de produção Cobb-Douglas:

$$y_i = \beta_0 x_{1i}^{\beta_1} x_{2i}^{\beta_2} + \varepsilon_i$$

Repare-se que, por causa do termo residual, não é possível linearizar este modelo tomando o logaritmo de ambos os membros.♦

Exemplo 8.5 Consideremos o modelo estatístico seguinte:

$$y_i = \beta_0 + \beta_1 x_{1i}^{\beta_2} + \varepsilon_i$$

que é um modelo não linear nos parâmetros a estimar.♦

O princípio dos mínimos quadrados pode, sem qualquer problema, ser usado para estimar parâmetros em modelos não lineares. A ideia é encontrar os valores dos estimadores que minimizam a soma dos quadrados dos resíduos. O que difere do modelo de regressão linear são os cálculos para encontrar esses estimadores. O que acontece é que por causa da não linearidade nos parâmetros, a função soma dos quadrados dos resíduos, $S(\mathbf{b})$, é mais complicada, podendo ter vários mínimos locais e, por isso, é mais díficil encontrar o valor dos estimadores que minimiza $S(\mathbf{b})$.

Na estimação de um modelo não linear usando mínimos quadrados é muitas vezes impossível obter uma solução analítica para o problema, ou seja o óptimo tem de ser encontrado numericamente. Por outro lado, é também frequente haver vários mínimos locais, e nesses casos tem que se identificar qual desses mínimos é o mínimo global.

A ideia na estimação numérica é a seguinte: começa-se por indicar «palpites» iniciais para o valor dos parâmetros. Com base nesses palpites iniciais é calculado o valor da função $S(\mathbf{b})$. O passo seguinte é alterar os «palpites» numa direcção que faça diminuir $S(\mathbf{b})$. Desta forma, um novo valor dos parâmetros é calculado e o processo repete-se até que se atinja um ponto em que não é possível com pequenas variações nos parâmetros alterar $S(\mathbf{b})$. Ou seja, até que o algorítmo convirja para um mínimo local.

Um problema na estimação númerica é que não há garantia que o mínimo local encontrado seja o mínimo global. A única forma de tentar aliviar este problema é repetir o processo descrito começando com «palpites iniciais» muito diferentes. Se ao fazermos isso se convergir sempre para o mesmo ponto, isso sugere que estamos na presença de um mínimo global. Se ao fazermos isso convergirmos para pontos diferentes, dependendo dos «palpites» iniciais, isso indica-nos que $S(\mathbf{b})$ tem vários mínimos locais e, como é óbvio devemos escolher o melhor desses pontos (aquele em que $S(\mathbf{b})$ é mais baixo). O problema é que será difícil termos a certeza que aquele é mesmo o máximo global.

8.6.1 Propriedades dos mínimos quadrados não lineares

O que podemos dizer sobre as propriedades dos estimadores dos mínimos quadrados não lineares? Em geral estes estimadores serão funções complicadas de y e, consequentemente, é muito difícil definir as suas propriedades em amostras limitadas. Mas é possível identificar as *propriedades assimptóticas*, para amostras de grande dimensão: *os estimadores dos mínimos quadrados não lineares são consistentes, seguem aproximadamente uma distribuição normal, e é possível calcular de forma aproximada a sua matriz de variâncias e covariâncias.*

Capítulo 9

Modelos com variáveis dependentes discretas

Muitas das decisões dos agentes económicos são decisões de natureza discretas: casar ou não, fazer um mestrado ou não, usar carro ou usar transportes públicos,... Nesta secção analisamos modelos económicos e estatísticos onde a avriável dependente é uma variável dummy, que toma o valor 1 se aquela decisão é tomada e toma o valor 0 no caso contrário.

9.1 Modelos económico e estatístico

9.1.1 Modelo económico

Consideremos um indivíduo que tem que escolher entre duas alternativas (ex: fazer um mestrado ou não fazer). Para cada indivíduo podemos observar qual é a sua escolha e definir a variável dummy y da seguinte forma:

$$y_i = \begin{cases} 1 & \text{se o indivíduo faz mestrado} \\ 0 & \text{se o indivíduo não faz mestrado} \end{cases}$$

Se admitirmos que cada indivíduo faz sempre a escolha que maximiza a sua utilidade, se designarmos por U_{i1} a utilidade do indivíduo i se fizer o mestrado, e por U_{i0} a sua utilidade se não fizer o mestrado temos que:

$$y_i = \begin{cases} 1 & \text{se } U_{i1} \geq U_{i0} \\ 0 & \text{se } U_{i1} < U_{i0} \end{cases}$$

Para completar o modelo poderíamos definir como é que a utilidade de cada indíviduo i para cada alternativa (fazer ou não mestrado) depende das características de cada indivíduo (idade, rendimento, profissão,...) e dos atributos do programa de mestrado (custo, qualidade,...).

9.1.2 Modelo estatístico

Na prática, é impossível prever com certeza a decisão que um indíivíduo escolhido ao acaso tomará. Ou seja, y_i é uma *variável aleatória discreta*. Se designarmos por p_i a probabilidade do indivíduo i escolher a alternativa 1, a função de probabilidades da v.a. y_i é dada por:

$$f(y_i) = p_i^{y_i}(1-p_i)^{(1-y_i)}$$

isto é equivalente a dizer que $f(1) = p_i$ e $f(0) = 1 - p_i$. Ou seja, a função densidade de probabilidade é uma Bernoulli, e é completamente descrita por p_i. O valor esperado e a variância de y_i são:

$$\begin{aligned} E[y_i] &= 1 \times p_i + 0 \times (1-p_i) = p_i \\ \text{var}[y_i] &= E\left[(y_i - p_i)^2\right] = (1-p_i)^2 p_i + (0-p_i)^2(1-p_i) = p_i(1-p_i) \end{aligned}$$

A ideia nos modelos estatísticos que vamos ver é relacionar a probabilidade p_i com várias variáveis explicativas, que incluem características dos indivíduos e características das alternativas.

9.2 O modelo de probabilidade linear

Este modelo usa as ideias básicas do modelo de regressão linear, em que y_i tem uma componente sistemática que está relacionada com o comportamento das variáveis explicativas e uma parte aleatória:

$$\begin{aligned} y_i &= E[y_i] + \varepsilon_i = p_i + \varepsilon_i \\ &= \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}}_{p_i} + \varepsilon_i \end{aligned}$$

Um primeiro problema com este modelo é que embora seja licíto assumir que o valor esperado dos resíduos é zero, não se pode admitir que os resíduos tenham distribuição normal. De facto, como y_i é discreto, os resíduos também são discretos. Para além disso, a variância dos resíduos não é a mesma para todo os indivíduos uma vez que:

$$\text{var}[\varepsilon_i] = \text{var}[y_i] = p_i(1-p_i)$$

Ou seja, o modelo apresenta heterocedasticidade.

Mas o problema mais grave do modelo anterior é que não há nada que garanta que os valores estimados para p_i pertençam ao intervalo $[0,1]$. É óbvio que isto não é satisfatório, uma medida de probabilidade abaixo de zero ou acima de 1 não fazem qualquer sentido!

9.3 O modelo probit

O modelo probit é um modelo não linear nos parâmetros que relaciona p_i com as variáveis explicativas, mas de uma forma que garante que $p_i \in [0,1]$.

O modelo probit tem por base o modelo de utilidade apresentado anteriormente. O indíce de utilidade do indíviduo i está relacionado com k variáveis explicativas:

$$U_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

U_i é uma medida da *diferença de utilidade* entre a alternativa $y_i = 1$ e $y_i = 0$. À partida U_i pode tomar qualquer valor. Quanto maior for o valor de U_i maior será a probabilidade do indíviduo i escolher $y_i = 1$. A questão é: qual é a relação entre p_i e U_i, será que é possível relacioná-los de tal forma que $p_i \in [0,1]$?

O que o modelo probit admite é que p_i é igual à probabilidade de uma variável aleatória normal estandartizada tomar um valor inferior ou igual a U_i. Ou seja, se designarmos por F a função de distribuição cumulativa da $N(0,1)$:

$$p_i = F(U_i) = P[z \leq U_i] = \int_{-\infty}^{U_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz.$$

Repare-se que desta forma está assegurado que $p_i \in [0,1]$, uma vez que o seu valor é retirado da função de distribuição da normal.

Uma questão interessante é: qual é a alteração na probabilidade de escolher a alternativa 1 se a variável explicativa x_j variar?

$$\frac{\partial p_i}{\partial x_{ij}} = \frac{dF}{dU_i}\frac{\partial U_i}{\partial x_{ij}} = f(U_i)\beta_j$$

onde $f(U_i)$ é a função densidade da $N(0,1)$ avaliada no ponto U_i. Repare-se que como $f(U_i)$ é sempre positivo o sinal de $\frac{\partial p_i}{\partial x_{ij}}$ é igual ao sinal de β_j (se β_j é positivo, qunado x_{ij} aumenta também aumenta o indíce de utilidade, e se U_i aumenta também aumenta a probabilidade de $y_i = 1$). Mas a magnitude do impacto de variações de x_{ij} em p_i depende do valor de U_i. Quando U_i é próximo de 0 (para indíviduos que estão na margem entre escolher $y_i = 1$ ou $y_i = 0$), é quando as variações de x_{ij} tem um maior impacto em p_i.

9.3.1 Estimação dos parâmetros no modelo probit

A estimação do modelo probit é feita usando o método de máxima verosimilhança. A razão para o uso deste método prende-se com as características dos resíduos.

O método de máxima verosimilhança é baseado na função de densidade conjunta das n observações de que dispomos. Admitindo que as observações são v.a. independentes a função de densidade conjunta é igual ao produto das funções densidade marginais, ou seja.

$$\begin{aligned} f(y_1, y_2, \cdots, y_n) &= f(y_1) \cdot f(y_2) \cdot \cdots \cdot f(y_n) = \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} \\ &= \prod_{i=1}^{n} F(U_i)^{y_i} \left[1 - F(U_i)\right]^{1-y_i} \\ &= \prod_{i=1}^{n} F(\mathbf{x}_i'\boldsymbol{\beta})^{y_i} \left[1 - F(\mathbf{x}_i'\boldsymbol{\beta})\right]^{1-y_i} \end{aligned}$$

Repare-se que na realidade não sabemos qual é a função de densidade conjunta, porque os parâmetros $\boldsymbol{\beta}$ não são conhecidos.Qual é a ideia do método de máxima verosimilhança? Os estimadores da máxima verosimilhança, são os valores dos parâmetros que maximizam a probabilidade de se obter a amostra que foi de facto observada.

Se considerarmos que na amostra se conhecem os valores de y_i e das variáveis explicativas $\mathbf{x}_i$ a função de densidade acima indicada depende apenas do vector de parâmetros $\boldsymbol{\beta}$, ou seja:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} F(\mathbf{x}_i'\boldsymbol{\beta})^{y_i} \left[1 - F(\mathbf{x}_i'\boldsymbol{\beta})\right]^{1-y_i}$$

A função $L(\boldsymbol{\beta})$ designa-se por *função de verosimilhança*. Os estimadores da máxima verosimilhança, $\mathbf{b}$, são os valores que maximizam a função de verosimilhança. Este problema é de díficil solução analítica, mas a solução pode ser encontrada usando optimização numérica.

9.3.2 Propriedades dos estimadores de ML no modelo probit

As propriedades dos estimadores de ML só podem ser, em geral, determinadas para amostras de grande dimensão. Neste caso, os estimadores são não enviesados e têm uma distribuição aproximadamente normal:

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \left(X'DX\right)^{-1})$$

onde D é uma matriz diagonal cujo elemento d_i é dado por:

$$d_i = \frac{[f(\mathbf{x}_i'\boldsymbol{\beta})]^2}{F(\mathbf{x}_i'\boldsymbol{\beta})\,[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]}$$

9.4 O modelo logit

O modelo logit difere do modelo probit unicamente num aspecto: na função de distribuição cumulativa que é usada para definir p_i a partir de U_i. No caso do modelo probit usamos a função normal, no modelo logit usa-se a função de distribuição cumulatiav logística:

$$\begin{aligned} P_i &= F(\mathbf{x}_i'\boldsymbol{\beta}) \\ &= \frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}} \end{aligned}$$

A função de densidade da logística é simétrica em torno do zero e tem um comportamento em sino, mas apresenta maior densidade que a normal para valores afastados do zero (mais peso nas «caudas»).

Para estimar o modelo logit usa-se, mais uma vez, o método de máxima verosimilhança. Definindo a função de verosimilhança:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right]^{y_i} \left[1 - \frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\right]^{1-y_i}$$

podemos encontrar os valores que maximizam esta função.

Capítulo 10

Análise de variância

10.1 Análise de variância com um factor

A ideia aqui é generalizar o teste da igualdade das médias de duas distribuições normais com a mesma variância. Por exemplo, podemos estar interessados em testar as diferenças no resultados de três ou mais metodos de ensino, consumo de gasolina de três ou mais modelos de automóveis.

Suponhamos, por exemplo, que temos 3 tipos de automóveis: A, B e C e o nosso objectivo é comparar o consumo médio dos três modelos. Em três amostras independentes: $n_A = 10$, $n_B = 10$ e $n_C = 10$ obtiveram-se os resultados seguintes $\overline{x}_A = 6.5$, $\overline{x}_B = 5.5$, $\overline{x}_C = 6.2$. A pergunta que fazemos é: será que a diferença observada nas médias nestas amostras e devida ao acaso ou, pelo contrário, há evidência de que o consumo médio nos três modelos é de facto diferente?

Intuição para a respostas: depende muito da variabilidade. Se a variabilidade em torno das médias nas amostras for pequena em relação a variabilidade entre as tres médias a evidência suporta mais a hipótese de que as médias sao diferentes. Se a variabilidade em torno das médias das amostras é grande em relação a variabilidade entre as três médias a hipótese nula não deve pode ser rejeitada.

$\Rightarrow$ Analise de variância

Se houver k grupos a hipótese nula é

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

Recolhendo k amostras independentes com dimensões $n_1, n_2, \cdots, n_k$ com $\sum_i n_i = n$. Designemos por X_{ij} a $j-$ésima observação do grupo i. A média da amostra do grupo i é

$$\overline{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

A média global é dada por

$$\overline{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{n} \quad \text{ou} \quad \overline{X} = \frac{\sum_{i=1}^{k} n_i \overline{X}_i}{n}$$

Como foi sugerido o teste da igualdade das médias e baseado na comparação entre variância nos grupos (em torno da media em cada grupo) versus variância entre os grupos (entre as médias dos diferentes grupos). Pode mostrar-se que a variância total pode ser decomposta na soma destas duas componentes

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i + \overline{X}_i - \overline{X}\right)^2$$

ora isto é equivalente a

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\overline{X}_i - \overline{X}\right)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)\left(\overline{X}_i - \overline{X}\right)$$

mas

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\overline{X}_i - \overline{X}\right)^2 = \sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}\right)^2$$

e

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)\left(\overline{X}_i - \overline{X}\right) = \sum_{i=1}^{k} \left(\overline{X}_i - \overline{X}\right) \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right) = \sum_{i=1}^{k} \left(\overline{X}_i - \overline{X}\right)\left(n_i\overline{X}_i - n_i\overline{X}_i\right) = 0$$

logo

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2 + \sum_{i=1}^{k} n_i\left(\overline{X}_i - \overline{X}\right)^2 = SSW + SSG$$

Se definirmos o desvio quadrado médio nos grupos como $\frac{SSW}{n-k}$ pode mostrar-se que este é um estimador não enviesado da variância na população (intuição para $n-k$?). Por outro lado, se a hipótese nula for verdadeira o erro quadrado médio entre os grupos é também um estimador não enviesado de σ^2. Mas, se a hipótese nula não for verdadeira $\frac{SSG}{k-1}$ é um estimador enviesado da variância, $\frac{SSG}{k-1}$ sobrestima a variância porque também contém informação sobre o quadrado das diferenças das médias das k populações.

O teste é baseado no rácio

$$\frac{\frac{SSG}{k-1}}{\frac{SSW}{n-k}}$$

Este rácio deve ser proximo de 1 se H_0 for verdadeiro. Se o rácio for muito maior que 1 devemos rejeitar a hipótese nula. O que vamos ver de seguida e que, se H_0 for verdadeira a estatistica acima tem distribuição F com $(k-1, n-k)$ graus de liberdade.

Se H_0 for verdadeira podemos olhar para X_{ij} com $i = 1, \cdots, k$ e $j = 1, 2, \cdots, n_i$ como uma amostra aleatoria de dimensao $\sum_i n_i = n$ de uma distribuição $N(\mu, \sigma^2)$ (as observacoes vem todas da mesma população). Neste caso sabemos ja que $\frac{SST}{n-1} = S^2$ e um estimador não enviesado de σ^2 e para alem disso $\frac{SST}{\sigma^2}$ tem distribuição $\mathcal{X}^2_{n-1}$ (**note-se** que H_0 tem que ser verdadeira para isto se verificar).

Um estimador não enviesado de σ^2 baseado so na amostra do grupo i e

$$S_i^2 = \frac{\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2}{n_i - 1}$$

e sabemos que

$$\frac{(n_i-1)S_i^2}{\sigma^2} = \frac{\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2}{\sigma^2} \sim \mathcal{X}^2_{n_i-1}$$

mas isto implica que

$$\frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{\sigma^2} = \frac{SSW}{\sigma^2} \sim \mathcal{X}^2_{n-k}$$

uma vez que a soma de qui-quadrados independentes e qui-quadrado com graus de liberdade igual a soma dos graus de liberdade.

Ora

$$\frac{SST}{\sigma^2} = \frac{SSW}{\sigma^2} + \frac{SSG}{\sigma^2}$$

e sabemos que $\frac{SST}{\sigma^2} \sim \mathcal{X}^2_{n-1}$ e $\frac{SSW}{\sigma^2} \sim \mathcal{X}^2_{n-k}$. há um teorema que garante $\frac{SSW}{\sigma^2}$ e $\frac{SSG}{\sigma^2}$ sao independentes e logo $\frac{SSG}{\sigma^2} \sim \mathcal{X}^2_{k-1}$.

Os resultados anteriores implicam que, se H_0 for verdadeira

$$\frac{\frac{SSG}{\sigma^2(k-1)}}{\frac{SSW}{\sigma^2(n-k)}} = \frac{\frac{SSG}{k-1}}{\frac{SSW}{n-k}} \sim F_{k-1,n-k}$$

Mostrar que se H_0 não for verdadeira o erro quadrado médio entre grupos e um estimador enviesado da variância na população.

$$E[SSG] = E\left[\sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}\right)^2\right] = E\left[\sum_{i=1}^{k} n_i \overline{X}_i^2 - n\overline{X}^2\right] = \sum_{i=1}^{k} n_i E\left[\overline{X}_i^2\right] - nE\left[\overline{X}^2\right]$$

usando o facto de $Var[X] = E\left[X^2\right] - [E(X)]^2$ a expressao anterior pode reescrever-se:

$$E[SSG] = \sum_{i=1}^{k} n_i \left[Var\left(\overline{X}_i\right) + E\left(\overline{X}_i\right)^2\right] - n\left[Var\left(\overline{X}\right) + E\left(\overline{X}\right)^2\right]$$

ou seja

$$E[SSG] = \sum_{i=1}^{k} n_i \left[\frac{\sigma^2}{n_i} + \mu_i^2\right] - n\left[\frac{\sigma^2}{n} + \mu^2\right] = (k-1)\sigma^2 + \sum_{i=1}^{k} n_i \left(\mu_i - \mu\right)^2$$

se a hipótese nula for verdadeira o ultimo termo e nulo e $E[SSG/(k-1)] = \sigma^2$, mas caso contrario o estimador tem enviesamento positivo.

10.1.1 Quadro da análise de variância

Esta e uma forma de sumarizar a informação usada no teste da igualdade das médias

Fonte da Variacao	Soma dos Quadrados	Graus de Liberdade	Erro Quadrado Medio	F
Entre Grupos	SSG	$k-1$	$MSG = \frac{SSG}{k-1}$	$fracMSGMSW$
Nos Grupos	SSW	$n-k$	$MSW = \frac{SSW}{n-k}$	
Total	SST	$n-1$		

Exemplo: No exemplo dos três tipos de automóveis as amostras recolhidas foram as seguintes:

A	B	C
6.6	6.2	5.1

10.1.2 Modelo de Análise de Variância de um Factor

Vamos olhar para o modelo de variância de uma forma ligeiramente diferente. Seja a v.a. X_{ij} a observação j do grupo i. X_{ij} pode ser visto como a soma de duas componentes: a media no grupo i mais uma v.a. com media zero

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

como estamos a assumir que as k amostras sao independentes isso implica que as v.a. ε_{ij} sao independentes. Por outro lado a hipótese de igual variância implica que todos os ε_{ij} tem a mesma variância.

Se designarmos a media global por μ e designarmos por G_i a diferença entre a media no grupo i e a media geral, $G_i = \mu_i - \mu$, podemos ainda escrever o modelo acima

$$X_{ij} = \mu + G_i + \varepsilon_{ij}$$

ou seja, a observação X_{ij} e igual a soma da media global com um termo G_i que e espicifico do grupo i com um residuo aleatorio. A hipótese nula de que as médias sao todas iguais pode ser reescrita

$$H_0 : G_0 = G_1 = \cdots = G_k = 0$$

Esta forma de olhar para o modelo ajuda a perceber o porque de analise de variância com mais de um factor.

Notar ainda que estimador de μ é $\overline{X}$, que o estimador de G_i é $\overline{X}_i - \overline{X}$ e por último o estimador de ε_{ij} obtem-se fazendo a diferença $X_{ij} - (\overline{X}_i - \overline{X}) - \overline{X} = X_{ij} - \overline{X}_i$ (é po isso que SSW também é designado como soma dos quadrados dos erros).

10.2 Análise de variância dois factores, uma observação por cela

Muitas vezes há mais do que um factor importante que pode afectar o resultado de um dado fenómeno. Por exemplo, o consumo médio de gasolina depende do tipo de automóvel

mas pode também depender do condutor, do tipo de gasolina,... Se quisermos estudar a influência de dois factores podemos usar a metodologia que se segue, a generalização para mais de dois factores fica para vocês.

Vamos chamar o primeiro factor A e o segundo B. Para cada um dos factores há vários grupos (a grupos no primeiro factor e b no segundo) , daqui resulta uma tabela de combinações possíveis dos dois factores ($a \times b$ celas). Um aspecto que é importante no tipo de análise que podemos fazer é o número de observações por cela (se há uma ou mais que uma observação por cela).

Talvez seja interessante vermos qual o modelo da população que está a ser assumido. Designemos por X_{ij} a observação no grupo i do factor A e no grupo j do factor B, com $i = 1, \cdots, a$ e $j = 1, \cdots, b$. Esta variável aleatória pode ser vista como a soma de quatro componentes: a média global, um parâmetro α_i que é específico do grupo i do factor A (que mede a diferença entre média global e média no grupo i), um parâmetro β_j específico do grupo j do factor B e uma v.a. ε_{ij} que representa aquilo que não é explicado por nenhum dos factores e que se assume $N(0, \sigma^2)$, ou seja:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

As médias na amostra para cada grupo e global podem ser utilizadas para estimar cada um dos parâmetros acima.

Podemos testar a hipótese da igualdade das médias nos a grupos do factor A ou dos b grupos do factor B

$$H_A : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \text{ e } H_B : \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

Esse teste é baseado na decomposição dos quadrados dos desvios total em três componentes: variação entre grupos do primeiro factor, variação entre grupos do segundo factor, variação nos grupos (soma dos quadrados dos erros):

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \overline{X}\right)^2 = b\sum_{i=1}^{a}\left(\overline{X}_{i\,.} - \overline{X}\right)^2 + a\sum_{j=1}^{b}\left(\overline{X}_{.\,j} - \overline{X}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \overline{X}_{i\,.} - \overline{X}_{.\,j} + \overline{X}\right)^2$$

ou seja, $SST = SSA + SSB + SSE$.

Tal como no caso anterior é possível mostrar que SSE/σ^2 segue uma distribuição qui-quadrado independentemente das hipóteses H_A e H_B serem ou não verdadeiras. Por outro lado, se H_A e H_B forem verdadeiras SSA/σ^2, SSB/σ^2 e SSE/σ^2 são qui-quadrados

independentes. SSA/σ^2 tem $(a-1)$ graus de liberdade, SSB/σ^2 tem $(b-1)$ graus de liberdade e SSE/σ^2 tem $(a-1)(b-1)$ graus de liberdade, isto vem de $(ab-1)-(a-1)-(b-1)$.

A partir daqui é fácil, para testar H_A usamos o rácio MSA/MSE, que segue uma F. Para testar H_B usamos MSB/MSE. Resumindo:

Fonte da Variacao	Soma dos Quadrados	Graus de Liberdade	Erro Quadrado Medio	F
Factor A	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Factor B	SSB	$n-k$	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
Erros	SSE	$(a-1)(b-1)$	$MSE = \frac{SSE}{(a-1)(b-1)}$	
Total	SST	$ab-1$		

Exemplo: Três tipos de carros são conduzidos com quatro tipos diferentes de gasolina. O consumo médio de cada carro com cada tipo de gasolina é descrito na seguinte tabela de 3×4:

Carro	1	2	3	4	$\bar{X}_{i.}$
1	6.2	6.0	6.1	5.8	
2	5.3	5.2	5.0	4.9	
3	5.7	5.8	5.6	5.5	
$\bar{X}_{.j}$					

10.3 Análise de variância dois factores, várias observações por cela

Se tivermos mais do que uma observação por cela obteremos estimadores mais precisos, mas para além disso vai ser possível isolar uma outra fonte de variabilidade: a **interacção** entre os dois factores (a ideia é que os dois factores podem não actuar independentemente, por exemplo um dos tipos de automóveis pode ser mais eficiente, mas a sua eficiência relativa pode ser diferente consoante o tipo de gasolina).

Agora em cada cela temos várias observações, para designar uma observação em particular usamos X_{ijl} a l-ésima observação do grupo i do factor A e grupo j do factor B. O modelo implicitamente usado aqui é

$$X_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}$$

Podemos calcular média global, em cada grupo (linha ou coluna) e em cada cela. A soma dos quadrados pode ser decomposta em quatro componentes

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{c}(X_{ijl}-\overline{X})^2 = bc\sum_{i=1}^{a}(\overline{X}_{i..}-\overline{X})^2 + ac\sum_{j=1}^{b}(\overline{X}_{.j.}-\overline{X})^2 + c\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij.}-\overline{X}_{i..}-\overline{X}_{.j.}+\overline{X})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{c}(X_{ijl}-\overline{X}_{ij.})^2$$

ou seja $SST = SSA+SSB+SSAB+SSE$. Se as hipóteses de que os α_i, os β_j e os γ_{ij} são todos zero forem verdadeiras $SSA/\sigma^2, SSB/\sigma^2, SSAB/\sigma^2$ e SSE/σ^2 são quiquadrados independentes. SST/σ^2 tem $(abc-1)$ graus de liberdade, SSA/σ^2 tem $(a-1)$ graus de liberdade, SSB/σ^2 tem $(b-1)$ graus de liberdade. Como SSE/σ^2 tem $ab(c-1)$, porque se fizermos a soma dos quadrados só numa cela os graus de liberdade são $(c-1)$ e há ab celas, é possível concluir que os graus de liberdade de $SSAB/\sigma^2$ são $(a-1)(b-1)$. Resumindo:

Fonte da Variacao	Soma dos Quadrados	Graus de Liberdade	Erro Quadrado Medio	F
Factor A	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Factor B	SSB	$n-k$	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
Factor AB	$SSAB$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{MSE}$
Erro	SSE	$ab(c-1)$	$MSE = \frac{SSE}{ab(c-1)}$	
Total	SST	$abc-1$		

Exemplo: O exercício 61 do cap. 15 de Paul Newbold.

Capítulo 11

Teste de Modelos Probabilísticos e Tabelas de Contigência

11.1 Teste de Modelos Probabilísticos, Parâmetros Conhecidos

Neste capítulo estudamos teste que usam estatísticas com distribuição qui-quadrado (Karl Pearson, 1900).

O teste do qui-quadrado mais elementar é baseado nas v.a. $Y_1, Y_2, ..., Y_k$ que tem uma distribuição multinominal com parâmetros n e $p_1, p_2, \cdots, p_k$. Para percebermos o porquê da estatística qui-quadrado suponhamos que a distribuição é binomial, $Y_1 \sim b(n, p_1)$. Neste caso, pelo teorema do limite central sabemos que

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

tem uma distribuição aproximadamente normal para n elevado ($np_1 > 5$ e $n(1-p_1) > 5$). Mas, nesse caso, $Q_1 = Z_1^2$ tem uma distribuição qui-quadrado com 1 grau de liberdade. Se definirmos $Y_2 = n - Y_1$ e $p_2 = 1 - p_1$ podemos mostrar que Q_1 se pode escrever

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

Em geral, para uma multinomial, Q_1 pode ser escrito

$$Q_1 = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$$

e Q_1 tem distribuição qui-quadrado com $k-1$ graus de liberdade.

Repare-se que $E(Y_i) = np_i$. Logo $Y_i - np_i$ mede a diferença entre o valor observado de ocorrências do atributo i e o valor esperado de ocorrências em n experiências. A estatística

do qui-qaudrado mede a proximidade dos valores observados em relação aos esperados. Se o valor da estatística for pequeno não rejeitamos a hipótese nula de que a distribuição de $Y_1, \cdots, Y_k$ é uma multinominal com parâmetros n e $p_1, p_2, \cdots, p_k$. Se, pelo contrário, o valor do teste for muito elevado o modelo probabilístico assumido é rejeitado.

Exemplo 1: Testar se uma amostra é aleatória Se tomarmos uma sequência aleatória de digítos aleatórios (números de 0 a 9), a probabilidade de dois dígitos consequtivos serem iguais é $\frac{1}{10}$, a probabilidade de dois dígitos consequtivos divergirem de 1 número (assumindo 0 e 9 são números consequtivos) é $\frac{2}{10}$ e todas as outras hipóteses têm probabilidade $\frac{7}{10}$.

Consideremos a sequência de 51 dígitos seguintes:

5	8	3	1	9	4	6	7	9	2
6	3	0	8	7	5	1	3	6	2
1	9	5	4	8	0	3	7	1	4
6	0	4	3	8	2	7	3	9	8
5	6	1	8	7	0	3	5	2	5
2									

dos três tipos de ocorrências possíveis verificaram-se (50 diferenças observadas):

	Freq	$E(Y_i)$
Mesmo	0	5
Dif 1	8	10
Dif >1	42	35
Total	50	50

Se calcularmos a estatística do qui-quadrado obtemos

$$\frac{(0-5)^2}{5} + \frac{(8-10)^2}{10} + \frac{(42-35)^2}{35} = 6.88 > 5.991 = \mathcal{X}^2_{0.05,2}$$

logo a hipótese nula de que os dígitos foram gerados aleatoriamente é rejeitada. Repare-se que este exemplo mostra que a ideia que em números gerados aleatoriamente não aparecem digítos seguidos iguais!

Exemplo 2: Testar se 4 moedas Seja X o número de caras que ocorrem quando quatro moedas são lançadas. Se as moedas forem todas *bem comportadas*, e independentes

a distribuição de X é $b(4, \frac{1}{2})$. Suponhamos que são feitas 100 repetições desta experiência com os resultados seguintes:

X	0	1	2	3	4
Freq	7	18	40	31	4
$E(Y)$	6.25	25	37.5	25	6.25

onde usamos as probabilidades de numa $b(4, \frac{1}{2})$ o valor de X ser 0,1,2,3 ou 4 para calcular os valores esperados nas 100 repetições. O valor da estatística qui-quadrado é 4.47. Usando $\alpha = 0.05$ o valor crítico da qui-quadrado com (5-1) graus de liberdade é 9.488. Logo, a hipótese nula não é rejeitada.

11.2 Teste de Modelos Probabilísticos, Parâmetros Desconhecidos

O que acontece se os parâmetros do modelo probabilístico que se pretende testar forem desconhecidos. Nesse caso teremos que estimar primeiro os parâmetros e só depois construir o teste do qui-quadrado. Uma forma de estimar é encontrar os estimadores que minimizam a estatística do qui-quadrado, **estimadores do qui-quadrado mínimio**. Neste caso, se for m o número de parâmetros a estimar a estatística do qui-quadrado segue uma qui-quadrado com $k - m - 1$ graus de liberdade.

Se usarmos outros métodos para estimar os parâmetros a estatística do qui-quadrado será em geral maior do que a que obtinhamos da minimização. Por esta razão, a probabilidade de rejeitar a hipótese nula será maior do que se se minimizasse a estatística do qui-quadrado.

Exemplo: Num estudo sobre jornais diários numa amostra de 262 blocos de texto (cada bloco com aproximadamente 200 palavras) conclui-se o número de ocorrências média da palavra "poder " foi de 0.66. A tabela seguinte mostra a frequência de 0, 1, 2, 3, e mais que 3 ocorrências

# de ocorr.	0	1	2	3 ou >
Freq	156	63	29	14

O nosso objectivo é testar a hipótese nula que as ocorrências daquela palavra seguem uma distribuição Poisson:

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

onde λ é o número médio de ocorrências. Como λ é desconhecido podemos usar a média na amostra para estimá-la $\widehat{\lambda} = 0.66$ e calcular, para este valor, qual a probabilidade de 0,1,2, e 3 ou mais ocorrências. Disto resulta:

# de ocorr.	0	1	2	3 ou >
Freq	156	63	29	14
Prob	.5169	.3412	.1126	.0293
Freq. Esp.	135	89	30	8

Calculando a estatística do qui-quadrado obtemos

$$\frac{(156-135)^2}{135} + \frac{(63-89)^2}{89} + \frac{(29-30)^2}{30} + \frac{(14-8)^2}{8} = 15.396$$

Como há quatro categorias e estimamos um parâmetro o número de graus de liberdade é 2. Para $\alpha = 0.005$ o valor de $\mathcal{X}^2_{0.005,2} = 10.6$. Logo a hipótese nula é rejeitada a um nível de significância de 0.5%.

11.3 Tabelas de Contingência

Consideremos um certo fenómeno aleatório cujos resulatdos podem ser classificados de acordo com dois atributos A e B (por exemplo, peso e altura). Vamos assumir que cada atributo tem um certo número de categorias $A_1, \cdots, A_r$ e $B_1, \cdots, B_c$ (categorias mutuamente excluxivas e exaustivas). Logo, temos um total de rc classificações possíveis. Os resultados da classificação podem ser representados numa tabela com r linhas e c colunas.

Seja $p_{ij} = P(A_i \bigcap B_j)$ e Y_{ij} o número de observações na cela da linha i e coluna j (frequência de $A_i \bigcap B_j$). Se a experiência for repetida n vezes a variável aleatória

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - np_{ij})^2}{np_{ij}} \sim \mathcal{X}^2_{rc-1}$$

Suponhamos que pretendemos testar a independência de A e B. Ou seja,

$$H_0 : P(A_i \bigcap B_j) = P(A_i)P(B_j)$$

e bastaria substituir p_{ij} por $p_{i.} \times p_{.j}$.

O problema é que nas aplicações raramente se conhecem as probabilidades $p_{i.}$ e $p_{.j}$. O que significa que vamos ter que estimá-las usando as frequências observadas

$$\widehat{p}_{i.} = \frac{y_{i.}}{n}, \text{ onde } y_{i.} = \sum_{j=1}^{c} y_{ij} \text{ e } \widehat{p}_{.j} = \frac{y_{.j}}{n}, \text{ onde } y_{.j} = \sum_{i=1}^{r} y_{ij}$$

no processo de estimação perdemos $(r + c - 2)$ graus de liberdade (isto porque probabilidades têm que somar 1, logo basta estimar $(r - 1) + (c - 1)$ parâmetros.

Usando como estimadores as frequências observadas na amostra e o facto de

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - n(Y_{i.}/n)(Y.j/n))^2}{n(Y_{i.}/n)(Y.j/n)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(Y_{ij} - \frac{Y_i\, Y.j}{n}\right)^2}{\frac{Y_i\, Y.j}{n}} \sim \mathcal{X}^2_{rc-1-(r+c-2)}$$

podemos efectuar o teste de independência de dois atributos de uma classificação. Se o valor da estatística exceder $\mathcal{X}^2_{rc-1-(r+c-2),\alpha}$ rejeitamos a hipótese nula ao nível de significância $\alpha \times (100)\%$.

Exemplo: Queremos testar se o tipo de curso escolhido é independente do sexo do estudante. A tabela seguinte apresenta os resultados de uma amostra de 400 estudantes classificados de acordo com sexo e curso

	Gestão	Engenh.	Artes & L.	Medicina	Outros	Total
Masc.	21	16	145	2	6	190
	(16.25)	(9.5)	(152)	(7.125)	4.75	
Femin.	14	4	175	13	4	210
	(18.375)	(10.5)	(168)	(7.875)	(5.25)	
Total	35	20	320	15	10	400

onde os valores em parenteses são os valores esperados na hipótese de independência. O valor da estatística do qui-quadrado é $q = 18.93$ e a nível de significância de 1% a hipótese nula de independência seria rejeitada ($\mathcal{X}^2_{4,0.01} = 13.28$). É interessante notar que se analisarmos os termos que contribuem mais para o valor de q eles correspondem a engenharia e medicina. É também de notar que um dos valores esperados é inferior a 5, mas não há grande problema porque a categoria outros contribui pouco para o valor de q.