I.4

CÁLCULO DAS PROBABILIDADES

Página em branco

ADVERTÊNCIA PRÉVIA

Com o capítulo relativo a populações finitas, inicia-se a publicação das nossas lições de Cálculo das Probabilidades no Instituto Superior de Agronomia. Os capítulos seguintes serão publicados à medida que a experiência nos indicar qual a melhor orientação a seguir no ensino desta matéria, tendo sempre em conta a finalidade prática desse ensino e as condições especiais em que tem de ser realizado.

Na preparação das nossas lições serviram-nos de base, principalmente, as seguintes obras:

G. CASTELNUOVO - *Calcolo delle probabilita*, Zanichelli, Bologna, 1933.

D. J. FINNEY - *An introduction to statistical science in Agriculture*, John Wiley Sons, New York, 1953.

CRAMÉR - *Mathematical methods of Statistics*, Princeton University Press, Princeton, 1951.

G. UDNY YULE and M. G. KENDALL - *An introduction to the Theory of Statistics*, Charles Griffin, Londres, 1937.

P. de VARENNES E MENDONÇA - *Noções de Cálculo das Probabilidades*, Instituto Superior de Agronomia, Lisboa, 1950.

Convirá, no entanto, precisar, desde já, que tanto a orientação geral do curso, como muitos dos pormenores didácticos têm carácter pessoal.

Lisboa, Maio de 1955

J. Sebastião e Silva

Página em branco

I.4.3

ADITAMENTO ÀS LIÇÕES DE CÁLCULO DAS PROBABILIDADES

A – Regressões. Ajustamentos. Correlação

1. Formulação geral do problema

Consideremos uma distribuição de frequência absoluta $\nu(x, y)$ de duas variáveis casuais x, y, que tomem um número finito de pares de valores (x_i, y_k), $i=1, 2, \ldots, R$, $k=1, 2, \ldots, S$. Cada par (x_i, y_k) terá a frequência absoluta $\nu(x_i, y_k)$, que pode, em particular, ser nula, o que significa simplesmente que esse par não chegou a ser observado. Daqui se deduzem as *frequências marginais*

$$\nu(x_i) = \sum_{k=1}^{S} \nu(x_i, y_k), \quad \nu(y_k) = \sum_{i=1}^{R} \nu(x_i, y_k)$$

e o *número total de pares observados*, $N = \sum \nu(x_i) = \sum \nu(y_k)$.

A distribuição pode ser representada por uma tabela de contingência, que neste caso se chama, mais vulgarmente, *tabela de correlação*:

y \ x	x_1		x_R	Total
y_1	$\nu(x_1, y_1)$	...	$\nu(x_R, y_1)$	$\nu(y_1)$
y_2	$\nu(x_1, y_2)$	...	$\nu(x_R, y_2)$	$\nu(y_2)$
	...	...	...	...
y_S	$\nu(x_1, y_S)$	...	$\nu(x_R, y_S)$	$\nu(y_S)$
Total	$\nu(x_1)$	...	$\nu(x_R)$	N

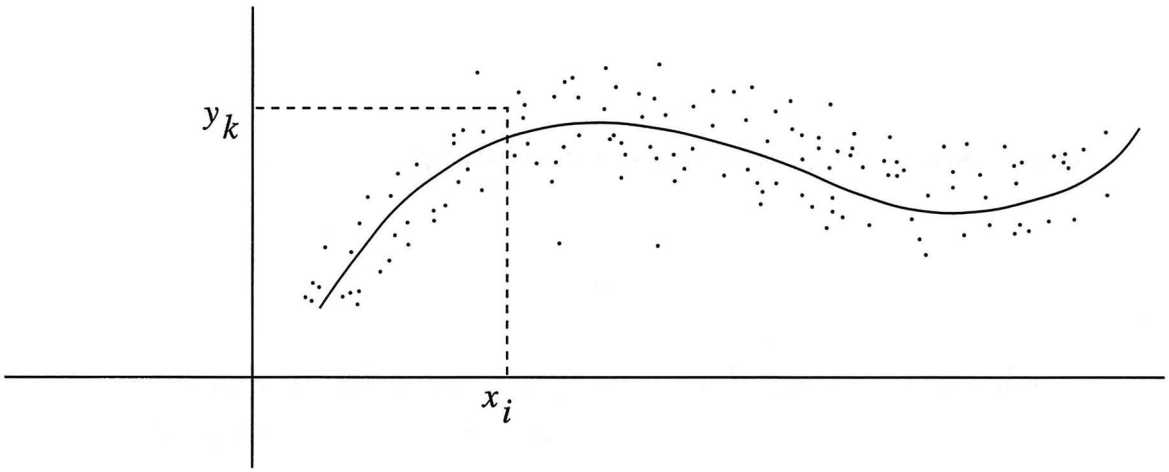


Fig. 1

ou por meio dum gráfico (Fig. 1), em que se marquem os pontos representativos dos pares observados (x_i, y_k) e se aponte, junto de cada um deles, a respectiva frequência $\nu(x_i, y_k)$ (serão omitidos, naturalmente, os pares de frequência *não observados*).

Em vez das frequências absolutas, poderão usar-se, também, as frequências relativas dadas pelas fórmulas

$$\mathrm{fr}(x_i, y_k) = \frac{\nu(x_i, y_k)}{N}, \; i = 1, \ldots, R, \; k = 1, \ldots, S.$$

Serão estas as frequências preferidas nas considerações teóricas. Por sua vez, as frequências absolutas são usadas, de preferência, nas aplicações práticas.

y \ x	1,56	1,58	1,60	1,62	1,64	1,66	1,68	1,70	1,72	1,74	1,76	1,78	1,80	1,82	Total
1,56			1			1									2
1,58			1	1	1	1	1	1							6
1,60			1	2	2	3	2	1	1						12
1,62		1	2	3	2	4	2	3	1	1					19
1,64		1	2	2	2	5	5	4	4	2					27
1,66			1	2	4	5	6	3	2	2	1				26
1,68				1	3	4	5	5	4	2	1	1			26
1,70				1	1	2	6	6	4	3	2	1			26
1,72						3	3	4	4	4	2				20
1,74						1	2	2	3	2	2	1	1	1	15
1,76							2	1	1	2	1	1			8
1,78									1	1	1	1	1		5
Total	0	2	8	12	15	29	34	30	25	19	10	5	2	1	192

Tabela 1 – Alturas x, y de pai e de filho, em 192 pares de pessoas observadas. As alturas são agrupadas em classes de comprimento 0,02 m.

Inúmeros são os exemplos de tais distribuições de frequência que se apresentam na prática.

As variáveis x e y podem ser, por exemplo, o comprimento e a largura das folhas numa dada espécie na variedade de plantas, a produtividade do trigo (em unidades de massa por unidade de área) e o respectivo teor em proteína ou em hidratos de carbono, a altura dos pais e dos respectivos filhos numa dada população (ver Tabela 1), etc., etc.

Quando os pares de valores observados são bastante numerosos, o gráfico (Fig. 1) apresenta-se com o aspecto duma "nebulosa" de pontos. Muitas vezes, nenhuma ordem, nenhum esboço de lei se vislumbra nesse aglomerado de pontos, que aparece, então, como um "caos". Outras vezes, porém, a "nebulosa" é mais densa em certas zonas do que noutras, de tal modo que os pontos parecem acumular-se de preferência à volta de uma curva ou de certas curvas privilegiadas do plano. Tal circunstância sugere naturalmente, com maior ou menor intensidade, a existência de uma *lei* ou *relação funcional aproximada*, $y = f(x)$, entre as variáveis x e y, e até, *algumas vezes*, a hipótese duma relação de causa a efeito entre ambas. Essa função $f(x)$ terá por gráfico, evidentemente, uma das referidas curvas privilegiadas, à volta das quais se adensam os pontos do gráfico.

Pois bem, um dos problemas centrais da Estatística consiste em determinar uma tal função e de avaliar em que medida ela se *ajusta* aos pares de pontos observados: chama-se *regressão* precisamente essa redução do conjunto de pares (x_i, y_k) a uma espécie de função central, $y = f(x)$, que traduza aproximadamente, num traço dominante, o aspecto geral da distribuição dos pontos representativos[1].

Convém, desde já, notar que o simples método de interpolação, tal como foi estudado, não resolve o problema, pois não corresponde à sua natureza. Basta notar, por exemplo, que um mesmo valor x_i de x pode aparecer associado a diversos valores, y_h, y_k, ..., de valores de y, em pares (x_i, y_h), (x_i, y_k), ..., de frequências não nulas (isto é, geometricamente, pode haver vários pontos representativos, com uma mesma abcissa x_i); nestas condições, pelo método da interpolação, a função $f(x)$ não poderia ser unívoca. Dizer que $f(x)$ se "ajusta bem" aos pares de valores observados não significa, de modo nenhum, que, para cada valor x_i de x, o valor $f(x_i)$ da função seja um valor de y observado com x_i (isto é, não significa que o gráfico

(1) – O termo "regressão" foi introduzido por GALTON, que, tendo estudado a correlação entre alturas de pais e de filhos, enunciou a célebre *"lei de regressão": a estatura dos filhos tende a regressar à estatura média da raça* (apesar da forte influência hereditária dos pais). Mais precisamente, se a altura média de um extenso grupo de pais se afasta δ cm da média da raça, a altura média dos filhos afasta-se só $(2/3)\delta$ cm da média da raça.

da função passe exactamente pelos pontos representativos dos pontos observados), mas, apenas, que os desvios $y_k - f(x_i)$ são "pequenos", podendo os desvios não nulos ser atribuídos a *erros* ou a *factores casuais da perturbação*.

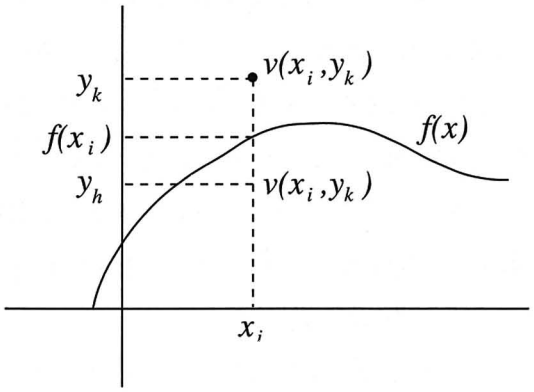


Fig. 2

Assim, o que se pretende no problema da regressão é conciliar as duas seguintes condições:

1) – que a função $f(x)$ seja tão simples quanto possível;

2) – que os desvios $y_k - f(x_i)$ entre os valores y_k de y *observados* e os valores $f(x_i)$ *calculados* (valores *teóricos* ou valores *esperados*) sejam, no seu conjunto, tão pequenos quanto possível.

É claro, desde já, que, quanto mais atendermos a uma destas condições, mais nos afastaremos, em geral, da outra e, portanto, do objectivo em vista. Além disso, as duas condições ainda não têm um enunciado matemático preciso; há em ambas uma grande margem de subjectividade, que autoriza várias interpretações. Se, por exemplo, restringirmos a função $f(x)$ à classe dos polinómios, a condição 1) adquire um significado preciso: a função $f(x)$ será tanto mais simples quanto menor for o grau do polinómio. Mas, se dos polinómios passarmos para as funções de tipo exponencial, logarítmico, etc., já não dispomos de um critério tão seguro; em todo o caso, uma função do tipo $Ce^{\alpha x}$, com C e α constantes, será mais simples que um polinómio (completo) de grau elevado e pode ser que se ajuste muito melhor ao conjunto de pares observados.

Por outro lado, a condição 2) pode receber várias interpretações precisas: pretende-se que seja mínima a *soma dos módulos* $|y_k - f(x_i)|$ dos desvios, ou *a soma dos quadrados* $[y_k - f(x_i)]^2$ dos desvios, ou ainda outra função adequada dos desvios; subentendendo-se que, nas referidas somas, cada parcela seja multiplicada pela frequência absoluta do par (x_i, y_k) a que corresponde. Se, em vez da frequência absoluta, utilizarmos a frequência relativa, a soma dos módulos dos desvios, a soma dos quadrados dos desvios, etc., virão substituídas pelos correspondentes valores médios, que se obtêm dividindo essas somas pelo número total N de pares observados. Em geral, é a soma (ou a média) dos quadrados dos desvios que se procura minimizar (*método dos mínimos quadrados*).

Assim restringido, o problema da regressão pode enunciar-se nos seguintes termos precisos:

Escolhida previamente uma classe $\mathcal{F}$ de funções, determinar uma função f desta classe, de modo que seja mínimo o valor médio dos quadrados dos desvios, dado pela fórmula:

$$M\{[y - f(x)]^2\} = \sum_{i,\,k} [y_k - f(x_i)]^2 \,\mathrm{fr}(x_i, y_k).$$

Diz-se, então, que se trata de *ajustar* uma função da classe $\mathcal{F}$ ao conjunto dos pares de valores observados.

Em particular, $\mathcal{F}$ pode ser a classe das funções lineares: neste caso, a regressão diz-se *linear*. Outras vezes, é necessário recorrer a funções não lineares, cujos gráficos são curvas, e por isso se diz que a regressão é curvilínea.

2. Ajustamentos pelo método dos mínimos quadrados

O método dos mínimos quadrados aplica-se comodamente a uma classe $\mathcal{F}$ qualquer de funções que se possam apresentar sob a forma de combinações lineares de funções dadas, isto é, sob a forma

$$y = f(x) = a_0 u_0 + a_1 u_1 + \cdots + a_n u_n, \tag{1}$$

sendo $u_0, u_1, \ldots, u_n$ funções de x dadas (por exemplo, potências de x, exponenciais, logaritmos, etc., etc.) e $a_0, a_1, \ldots, a_n$ coeficientes a determinar pela condição de mínimo atrás enunciada.

Seja então:

$$u_p = \varphi_p(x), \quad p = 0, 1, \ldots, n, \tag{2}$$

e convencionemos designar por U_{pi} o valor que a variável u_p toma para $x = x_i$, isto é, ponhamos

$$U_{pi} = \varphi_p(x_i) \quad p = 0, 1, \ldots, n, \quad i = 1, 2, \ldots, R. \tag{3}$$

Nestas condições, cada desvio $y_k - f(x_i)$ entre o valor *observado* y_k de y e o valor *calculado* $f(x_i)$ de y será, em virtude de (1), (2) e (3):

$$y_k - f(x_i) = y_k - a_0 U_{0i} - a_1 U_{1i} - \cdots - a_n U_{ni},$$
$$i = 1, 2, \ldots, \quad R, \quad k = 1, 2, \ldots, S.$$

Então, visto que $[y_k - f(x_i)]^2 = [f(x_i) - y_k]^2$, o valor médio dos quadrados dos desvios, que representamos por Q, será $Q = M\{[f(x) - y]^2\}$, ou seja, por extenso·

$$Q = \sum_{i,k} (a_0 U_{0i} + a_1 U_{1i} + \cdots + a_n U_{ni} - y_k)^2 \, \mathrm{fr}(x_i, y_k), \tag{4}$$

em que $i = 1, 2, \ldots R$, $k = 1, 2, \ldots, S$. O valor médio Q é visivelmente uma função de $a_0, a_1, \ldots, a_n$; pode mesmo reconhecer-se que o desenvolvimento do 2.º membro conduz a um polinómio do 2.º grau em $a_0, a_1, \ldots, a_n$. O método dos mínimos quadrados consiste pois, aqui, em determinar os coeficientes $a_0, a_1, \ldots, a_n$, de modo que o valor de Q seja mínimo. Ora, para isso, devemos, segundo a teoria dos máximos e mínimos, começar por achar os possíveis pontos de estacionaridade da função Q, isto é, os sistemas de valores de $a_0, \ldots, a_n$ que anulam todas as derivadas parciais $\dfrac{\partial Q}{\partial a_p}$, $p = 0, 1, \ldots, n$.

Atendendo a que os valores de $\mathrm{fr}(x_i, y_k)$ são constantes, e a que, no quadrado que figura em (4), o coeficiente de a_p é U_{pi}, virá, aplicando as regras de derivação:

$$\frac{\partial Q}{\partial a_p} = 2 \sum_{i,\,k} (a_0 U_{0i} + \cdots + a_n U_{ni} - y_k) U_{pi} \,\mathrm{fr}(x_i, y_k)$$
$$= 2[a_0 M\{u_0 u_p\} + \cdots + a_n M\{u_n u_p\} - M\{y u_p\}] ,$$

visto que:

$$M\{u_0 u_p\} = \sum_{i,\,k} U_{0i} U_{pi} \,\mathrm{fr}(x_i, y_k) ,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$M\{u_n u_p\} = \sum_{i,\,k} U_{ni} U_{pi} \,\mathrm{fr}(x_i, y_k) ,$$

$$M\{y u_p\} = \sum_{i,\,k} y_k U_{pi} \,\mathrm{fr}(x_i, y_k) ,$$

onde $p = 0, 1, \ldots, n$, $i = 1, 2, \ldots, R$, e $k = 1, 2, \ldots, S$.

Por conseguinte, o sistema de equações em $a_0, a_1, \ldots, a_n$

$$\frac{\partial Q}{\partial a_p} = 0, \quad p = 0, 1, \ldots, n ,$$

cujas soluções são os pontos de estacionaridade procurados, será equivalente ao sistema de equações:

$$M\{u_0 u_0\} a_0 + M\{u_1 u_0\} a_1 + \cdots + M\{u_n u_0\} a_n = M\{y u_0\}$$
$$M\{u_0 u_1\} a_0 + M\{u_1 u_1\} a_1 + \cdots + M\{u_n u_1\} a_n = M\{y u_1\}$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$M\{u_0 u_n\} a_0 + M\{u_1 u_n\} a_1 + \cdots + M\{u_n u_n\} a_n = M\{y u_n\}$$

chamadas *equações normais*. É claro que, por ser $u_p u_q = u_q u_p$, quaisquer que sejam p e q, a matriz deste sistema é simétrica.

Notemos, por outro lado, que se tem

$$\frac{\partial^2 Q}{\partial a_p \partial a_q} = 2\, M\{u_p u_q\}, \text{ para } p, q = 0, 1, \ldots, n,$$

e que os valores médios $M\{u_p u_q\}$ são, precisamente, os coeficientes da forma quadrática em $a_0, a_1, \ldots, a_n$, que se obtem desenvolvendo

$$\sum_{i,\,k} (a_0 U_{0i} + a_1 U_{1i} + \cdots + a_n U_{ni})^2 \,\mathrm{fr}(x_i, y_k).$$

Mas esta só pode tomar valores não negativos, e será mesmo, *em geral*, uma forma definida positiva, quando $R > n$[1]. Neste caso, como o seu discriminante é, precisamente, o determinante do anterior sistema de equações normais, segue-se que este é um sistema de CRAMER, cuja solução (ponto de estacionaridade) é um ponto de mínimo local e, *portanto, de mínimo absoluto* (por ser único).

Designando por $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ essa solução, a função procurada será, pois:

$$y = \alpha_0 u_0 + \alpha_1 u_1 + \cdots + \alpha_n u_n .$$

Reciprocamente, é fácil ver que, se o determinante do sistema (discriminante da forma) é $\neq 0$, o ponto de estacionaridade (único) é um ponto de mínimo absoluto.

3. Outra forma das equações normais

Muitas vezes, na prática, em vez dos *valores médios* $M\{u_p u_q\}$ e $M\{y u_p\}$, introduzem-se nas equações normais as *somas* dos valores de cada uma das variáveis $u_p u_q$ e $y u_p$ ($p, q = 0, 1, \ldots, n$). É claro que

---

(1) – Recordemos que uma forma quadrática $\sum_{i,\,k=1}^{n} c_{ik}\xi_i\xi_k$ nas variáveis $\xi_1, \ldots, \xi_n$ (com $c_{ik} = c_{ki}$) se diz *definida positiva,* quando, para todo o sistema de valores de $\xi_1, \ldots, \xi_n$ não simultaneamente nulos, toma valor > 0. Chama-se *discriminante* da forma o determinante $|c_{ik}|$, $i, k = 1, \ldots, n$. A forma será definida positiva, se e só se forem positivos todos os termos duma cadeia própria de menores do discriminante.

isto equivale a trabalhar com frequências absolutas $\nu(x_i, y_k)$, em vez de frequências relativas, $\mathrm{fr}(x_i, y_k)$. Como se tem

$$\mathrm{fr}(x_i, y_k) = \frac{\nu(x_i, y_k)}{N}, \quad \text{para } i = 1, \ldots, R, \quad k = 1, \ldots, S,$$

se designarmos por $[u_p u_q]$, em geral, a soma dos valores da variável $u_p u_q$, isto é, se pusermos

$$[u_p u_q] = \sum_{i,\,k} U_{pi} U_{qi}\, \nu(x_i, y_k) = \sum_{i} U_{pi} U_{qi}\, \nu(x_i)$$

para $p, q = 0, 1, \ldots, n$, e, analogamente:

$$[u_p y] = \sum_{i,\,k} U_{pi}\, y_k \nu(x_i, y_k),$$

será:

$$M\{u_p u_q\} = \frac{[u_p u_q]}{N}, \quad M\{y u_p\} = \frac{[y u_p]}{N}.$$

Então, multiplicando ambos os membros de cada equação normal por N, os coeficientes $M\{u_p u_q\}$ e os termos independentes $M\{y u_p\}$ resultam substituídos por $[u_p u_q]$ e $[u_p y]$, respectivamente, e é agora evidente que o sistema obtido é equivalente ao primeiro.

4. Regressão polinomial

Em particular, as funções $u_0, u_1, \ldots, u_n$ de x podem ser as próprias potências de x:

$$u_0 = x^0 = 1,\ u_1 = x^1 = x,\ u_2 = x^2, \ldots, u_n = x^n.$$

Neste caso, a função $f(x) = \sum a_p u_p$ reduz-se ao polinómio:

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

cujos coeficientes serão determinados pelo sistema de equações

$$\begin{aligned}
&a_0 + M\{x\}a_1 + \cdots + M\{x^n\}a_n = M\{y\}\\
&M\{x\}a_0 + M\{x^2\}a_1 + \cdots + M\{x^{n+1}\}a_n = M\{xy\}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&M\{x^n\}a_0 + M\{x^{n+1}\}a_1 + \cdots + M\{x^{2n}\}a_n = M\{x^n y\}
\end{aligned}$$

visto que $x^0 = 1$ e $M\{1\} = 1$.

Pelo que se disse no n.° anterior, estas equações normais também podem ser escritas sob a forma

$$\begin{aligned}
&Na_0 + [x]a_1 + \cdots + [x^n]a_n = [y]\\
&[x]a_0 + [x^2]a_1 + \cdots + [x^{n+1}]a_n = [xy]\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&[x^n]a_0 + [x^{n+1}]a_1 + \cdots + [x^{2n}]a_n = [x^n y]\,.
\end{aligned}$$

5. Regressão linear. Correlação

Mais particularmente, ainda pode ter-se $n=1$, $u_0=1$ e $u_1=x$. Então, $f(x)$ é uma função linear

$$y = a + bx, \tag{5}$$

onde, para simplificar, pusemos $a_0=a$ e $a_1=b$.

Os coeficientes a e b serão determinados pelo sistema

$$\begin{aligned}
&a + M\{x\}b = M\{y\}\\
&M\{x\}a + M\{x^2\}b = M\{xy\}\,.
\end{aligned} \tag{6}$$

Eliminando a entre as duas equações pelo método de redução, obtem-se

$$b = \frac{M\{xy\} - M\{x\}M\{y\}}{M\{x^2\} - (M\{x\})^2},$$

ou seja,

$$b = \frac{C\{x, y\}}{V\{x\}}, \tag{7}$$

em virtude de propriedades conhecidas da covariância $C\{x, y\}$ e da variância $V\{x\}$.

Por outro lado, a primeira equação do sistema (6) dá

$$a = M\{y\} - M\{x\}b,$$

donde, substituindo em (5) e pondo, para simplificar, $M\{x\} = \bar{x}$, $M\{y\} = \bar{y}$:

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

ou seja,

$$y - \bar{y} = b(x - \bar{x}). \tag{8}$$

Substituindo finalmente b pelo valor dado por (7) (*coeficiente de regressão*), obtem-se a *equação de regressão* procurada. O seu gráfico é, manifestamente, uma recta (*recta de regressão*) que passa pelo ponto $(\bar{x}, \bar{y})$, *centro* da distribuição considerada, visto serem $\bar{x}$ e $\bar{y}$ os valores médios de x e de y.

À fórmula (7) pode dar-se um outro aspecto, que interessa particularmente na prática. Chama-se *coeficiente de correlação* das duas variáveis casuais x, y e representa-se por $\rho_{x,y}$, ou simplesmente por ρ, à covariância dos respectivos desvios reduzidos

$$h = \frac{x - \bar{x}}{\sigma_x} \text{ e } k = \frac{y - \bar{y}}{\sigma_y},$$

onde σ_x e σ_y designam, respectivamente, o desvio padrão de x e o desvio padrão de y. Será, pois,

$$\rho = C\left\{\frac{x-\bar{x}}{\sigma_x}, \frac{y-\bar{y}}{\sigma_y}\right\} = \frac{1}{\sigma_x \sigma_y} C\{x-\bar{x}, y-\bar{y}\},$$

portanto,

$$\rho = \frac{C\{x, y\}}{\sigma_x \sigma_y} = \frac{C\{x, y\}}{\sqrt{V\{x\}V\{y\}}}. \tag{9}$$

O coeficiente de correlação mede o grau de associação das duas variáveis. Como sabemos, tem-se $C\{x, y\} = 0$ se as variáveis casuais x e y são independentes, isto é, se $\mathrm{fr}(x, y) = \mathrm{fr}(x)\ \mathrm{fr}(y)$; mas a recíproca não é verdadeira. Daqui e da fórmula (9) deduz-se que:

O coeficiente de correlação de duas variáveis casuais é nulo, quando as variáveis são independentes.

Porém, a recíproca não é verdadeira: correlação nula não significa, necessariamente, independência casual.

Posto isto, deduz-se de (9) que

$$C\{x, y\} = \rho \sigma_x \sigma_y, \tag{10}$$

o que, por substituição em (7), atendendo a que $V\{x\} = \sigma_x^2$, dá

$$b = \rho \frac{\sigma_y}{\sigma_x}, \tag{11}$$

fórmula esta que permite calcular o *coeficiente de regressão* b a partir do *coeficiente de correlação* ρ.

Observemos, agora, que, para cada valor x_i de x, o valor de y calculado pela equação (8) de regressão é

$$Y_i = \bar{y} + b(x_i - \bar{x})$$

e, portanto, o valor médio dos quadrados dos desvios

$$y_k - Y_i = y_k - \bar{y} - b(x_i - \bar{x})$$

entre os *valores observados* y_k de y e os *valores calculados* Y_i será

$$\begin{aligned} Q &= M\{[y-\bar{y}-b(x-\bar{x})]^2\} \\ &= M\{(y-\bar{y})^2 + b^2(x-\bar{x})^2 - 2b(x-\bar{x})(y-\bar{y})\} \\ &= M\{(y-\bar{y})^2\} + b^2 M\{(x-\bar{x})^2\} - 2bM\{(x-\bar{x})(y-\bar{y})\} \\ &= V\{y\} + b^2 V\{x\} - 2bC\{x, y\}, \end{aligned}$$

donde, atendendo a (10) e (11):

$$Q = V\{y\} + \rho^2\sigma_y^2 - 2\rho^2\sigma_y^2,$$

ou seja, visto que $\sigma_y^2 = V\{y\}$:

$$Q = V\{y\}(1-\rho^2). \tag{12}$$

Como é sempre $Q \geq 0$ e $V\{y\} \geq 0$, daqui se deduz logo que também será sempre $1-\rho^2 \geq 0$, ou seja,

$$-1 \leq \rho \leq 1.$$

Quando se tem $\rho = 1$ ou $\rho = -1$, será $Q = 0$, o que significa que *são nulos todos os desvios* $y_k - Y_i$ *entre os valores observados e os valores calculados por meio de* (8) *e* (11). Por conseguinte:

Quando $|\rho| = 1$, *todos os pares de valores observados* (x_i, y_k) *(com frequência não nula) representam pontos situados sobre a recta de regressão.*

Diz-se, neste caso, que as variáveis x e y estão *completamente correlacionadas*. Mas este é um caso limite que, em geral, não se verifica rigorosamente na prática: as variáveis apresentam-se, apenas, mais ou menos correlacionadas, *positivamente* se $\rho > 0$, *negativamente* se $\rho < 0$. O caso oposto é aquele em que $\rho = 0$ (*variáveis não correlacionadas*), de que é um caso particular, como vimos, o das variáveis casualmente independentes.

Notemos ainda que, geralmente, os pares (x_i, y_k) constituem uma *amostra* de pares de valores de duas variáveis *x, y, contínuas*. Assim, o coeficiente ρ determinado e a equação de regressão estabelecida referem-se a essa amostra e não à totalidade dos pares de valores possíveis. Para se ter uma ideia justa do significado de ρ relativamente a essa totalidade, efectua-se sobre este coeficiente uma *prova de significação* (ou *teste de significância*), em que se toma para número de graus de liberdade, precisamente, o número N de pares observados diminuído de 1. (Sobre este ponto e sobre um exemplo concreto de regressão linear, ver as folhas anteriores).

6. Segunda recta de regressão

É claro que, assim como procurámos exprimir y como função linear de x (*regressão linear de y sobre x*), assim, também, poderíamos procurar exprimir x como função linear de y (*regressão de x sobre y*). Trocando os papéis de x e de y, imediatamente se acha a equação de regressão de x sobre y:

$$x-\bar{x} = b_{xy}(y-\bar{y}),$$

em que b_{xy} (*coeficiente de regressão de x sobre y*) é dado pela fórmula

$$b_{xy} = \rho \frac{\sigma_x}{\sigma_y}.$$

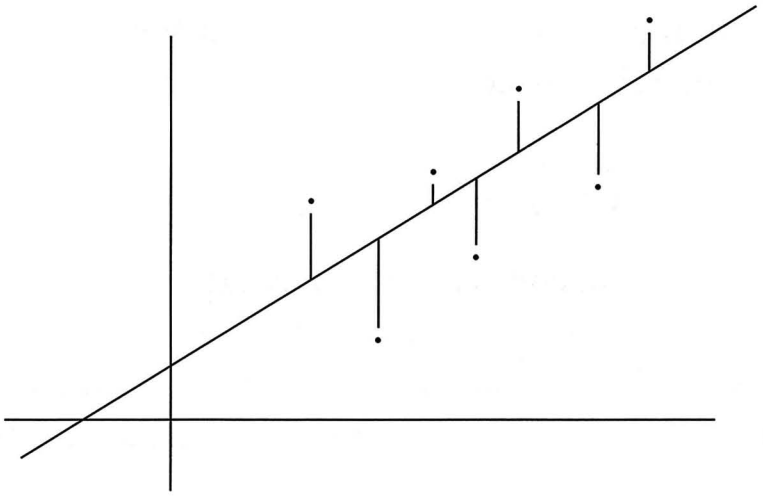
Para evitar confusões, o coeficiente de regressão de y sobre x passará a ser designado por b_{yx}, tendo-se, como vimos,

$$b_{yx} = \rho \frac{\sigma_y}{\sigma_x}.$$

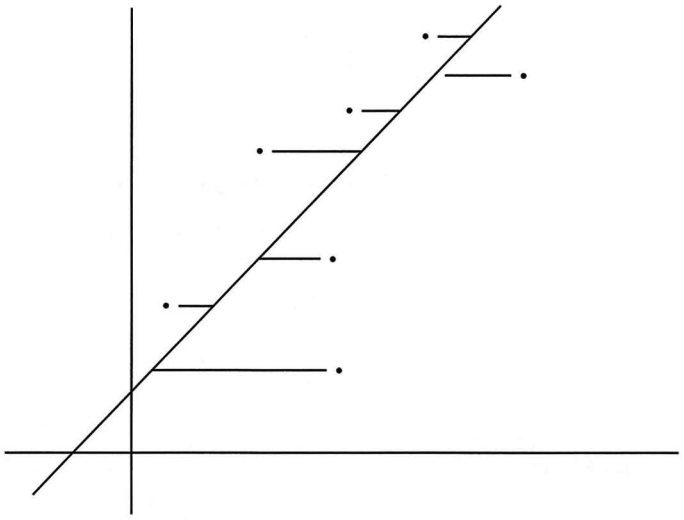
Mostram estas fórmulas que, se for $\rho \geq 0$, será também $b_{yx} \geq 0$ e $b_{xy} \geq 0$: y cresce com x e x cresce com y (*correlação directa ou positiva*); se for $\rho \leq 0$, será $b_{yx} \leq 0$ e $b_{xy} \leq 0$ (*correlação inversa ou negativa*).

É claro que a segunda equação de regressão é a que torna mínima a soma dos quadrados dos desvios $x_i - X_k$ entre os valores de x observados e os valores de x calculados.

Como ambas as rectas de regressão passam pelo ponto $(\bar{x}, \bar{y})$, centro da distribuição de x e y, é fácil ver que tais rectas coincidem, se e só se for $|\rho| = 1$, caso em que, segundo vimos, todos os pontos representativos estão sobre a primeira (e, portanto, sobre a segunda) recta de regressão (*correlação completa*). Excluído este caso, o ângulo das rectas será tanto maior quanto menor for $|\rho|$, sendo igual a 90° se $\rho = 0$ (rectas paralelas aos eixos).



1.ª recta de regressão $y - \bar{y} = b_{yx}(x - \bar{x})$



2.ª recta de regressão $x - \bar{x} = b_{xy}(y - \bar{y})$

7. Organização prática dos cálculos

Visto que se tem $C\{x, y\} = M\{a, y\} - M\{x\}\, M\{y\}$, $V\{x\} = M\{x^2\} - -(M\{x\})^2$, $V\{y\} = M\{y^2\} - (M\{y\})^2$, o coeficiente de correlação pode ser calculado pela fórmula

$$\rho = \frac{\frac{1}{N}\sum_{i,k} n_{ik} x_i y_k - \bar{x}\bar{y}}{\sqrt{\frac{1}{N}\sum_i n_i x_i^2 - \bar{x}^2}\ \sqrt{\frac{1}{N}\sum_k m_k y_k^2 - \bar{y}^2}},$$

em que, para simplificar, pusemos

$$n_{ik} = \nu(x_i, y_k),\ n_i = \nu(x_i),\ m_k = \nu(y_k).$$

Muitas vezes, é aconselhável fazer uma mudança de origem e uma mudança de unidade de medida, para simplificar os cálculos, de modo análogo ao que se indicou para o cálculo do valor médio e da variância. Se pusermos

$$x = \alpha u + X_0, \qquad y = \beta v + Y_0,$$

sendo α, β, X_0 e Y_0 constantes (X_0 e Y_0 chamadas *médias arbitrárias*), virá $x - \bar{x} = \alpha(u - \bar{u})$, $y - \bar{y} = \beta(v - \bar{v})$, donde

$$C\{x, y\} = \alpha\beta C\{u, v\}, \qquad V\{x\} = \alpha^2 V\{u\},$$
$$V\{y\} = \beta^2 V\{v\}$$

e, portanto, $\bar{v}^2$

$$\rho = \frac{\frac{1}{N}\sum_{i,k} n_{ik} u_i v_k - \bar{u}\,\bar{v}}{\sqrt{\frac{1}{N}\sum_i n_i u_i^2 - \bar{u}^2}\ \sqrt{\frac{1}{N}\sum_k m_k v_k^2 - \bar{v}^2}}. \tag{13}$$

Por exemplo, se quisermos aplicar estes resultados à tábua de correlação apresentada no n.° 1, podemos tomar para *médias arbitrárias* os valores

$$X_0 = x_8 = 1{,}70, \qquad Y_0 = y_7 = 1{,}68.$$

Então, pondo $\alpha = \beta = 1$, a mudança de variáveis será

$$x = u + 1{,}70, \qquad y = v + 1{,}68.$$

Posto isto, deverão calcular-se sucessivamente, por um lado, os valores de

$$n_i,\ u_i,\ n_i u_i,\ n_i u_i^2,\ u_i \sum_k n_{ik} v_k,$$

bem como as respectivas somas, e, por outro lado,

$$m_k,\ v_k,\ m_k v_k,\ m_k v_k^2,\ v_k \sum_i n_{ik} u_i.$$

É claro que deve ter-se

$$\sum_i u_i \sum_k n_{ik} v_k = \sum_k v_k \sum_i n_{ik} u_i = \sum_{i,\,k} n_{ik} u_i v_k,$$

o que fornece uma verificação. Feitos estes cálculos, resta só aplicar a fórmula (13).

Note-se como, por este processo, ficam calculados $\bar{u}$, $\bar{v}$, $V\{u\}$, $V\{v\}$, o que permite achar rapidamente $\bar{x} = \alpha\bar{u} + X_0$, $\bar{y} = \beta\bar{v} + Y_0$, $V\{x\} = \alpha^2 V\{u\}$, $V\{y\} = \beta^2 V\{v\}$ e, portanto, as rectas de regressão.

Para a regressão polinomial pode seguir-se um processo análogo. Suponhamos, por exemplo, que se pretende ajustar um polinómio ao seguinte conjunto de pares:

x	1,0	1,5	2,0	2,5	3,0	3,5	4,0
y	1,1	1,3	1,6	2,0	2,7	3,4	4,1

Começaremos, então, por representar estes pares graficamente e observar qual o tipo de parábola (isto é, o grau de polinómio) que convém escolher. Neste caso, o gráfico sugere uma parábola do 2.º grau,

$$y = a_0 + a_1 x + a_2 x^2.$$

Para achar os seus coeficientes, convém fazer a mudança de variável $u = 2x - 5$, que dá, para os valores de x atrás indicados, os valores de u

$$-3, \quad -2, \quad -1, \quad 0, \quad 1, \quad 2, \quad 3.$$

Os cálculos dispõem-se no seguinte quadro:

x	u	y	u^2	u^4	uy	u^2y
1,0	−3	1,1	9	81	−3,3	9,9
1,5	−2	1,3	4	16	−2,6	5,2
2,0	−1	1,6	1	1	−1,6	1,6
2,5	0	2,0	0	0	0,0	0,0
3,0	1	2,7	1	1	2,7	2,7
3,5	2	3,4	4	16	6,8	13,6
4,0	3	4,1	9	81	12,3	36,9
	0	16,2	28	196	14,3	69,9

Daqui, para achar $y = b_0 + b_1 u + b_2 u^2$, deduzem-se as equações normais em b_0, b_1, b_2:

$$7b_0 + 0b_1 + 28b_2 = 16{,}2$$

$$0b_0 + 28b_1 + 0b_2 = 14{,}3$$

$$28b_0 + 0b_1 + 196b_2 = 69{,}9$$

Obtém-se, então,

$$y = 2{,}07 - 0{,}511u + 0{,}061u^2,$$

donde, passando à variável inicial, x:

$$y = 6{,}15 - 2{,}24x + 0{,}24x^2.$$

8. Ajustamentos com mudanças não lineares de variáveis

Muitas vezes, o gráfico dos pares de valores observados, ou o conhecimento que se tem *a priori* do fenómeno a estudar, aconselham um tipo de funções que não se exprime como combinação linear de funções conhecidas $u_0, u_1, \ldots, u_n$ (cf. n.º 2), mas que se converte numa função desse tipo por conveniente passagem a logaritmos.

Tal é, por exemplo, o caso das funções do tipo

$$y = Ca^x \quad \text{(exponencial)}$$

com C e a constantes. Passando a logaritmos e pondo $a_0 = \log C$, $a_1 = \log a$, virá

$$\log y = a_0 + a_1 x.$$

Poderá, então, aplicar-se o método dos mínimos quadrados a $\log y$, em vez de o fazer para y. Note-se que, mediante esta transformação, a função foi *linearizada*. Na prática, usa-se nestes casos papel *semi-logarítmico*, com *escala logarítmica* no eixo dos yy e *escala natural*

no eixo dos xx, começando por fazer a marcação dos pontos neste papel: se os pontos se apresentarem *aproximadamente* em linha recta, é recomendável a regressão linear para $\log y$. Pode acontecer que o gráfico no papel semi-logarítmico aconselhe uma parábola de grau igual ou superior a 2, imagem dum polinómio $P(x)$. É claro que, neste caso, o ajustamento de

$$\log y = \log C + P(x) \log a$$

equivale ao de

$$y = Ca^{P(x)}.$$

Além do papel semi-logarítmico, pode também utilizar-se *papel logarítmico* (com escalas logarítmicas em ambos os eixos). Então, se os pontos marcados estiverem *aproximadamente* em linha recta, torna-se aconselhável para y uma expressão do tipo

$$y = Cx^{\alpha} \qquad \text{(potência de expoente real } \alpha),$$

pois que, por logaritmização, se passa a

$$\log y = \log C + \alpha \log x,$$

relação linear entre $\log y$ e $\log x$. O método dos mínimos quadrados será pois, neste caso, aplicado às variáveis $\log x$, $\log y$, e não às variáveis x, y.

Outras mudanças de variáveis poderão ainda impor-se em diversos casos. Seja, por exemplo, uma função do tipo

$$y = \frac{1}{a + bx},$$

cujo gráfico, como sabemos, é uma hipérbole de assimptotas $y=0$, $x=-a/b$, visto tratar-se duma *função homográfica*. Neste caso, por ser

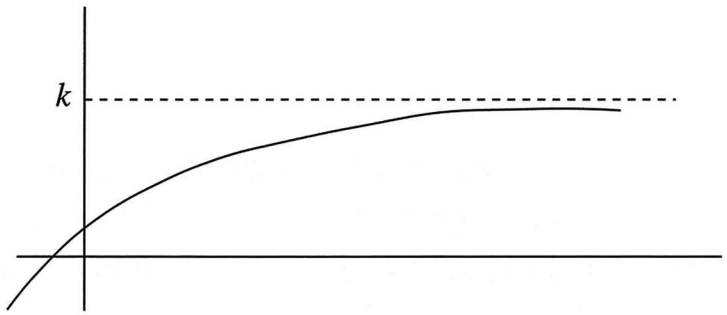
$$\frac{1}{y} = a + bx,$$

o método dos mínimos quadrados poderá ser aplicado às variáveis x e $1/y$, para determinação de a e b.

Consideremos, ainda, uma função do tipo

$$y = k - Ce^{-\alpha x}, \quad \text{com } \alpha > 0,$$

sendo k uma constante conhecida e C, α constantes a determinar. Como se tem $\lim_{x \to +\infty} y = k$,

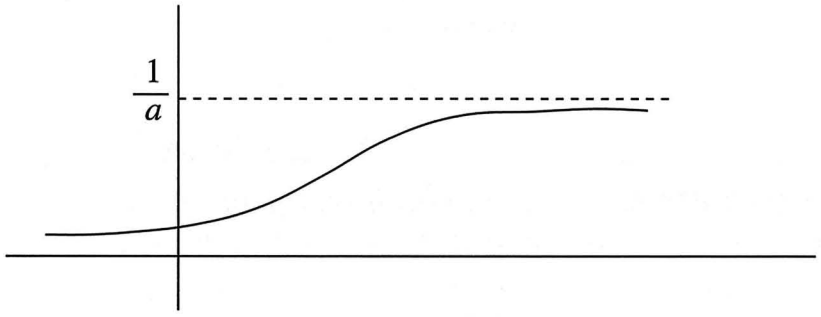


a recta $y=k$ é uma assimptota da curva geralmente conhecida *a priori* (este tipo de função é comum em fenómenos biológicos de crescimento). Ora, por ser

$$\log(k-y) = \log C - \alpha x,$$

o que está indicado, neste caso, é uma regressão linear entre as variáveis x e $\log(k-y)$, tornando-se, ainda aqui, aconselhável o uso do papel semi-logarítmico.

Seja, finalmente, uma função do tipo $y = \dfrac{1}{a + Ce^{-\alpha x}}$, sendo a uma constante conhecida, C e α constantes a determinar ($\alpha > 0$).



Como

$$\lim_{x\to+\infty} y = \frac{1}{a}, \quad \lim_{x\to-\infty} y = 0,$$

a curva tem por assimptotas as rectas $y = 1/a$ e $y = 0$; é fácil ver ainda que apresenta um ponto de inflexão: dá-se-lhe o nome de *curva logística*, e é especialmente indicada para a interpretação de certos fenómenos. Uma curva com análoga configuração (de S deformado), tendo por assimptotas as rectas $y = 0$ e $y = 1$, é a que representa a cumulante da distribuição normal; porém, a sua expressão analítica é diversa. Por ser neste caso

$$\log\left(\frac{1}{y} - a\right) = \log C - \alpha x,$$

o que haverá a fazer é aplicar o método dos mínimos quadrados às variáveis x e $\log\left(\frac{1}{y} - a\right)$, podendo ainda usar-se, com vantagem, o papel semi-logarítmico.

9. Regressão múltipla. Índice de correlação, em geral

Suponhamos agora que, em vez de duas, se trata, em geral, de $m + 1$ variáveis casuais

$$x_1, x_2, \ldots, x_m, y,$$

das quais se observaram N sistemas de valores

$$(x_{1,i_1}, x_{2,i_2}, \ldots, x_{m,i_m}, y_k),$$

tendo cada um deles uma determinada frequência (absoluta ou relativa) não nula, e que se pretende exprimir aproximadamente y como função de $x_1, x_2, \ldots, x_m$. Se essa função for da forma

$$y = a_0 u_0 + a_1 u_1 + \cdots + a_n u_n,$$

sendo $u_0, u_1, \ldots, u_n$ funções conhecidas de $x_1, \ldots, x_m$ e $a_0, a_1, \ldots, a_n$ parâmetros a determinar, continua aplicável, *mutatis mutandis*, tudo o que foi dito no n.º 2.

Em particular, pode ter-se, precisamente, $m = n$ e

$$u_0 = 1,\ u_1 = x_1,\ u_2 = x_2, \ldots,\ u_n = x_n.$$

A função a ajustar será então uma função linear das n variáveis $x_1, \ldots, x_n$ (*regressão linear múltipla*):

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n,$$

em que os coeficientes $a_0, a_1, \ldots, a_n$ serão determinados pelo método dos mínimos quadrados, que, neste caso, conduz às equações normais:

$$\begin{aligned}
&Na_0 + [x_1]a_1 + [x_2]a_2 + \cdots + [x_n]a_n = [y]\\
&[x_1]a_0 + [x_1^2]a_1 + [x_1x_2]a_2 + \cdots + [x_1x_n]a_n = [x_1y]\\
&[x_2]a_0 + [x_1x_2]a_1 + [x_2^2]a_2 + \cdots + [x_2x_n]a_n = [x_2y]\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&[x_n]a_0 + [x_nx_1]a_1 + [x_nx_2]a_2 + \cdots + [x_n^2]a_n = [x_ny].
\end{aligned}$$

Pode ainda, com vantagem, recorrer-se na prática a mudanças de variáveis que se traduzam por mudança de origem e mudanças de unidade nos diversos eixos.

A imagem geométrica da equação de regressão é um *hiperplano* (um plano, se $n=3$). Prova-se facilmente que o hiperplano de regressão passa pelo centro da distribuição dada, isto é, pelo ponto

$$(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n, \bar{y}),$$

cujas coordenadas em $\mathbf{R}^{n+1}$ são os valores médios das variáveis $x_1, x_2, \ldots, x_n, y$.

Em vez da regressão linear (múltipla), é muitas vezes necessário considerar uma regressão não linear, polinomial ou não. Por exemplo, no caso de três variáveis x, y, z, pode tornar-se aconselhável ajustar, aos sistemas de valores observados, um polinómio do tipo

$$z = a + bx + cy + dx^2 + exy + fy^2, \tag{14}$$

cuja imagem é um paraboloide, elíptico ou hiperbólico. A determinação dos coeficientes será, então, feita pelas seis equações normais seguintes:

$$Na + [x]b + [y]c + [x^2]d + [xy]e + [y^2]f = [z]$$

$$[x]a + [x^2]b + [xy]c + [x^3]d + [x^2y]e + [xy^2]f = [xz]$$

$$[x^2]a + [x^3]b + [x^2y]c + [x^4]d + [x^3y]e + [x^2y^2]f = [x^2z]$$

$$[y]a + [xy]b + [y^2]c + [x^2y]d + [xy^2]e + [y^3]f = [yz]$$

$$[y^2]a + [xy^2]b + [y^3]c + [x^2y^2]d + [xy^3]e + [y^4]f = [y^2z]$$

$$[xy]a + [x^2y]b + [xy^2]c + [x^3y]d + [x^2y^2]e + [xy^3]f = [xyz]$$

visto que, neste caso, se pode tomar

$$u_0 = 1,\ u_1 = x,\ u_2 = x^2,\ u_3 = y,\ u_4 = y^2,\ u_5 = xy.$$

Note-se como as equações normais se deduzem de (14) segundo uma lei simples, que se torna variável reparando primeiro nos segundos membros das equações escritas. Convem ainda lembrar, aqui, que os colchetes com uma só variável representam somatórios simples, os colchetes com duas variáveis, somatórios duplos, os colchetes com três variáveis, somatórios triplos, etc.

O coeficiente de correlação ρ foi definido no n.° 5 apenas para o caso da regressão linear simples; mas vimos que se tem

$$Q = V\{y\}(1-\rho^2) = \sigma_y^2(1-\rho^2),$$

sendo Q o valor médio dos *quadrados residuais*, isto é, dos quadrados dos desvios entre os valores de y observados e os valores de y calculados. Desta fórmula se deduz

$$\rho^2 = 1 - \frac{Q}{\sigma_y^2},$$

donde a ideia de tomar para *índice de correlação*, no caso geral (regressão linear ou não linear, simples ou múltipla), o número R dado pela fórmula

$$R^2 = 1 - \frac{S_y^2}{\sigma_y^2}, \text{ com } S_y = \sqrt{Q},$$

em que $S_y^2 (\leqslant \sigma_y^2)$ representa *a parte da variância*, σ_y^2, *de y, que não pode ser explicada pela regressão, isto é, pela relação funcional estabelecida entre as variáveis, e que poderá atribuir-se a factores casuais da perturbação* (no caso de uma correlação elevada) *ou a outras variáveis que possam influir significativamente em y e que não foram consideradas*.

Muitas vezes, ao estudar a correlação (linear ou não linear) entre três ou mais variáveis $x_1, x_2, \ldots, x_m, y$, considera-se não só a *correlação total*, mas também as *correlações parciais* destas variáveis duas a duas, três a três, etc., para avaliar a influência das variáveis $x_1, \ldots, x_m$ sobre y (variável que se pretende exprimir como função das primeiras) e daquelas entre si. Pode acontecer que y seja "praticamente" independente de alguma ou algumas das variáveis $x_1, \ldots, x_m$, ou algumas destas se exprimam significativamente como função das restantes, o que tornará aconselhável a supressão de tais variáveis ou a sua substituição por funções das outras, na fórmula final.

Note-se que existem *provas de significação*, não só para os coeficientes de correlação, como ainda para os de regressão, atendendo a que os sistemas de valores observados constituem, apenas, amostras de populações, nos casos correntes da prática.

Pode, finalmente, acontecer que a função a ajustar não seja do tipo geral

$$y = a_0 u_0 + a_1 u_1 + \cdots + a_n u_n,$$

atrás considerado. Neste caso, podem ainda ensaiar-se mudanças de variáveis, tais como as que foram apresentadas em exemplos no n.º anterior.

10. Nota sobre as notações

Como se disse há pouco, os sistemas de valores observados nos casos correntes da prática constituem apenas amostras duma população base. É então aconselhável representar os diversos parâmetros por letras latinas, para os distinguir dos valores dos parâmetros na população, designados pelas letras gregas correspondentes; por exemplo, s_x (em vez de σ_x), para o desvio padrão de x; r (em vez de ρ), para o coeficiente de correlação, etc.

B – Distribuições de STUDENT e de FISHER. Suas aplicações

1. A melhor estimativa do desvio padrão deduzida duma amostra

Consideremos n valores casuais independentes

$$x_1, x_2, \ldots, x_n,$$

(possivelmente repetidos) de uma variável x, normalmente distribuída com valor médio μ e desvio padrão σ. O sistema $(x_1, x_2, \ldots, x_n)$ constitui, pois, uma *amostra casual* de uma população normal $N(\mu, \sigma)$ (essa população pode ser constituída, nos casos da prática, pelas árvores dum povoamento homogéneo, pelas percentagens da gordura do leite numa raça de vacas, etc., etc.).

Considerando a amostra como nova variável (no espaço $\mathbf{R}^n$ das amostras de *tamanho n*), as variáveis casuais $x_1, \ldots, x_n$ terão todas a distribuição de x, isto é, serão variáveis $N(\mu, \sigma)$. Então, segundo a propriedade reprodutiva da distribuição normal, a média

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} x_1 + \frac{1}{n} x_2 + \cdots + \frac{1}{n} x_n,$$

(*função linear* de $x_1, \ldots, x_n$) será também normalmente distribuída, com o valor médio

$$M\{\bar{x}\} = \frac{1}{n} M\{x_1\} + \frac{1}{n} M\{x_2\} + \cdots + \frac{1}{n} M\{x_n\}$$

$$= \frac{1}{n} nM\{x\} = M\{x\} = \mu,$$

visto que $M\{x_1\} = \cdots = M\{x_n\} = M\{x\}$, e com a variância

$$V\{\bar{x}\} = \frac{1}{n^2} V\{x_1\} + \cdots + \frac{1}{n^2} V\{x_n\}$$

$$= \frac{1}{n^2} nV\{x\} = \frac{\sigma^2}{n},$$

donde o desvio padrão

$$\sigma_{\bar{x}} = \sqrt{V\{\bar{x}\}} = \frac{\sigma}{\sqrt{n}}.$$

A distribuição de $\bar{x}$ será, pois,

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

e o desvio reduzido da variável $\bar{x}$ será

$$\tau = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma} \sqrt{n},$$

com a distribuição normal estandardizada, N(0, 1).

Em muitas questões da prática é desconhecido (ou até hipotético) o desvio padrão, σ, da população base, e é-se tentado a substituí-lo pelo desvio padrão, s, duma amostra $(x_1, \ldots, x_n)$ da população. Tem-se, então,

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

onde, como vimos, $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$. Porém, um cálculo simples mostra que a esperança matemática (ou valor médio) de s^2 é

$$E\{s^2\} = \frac{n-1}{n} \sigma^2.$$

Assim, o *valor esperado* de s^2 não é a variância σ^2 da população. Por isso, como se tem

$$E\left\{\frac{n}{n-1} s^2\right\} = \sigma^2,$$

toma-se como a *melhor estimativa* de σ^2 (na amostra considerada) o valor

$$\frac{n}{n-1}s^2 = \frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

e, portanto, como a *melhor estimativa* de σ (na amostra), o valor

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

Dum modo geral, a melhor estimativa dum parâmetro da população, deduzida de uma ou mais amostras, é representada pela letra grega que designa esse parâmetro, encimada dum acento circunflexo. Será, pois,

$$\hat{\sigma} = s\sqrt{\frac{n}{n-1}}.$$

Este factor $\sqrt{\frac{n}{n-1}}$, que permite passar de s para a melhor estimativa de σ, é chamado de *correcção de* BESSEL, a qual pode ser dispensada quando n é bastante grande, por ser então praticamente igual a 1.

2. Distribuição de *t* de STUDENT

Como vimos atrás, o desvio reduzido de $\bar{x}$, ou seja,

$$\tau = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}},$$

é uma variável N(0, 1). Porém, se substituirmos σ pela sua melhor estimativa, calculada na amostra $(x_1, x_2, \ldots, x_n)$, obtém-se a seguinte variável, estimativa de τ:

$$t = \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{x}-\mu}{s/\sqrt{n-1}},$$

que depende de $x_1, \ldots, x_n$, não só por intermédio de $\bar{x}$, como também por intermédio de s e que, por isso, já não segue a distribuição normal, embora desta se aproxime, com valor médio 0 e desvio padrão 1, quando n é bastante elevado. Foi STUDENT quem primeiro abordou e resolveu o problema da distribuição exacta da referida variável t, função das n variáveis $x_1, \ldots, x_n$, todas $N(\mu, \sigma)$; obteve, assim, a chamada *distribuição de STUDENT com* $n-1$ *graus de liberdade*, cuja função de densidade é

$$S_{n-1}(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}$$

que, como se vê, não depende dos parâmetros da população inicial. (Note-se como aqui intervem a função Γ de EULER, o que sucede em várias distribuições estudadas em Estatística).

Como era de esperar, o gráfico de $S_n(t)$ assemelha-se à curva de GAUSS; é, como esta, simétrica em relação ao eixo das ordenadas, mas um pouco mais achatada, tanto mais quanto menor for n.

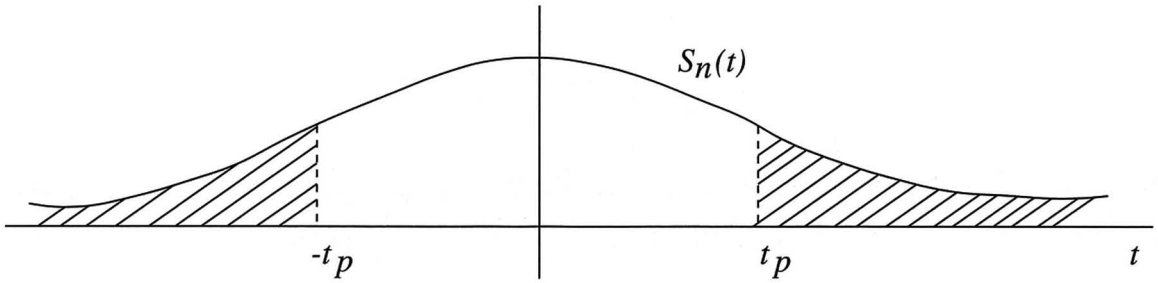
No que se segue, designaremos genericamente por P uma probabilidade e por p o número $100P$. Será, assim,

$$P = \frac{p}{100} = p\%\,.$$

Ora, a probabilidade P de o desvio t exceder em módulo um certo limite t_p (onde $p = 100\,P$) é dada pela fórmula

$$P = Pr(|t| \geqslant t_p) = 2\int_{t_p}^{+\infty} S_n(t)\,\mathrm{d}t,$$

visto que o gráfico de $S_n(t)$ é simétrico em relação ao eixo das ordenadas. Este valor de P representa, com efeito, a área do domínio ilimitado que se indica a tracejado na figura:



No final destes apontamentos é dada uma tabela em que, para diferentes valores de ν (número de graus de liberdade) se indicam os valores de t_p correspondentes às probabilidades $P=0{,}05$ (5%), $P=0{,}01$ (1%), $P=0{,}001$ (0,1%). A tabela está, pois, construída para a *função inversa da anterior.*

3. A melhor estimativa de σ deduzida a partir de várias amostras

Suponhamos agora que, em vez de uma, se trata de k amostras de uma mesma população N(μ, σ). Sejam $n_1, n_2, \ldots, n_k$, os tamanhos dessas amostras, e $s_1, s_2, \ldots, s_k$, os respectivos desvios padrão. Prova-se então, como para o caso duma só amostra, que a melhor estimativa de σ baseada nessas amostras é dada pela fórmula

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + \cdots + n_k s_k^2}{n_1 + n_2 + \cdots + n_k - k},$$

querendo-se com isto dizer que o valor esperado de $\hat{\sigma}^2$ é σ^2, isto é, que

$$\mathrm{E}\{\hat{\sigma}^2\} = \sigma^2.$$

4. Distribuição da diferença entre duas médias

Consideremos duas amostras casuais independentes

$$X = (x_1, x_2, \ldots, x_m) \text{ e } Y = (y_1, y_2, \ldots, y_n)$$

de uma mesma população normal $N(\mu, \sigma)$, e sejam $\bar{x}$, $\bar{y}$ as respectivas médias. O valor esperado para $\bar{x} - \bar{y}$ será, então,

$$M\{\bar{x} - \bar{y}\} = M\{\bar{x}\} - M\{\bar{y}\} = \mu - \mu = 0.$$

Por outro lado, já sabemos que serão $\sigma/\sqrt{m}$ e $\sigma/\sqrt{n}$ os desvios padrão, respectivamente, de $\bar{x}$ e $\bar{y}$, donde

$$V\{\bar{x}\} = \frac{\sigma^2}{m}, \quad V\{\bar{y}\} = \frac{\sigma^2}{n}$$

e, portanto, visto que $\bar{x}$ e $\bar{y}$ são independentes,

$$\begin{aligned} V\{\bar{x} - \bar{y}\} &= V\{\bar{x} + (-1)\bar{y}\} \\ &= V\{\bar{x}\} + (-1)^2 V\{\bar{y}\} \\ &= \frac{\sigma^2}{m} + \frac{\sigma^2}{n}. \end{aligned}$$

O desvio padrão de $\bar{x} - \bar{y}$ será, pois,

$$\sigma\{\bar{x} - \bar{y}\} = \sqrt{V\{\bar{x} - \bar{y}\}} = \sigma\sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Notemos, ainda, que por ser

$$\bar{x} - \bar{y} = \frac{1}{m}(x_1 + \cdots + x_m) - \frac{1}{n}(y_1 + \cdots + y_n)$$

(função linear das variáveis normais $x_1, x_2, \ldots, x_m, y_1, \ldots, y_n$), será também $\bar{x} - \bar{y}$ uma variável normal, cujo desvio reduzido,

$$(1) \qquad \tau = \frac{(\bar{x}-\bar{y}) - M(\bar{x}-\bar{y})}{\sigma\{\bar{x}-\bar{y}\}} = \frac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

terá, portanto, a distribuição N(0, 1).

Desconhecendo-se o valor de σ, é-se levado a substituir σ pela sua melhor estimativa, $\hat{\sigma}$, baseada nas duas amostras consideradas. Tem-se, pelo que vimos no n.° anterior,

$$\hat{\sigma}^2 = \frac{ms_1^2 + ns_2^2}{m+n-2},$$

sendo s_1 e s_2 os desvios padrão de cada uma das amostras. Substituindo, então, σ por $\hat{\sigma}$ em (1), o desvio reduzido τ resulta substituído pelo estatístico

$$t = \frac{\bar{x}-\bar{y}}{\hat{\sigma}\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{\bar{x}-\bar{y}}{\sqrt{ms_1^2+ns_2^2}} \cdot \sqrt{\frac{m+n-2}{\frac{1}{m}+\frac{1}{n}}}.$$

Pois bem, demonstra-se que a distribuição desta variável é ainda a distribuição de STUDENT *com* $m+n-2$ *graus de liberdade.*

5. Prova do *t* (de significação)

Os importantes resultados anteriores aplicam-se em provas de significação, correntemente usadas na prática e das quais distinguiremos dois tipos:

a) – Determinou-se a média $\bar{x}$ duma amostra de n valores duma variável normal x e pretende-se saber se, em face desse resultado, é ou não aceitável a hipótese de que a média μ na população base tem um certo valor μ_0. A "hipótese nula" consiste, pois, neste caso, em supor $\mu = \mu_0$. Para aplicar a prova do *t,* há que fixar, previamente, um *nível de significação p%* (geralmente 5%, 1% ou 0,1%).

Ora, já sabemos que o estatístico[1]

(1) – Convem ter presente que *t* é uma estimativa do desvio reduzido τ.

(2) $$t = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

tem a distribuiçao de STUDENT com $n-1$ graus de liberdade. A prova de significação consiste, então, em substituir μ por μ_0 em (2), calcular o valor de t correspondente,

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

e procurar, na tabela do t de STUDENT, para $n-1$ graus de liberdade [1] o valor t_p correspondente ao nível $p\%$ escolhido ($P = p/100$). Já sabemos que é, então,

$$Pr(|t| \geqslant t_p) = \frac{p}{100}.$$

Portanto, se o valor de t calculado é superior a t_p, rejeita-se a hipótese nula ao nível de $p\%$ escolhido (visto que a probabilidade de um desvio $\geq t$ em módulo é tanto menor quanto maior for t). Se o valor de t calculado for inferior a t_p, aceita-se a hipótese nula ao nível de $p\%$ ou aguarda-se ulterior informação.

Quando a amostra é bastante grande, a distribuição de t aproxima-se da normal, podendo, então, ser substituída por esta.

Exemplo – Uma amostra de 9 homens de uma grande cidade deu, para as suas alturas, uma média de 1,72 m, e uma variância corrigida ($\hat{\sigma}^2$) de 0,13 m^2. Deseja-se saber se este resultado é compatível, ao nível de 5%, com a hipótese de que a média na cidade é 1,70 (admitindo que a distribuição das alturas na cidade é sensivelmente normal).

Então:

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}} \sqrt{n} = \frac{0{,}02 \times 3}{0{,}36} = 0{,}17.$$

---

(1) – Ver tabela final. *Não esquecer que nesta tabela* ν *indica o número de graus de liberdade, que no caso da fórmula é* $n-1$.

Ora, o valor t_5 da t correspondente ao nível de 5%, para $9-1=8$ graus de liberdade, é 2,306; como o valor de t calculado (0,17) é muito inferior a t_5, a hipótese é confirmada ao nível de 5% (visto que a probabilidade de um valor casual de t igual ou superior em módulo a 0,17 é bastante superior a 0,05).

b) – Determinaram-se as médias $\bar{x}$ e $\bar{y}$ de duas amostras, de tamanhos m e n de populações normais, e pretende-se saber se estas médias são significativamente diversas, isto é, se são, na realidade, diferentes as médias μ_1 e μ_2 das respectivas populações.

A hipótese nula consiste em supor $\mu_1=\mu_2$, o que, supondo também iguais os desvios padrão σ_1 e σ_2, equivale a supor que as amostras foram extraídas da mesma população normal.

Então, pelo que vimos no número anterior, basta calcular

$$t=\frac{\bar{x}-\bar{y}}{\hat{\sigma}\sqrt{\frac{1}{m}+\frac{1}{n}}}=\frac{\bar{x}-\bar{y}}{\sqrt{ms_1^2+ns_2^2}}\sqrt{\frac{m+n-2}{\frac{1}{m}+\frac{1}{n}}}$$

e procurar na tabela o valor de t_p correspondente ao nível de significação de $p\%$ escolhido. Se t for superior ou igual a t_p, rejeita-se a hipótese nula, se não, aceita-se a hipótese nula ou aguarda-se nova informação.

Exemplo – Uma amostra das alturas de 9 habitantes de uma grande cidade deu os valores $\bar{x}=1{,}70\,\text{m}$ e $s_1^2=90\,\text{cm}^2$. Outra amostra de 10 alturas de outra grande cidade deu $\bar{y}=1{,}69\,\text{m}$ e $s_2^2=105\,\text{cm}^2$. Pretende-se saber se é aceitável a hipótese de que nas duas cidades a estatura média é sensivelmente a mesma (admitindo que a distribuição das alturas nas duas cidades é normal, com iguais desvios padrão).

Neste caso, o valor de t calculado é 0,208 e, como o valor t_5 de t, correspondente a 5% para 17 graus de liberdade ($9+10-2=17$), é 2,110, bastante superior a 0,208, segue-se que a hipótese é admissível ao nível estabelecido.

OBSERVAÇÃO IMPORTANTE. Nem sempre é necessário que as variáveis x, y, ... consideradas nas questões práticas sejam normais para que se possam aplicar as provas de significação anteriores. Com efeito, há um teorema muito importante do Cálculo das Probabilidades, chamado *teorema central do limite*, do qual se deduz como corolário o seguinte: *Qualquer que seja a distribuição duma variável casual x, se for* μ o seu valor médio e σ *o seu desvio padrão, a distribuição da variável*

$$\xi = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

converge para a distribuição normal estandardizada quando $n \to \infty$.

Na prática, basta que a amostra seja de tamanho $n > 30$ para que a distribuição de ξ se possa considerar, sem erro apreciável, idêntica a N(0, 1).

6. Intervalos de tolerância e intervalos de confiança

Suponhamos que *é conhecido o desvio padrão* σ duma variável normal x e que se pretende saber se é legítimo ou não tomar para valor médio μ de x um dado número μ_0. Consideremos, então, uma amostra casual

$$X = (x_1, x_2, \ldots, x_n)$$

de valores independentes de x; já sabemos (n.° 1) que a média $\bar{x} = \frac{1}{n}\sum x_i$ é uma variável $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Portanto, na hipótese $\mu = \mu_0$ (*hipótese nula*), o estatístico

$$\tau = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{3}$$

deverá ser uma variável N(0, 1), e, assim, a probabilidade P de que $|\tau|$ exceda um certo limite τ_p (onde $p = 100\,P$) é dada pela fórmula

$$P = \Pr(|\tau| \geqslant \tau_p) = 2\int_{\tau_p}^{+\infty} \varphi(x)\,dx\,,$$

onde

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

(função de densidade da distribuição normal estandardizada).

Por conseguinte, a hipótese $\mu = \mu_0$ será admitida ao nível de $p\%$, se e só se $|\tau| < \tau_p$, isto é, se

$$-\tau_p < \tau < \tau_p . \tag{4}$$

Ora, como de (3) se deduz

$$\bar{x} = \mu_0 + \tau \frac{\sigma}{\sqrt{n}},$$

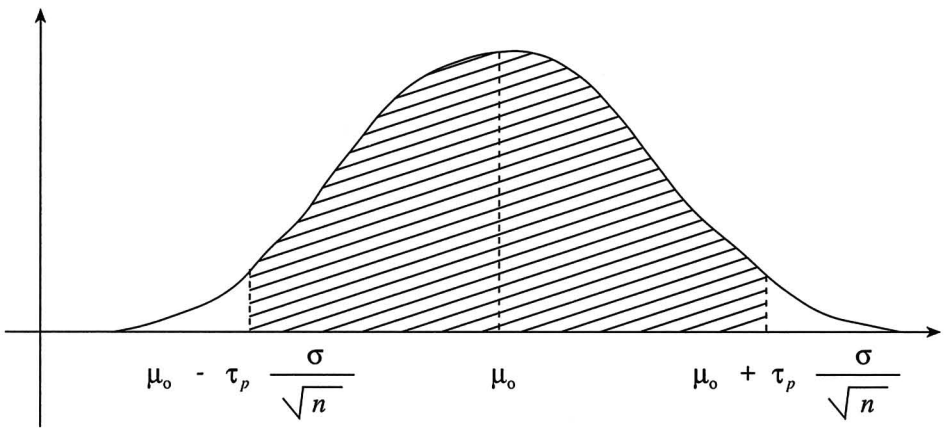
podemos concluir de (4) que os valores de $\bar{x}$ *tolerados* pela hipótese nula são os que verificam a condição

$$\mu_0 - \tau_p \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu_0 + \tau_p \frac{\sigma}{\sqrt{n}}, \tag{5}$$

isto é, os valores do intervalo

$$\left]\mu_0 - \tau_p \frac{\sigma}{\sqrt{n}}, \mu_0 + \tau_p \frac{\sigma}{\sqrt{n}}\right[,$$

chamado *intervalo de tolerância*. A probabilidade de que $\bar{x}$ esteja *fora* deste intervalo é $P = p/100$ e, portanto, a probabilidade de que $\bar{x}$ esteja *dentro* deste intervalo é $1 - P$ (área do domínio tracejado na figura).



Mas a questão pode ainda pôr-se da maneira inversa, embora equivalente. A fórmula (5) mostra que os valores μ_0 que merecem confiança ao nível considerado, em face da média $\bar{x}$ achada na amostra, são todos os que verificam a condição $\bar{x} - \tau_p \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + \tau_p \frac{\sigma}{\sqrt{n}}$, isto é, são os valores μ_0 do intervalo $\left]\bar{x} - \tau_p \frac{\sigma}{\sqrt{n}}, \bar{x} + \tau_p \frac{\sigma}{\sqrt{n}}\right[$ chamado *intervalo de confiança*. Neste caso, sendo $\bar{x}$ variável, $1-P$ dá-nos a probabilidade de que o intervalo de confiança contenha o verdadeiro valor médio μ de x e recebe o nome de *grau de confiança* desse intervalo [1].

Na prática, toma-se geralmente para nível de significação nestas questões o de 5% ($p=5$) e, portanto, para grau de confiança, o de 95%. Ora, como se pode ver numa tabela relativa à distribuição N(0, 1), tem-se

$$\tau_5 = 1{,}96,$$

valor próximo de 2. Assim, a probabilidade de um desvio superior em módulo ao dobro do desvio padrão é um pouco menor que 5% (já sabemos que a probabilidade de um desvio superior em módulo ao triplo do desvio padrão é cerca de $0{,}003 = 0{,}3\%$).

Exemplo – Um fabricante produz lâmpadas eléctricas cuja duração média μ é de 2000 kw/h, com o desvio padrão $\sigma = 300$ kw/h. Examinando uma amostra de 100 lâmpadas obtidas por um novo método de fabrico, encontra a média $\bar{x} = 2080$ kw/h. Admitindo que o novo método de fabrico não altera o desvio padrão desta variável, achar o intervalo de confiança correspondente à média obtida (com o grau $1-P=0{,}95$) e compará-lo com o primeiro valor médio.

Neste caso será (ver OBSERVAÇÃO IMPORTANTE do n.° 5)

$$\tau_p \frac{\sigma}{\sqrt{n}} = 1{,}96 \frac{300}{\sqrt{100}} \approx 60,$$

(1) – Não seria correcto dizer que $1-P$ é a probabilidade de μ estar naquele intervalo, visto que μ é fixo.

donde os extremos do intervalo de confiança:

$$2080-60=2020, \quad 2080+60=2140.$$

O intervalo de confiança pedido será, pois,

$$]2020\,,\,2140[\,.$$

Vê-se assim que o primeiro valor médio não cai neste intervalo. Também se vê que, por exemplo, é razoável admitir como duração média das novas lâmpadas o número redondo 2100 kw/h.

7. Intervalos de confiança quando não é conhecido o σ da população

As considerações anteriores foram desenvolvidas na hipótese de ser conhecido o desvio padrão σ da população base. Se este não for conhecido, poderá ser substituído pela sua melhor estimativa, $\hat{\sigma}$, deduzida da amostra. Mas então, em vez do desvio reduzido τ, só podemos utilizar a sua estimativa t dada pela fórmula

$$t=\frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}}\,.$$

Portanto, uma vez fixado um nível $P=p\%$, o valor t_p correspondente é, como vimos, dado pela tabela da distribuição de STUDENT para $n-1$ graus de liberdade. Somos assim, naturalmente, induzidos a chamar *intervalo de confiança* para μ, com o grau $1-P=100-p\%$, ao intervalo

$$\left]\bar{x}-t_p\frac{\hat{\sigma}}{\sqrt{n}}\,,\,\bar{x}+t_p\frac{\hat{\sigma}}{\sqrt{n}}\right[\,.$$

Obtêm-se, deste modo, com um mesmo nível $p\%$, intervalos de confiança mais largos (logo, menos precisos) do que no caso anterior. Mas não esqueçamos que, *quando n é muito grande, a distribuição de* STUDENT *confunde-se praticamente com a normal.*

Exemplo – Retomemos o primeiro exemplo do número 5. Trata-se duma amostra de alturas de 9 homens ($n=9$), com a média $\bar{x}=1{,}72$ e o desvio padrão corrigido $\hat{\sigma}=0{,}36$. Fixado o nível de significação 5%, acha-se, para 8 ($=9-1$) graus de liberdade, o valor

$$t_5 \approx 2{,}31 \text{ (superior a } \tau_5 = 1{,}96).$$

Portanto, os extremos do intervalo de confiança com o grau 95% serão

$$1{,}72 \pm 2{,}31 \times \frac{0{,}36}{3} = 1{,}72 \pm 0{,}28.$$

Como se vê, o valor 1,70 proposto no n.° 5 para média da população está perfeitamente incluído neste intervalo.

É claro que todas estas considerações se podem aplicar, mutatis mutandis, ao caso da medição de grandezas (TEORIA DOS ERROS).

8. Aplicações agronómicas

A prova do *t* é de uso correntíssimo na prática agronómica. Pode usar-se, por exemplo, para comparar a percentagem média da gordura do leite em duas espécies, raças ou sub-raças de mamíferos, a produtividade média do trigo em duas variedades deste cereal, etc., etc. Assim, o dizer-se que *o leite de cabra é (em média) mais gordo que o leite de vaca* é uma afirmação cujo valor só pode ser avaliado estatisticamente usando, por exemplo, a prova do *t*. De resto, já o dizer que *a percentagem média da gordura em tal espécie ou tal raça é um certo número* μ é uma afirmação que, para ser verdadeiramente útil, deve vir acompanhada da indicação do desvio padrão ou ser substituída pela indicação dum intervalo de confiança (no caso de se conhecer, apenas, uma amostra pequena).

Importa ainda salientar o seguinte: nem sempre é razoável admitir que a distribuição duma variável biométrica, numa dada população, é normal; mas, neste caso, bastará ter presente a OBSERVAÇÃO IMPORTANTE do n.° 5.

9. Distribuição de F e de z de FISHER

Já sabemos que, dadas m variáveis independentes e normais estandardizadas, $x_1, x_2, \ldots, x_m$, sendo $m > 1$, a variável

$$\chi^2 = x_1^2 + x_2^2 + \cdots + x_m^2$$

não segue a distribuição normal, mas sim a distribuição do χ^2 de PEARSON, cuja função de densidade já foi indicada neste curso.

Consideremos agora, mais geralmente, $m+n$ variáveis independentes e normais estandardizadas, $x_1, x_2, \ldots, x_m$, $y_1, y_2, \ldots, y_n$, e ponhamos

$$\chi_1^2 = x_1^2 + x_2^2 + \cdots + x_m^2, \quad \chi_2^2 = y_1^2 + y_2^2 + \cdots + y_n^2.$$

Então, a variável

$$F = \frac{n}{m} \frac{\chi_1^2}{\chi_2^2}$$

seguirá uma nova distribuição, chamada por SNEDECOR *distribuição de F para (m, n) graus de liberdade*, e tabelada por aquele estatístico para os níveis de 5% e 1% e para diferentes pares de valores (m, n).

A letra F foi escolhida em homenagem a FISHER, que primeiramente tinha estudado a distribuição da variável

$$z = \frac{1}{2} \log F$$

deduzindo matematicamente a expressão analítica da função da densidade dessa distribuição, expressão que nos abstemos de apresentar aqui.

Estas distribuições intervêm essencialmente na *análise de variância*, método estatístico criado por FISHER, de grande importância em investigações agronómicas, usando-se, por exemplo, para comparar simultaneamente diversas variedades de trigo, os efeitos de diversos factores fertilizantes ou tratamentos de plantas, etc., etc.

Infelizmente, não podemos, sequer, abordar o estudo deste assunto, relativo ao DELINEAMENTO E ANÁLISE DE EXPERIÊNCIAS, o grande problema central da Estatística Agronómica.

TÁBUA DA DISTRIBUIÇÃO NORMAL

$$P = \Pr(|x - \mu| > \tau_p) = \frac{2}{\sqrt{2\pi}} \int_{\tau_p}^{\infty} e^{-\frac{t^2}{2}}\, dt$$

τ_p como função de p		p como função de τ_p	
$p = 100P$	τ_p	τ_p	$p = 100P$
100	0,0000	0,0	100,000
95	0,0627	0,2	84,148
90	0,1257	0,4	68,916
85	0,1891	0,6	54,851
80	0,2533	0,8	42,371
75	0,3186	1,0	31,731
70	0,3853	1,2	23,014
65	0,4538	1,4	16,151
60	0,5244	1,6	10,960
55	0,5978	1,8	7,186
50	0,6745	2,0	4,550
45	0,7554	2,2	2,781
40	0,8416	2,4	1,640
35	0,9346	2,6	0,932
30	1,0364	2,8	0,511
25	1,1603	3,0	0,270
20	1,2816	3,2	0,137
15	1,4395	3,4	0,067
10	1,6449	3,6	0,032
5	1,9600	3,8	0,014
1	2,5758	4,0	0,006
0,1	3,2905		
0,01	3,8906		

TÁBUA DA DISTRIBUIÇÃO DE *t* DE STUDENT (1)

$$P = \frac{p}{100} = \Pr(|t| > t_p) = 2\int_{t_p}^{+\infty} S_n(x)dx$$

ν	$p = 5$ $t_p =$	$p = 1$ $t_p =$	$p = 0{,}1$ $t_p =$	ν	$p = 5$ $t_p =$	$p = 1$ $t_p =$	$p = 0{,}1$ $t_p =$
1	12,706	63,657	636,619	26	2,056	2,779	3,707
2	4,303	9,925	31,598	27	2,052	2,771	3,690
3	3,182	5,841	12,941	28	2,048	2,763	3,674
4	2,776	4,604	8,610	29	2,045	2,756	3,659
5	2,571	4,032	6,859	30	2,042	2,750	3,646
6	2,447	3,707	5,959	35	2,030	2,724	3,592
7	2,365	3,499	5,405	40	2,021	2,704	3,551
8	2,306	3,355	5,041	45	2,014	2,689	3,521
9	2,262	3,250	4,781	50	2,008	2,678	3,496
10	2,228	3,169	4,587	60	2,000	2,660	3,460
11	2,201	3,106	4,437	70	1,994	2,648	3,435
12	2,179	3,055	4,318	80	1,990	2,638	3,416
13	2,160	3,012	4,221	90	1,987	2,631	3,402
14	2,145	2,977	4,140	100	1,984	2,626	3,390
15	2,131	2,947	4,073	120	1,980	2,617	3,373
16	2,120	2,921	4,015	140	1,977	2,611	3,361
17	2,110	2,898	3,965	160	1,975	2,607	3,352
18	2,101	2,878	3,922	180	1,973	2,603	3,346
19	2,093	2,861	3,883	200	1,972	2,601	3,340
20	2,086	2,845	3,850	300	1,968	2,592	3,324
21	2,080	2,831	3,819	400	1,966	2,588	3,315
22	2,074	2,819	3,792	500	1,965	2,586	3,310
23	2,069	2,807	3,767				
24	2,064	2,797	3,745	1000	1,962	2,581	3,300
25	2,060	2,787	3,725	∞	1,960	2,576	3,291

ν = número de graus de liberdade.

(1) – Comparar t_p com τ_p na tábua da distribuição normal, para valores pequenos e para valores grandes de ν (cf. $\nu = \infty$).

TÁBUA DA DISTRIBUIÇÃO DO χ^2 DE PEARSON

	PROBABILIDADE (P)										
n	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,016	0,064	0,15	0,46	1,07	1,64	2,71	3,84	5,41	6,64	10,83
2	0,21	0,45	0,71	1,39	2,41	3,22	4,61	5,99	7,82	9,21	13,82
3	0,58	1,01	1,42	2,37	3,67	4,64	6,25	7,82	9,84	11,34	16,27
4	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	11,77	13,28	18,47
5	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	13,39	15,09	20,52
6	2,20	3,07	3,83	5,35	7,23	8,56	10,65	12,59	15,03	16,81	22,46
7	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	16,62	18,48	24,32
8	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	18,17	20,09	26,13
9	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,68	21,67	27,88
10	4,87	6,18	7,27	9,34	11,78	13,44	15,99	18,31	21,16	23,21	29,59
12	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,03	24,05	26,22	32,91
14	7,79	9,47	10,82	13,34	16,22	18,15	21,06	23,69	26,87	29,14	36,12
16	9,31	11,15	12,62	15,34	18,42	20,47	23,54	26,30	29,63	32,00	39,25
18	10,87	12,86	14,44	17,34	20,60	22,76	25,99	28,87	32,35	34,81	42,31
20	12,44	14,58	16,27	19,34	22,78	25,04	28,41	31,41	35,02	37,57	45,32
22	14,04	16,31	18,10	21,34	24,94	27,30	30,81	33,92	37,66	40,29	48,27
24	15,66	18,06	19,94	23,34	27,10	29,55	33,20	36,42	40,27	42,98	51,18
26	17,29	19,82	21,79	25,34	29,25	31,80	35,56	38,89	42,86	45,64	54,05
28	18,94	21,59	23,65	27,34	31,39	34,03	37,92	41,34	45,42	48,28	56,89
30	20,60	23,36	25,51	29,34	33,53	36,25	40,26	43,77	47,96	50,89	59,70

INDICAÇÕES BIBLIOGRÁFICAS

São particularmente recomendáveis, aos alunos do Instituto Superior de Agronomia, as três seguintes obras, além das que já foram anteriormente indicadas.

M. LAMOTTE – *Initiation aux Méthodes Statistiques en Biologie*. Masson & C^{ie}. Paris, 1957.

M. J. MORONEY – *Facts from Figures*. Penguin Books, Londres, 1954.

SIXTO RIOS – *Métodos de la Estatística*. Madrid, 1952.

As duas primeiras fornecem uma excelente e agradável iniciação nos métodos estatísticos, com grande número e variedade de exemplos de aplicação.

A terceira é uma obra de nível mais elevado, com larga informação que inclui os aspectos mais modernos da Estatística. Desse livro foram extraídos vários dos exemplos que figuram nestes apontamentos.

ÍNDICE

CÁLCULO DAS PROBABILIDADES

ADVERTÊNCIA PRÉVIA 317

I.4.1 INTRODUÇÃO AO CÁLCULO DAS PROBABILIDADES: POPULAÇÕES FINITAS 319

A – Frequências 319

1. Primeiros exemplos 319
2. Populações. Álgebra dos atributos 322
3. Álgebra dos acontecimentos 325
4. Acontecimentos expressos em forma proposicional 326
5. Frequência dum atributo numa população 328
6. Frequência de um acontecimento numa série de provas 330
7. Partições 331
8. Corpos de conjuntos, corpos de atributos, corpos de acontecimentos 333
9. Distribuição em universos finitos 337
10. Soma de conjuntos não disjuntos (atributos ou acontecimentos compatíveis) 339
11. Atributos quantitativos 341
12. Representação gráfica das distribuições: histogramas e polígonos de frequência 345
13. Independência e associação de atributos. Distribuições de duas ou mais variáveis 348

14. Associação e independência de partições múltiplas. Tábuas de contingência 355
15. Associações parciais de atributos. Independência de atributos no caso em que o seu número é superior a dois 359
16. Interpretação de uma tábua de contingência. Testes de significância 363

B – Probabilidades 373

1. Lógica indutiva 373
2. Lógica dedutiva 376
3. Conceito natural de probabilidade 378
4. Axiomatização do conceito de probabilidade 382
5. Alguns exemplos de cálculo de prodabilidades *a priori* 385
6. Independência e associação de acontecimentos 396
7. Sistema de duas experiências 398
8. Sistema de várias experiências 403
9. Distribuição binomial ou de Bernoulli 404
10. Conceito de moda. Caso da distribuição normal 409
11. Distribuição polinomial. Amostras casuais 411

BIBLIOGRAFIA 414

I.4.2 APONTAMENTOS DE CÁLCULO DAS PROBABILIDADES 415

A – Distribuições de uma variável contínua real 415

B – Valores médios para distribuições de uma variável real ... 425

C – Valores médios para distribuições de mais de uma variável real 438

D – Aplicação à distribuição binomial. Teorema de BERNOULLI 451

E – Distribuição normal 455

F – Convergência de distribuições. Relação entre as distribuições normal e binomial 464

G – A distribuição de χ^2 de PEARSON 465

NOTA SOBRE A AVALIAÇÃO DA VARIÂNCIA 469

I.4.3 ADITAMENTO ÀS LIÇÕES DE CÁLCULO DE PROBABILIDADES 471

A – Regressões. Ajustamentos. Correlação 471

1. Formulação geral do problema 471
2. Ajustamentos pelo método dos mínimos quadrados 476
3. Outra forma das equações normais 479
4. Regressão polinomial 480
5. Regressão linear. Correlação 481
6. Segunda recta de regressão 485
7. Organização prática dos cálculos 487
8. Ajustamentos com mudanças não lineares de variáveis 490
9. Regressão múltipla. Índice de correlação em geral 493
10. Nota sobre as notações 497

B – Distribuições de STUDENT e de FISHER. Suas aplicações 497

1. A melhor estimativa do desvio padrão deduzida duma amostra 497
2. Distribuição *t* de STUDENT 499
3. A melhor estimativa de σ deduzida a partir de várias amostras 501
4. Distribuição da diferença entre duas médias 502
5. Prova do *t* (de significação) 503
6. Intervalos de tolerância e intervalos de confiança 506
7. Intervalos de confiança quando não é conhecido o σ da população 509
8. Aplicações agronómicas 510
9. Distribuição de *F* e *z* de FISHER 511

TÁBUA DA DISTRIBUIÇÃO NORMAL 512

TÁBUA DA DISTRIBUIÇÃO DE *t* DE STUDENT 513

TÁBUA DA DISTRIBUIÇÃO DE χ^2 DE PEARSON 514

INDICAÇÕES BIBLIOGRÁFICAS 515