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.4.1

INTRODUÇÃO AO CÁLCULO DAS PROBABILIDADES: POPULAÇÕES FINITAS

A – Frequências

1. Primeiros exemplos

Comecemos por um exemplo que nos é bem familiar. Suponhamos que, numa dada época de exames finais duma cadeira, se apresentaram a exame 189 alunos e os resultados foram os que constam da seguinte tabela:

TABELA N.° 1

Classificações	N.° de Alunos	Classificações	N.° de Alunos
0	0	11	45
1	0	12	28
2	0	13	17
3	0	14	9
4	2	15	12
5	1	16	7
6	5	17	5
7	11	18	2
8	7	19	1
9	0	20	0
10	37		

Diremos então que, nestas provas, a *frequência* da classificação 3 foi 0, a *frequência* da classificação 10 foi 37, etc.

Todavia, para ter uma ideia de conjunto destes resultados, nas suas relações mútuas, é preferível indicar a frequência de cada classificação em *percentagem,* tal como se observa na seguinte tabela deduzida da precedente:

TABELA N.° 2

Classificações	Percentagens	Classificações	Percentagens
0	0,0%	10	19,6%
1	0,0%	11	23,8%
2	0,0%	12	14,8%
3	0,0%	13	9,0%
4	1,1%	14	4,8%
5	0,5%	15	6,3%
6	2,6%	16	3,8%
7	5,8%	17	2,6%
8	3,7%	18	1,1%
9	0,0%	19	0,5%
		20	0,0%

No primeiro caso, diz-se que se trata de *frequências absolutas* e, no segundo, de *frequências relativas*. Assim, a frequência relativa da classificação 2 foi 0%, a frequência relativa da classificação 10 foi 19,6%, etc.

É claro que, designando por n o número total de alunos e por ν o número de alunos que obtiveram uma dada classificação x (frequência absoluta de x), a frequência relativa de x, *em percentagem*, será o número $\mathrm{fr}(x)$ dado pela fórmula

$$\mathrm{fr}(x) = \frac{100\,\nu}{n}\,\%\,.$$

Mas uma percentagem não é mais do que uma maneira especial, usada na prática, de designar um número, geralmente compreendido entre 0 e 1. Em considerações puramente teóricas será preferível

designar esse número à maneira usual, e então, a fórmula que dá a frequência relativa toma o aspecto:

$$\mathrm{fr}(x) = \frac{\nu}{n}.$$

Observe-se ainda que, na prática, o cálculo das percentagens conduz geralmente a dízimas, das quais não interessam todos os algarismos decimais. Escolhem-se valores aproximados desses números, por defeito ou por excesso, com o grau de aproximação que se considere suficiente. Mas, por coerência lógica, convirá escolher esses valores de modo que a sua soma seja exactamente 100. (No exemplo anterior tomaram-se valores aproximados a menos de 0,1).

Os conceitos de frequência absoluta e de frequência relativa são usados a cada passo na vida corrente. Indicaremos mais alguns:

1) – Numa dada cidade com 23.528 habitantes, 9.253 pertencem ao sexo masculino e os restantes 14.275 ao sexo feminino. Estes dois últimos números dão, pois, as frequências absolutas do sexo masculino e do sexo feminino entre os habitantes da referida cidade. As frequências relativas dos mesmos atributos, serão, respectivamente, de 39% e de 61%.

2) – Num dado país, 82% dos habitantes têm os cabelos castanhos ou pretos, 14% têm os cabelos louros e 4% têm os cabelos ruivos. São estas, pois, as frequências relativas daqueles atributos entre os habitantes do referido país.

3) – Numa série de 5.000 viagens efectuadas por uma dada companhia de aviação, houve desastres em 2 viagens. Nesta série, a frequência relativa dos desastres foi, portanto, de 0,04%, ou seja, de 0,0004.

4) – Numa série de competições, um dado clube desportivo teve 10 vitórias, 2 derrotas e 0 empates. Nesta série, as frequências relativas das vitórias, derrotas e empates foram, pois, respectivamente, iguais a 0,8, a 0,2 e a 0 (a menos de uma décima).

2. Populações. Álgebra dos atributos

Entre os exemplos anteriores, distinguem-se dois grupos principais. Num dos grupos apresentam-se-nos frequências de *atributos*; no outro, aparecem-nos frequências de *acontecimentos*. Em qualquer dos casos as considerações restringem-se a um determinado conjunto, que pode ser constituído por entidades das mais diversas naturezas (seres humanos, competições desportivas, viagens de aviação, etc.). Este conjunto fundamental, a que se refere um dado inquérito estatístico, é chamado, conforme os autores, *população*, *universo lógico*, *universo do discurso* ou, simplesmente, *universo*; os seus elementos são chamados os *indivíduos* (desse universo)[1].

Consideremos uma dada população U. Em U podem estar definidos diferentes *atributos* (em vez do termo "atributo" usam-se, ainda, como sinónimos, "propriedade", "predicado", "carácter", etc.). A cada atributo α corresponde um determinado *conjunto* ou *classe* A, constituída por todos os indivíduos (elementos de U), que possuem o atributo α.

A presença simultânea de dois atributos α, β num mesmo indivíduo constitui um novo atributo, que se chama *conjunção* (ou *produto lógico*) dos primeiros e se representa por $\alpha \cap \beta$ ou, simplesmente, por $\alpha\beta$. Claro está que, se for A o conjunto dos indivíduos com o atributo α, e se for B o conjunto dos indivíduos com o atributo β, será $A \cap B$ (*intersecção* de A com B) o conjunto dos indivíduos com o atributo $\alpha\beta$ (ler "α e β"). Por exemplo, pode acontecer que, numa dada população de seres humanos, o atributo "masculino" e o atributo "ruivo" coexistam num mesmo indivíduo: a totalidade dos indivíduos que os possuem simultaneamente será a intersecção do conjunto dos indivíduos do sexo masculino com o conjunto dos indivíduos ruivos; mas também pode acontecer que, numa outra população, esta intersecção seja o conjunto vazio, isto é, que não haja nessa população homens ruivos.

O facto de um indivíduo possuir *um, pelo menos*, dos atributos α, β (podendo ter ambos ao mesmo tempo) exprime-se por um novo

---

(1) – Note-se que as designações *universo lógico* e *universo do discurso* já tinham sido adoptadas em lógica matemática com um sentido equivalente. De resto, as noções de lógica matemática introduzidas em Matemáticas Gerais vão ser também aqui utilizadas.

atributo, que se chama *disjunção* (ou *soma lógica*) dos primeiros e se representa por $\alpha \cup \beta$ ou por $\alpha + \beta$ (ler "α ou β"). Se forem A e B os conjuntos correspondentes a α e β, será $A \cup B$ (reunião de A com B) o conjunto correspondente a $\alpha \cup \beta$. Exemplo: entre os seres humanos, o atributo "casado ou viúvo" é a disjunção dos atributos "casado" e "viúvo".

De modo análogo, se define a conjunção e a disjunção de *vários* atributos α, β, γ, ... Por exemplo, o símbolo $\alpha\beta\gamma$ significa a presença simultânea dos atributos α, β, γ num indivíduo: designa, pois, a *conjunção* desses atributos.

Note-se que as operações da conjunção e disjunção são ambas associativas e comutativas (mas não reversíveis) e qualquer delas é distributiva em relação à outra, isto é, tem-se

$$\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, \qquad \alpha + \beta\gamma = (\alpha + \beta)\,(\alpha + \gamma).$$

A ausência dum atributo α num dado indivíduo constitui um novo atributo, que se chama *contrário* (ou *negação*) de α e se representa por $\sim\alpha$ ou por $\tilde{\alpha}$ (ler "não α"). É claro que o conjunto dos indivíduos com o atributo $\tilde{\alpha}$ é o *complementar* do conjunto dos indivíduos com o atributo α. Por exemplo, nos resultados dum exame, o atributo "aprovado" é contrário do atributo "reprovado".

Tem-se, ainda, como é fácil ver:

$$\sim(\sim\alpha) = \alpha \quad \text{(Lei da dupla negação)}$$

$$\sim(\alpha\beta) = (\sim\alpha) + (\sim\beta), \quad \sim(\alpha + \beta) = (\sim\alpha)(\sim\beta) \quad (1^{a} \text{ lei de Morgan}).$$

Um atributo diz-se *universal*, quando se verifica em todos os indivíduos; o conjunto correspondente é pois o universo considerado, U. Um atributo diz-se *impossível*, quando não se verifica em nenhum indivíduo (elemento de U); o conjunto correspondente é o conjunto *vazio*.

Dois atributos α e β dizem-se *incompatíveis* quando a sua conjunção é um atributo impossível, isto é, quando os conjuntos correspondentes são *disjuntos*. Assim, por exemplo, no universo dos seres humanos, o atributo "solteiro" é incompatível com o atributo "casado".

Dois atributos α e β contrários são sempre incompatíveis. Mas dois atributos podem ser incompatíveis sem serem contrários; exemplos: os atributos "solteiro" e "casado" entre seres humanos, os atributos "medíocre" e "bom" nos resultados dum exame, etc. Para dois atributos serem contrários, é necessário e suficiente que a sua conjunção seja um atributo impossível e a sua disjunção um atributo universal.

Dados dois atributos α, β, diz-se que α *implica* β e escreve-se $\alpha \subset \beta$, quando todos os indivíduos que possuem α também possuem β. Exemplos: "medíocre" implica "reprovado", "casado" implica "não solteiro". Os atributos α e β dizem-se *equivalentes* quando $\alpha \subset \beta$ e $\beta \subset \alpha$; exemplo: "medíocre ou mau" equivale a "reprovado".

De tudo o que precede ressalta que a álgebra dos atributos é perfeitamente análoga à álgebra das funções proposicionais (numa variável) e ambas se traduzem directamente na álgebra dos conjuntos. É que, a bem dizer, entre atributos e funções proposicionais existe apenas uma diferença de *forma*. Seja, por exemplo, a função proposicional

"*x* é agrónomo ou silvicultor"

definida entre indivíduos portugueses; trata-se, como se vê, duma fórmula que se converte em proposição verdadeira para certas determinações de *x* e em proposição falsa para outras determinações de *x*. Exprime, pois, um *atributo* que se concretiza na reunião de duas classes: a dos engenheiros agrónomos e a dos engenheiros silvicultores (classes não necessariamente disjuntas).

Uma classe pode ser determinada de dois modos diversos: – ou dando um sistema de atributos ou regras que a caracterizem completamente (*ponto de vista da compreensão*); – ou indicando um por um os indivíduos que a compõem (*ponto de vista de extensão*). Conforme o ponto de vista adoptado, assim o conceito de classe se aproxima do conceito de atributo ou do conceito de conjunto. A distinção entre tais conceitos tem sempre um carácter mais ou menos subjectivo.

Note-se que, na linguagem comum, os *adjectivos* exprimem normalmente atributos, enquanto os *substantivos comuns* se referem a *classes* e os *substantivos próprios* a indivíduos.

Convém, ainda, lembrar que os atributos (ou as classes) que a Natureza nos apresenta nunca são nitidamente definidos, como os conceitos abstractos da Matemática. Entre um atributo e os atributos vizinhos há geralmente zonas fronteiriças, em que a distinção acaba sempre por ser feita de modo mais ou menos artificial ou arbitrário. Por exemplo, ao classificar os habitantes duma cidade segundo a cor dos cabelos, muitas vezes se hesita entre os atributos "castanho" e "loiro" ou "ruivo", para um dado indivíduo.

3. Álgebra dos acontecimentos

Recordemos que, entre os exemplos do n.º 1, alguns se referem, não propriamente a atributos, mas sim a *acontecimentos*. Não pretendemos aqui definir "acontecimento", assim como não tentámos definir "atributo": tratam-se de conceitos psicologicamente primitivos. O que será possível e conveniente é esclarecer tais conceitos e precisar, tanto quanto possível, a terminologia que lhes diz respeito.

Comecemos por notar que, em vez de "acontecimento", se usam com significado semelhante os termos "facto", "fenómeno", "eventualidade", etc.

Imaginemos uma prova ou experiência, que se possa repetir várias vezes em condições idênticas, conduzindo, de cada vez, a um ou mais resultados, entre vários que são possíveis. É a cada um desses resultados que, em cálculo das probabilidades e em estatística, se costuma dar o nome *acontecimento*. Assim, antes de efectuar a prova, vários acontecimentos são de esperar: cada um deles é, então, apenas uma *eventualidade* ou *hipótese* (*acontecimento em potência*); efectuada a prova, apenas alguma ou algumas dessas hipóteses se realizam (*acontecimento em acto*).

São inúmeros os exemplos que, neste sentido, se podem apresentar. Por enquanto, recordaremos apenas os que foram citados no n.º 1: desastre ou ausência de desastre numa viagem aérea, resultados duma prova desportiva, etc.

Mas convém desde já observar que, muitas vezes, os acontecimentos se exprimem por meio de atributos (e vice-versa). Tal é, por exemplo, o caso dos resultados dum exame (ou o caso análogo duma competição desportiva); assim, ao atributo "reprovado" corresponde o acontecimento "ficar reprovado", ao atributo "vitorioso" corresponde

o acontecimento "vencer" (na linguagem comum, os acontecimentos são geralmente expressos por *verbos*, enquanto os atributos são expressos por *adjectivos*).

Ainda aqui, portanto, há, de certo modo, uma questão de ponto de vista psicológico. Podemos, portanto, desde já, prever que para os acontecimentos, exista uma álgebra perfeitamente análoga à dos atributos. Com efeito, sejam α e β dois acontecimentos a esperar numa dada prova $\mathcal{P}$:

1) - Chama-se *conjunção* (ou *produto lógico*) de α e β, e representa-se por $\alpha\beta$, o acontecimento que consiste na realização simultânea de α e de β.
2) - Chama-se *disjunção* (ou *soma lógica*) de α e β, e representa-se por $\alpha + \beta$, o acontecimento que consiste em se realizar *um, pelo menos*, dos acontecimentos α e β.
3) - Chama-se *contrário* (ou *negação*) de α, e representa-se $\sim\alpha$ ou por $\tilde{\alpha}$, o acontecimento que consiste em não se realizar α.
4) - Diz-se que α *implica* β, e escreve-se $\alpha \subset \beta$, quando β se realiza todas as vezes que α se realiza.
5) - Dois acontecimentos α, β dizem-se *equivalentes*, e escreve-se $\alpha \equiv \beta$, quando $\alpha \subset \beta$ e $\beta \subset \alpha$.
6) - Um acontecimento diz-se *certo* quando já se sabe *a priori*, com certeza absoluta, que se realizará na prova $\mathcal{P}$, todas as vezes que esta for efectuada. Um acontecimento diz-se *impossível*, quando é certo o seu contrário. Por exemplo, no exame final dum aluno é certo o acontecimento "aprovação ou reprovação" (excluem-se os casos de falta ou desistência) e é impossível o acontecimento "obter 21 valores" (em escolas portuguesas).
7) - Dois acontecimentos dizem-se *incompatíveis*, quando a sua conjunção é um acontecimento impossível.

As operações de conjunção, disjunção e negação entre acontecimentos gozam das mesmas propriedades que entre atributos, tornando-se, pois, supérfluo mencioná-las aqui.

4. Acontecimentos expressos em forma proposicional

A maior parte dos acontecimentos estudados em cálculo das probabilidades apresentam-se, naturalmente, sob a forma de funções proposicionais. Vejamos dois exemplos:

a) Imaginemos uma urna que contenha bolas brancas e bolas pretas. Designando por U o universo das bolas contidas na urna, por B o conjunto das bolas brancas e por P o conjunto das bolas pretas, teremos duas funções proposicionais

$$x \in B, \quad x \in P$$

definidas em U. A variável x toma, pois, cada um dos seus valores em U. Ora, o valor de x pode ser determinado, extraindo *ao acaso* uma bola de U. Nesta prova (extracção da bola) realiza-se, então, um dos seguintes *acontecimentos contrários*: $x \in B$, (a bola x que sai é branca), $x \in P$ (a bola x que sai é preta). Vê-se, pois, como as funções proposicionais consideradas passam a exprimir acontecimentos.

b) Seja x a classificação dum aluno numa prova $\mathcal{P}$ a realizar. Os resultados "mau", "medíocre", "suficiente", "bom", "distinto" e "muito bom", segundo convenções frequentemente adoptadas, são expressos, respectivamente, pelas funções proposicionais seguintes:

$$x < 6, \quad 6 \leq x < 10, \quad 10 \leq x < 14, \quad 14 \leq x < 16,$$
$$16 \leq x < 18, \quad 18 \leq x.$$

O resultado "aprovado", soma lógica de "suficiente", "bom", "distinto" e "muito bom", é expresso pela função proposicional $x \geq 10$.

Assim, nestes casos, os acontecimentos a que pode dar lugar a prova $\mathcal{P}$ considerada aparecem directamente sob a forma de funções proposicionais $\alpha(x)$ definidas num dado universo U, sendo o valor da variável x (elemento de U) determinado em cada realização da prova $\mathcal{P}$. A variável x recebe, nos casos como o considerado em a), o nome de *variável casual* ou *variável aleatória*, atendendo a que na sua determinação, intervém, mais ou menos, o *acaso*[1]. É claro

---

(1) – É impossível definir logicamente o termo "acaso". Numa primeira aproximação, poderíamos dizer que este termo designa ausência de *causa*, ou, pelo menos, de *causa conhecida*. Para certos autores, o acaso consistiria na acumulação de um grande número de pequenas causas desconhecidas que actuam em diversos sentidos, tornando praticamente impossível a previsão do efeito global. Sobre este ponto, aconselhamos a leitura do prefácio do livro de CASTELNUOVO, citado na Bibliografia.

que a conjunção de dois acontecimentos α e β será expressa pela conjunção das funções proposicionais que exprimem α e β e analogamente para a disjunção, para a negação, etc.

Deste modo, a cada acontecimento $\alpha(x)$ ficará a corresponder um determinado subconjunto A de U: o conjunto de indivíduos que verificam $\alpha(x)$; reciprocamente, a cada conjunto A de U corresponde o acontecimento expresso pela função proposicional $x \in A$. (A dois acontecimentos corresponde o mesmo conjunto, quando, e só quando, os acontecimentos são equivalentes). *A álgebra dos acontecimentos poderá, pois, traduzir-se na álgebra dos conjuntos.*

Mesmo no caso geral, a redução dos acontecimentos à forma de função proposicional é sempre possível, pelo menos de modo indirecto.

5. Frequência dum atributo numa população

Chama-se *frequência absoluta* dum atributo α (num dado universo U) ao número de indivíduos que possuem esse atributo, isto é, ao número de elementos do conjunto A correspondente ao atributo α. Designaremos esse número por (α).

Chama-se *frequência relativa* do atributo α ao quociente de (α) pelo número total de indivíduos (elementos de U). Designaremos por N esse número e por $\mathrm{fr}(\alpha)$ a frequência relativa de α. Ter-se-á, pois, por definição:

$$\mathrm{fr}(\alpha) = \frac{(\alpha)}{\mathrm{N}}.$$

A frequência relativa costuma também ser expressa em *tantos por cento* e, algumas vezes, em *tantos por mil*, devendo usar-se, então, respectivamente, as fórmulas:

$$\mathrm{fr}(\alpha) = \frac{100 \cdot (\alpha)}{\mathrm{N}}\ \%, \quad \mathrm{fr}(\alpha) = \frac{1.000 \cdot (\alpha)}{\mathrm{N}}\ ‰.$$

Já no n.º 1 demos vários exemplos de frequências absolutas e de frequências relativas.

Resulta logo da definição que se tem $0 \leq \mathrm{fr}(\alpha) \leq 1$, qualquer que seja o atributo α.

Em particular, será $\mathrm{fr}(\alpha) = 1$ [isto é, $(\alpha) = \mathrm{N}$], quando é atributo universal, e $\mathrm{fr}(\alpha) = 0$ [ou seja $(\alpha) = 0$], quando é atributo impossível. Além disso, tem-se o

TEOREMA DA SOMA. *Dados p atributos* $\alpha_1, \alpha_2, \ldots, \alpha_p$, *incompatíveis dois a dois, a frequência absoluta* [*resp. relativa*] *da soma lógica desses atributos é sempre igual à soma das frequências absolutas* [*resp. relativas*] *dos mesmos atributos; isto é, em fórmulas:*

$$(\alpha_1 + \alpha_2 + \cdots + \alpha_p) = (\alpha_1) + (\alpha_2) + \cdots + (\alpha_p),$$

$$\mathrm{fr}(\alpha_1 + \alpha_2 + \cdots + \alpha_p) = \mathrm{fr}(\alpha_1) + \mathrm{fr}(\alpha_2) + \cdots + \mathrm{fr}(\alpha_p). \tag{5.1}$$

Com efeito, sendo cada um dos atributos α_i incompatível com qualquer dos outros, os conjuntos correspondentes são disjuntos dois a dois e, portanto, a sua reunião, correspondente ao atributo $\alpha_1 + \alpha_2 + \cdots + \alpha_p$, tem um número de elementos igual à soma dos elementos dos primeiros, isto é,

$$(\alpha_1 + \alpha_2 + \cdots + \alpha_p) = (\alpha_1) + (\alpha_2) + \cdots + (\alpha_p)$$

e, portanto,

$$\frac{(\alpha_1 + \alpha_2 + \ldots + \alpha_p)}{\mathrm{N}} = \frac{(\alpha_1)}{\mathrm{N}} + \frac{(\alpha_2)}{\mathrm{N}} + \cdots + \frac{(\alpha_p)}{\mathrm{N}},$$

que é, precisamente, o que exprime a fórmula (5.1).

Note-se que, na aplicação deste teorema, *é essencial que esses atributos considerados sejam incompatíveis, dois a dois*. Por exemplo, suponhamos que, numa certa cidade, 8% dos habitantes são empregados do Estado e 34% são empregados de empresas particulares; não se pode daí concluir, sem mais, que 42% dos habitantes são empregados (do Estado ou de empresas particulares), pois pode haver habitantes que sejam ao mesmo tempo empregados do Estado e de empresas particulares. Vemos mais adiante como proceder, de modo sistemático, em casos tais.

Entretanto, observemos que do teorema anterior (ou mesmo da definição) se deduz logo o

COROLÁRIO. *Se for* fr *a frequência relativa dum dado atributo, será* $1-\text{fr}$ *a frequência relativa do atributo contrário; isto é, em fórmula*

$$\text{fr}(\tilde{\alpha}) = 1 - \text{fr}(\alpha).$$

Com efeito, os atributos α, $\tilde{\alpha}$ são incompatíveis e, como a sua soma lógica é atributo universal, tem-se

$$\text{fr}(\alpha) + \text{fr}(\tilde{\alpha}) = \text{fr}(\alpha + \tilde{\alpha}) = 1, \quad \text{donde,} \quad \text{fr}(\tilde{\alpha}) = 1 - \text{fr}(\alpha).$$

6. Frequência dum acontecimento numa série de provas

Consideremos n realizações, $\mathcal{P}_1$, $\mathcal{P}_2$, ..., $\mathcal{P}_n$, duma prova genérica $\mathcal{P}$ e seja α um acontecimento a provar em cada realização de $\mathcal{P}$. Suponhamos que o acontecimento α se realiza em ν destas provas e que, portanto, o seu contrário se realiza nas $n-\nu$ provas restantes. Diremos, então, que ν é a *frequência absoluta* do acontecimento α na série de provas consideradas. Por outro lado, chamaremos *frequência relativa* de α na referida série ao cociente de ν por n, isto é, ao número f dado pela fórmula

$$f = \frac{\nu}{n},$$

número este que depende não só do acontecimento, como, ainda, do número n das provas. É claro que se terá sempre

$$0 \leq f \leq 1.$$

Mas o caso $f = 1$ *não* habilita em absoluto a concluir que o acontecimento α é *certo*, assim como o caso $f = 0$ *não* habilita em absoluto concluir que o acontecimento é *impossível*.

Entretanto, importa observar que o *teorema da soma, bem como o respectivo corolário, subsistem para os acontecimentos, bastando substituir nos enunciados a palavra "atributo" pela palavra "acontecimento"*. As demonstrações são perfeitamente análogas às anteriores.

NOTA. Se tomarmos para universo lógico o conjunto $\{\mathcal{P}_1, \mathcal{P}_2, \ldots \mathcal{P}_n\}$, a função proposicional de

"$\mathcal{P}$ *é uma prova em que se realizou* α"

exprime visivelmente um atributo, cuja frequência (absoluta ou relativa) é por definição a frequência (absoluta ou relativa) do acontecimento α na série considerada.

7. Partições

Chama-se *partição* (ou *classificação*) dum universo U a toda a decomposição de U num conjunto $\{C_1, C_2, \ldots, C_m\}$ de conjuntos disjuntos dois a dois, cuja reunião seja U, isto é, de conjuntos C_i tais que

$$C_i C_k = \varnothing \text{ para } i \neq k \text{ e } \sum_{i=1}^{m} C_i = U.$$

Os conjuntos C_i dir-se-ão *elementos* ou *células* da partição [1].

Paralelamente, chamaremos *partição em atributos* a todo o conjunto $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ de atributos incompatíveis, dois a dois, cuja soma lógica seja o atributo universal.

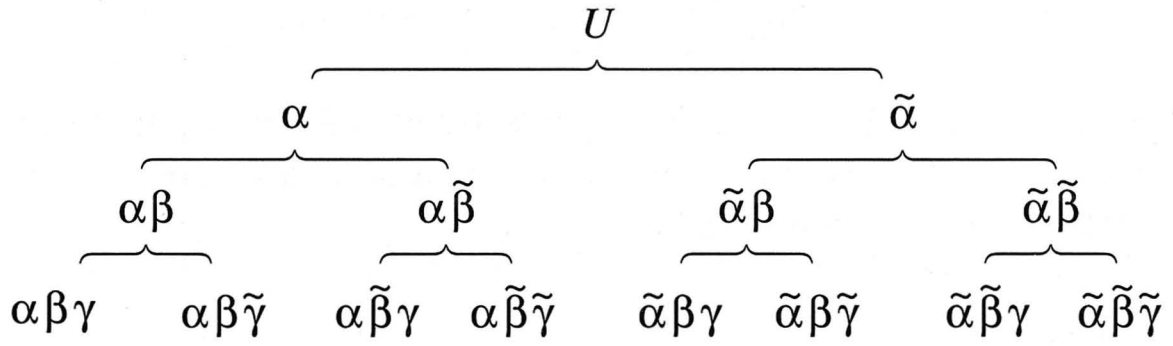
Uma partição diz-se *dupla, tripla*, etc., conforme o número m dos seus elementos for 2, 3,.... Sempre que esse número é superior a dois, a partição diz-se *múltipla* (excluimos o caso sem interesse $m=1$).

O caso mais simples será o das classificações duplas, também chamadas *classificações dicotómicas ou dicotomias*. Estas podem sempre ser determinadas em U, por um atributo α e pelo seu contrário $\tilde{\alpha}$ (é claro que não interessa o caso em que α é impossível ou universal).

Dados vários atributos α, β, γ, ..., sobre U, as operações de negação e de conjunção efectuadas sobre esses atributos e sobre os resultados obtidos conduzem necessariamente a uma partição em atributos. Ora, essa partição pode sempre ser atingida por sucessivas dicotomias, como vamos ver. Para fixar ideias, limitemo-nos ao caso de 3 atributos α, β, γ. As sucessivas dicotomias podem ser esquematizadas do seguinte modo:

---

(1) – Supõe-se nesta definição que o universo é finito, mas é claro que a definição se pode estender imediatamente ao caso dos universos infinitos e das partições com uma infinidade de células.

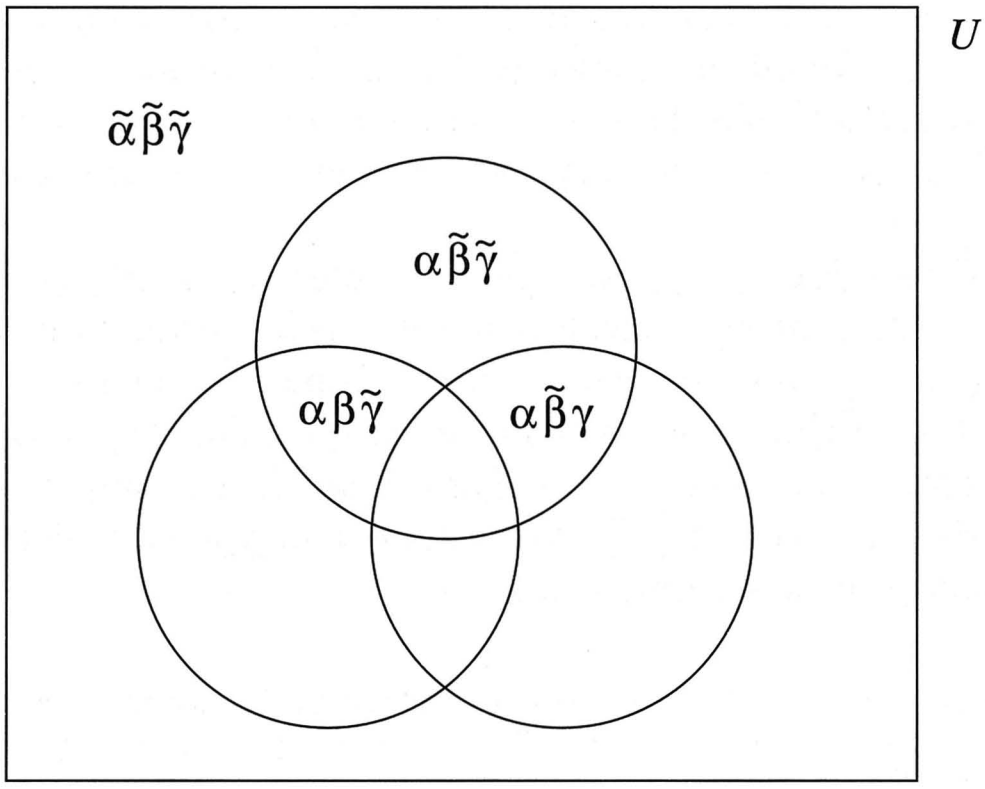


Obtivemos assim, como é fácil verificar, uma partição constituída pelos 2^3 atributos

$$\alpha\beta\gamma,\ \alpha\tilde{\beta}\gamma,\ \alpha\tilde{\beta}\tilde{\gamma},\ \tilde{\alpha}\beta\gamma,\ \tilde{\alpha}\beta\tilde{\gamma},\ \tilde{\alpha}\tilde{\beta}\gamma,\ \tilde{\alpha}\tilde{\beta}\tilde{\gamma},\ \alpha\beta\tilde{\gamma},$$

entre os quais, porém, pode haver algum repetido ou impossível, o que reduz o seu efectivo.

Estes resultados estão sugestivamente figurados no diagrama junto, em que as classes correspondentes aos atributos α, β, γ são representadas por círculos e o universo por um rectângulo.



Suponhamos, por exemplo, que o universo U é constituído pelos portugueses, sendo α o atributo "casado", β o atributo "empregado" e γ o atributo "com menos de 35 anos". Então, $\alpha\beta\gamma$ será o atributo "casado, desempregado, com menos de 35 anos", $\tilde{\alpha}\beta\tilde{\gamma}$ o atributo "não casado, empregado, com idade não inferior a 35 anos", etc.

Toda a partição U^* dum universo U (em conjuntos ou em atributos) pode ser tomada para novo universo, dando-se, por assim dizer, uma *condensação* do universo primitivo. Os novos indivíduos serão, é claro, as células da partição U^*. Este processo de condensação é muitas vezes usado em Estatística para resumir dados. Por exemplo, na Tabela n.º 1 fez-se uma partição dos alunos em classes, conforme as notas obtidas no exame, substituindo-se o universo dos alunos pelo universo das notas; este, por sua vez, pode ser condensado no universo constituído pelos atributos "mau", "medíocre", "suficiente", "bom", "distinto" e "muito bom", o qual, por sua vez, pode ser substituído pelo universo {"aprovado", "reprovado"}.

Estas considerações aplicam-se *mutatis mutandis* ao caso dos acontecimentos. Chamaremos *partição dum acontecimento certo* a todo o sistema de acontecimentos incompatíveis, dois a dois, cuja soma lógica seja um acontecimento certo. (Esta definição, torna-se mais sugestiva, se em vez do termo "acontecimento" usarmos o termo "eventualidade").

8. Corpos de conjuntos, corpos de atributos, corpos de acontecimentos

Consideremos um universo U (finito). Diz-se que uma família $\{C_1, C_2, \ldots, C_r\}$ de conjuntos de elementos de U é um *corpo*, quando se verificam as duas seguintes condições: 1) – a soma ou o produto lógico de dois (ou mais) conjuntos da família ainda é um conjunto da família; 2) – o complemento de cada conjunto da família ainda pertence à mesma.

Analogamente, se define "corpo de atributos" e "corpo de acontecimentos". Neste último caso, devemos supor que se trata dum sistema *finito* de acontecimentos a prever numa dada prova $\mathcal{P}$.

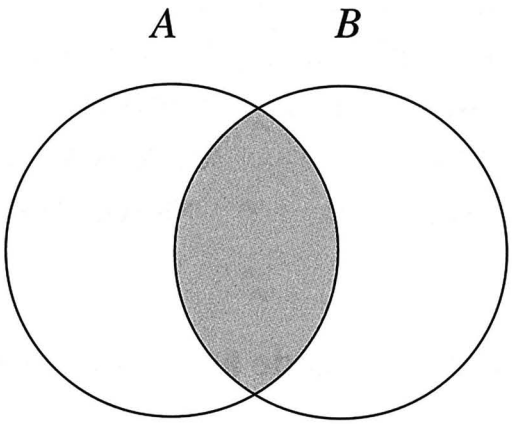
A família de todos os possíveis conjuntos de elementos de U é, manifestamente, um corpo de conjuntos. Mas há outros exemplos de corpos.

Consideremos um corpo $\mathcal{R}$ de conjuntos e seja C um desses conjuntos. Então, $\tilde{C}$ também pertence a $\mathcal{R}$. Logo, também pertencem a $\mathcal{R}$ o conjunto $C + \tilde{C} = U$ e o conjunto $C\tilde{C} = \varnothing$; isto é, *entre os conjuntos dum corpo figuram sempre o universo e o conjunto vazio.*

Chamam-se *células* do corpo $\mathcal{R}$ os conjuntos não vazios de $\mathcal{R}$ que não contêm nenhum outro conjunto de $\mathcal{R}$, a não ser o conjunto vazio. Dado um indivíduo a, a intersecção de todos os conjuntos de $\mathcal{R}$ que contêm a é manifestamente uma célula de $\mathcal{R}$; o menor conjunto de $\mathcal{R}$ que contém a. Tem-se, além disso, o seguinte

TEOREMA. *Sejam A e B duas células de $\mathcal{R}$. Então, de duas uma: ou os conjuntos A e B coincidem ou são disjuntos.*

Com efeito, suponhamos que os conjuntos A, B não coincidem, isto é, que $A \neq B$, mas que a intersecção AB não é vazia.



Então, visto ser $A \neq B$, podemos garantir que um, pelo menos, destes conjuntos não estará contido no outro: seja, por exemplo, A esse conjunto. Nesta hipótese, tanto AB como $A\tilde{B}$ são conjuntos de $\mathcal{R}$ diferentes de A e contidos em A e, como $A = AB + A\tilde{B}$, nenhum deles será vazio. Mas isto é impossível, porque então não seria A uma célula. Logo, ou A e B são disjuntos (isto é, $AB = \varnothing$) ou coincidem (isto é, $A = B$).

Daqui se deduzem logo os seguintes corolários:

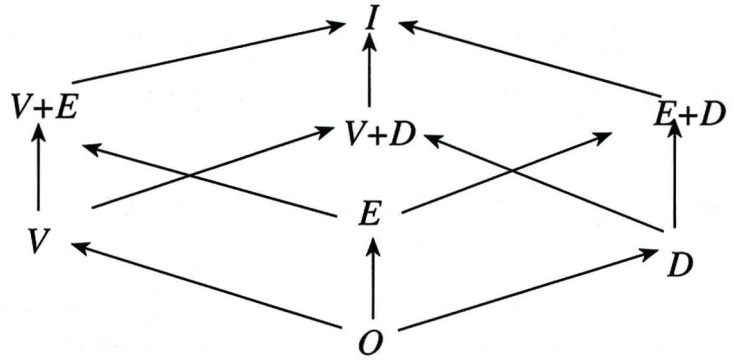
COROLÁRIO I. *As células dum corpo de conjuntos formam uma partição do universo.*

COROLÁRIO II. *Para todo o conjunto A pertencente a um dado corpo $\mathfrak{R}$ de conjuntos, existe um e só um conjunto de células de $\mathfrak{R}$, de que A é soma lógica.*

Posto isto, consideremos uma classe $\mathfrak{F}$ qualquer de subconjuntos de *P*. Efectuando operações de intersecção, reunião e passagem ao complementar, de todos os modos possíveis, a partir dos conjuntos de $\mathfrak{F}$, os resultados obtidos constituem um corpo de conjuntos; ao menor corpo que contem $\mathfrak{F}$, dá-se-lhe o nome de *corpo de conjuntos gerado por* $\mathfrak{F}$. É claro que as células deste corpo são as células da partição determinada por $\mathfrak{F}$, a qual se pode obter por sucessivas dicotomias como foi indicado no n.° anterior. Os restantes conjuntos do corpo (além do conjunto vazio) são reuniões de células, como indica o Corolário II.

Todas estas considerações se aplicam, *mutatis mutandis*, aos corpos de atributos e de acontecimentos.

Consideremos, por exemplo, entre os resultados normais dum desafio de futebol entre dois grupos *A* e *B*, os dois seguintes: *vitória de A* (que designaremos por *V*) e *empate* (que designaremos por *E*). O corpo de acontecimentos gerado por *V* e *E* é formado pelos seguintes elementos: $V + E$ (vitória de *A* ou empate), $\widetilde{V+E}$ (derrota de *A* – acontecimento que designaremos por *D*), $V + D$ (vitória ou derrota de *A*), $E + D$ (empate ou derrota de *A*), $V + D + E$ (vitória ou derrota de *A*, ou empate – acontecimento certo, que designaremos por *I*), *VD* (vitória e derrota de *A* – acontecimento impossível, que designaremos por *O*). Este corpo de eventualidades pode ser representado pelo seguinte esquema, em que usamos a seta como símbolo de implicação:



É claro que *V, E, D* (células do corpo considerado) ainda podem exprimir-se como somas de acontecimentos elementares (indecomponíveis), *não pertencentes a este corpo*. Com efeito, todo o resultado dum jogo se pode exprimir por um par ordenado (m, n) de números inteiros não negativos, sendo m o número de golos marcados por A e n o número de golos marcados por B. Então, por exemplo, o acontecimento V será a soma dos acontecimentos representados pelos pares (m, n) com $m>n$. Por sua vez, os acontecimentos E e D são expressos, respectivamente, pelas funções proposicionais $m=n$ e $m<n$.

Dum modo geral, a totalidade dos acontecimentos a considerar numa dada prova ou experiência $\mathcal{P}$ constitui um corpo, desde que se convencione (como é uso) incluir nessa totalidade o acontecimento impossível[1].

Esta convenção é adoptada para comodidade de linguagem. Na mesma ordem de ideias se convenciona que *o acontecimento impossível implica qualquer outro acontecimento* – convenção, de resto, natural, pois que, só admitindo-a, será inteiramente válida a seguinte propriedade intuitiva:

$$\text{Se } \alpha \subset \beta, \text{ então } \tilde{\beta} \subset \tilde{\alpha}$$

Com efeito, designando por I o acontecimento certo, tem-se $\alpha \subset I$, qualquer que seja α, e portanto, $\tilde{I} \subset \tilde{\alpha}$, qualquer que seja $\tilde{\alpha}$.

Se isto não fosse verdade, a propriedade anterior não seria válida.

Consideremos agora um corpo de atributos e sejam $\alpha_1, \alpha_2, \ldots, \alpha_m$ as suas células. Qualquer outro atributo α do corpo se exprime como soma de células. Ora, como estas são incompatíveis duas a duas, deduz-se do teorema da soma, este outro

TEOREMA. *A frequência (absoluta ou relativa) de qualquer atributo α dum corpo $\mathcal{R}$ de atributos é igual à soma das frequências (absolutas ou relativas) das células de $\mathcal{R}$ cuja soma é α.*

Conclusão análoga para os corpos de acontecimentos.

---

(1) – Subentende-se que a totalidade é finita. Note-se, porém, que o conceito de corpo se pode generalizar ao caso dos conjuntos infinitos (de conjuntos, de atributos ou de acontecimentos).

NOTA. É preciso não confundir um *conjunto de conjuntos* com a *reunião* desses conjuntos. Por exemplo, o conjunto V dos vertebrados está classificado em vários conjuntos parciais: mamíferos, aves, etc.; porém, V não é o *conjunto* desses conjuntos, mas sim a sua *reunião* ou *soma*.

Ora, vimos atrás que, para todo o conjunto C pertencente a um corpo $\mathcal{R}$ de conjuntos, existe um, e um só, conjunto C^* de células de $\mathcal{R}$ cuja soma é C; reciprocamente, todo o conjunto C^* de células de $\mathcal{R}$ determina um conjunto $C \in \mathcal{R}$. Portanto, se designarmos por U^* a partição de U constituída pelas células de $\mathcal{R}$, fica, assim, estabelecida uma correspondência biúnivoca $C \leftrightarrow C^*$ entre os conjuntos C de $\mathcal{R}$ e os subconjuntos de U^*. Além disso, é fácil ver que à soma $C_1 + C_2$ de dois conjuntos C_1, C_2 de $\mathcal{R}$ corresponde, deste modo, a soma $C_1^* + C_2^*$ dos subconjuntos correspondentes de U^*, ao produto $C_1 C_2$ corresponde o produto $C_1^*\ C_2^*$, e ao complementar de C corresponde o complementar de C^*. Exprime-se este facto dizendo que a referida correspondência é um *isomorfismo* entre o corpo $\mathcal{R}$ e o corpo dos subconjuntos de U^*; diz-se, também, que esses dois corpos são *isomorfos*.

Analogamente, se conclui o seguinte facto:

Se $\mathcal{S}$ é um corpo (finito) de acontecimentos e $\mathcal{E}$ o conjunto das células de $\mathcal{S}$, o corpo $\mathcal{S}$ é isomorfo a um corpo de conjuntos, que é precisamente o corpo dos subconjuntos de $\mathcal{E}$.

Por exemplo, o corpo gerado pelos acontecimentos V, E atrás considerados é isomorfo ao corpo dos subconjuntos de $\{V, E, D\}$, que são: este próprio conjunto, o conjunto vazio e, ainda, os conjuntos $\{V\}$, $\{E\}$, $\{D\}$, $\{V, E\}$, $\{V, D\}$, $\{E, D\}$.

Ficam assim completadas as considerações do n.° 3: o isomorfismo em questão traduz exactamente a álgebra dos acontecimentos na álgebra dos conjuntos.

9. Distribuição em universos finitos

Consideremos um corpo $\mathcal{R}$ de conjuntos (num universo U finito). Suponhamos que, a cada conjunto $C \subset \mathcal{R}$, se faz corresponder um determinado número real e designemos esse número por $\Phi(C)$: fica assim definida em $\mathcal{R}$ *a função* $\Phi(C)$, *do conjunto variável* C.

Pois bem, chamaremos *distribuição em* $\mathcal{R}$ toda a função de conjunto, $\Phi(C)$, definida em $\mathcal{R}$, que verifique as duas seguintes condições:

1) – $\Phi(C) \geq 0$, para todo o $C \in \mathcal{R}$.

2) – Se C_1, C_2 são conjuntos *disjuntos* de $\mathcal{R}$, então

$$\Phi(C_1 + C_2) = \Phi(C_1) + \Phi(C_2).$$

Em particular, se $\mathcal{R}$ é formado por todos os subconjuntos de U, diremos, então, que $\Phi(C)$ é uma *distribuição sobre U.*

É fácil reconhecer a veracidade da seguinte proposição:

TEOREMA. *Sejam* $C_1, C_2, \ldots, C_m$ *as células do corpo* $\mathcal{R}$. *Se associarmos a cada conjunto* C_i, *arbitrariamente, um número real* $\gamma_i \geq 0$, *existe sempre uma, e uma só, distribuição* $\Phi(A)$ *em* $\mathcal{R}$ *que, para cada conjunto* C_i, *toma o valor* γ_i $(i = 1, 2, \ldots, m)$. *O valor de* $\Phi(A)$ *para cada* $A \subset \mathcal{R}$ *é, então, a soma dos valores que a função toma nas células cuja soma é A.*

Em particular, se $\mathcal{R}$ é formado por todos os subconjuntos de U, as suas células reduzem-se aos indivíduos (elementos de U) e tem-se:

COROLÁRIO. *Toda a função não negativa definida em U é prolongável numa (e numa só) distribuição sobre U.*

Daqui o chamar-se também, algumas vezes, *distribuição em U* a qualquer função real não negativa definida em U.

É claro que o conceito de "distribuição num corpo de conjuntos" se estende imediatamente ao caso dos corpos de atributos e dos corpos de acontecimentos.

Exemplos de distribuições são-nos oferecidos pelas frequências, tais como atrás foram definidas, tendo em vista o teorema da soma. Mas, além das distribuições de frequência, apresentam-se ainda na prática muitos outros exemplos de distribuições: distribuições de probabilidade (que estudaremos mais adiante), distribuições de massa, distribuições de carga eléctrica, etc., etc.

Chamaremos *distribuição relativa* (num corpo $\mathcal{R}$) a toda a distribuição $\Phi(A)$ em $\mathcal{R}$ que verifique a condição suplementar seguinte:

3) – $\Phi(U) = 1$ (sendo U o universo, como já foi dito).

É fácil ver que, de cada distribuição $\Phi(A)$ em $\mathcal{R}$, se deduz uma distribuição relativa, $\phi(A)$, em $\mathcal{R}$, pondo

$$\phi(A) = \frac{\Phi(A)}{\Phi(U)}\text{, sendo } U \text{ o universo.}$$

Do teorema anterior deduz-se o seguinte

COROLÁRIO. *Se* $\Phi(A)$ *é uma distribuição relativa, tem-se sempre*

$$\Phi(\tilde{A}) = 1 - \Phi(A).$$

A dedução faz-se exactamente como no caso particular das frequências relativas (n.° 5).

10. Soma de conjuntos não disjuntos (atributos ou acontecimentos compatíveis)

Seja $\mu(A)$ uma distribuição definida num corpo $\mathcal{R}$ de conjuntos. Sendo A, B dois conjuntos de $\mathcal{R}$, o uso da fórmula

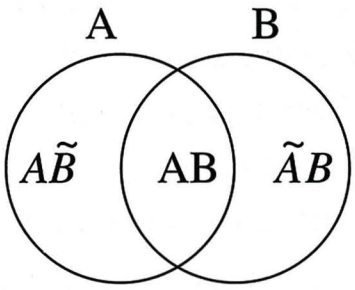
$$\mu(A + B) = \mu(A) + \mu(B)$$

exige a restrição de que A e B sejam disjuntos. Como proceder, porém, no caso geral? Já atrás (n.° 5) se nos pôs esta questão a propósito dum exemplo comezinho.

Basta notar que se tem, necessariamente,

$$A = A\tilde{B} + AB,\ B = AB + \tilde{A}B,\ A + B = A\tilde{B} + AB + \tilde{A}B.$$

Como os conjuntos $A\tilde{B}$, AB, $\tilde{A}B$ são disjuntos dois a dois, ter-se-á:



$$\mu(\mathrm{A}) = \mu(A\tilde{B}) + \mu(AB),\ \mu(B) = \mu(AB) + \mu(\tilde{A}B),$$
$$\mu(A + B) = \mu(A\tilde{B}) + \mu(AB) + \mu(\tilde{A}B)$$

donde $\mu(A\tilde{B}) = \mu(A) - \mu(AB)$, $\mu(\tilde{A}B) = \mu(B) - \mu(AB)$ e, portanto,

$$\mu(A + B) = \mu(A) + \mu(B) - \mu(AB).$$

Para o caso de 3 conjuntos *A, B, C,* bastará aplicar esta fórmula duas vezes:

$$\mu(A + B + C) = \mu[A + (B + C)] = \mu(A) + \mu(B + C) - \mu[A\ (B + C)]$$
$$= \mu(A) + \mu(B) + \mu(C) - \mu(BC) - \mu(AB + AC).$$

Como

$$\mu(AB + AC) = \mu(AB) + \mu(AC) - \mu(ABC),$$

virá, por último

$$\mu(A + B + C) = \mu(A) + \mu(B) + \mu(C) - \mu(AB) -$$
$$- \mu(AC) - \mu(BC) + \mu(ABC).$$

Consideremos agora, em geral, p conjuntos quaisquer, $A_1, A_2, \ldots, A_p$, de $\mathcal{R}$. Por indução, chega-se à fórmula seguinte:

$$\mu\left(\sum_i^p A_i\right) = \sum_i \mu(A_i) - \sum_{i<j} \mu(A_i A_j) + \sum_{i<j<k} \mu(A_i A_j A_k) - \cdots$$

$$\cdots + (-1)^{p-1}\, \mu(A_1 A_2 \ldots A_p)$$

que, no seu domínio inicial foi descoberta pelo matemático português DANIEL DA SILVA, que a indica no seu trabalho *"Propriedades gerais e resolução directa das congruências binómias: Introdução ao estudo da teoria dos números"*.

É claro que estes resultados se aplicam, *mutatis mutandis*, ao caso dos atributos ou acontecimentos compatíveis.

11. Atributos quantitativos

Entre os atributos considerados em inquéritos estatísticos, há que distinguir duas espécies principais – os que são grandezas e os que o não são. Os primeiros são traduzíveis em números, resultados de medições, enquanto para os segundos não faz sentido falar de medida. Estes são chamados *atributos qualitativos ou qualidades,* e aqueles, *atributos quantitativos ou quantidades* (ou, ainda, grandezas).

Como sempre sucede, quando se trata de entes do mundo empírico, a fronteira entre as duas espécies é imprecisa. Há casos intermédios, dúbios, em que se fala de *avaliação* ou *correcção* dum atributo (em vez de medição): tal é o caso da escala de Mohs relativa à dureza dos minerais, tal é, ainda, o caso das cotações que se atribuem aos alunos nos exames etc.

De resto, dada uma partição em atributos (finita), qualquer que ela seja, é sempre possível, com critério mais ou menos convencional, distinguir esses atributos por meio de números. Em particular, no caso da partição dicotómica $\{\alpha, \tilde{\alpha}\}$, formada por um atributo e pelo seu contrário, é natural designar $\tilde{\alpha}$ por 0 (ausência) e α por 1 (presença) ou por -1 e $+1$, etc.

Em correspondência às duas espécies de atributos, apresentam-se, naturalmente, duas modalidades de estatística: a *qualitativa* e a *quantitativa*. A primeira é, também, chamada *estatística de atributos* e a segunda, *estatística de variáveis*; mas estas designações não condizem com a nossa terminologia precedente.

Entre os atributos quantitativos (ou, melhor, entre as partições em tais atributos) há, ainda, que distinguir duas categorias: a dos que formam escalas contínuas (*variáveis contínuas*) e a dos que se dispõem conforme a sucessão de números naturais (*variáveis discretas*). Pertencem à primeira categoria, por exemplo, os comprimentos, as massas, as densidades, etc. Pertencem à segunda categoria por exemplo, os números de pétalas de uma flor, de grãos de uma espiga, etc., etc.

O caso das variáveis contínuas só poderá ser estudado devidamente nos capítulos seguintes, em que trataremos de universos infinitos. Todavia, numa primeira fase, que é a da recolha e classificação dos dados, trabalha-se ainda com universos finitos, tal como vamos ver.

Consideremos uma variável contínua x (variável real), atributo quantitativo definido numa dada população U finita: por exemplo, a *altura* dos indivíduos. Em vez de considerar a frequência de cada um dos valores de x em U (valores estes que são em número infinito), o que há a fazer, na prática, é agrupar esses valores em classes e considerar a frequência de cada uma dessas classes. Por outras palavras, o que há a fazer é uma *classificação* dos valores possíveis de x, ou seja, uma *partição* do conjunto dos números reais. De resto, sendo a população U finita, nem é necessário considerar toda a recta real, visto que, para além de certos limites, a frequência se torna nula. Por exemplo, sendo x a variável "altura" numa população de seres humanos e tomando para unidade o metro, não será necessário, geralmente, sair do intervalo [0, 2], a não ser em casos excepcionais de gigantismo.

Quanto às células da partição, convém que sejam, evidentemente, intervalos, em número finito. Tudo se resume, portanto, na partição dum intervalo limitado $[a, b]$ em intervalos, por meio dum número finito de pontos, $a < x_1 < x_2 < \ldots < x_{n-1} < b$. Todavia, para que se trate efectivamente duma partição, não podem os intervalos ser todos abertos ou todos fechados. Adoptaremos, em regra, intervalos fechados à esquerda e abertos à direita (excepto o último que poderá ser fechado). Assim, para o intervalo $[a, b]$, dividido pelos pontos $x_1, x_2, \ldots, x_n$, teremos a partição

$$[a, x_1[\ , [x_1, x_2[\ , \ldots\ldots, [x_{n-2}, x_{n-1}[\ , [x_{n-1}, b].$$

A tais intervalos dá-se, em Estatística, o nome de *intervalos de classe* ou simplesmente, *classes*. Haverá vantagem, é claro, em que todos esses intervalos tenham o mesmo comprimento e que os pontos de divisão correspondam a números inteiros da unidade escolhida ou dum submúltiplo decimal da mesma.

Posto isto, consideremos, de novo, o atributo quantitativo x relativo à população U. Supondo que se efectuou a medição dessa grandeza em cada um dos indivíduos, os resultados serão recolhidos numa primeira tabela, em que, à frente do número de ordem de cada indivíduo, se indica o valor correspondente de x, por exemplo, a altura desse indivíduo com o grau de aproximação adoptado.

Em seguida, feita a escolha dos intervalos de classe, as observações podem ser classificadas contando, para cada intervalo, o número total de indivíduos a que correspondem valores de x situados nesse intervalo (*frequência absoluta*). Organiza-se, deste modo, uma segunda tabela, que é já uma *tabela de frequências*. O universo U é, então, substituído pelo universo U^* das classes e a tabela de frequência define, manifestamente, uma *distribuição de frequência* sobre U^* (recorde-se o que, a este respeito, se disse no n.° 8). Para ilustrar estas considerações, citaremos um exemplo extraído do trabalho de Daniel Nagore, *Biometria, nociones sobre este método de investigacion en genetica*, Ministerio de Agricultura, Madrid, 1941. O exemplo refere-se a uma sub-variedade de trigo, denominado vulgarmente "Catalão compacto", classificado cientificamente entre os *Triticum vulgare crythrospermum* (Percival) e obtido em campos de estudo por selecção genealógica. Desse trigo foi constituída uma população de 400 espigas, na qual se determinou directamente, em cada indivíduo, *o número de grãos e o comprimento da espiga* (em milímetros). Destes caracteres biométricos, que designaremos, respectivamente por g e c, se deduziram os valores duma outra grandeza: *o número de grãos por 10 cm de comprimento*. Esta grandeza, denominada *densidade de espiga*, é definida em função das primeiras pela fórmula

$$d = \frac{100\, g}{c}\,.$$

Estes dados foram registados numa tabela (que não reproduzimos), na qual, à direita do número da ordem de cada espiga se indica, numa coluna, o valor de g, noutra, o valor de c, e na última, o valor de d, calculado este a partir dos primeiros a menos de 0,01. Em seguida, escolheram-se os intervalos de classe, e daquela primeira tabela deduziu-se a seguinte tabela de frequência:

TABELA N.º 3

Densidade	Frequência absoluta
$14{,}5 \leq d < 15$	1
$15 \leq d < 15{,}5$	1
$15{,}5 \leq d < 16$	1
$16 \leq d < 16{,}5$	1
$16{,}5 \leq d < 17$	1
$17 \leq d < 17{,}5$	4
$17{,}5 \leq d < 18$	5
$18 \leq d < 18{,}5$	5
$18{,}5 \leq d < 19$	6
$19 \leq d < 19{,}5$	14
$19{,}5 \leq d < 20$	6
$20 \leq d < 20{,}5$	44
$20{,}5 \leq d < 21$	31
$21 \leq d < 21{,}5$	34
$21{,}5 \leq d < 22$	50
$22 \leq d < 22{,}5$	43

Densidade	Frequência absoluta
$22{,}5 \leq d < 23$	52
$23 \leq d < 23{,}5$	41
$23{,}5 \leq d < 24$	19
$24 \leq d < 24{,}5$	14
$24{,}5 \leq d < 25$	6
$25 \leq d < 25{,}5$	12
$25{,}5 \leq d < 26$	2
$26 \leq d < 26{,}5$	2
$26{,}5 \leq d < 27$	1
$27 \leq d < 27{,}5$	1
$27{,}5 \leq d < 28$	1
$28 \leq d < 28{,}5$	0
$28{,}5 \leq d < 29$	1
$29 \leq d < 29{,}5$	0
$29{,}5 \leq d < 30$	1

Ao olhar para esta tábua pode haver um equívoco, que é preciso evitar. Por exemplo, à direita da função proposicional $19 \leq d < 19{,}5$ escreveu-se 14. Ora, isto não significa que, para todo o valor de d pertencente àquele intervalo, a frequência é 14, mas, sim, que é esta a frequência do atributo "*densidade maior ou igual a* 19 *e menor que* 19,5" na população considerada. O número 14 corresponde, pois, ao intervalo [19; 19, 5[e não a cada ponto deste intervalo. Por outras palavras: é uma função *de conjunto* – uma *distribuição* – e não uma função de *ponto* (isto é, uma função da variável d). Estas observações vêm, assim, esclarecer o conceito de distribuição introduzido no n.º 9.

No trabalho a que nos referimos faz-se, depois, para determinado efeito, uma condensação dos dados, substituindo os intervalos de comprimento 0,5 por intervalos de comprimento 1 e aplicando a propriedade aditiva das distribuições. Obtém-se, deste modo, a seguinte tabela:

TABELA N.º 4

Densidade	Frequência absoluta
$14 \leq d < 15$	1
$15 \leq d < 16$	2
$16 \leq d < 17$	2
$17 \leq d < 18$	9
$18 \leq d < 19$	11
$19 \leq d < 20$	20
$20 \leq d < 21$	75
$21 \leq d < 22$	84

Densidade	Frequência absoluta
$22 \leq d < 23$	95
$23 \leq d < 24$	60
$24 \leq d < 25$	20
$25 \leq d < 26$	14
$26 \leq d < 27$	3
$27 \leq d < 28$	2
$28 \leq d < 29$	1
$29 \leq d < 30$	1

Note-se que muitas vezes, para se designar cada intervalo de classe, se indica o ponto médio desse intervalo. Assim, na tábua anterior, o intervalo [14, 15[seria designado como a classe 14,5, etc. Nestes casos, convém escolher os intervalos de modo que os pontos médios sejam números inteiros de unidades ou de submúltiplos decimais da unidade. Recordemos que é este o critério usado no arredondamento das médias de notas de alunos; assim, quando se diz que a média é de 12 valores, está-se a indicar, resumidamente, que a média é um número x tal que $11{,}5 \leq x < 12{,}5$; analogamente, dizendo que a média é de 14,3, pretende-se dizer que a média verifica a condição $14{,}35 \leq x < 14{,}36$, etc.

O mesmo critério é ainda usado pelas companhias de seguros na atribuição da idade para o pagamento do prémio, em seguros de vida.

12. Representação gráfica das distribuições: histogramas e polígonos de frequência

Para ter uma visão mais directa e sugestiva duma distribuição de frequência, procede-se à sua representação gráfica, começando por fixar no plano um sistema de eixos coordenados rectangulares, com escalas geralmente diferentes. Uma vez marcados no eixo das abcissas os sucessivos intervalos de classe (não sendo indispensável que o ponto de abcissa 0 coincida com o cruzamento dos eixos), a representação gráfica pode, na prática, ser feita de dois modos diversos:

1.° processo – Sobre o segmento representativo de cada classe, constrói-se um rectângulo de altura (positiva) correspondente à frequência dessa classe. A reunião dos rectângulos assim obtidos é chamada o *histograma* ou *diagrama de colunas* de distribuição considerada. (Na Fig. 1 é dado o histograma correspondente à Tabela n.° 4).

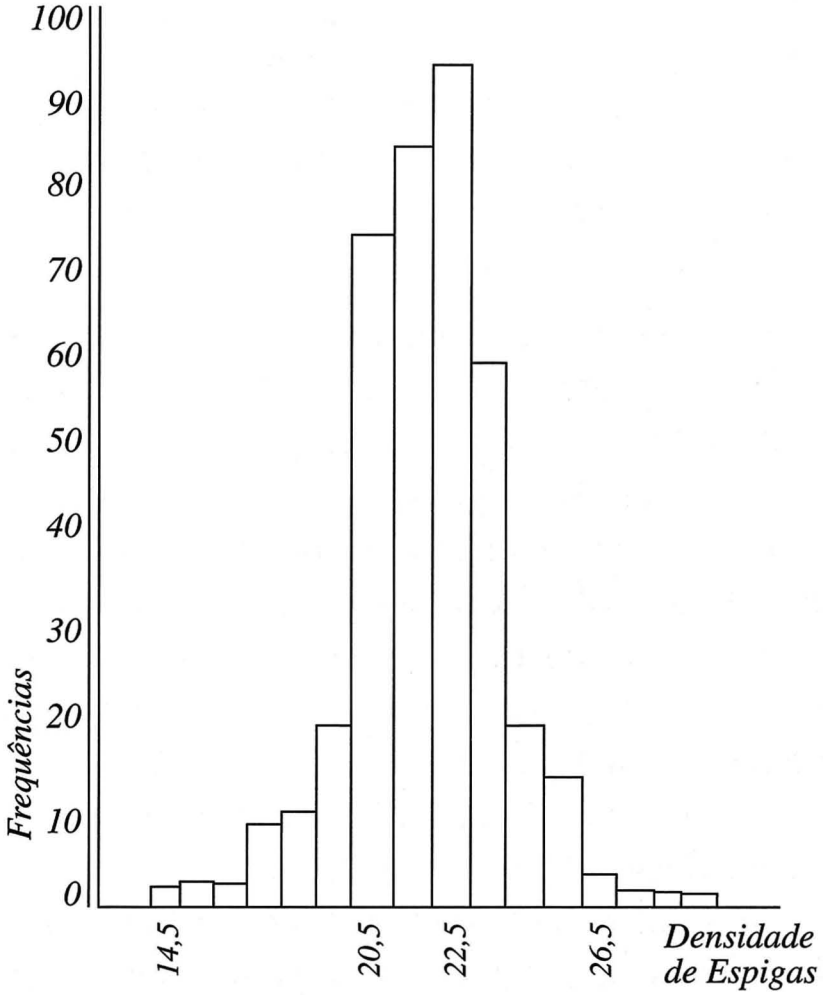


Fig. 1

É claro que as áreas dos diferentes rectângulos são proporcionais às frequências das respectivas classes, de modo que a área do histograma representa, na mesma escala, o número total de indivíduos. Em particular, o quociente da área dum rectângulo (ou duma reunião de rectângulos) pela área do histograma dá a frequência relativa da respectiva classe (ou da respectiva reunião de classes).

2.° processo – Pelo ponto médio da cada intervalo de classe conduz-se uma perpendicular ao eixo das abcissas, sobre a qual se marca o ponto de ordenada igual à frequência dessa classe. Traçam-se, depois, segmentos de recta, unindo entre si os pontos consecutivos assim obtidos e unindo o primeiro e o último destes pontos, respectivamente, com o primeiro e o último extremo dos intervalos marcados. Obtém-se, deste modo, uma linha poligonal, à qual se dá o nome de *polígono de frequência* ou *polígono de Johannsen* da distribuição considerada. Na Fig. 2 são dados os polígonos de frequência correspondentes à Tabela n.° 3 (a traço fino) e à Tabela n.° 4 (a traço cheio).

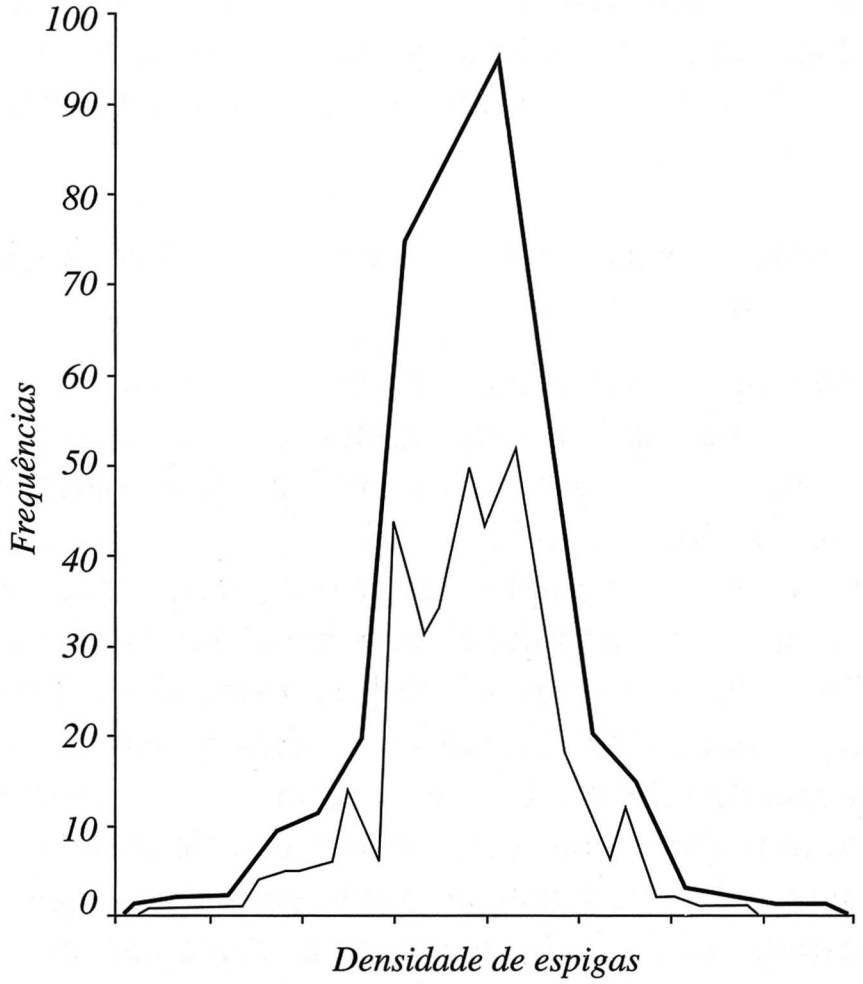


Fig. 2

Observe-se que a primeira poligonal é mais sinuosa, mais irregular do que a segunda.

Dum modo geral, não convém que os intervalos de classe sejam demasiado grandes, pois que, então, se obtém uma informação grosseira da distribuição. Mas também não convém que os intervalos sejam demasiado pequenos, porque, então, se tornam mais sensíveis às influências individuais, rompendo-se aquela *regularidade estatística*, aquele possível esboço duma lei simples que se colhe da visão dum conjunto, mas que se desvanece ao aproximarmo-nos dos indivíduos, dos casos isolados.

É, ainda, manifesto que, uma tal regularidade estatística será tanto maior quanto maior for o número de indivíduos, podendo, então, reduzir-se cada vez mais a amplitude dos intervalos. Deste modo, a poligonal acabará por se confundir, sensivelmente, com uma curva. E, assim, a intuição nos vai aproximando do conceito de distribuição duma variável contínua, que será estudado em rigor a partir do capítulo seguinte. Só depois disso poderemos tirar conclusões dos exemplos agora citados.

13. Independência e associação de atributos. Distribuições de duas ou mais variáveis

Quando se pretende averiguar em que medida dois atributos ou dois fenómenos andam ligados, numa população ou num tipo de experiências, o que há a fazer é um inquérito estatístico de que vamos indicar as formas preliminares.

Suponhamos, por exemplo, que se trata de saber se, entre seres humanos, o atributo "ser míope" está, de qualquer modo, associado ao atributo "ter olhos azuis". O que, ocorre desde logo, é indagar, numa população tão numerosa e variada quanto possível, qual a percentagem de míopes: 1.° – entre indivíduos com olhos azuis; 2.° – entre indivíduos com olhos não azuis. Se as duas percentagens são sensivelmente iguais, há razão para pensar que os atributos considerados são *independentes*; se as percentagens se afastam consideravelmente uma da outra, seremos inclinados a admitir que os dois atributos estão *associados* ou *correlacionados*: em sentido *positivo*, se a primeira percentagem é maior que a segunda, em sentido *negativo*, no caso oposto.

Analogamente se procederia para averiguar, por exemplo, se o fumar está ou não ligado com doenças de estômago, se um insecti-

cida é ou não eficaz no tratamento de certas plantas, etc., etc. Nestes exemplos, começa a desenhar-se a hipótese duma relação de causalidade entre dois fenómenos.

Estas considerações podem ser teorizadas do seguinte modo. Sejam α, β dois atributos definidos num dado universo U, composto de N indivíduos. Para indicar as frequências destes atributos e dos seus contrários, pode fazer-se uso duma tabela de duas entradas, do seguinte tipo:

	β	$\tilde{\beta}$	Total
α	$(\alpha\beta)$	$(\alpha\tilde{\beta})$	(α)
$\tilde{\alpha}$	$(\tilde{\alpha}\beta)$	$(\tilde{\alpha}\tilde{\beta})$	$(\tilde{\alpha})$
Total	(β)	$(\tilde{\beta})$	N

Para brevidade de linguagem diremos "os α" em vez de "os indivíduos com o atributo α", e, analogamente, para β.

A proporção dos α em U é $(\alpha)/N$ (frequência relativa de α). A proporção dos α entre os β será $(\alpha\beta)/(\beta)$; chamar-lhe-emos *frequência condicional* de α *relativa a* β e representá-la-emos por $\mathrm{fr}(\alpha|\beta)$. Ter-se-á, pois, por definição:

$$\mathrm{fr}(\alpha|\beta) = \frac{(\alpha|\beta)}{(\beta)} = \frac{\mathrm{fr}(\alpha\beta)}{\mathrm{fr}(\beta)}. \tag{13.1}$$

O atributo α dir-se-á *independente de* β *em* U se a proporção dos α entre os β é igual à proporção dos α entre os não β, isto é, se

$$\mathrm{fr}(\alpha|\beta) = \mathrm{fr}(\alpha|\tilde{\beta}), \text{ ou seja, se } \frac{(\alpha\beta)}{(\beta)} = \frac{(\alpha\tilde{\beta})}{(\tilde{\beta})}.$$

Mas, recordemos a seguinte propriedade das fracções:

$$Se\ \frac{a}{b} = \frac{c}{d},\ então\ \frac{a+c}{b+d} = \frac{a}{b},\ e\ reciprocamente.^{(1)}$$

Ter-se-á, portanto:

$$Se\ \frac{(\alpha\beta)}{(\tilde{\beta})} = \frac{(\alpha\tilde{\beta})}{\beta},\ então\ \frac{(\alpha\beta)+(\alpha\tilde{\beta})}{(\beta)+(\tilde{\beta})} = \frac{(\alpha\beta)}{(\beta)},\ e\ reciprocamente.$$

Mas $(\alpha\beta) + (\alpha\tilde{\beta}) = (\alpha)$ e $(\beta) + (\tilde{\beta}) = N$. Portanto, dizer que α é independente de β equivale a dizer que

$$\frac{(\alpha\beta)}{(\beta)} = \frac{(\alpha)}{N}, \tag{13.2}$$

isto é, que *a proporção dos* α *entre os* β *é igual à proporção dos* α *na população total.* A fórmula precedente, pode ainda, escrever-se

$$\mathrm{fr}(\alpha|\beta) = \mathrm{fr}(\alpha).$$

Logo, dizer que α é independente de β equivale a dizer que *a frequência condicional de* α *relativa a* β *é igual à frequência relativa de* α *(em U).*

Uma outra maneira de escrever (13.2) é a seguinte:

$$(\alpha\beta) = \frac{(\alpha)\,(\beta)}{N}, \tag{13.3}$$

fórmula esta que, pela sua simetria em α e β nos mostra imediatamente que:

---

(1) – Com efeito, se pusermos $\frac{a}{b} = \frac{c}{d} = k$, virá $a = bk$, $c = dk$, $a + c = (b + d)k$ e, portanto,

$$\frac{a+c}{b+d} = k = \frac{a}{b}.$$

Se α *é independente de* β, *também* β *é independente de* α.

Diremos, então, simplesmente, que os atributos α, β são *independentes* (em *U*).

Mas a fórmula (13.3) ainda se pode escrever com o aspecto

$$\frac{(\alpha\beta)}{N} = \frac{(\alpha)}{N} \cdot \frac{(\beta)}{N}, \text{ ou seja, } \mathrm{fr}(\alpha\beta) = \mathrm{fr}(\alpha) \cdot \mathrm{fr}(\beta),$$

donde o

TEOREMA DO PRODUTO. *Se dois atributos* α, β *são independentes, a frequência relativa de* $\alpha\beta$ *é igual ao produto das frequências relativas de* α *e de* β. *Reciprocamente, se esta condição se verifica, os dois atributos são independentes.*

Mais geralmente, da definição (13.1) de frequência condicional, deduz-se

$$\mathrm{fr}(\alpha\beta) = \mathrm{fr}(\beta) \cdot \mathrm{fr}(\alpha|\beta)$$

ou, ainda, trocando os papéis de α e de β

$$\mathrm{fr}(\alpha\,\beta) = \mathrm{fr}(\alpha) \cdot \mathrm{fr}(\beta|\alpha),$$

o que se exprime dizendo que:

A frequência relativa de $\alpha\beta$ *é igual ao produto da frequência relativa de* α *pela frequência condicional de* β *relativa a* α *ou (vice-versa).*

É claro que o teorema do produto se obtém como caso particular desta última proposição, pois que se tem $\mathrm{fr}(\beta|\alpha) = \mathrm{fr}(\beta)$, se e só se, α e β são independentes.

Suponhamos, agora, que α e β não são independentes em *U*. Então, de duas uma:

1) ou $\dfrac{(\alpha\beta)}{(\beta)} > \dfrac{(\alpha)}{N}$ (isto é, a proporção dos α é maior entre os β que no universo inteiro),

2) ou $\frac{(\alpha\beta)}{(\beta)} < \frac{(\alpha)}{N}$ (isto é, a proporção dos α é menor entre os β que no universo inteiro).

No primeiro caso, tem-se, mais simetricamente,

$$(13.4) \qquad (\alpha\beta) > \frac{(\alpha)\,(\beta)}{N} \text{ ou } \mathrm{fr}(\alpha\beta) > \mathrm{fr}(\alpha)\,\mathrm{fr}(\beta)$$

e diz-se que os atributos α, β estão *associados positivamente*, ou, simplesmente, *associados* (em U).

No segundo caso, também se pode escrever

$$(13.5) \qquad (\alpha\beta) < \frac{(\alpha)\,(\beta)}{N} \text{ ou } \mathrm{fr}(\alpha\beta) < \mathrm{fr}(\alpha)\,\mathrm{fr}(\beta),$$

e diz-se que α, β estão *associados negativamente ou desassociados* (em U).

Os casos extremos são aqueles em que a associação ou a desassociação é *completa*. Diremos que os atributos α, β estão *completamente associados* (em U) quando um deles, pelo menos, implica o outro, isto é, quando se tem:

$$\text{ou } \alpha \subset \beta \text{ (todo o } \alpha \text{ é um } \beta) \text{ ou } \beta \subset \alpha \text{ (todo o } \beta \text{ é um } \alpha),$$

podendo, em particular, verificar-se as duas hipóteses (será, então, α equivalente a β). É claro, que se tem $\alpha \subset \beta$ ou $\beta \subset \alpha$ conforme for $(\alpha) = (\alpha\beta)$ ou $(\beta) = (\alpha\beta)$. Portanto, dizer que os atributos α, β são completamente associados equivale a dizer que $(\alpha\beta)$ é igual ao menor dos números (α), (β).

Os atributos α, β dizem-se *completamente desassociados* em U quando se verifica uma, pelo menos, das seguintes condições:

$$\alpha\beta = \varnothing, \qquad \tilde{\alpha}\tilde{\beta} = \varnothing,$$

isto, é quando são incompatíveis os atributos α, β (nenhum α é β) ou os seus contrários (todo o $\tilde{\alpha}$ é β).

Deste modo, é natural considerar a diferença entre $(\alpha\beta)$ e o valor $(\alpha)\,(\beta)/N$ (chamado *valor de independência*) como um índice da

intensidade da associação ou desassociação. É costume designar por δ esta diferença e por $(\alpha\beta)_0$ o valor de independência. Tem-se, pois:

$$(\alpha\beta)_0 = \frac{(\alpha)\,(\beta)}{N}\,,\quad \delta = (\alpha\beta) - (\alpha\beta)_0\,.$$

No entanto, para ter uma ideia comparativa dos graus de associação, é necessário substituir δ por um índice *relativo*. Entre os vários índices que têm sido propostos para este efeito, o mais simples é o seguinte, chamado *coeficiente de associação*:

$$Q = \frac{(\alpha\beta)\,(\tilde{\alpha}\tilde{\beta}) - (\alpha\tilde{\beta})\,(\tilde{\alpha}\beta)}{(\alpha\beta)\,(\tilde{\alpha}\tilde{\beta}) + (\alpha\tilde{\beta})\,(\tilde{\alpha}\beta)} = \frac{N\,\delta}{(\alpha\beta)\,(\tilde{\alpha}\tilde{\beta}) + (\alpha\tilde{\beta})(\tilde{\alpha}\beta)}\,.$$

É claro que se tem $Q = 0$ se e só se os atributos são independentes (o que equivale a ser $\delta = 0$). Os valores extremos de Q são $+1$ e -1:

Tem-se $Q = +1$, se e só se os atributos estão completamente associados, pois que, só nesse caso um dos factores $(\alpha\tilde{\beta})$, $(\tilde{\alpha}\beta)$ se anula.

Tem-se $Q = -1$, se e só se os atributos estão completamente desassociados, pois que, só então, se tem $(\alpha\beta) = 0$ ou $(\tilde{\alpha}\tilde{\beta}) = 0$.

Nos dois números seguintes, estudaremos um outro índice de associação, de grande interesse estatístico.

É fácil ver, ainda, que a condição (13.4) de associação positiva (simétrica em α e β) é equivalente a qualquer das seguintes (assimétricas):

$$\frac{(\alpha\beta)}{(\beta)} > \frac{(\alpha\tilde{\beta})}{(\tilde{\beta})} \tag{13.6}$$

$$\frac{(\alpha\beta)}{(\alpha)} > \frac{(\tilde{\alpha}\beta)}{(\tilde{\alpha})} \tag{13.7}$$

e, analogamente, para a associação negativa. Ora, na prática, qualquer destas comparações é, geralmente, mais elucidativa do que a

primeira para dar uma ideia do tipo de associação. Segundo a disposição adoptada na tabela-tipo atrás considerada, a fórmula (13.6) corresponde a uma *comparação por colunas* e a fórmula (13.7) a uma *comparação por linhas*. Qual das duas deve ser preferida? Isso depende da hipótese que se tiver em mente ao pôr o problema. Se a relação suposta entre os atributos é assimétrica, próxima duma das implicações $\alpha \subset \beta$ ou $\beta \subset \alpha$, deve preferir-se a comparação por linhas ou por colunas, conforme for α ou β que se presume no papel de *antecedente ou causa*. Se, porém, a relação suposta for de reciprocidade, vizinha da equivalência $\alpha \equiv \beta$, é preferível a comparação entre $(\alpha\beta)$ e $(\alpha\beta)_0$, como se faz nas fórmulas (13.4) e (13.5).

Consideremos o seguinte exemplo indicado por YULE e KENDALL (obra citada, p. 40). Trata-se de investigar a *associação entre inoculação contra a cólera e a isenção do ataque*; obtiveram-se os seguintes resultados:

TABELA N.º 5

	Não atacados	Atacados	Total
Inoculados	276	3	279
Não inoculados	473	66	539
Total	749	69	818

Pressupõe-se, é claro, uma relação assimétrica, com α no papel de causa. Convirá, pois, fazer uma comparação por linhas:

1) – Percentagem de inoculados que não foram atacados:

$$\frac{(\alpha\beta)}{(\alpha)} = \frac{276}{279} = 98{,}9\%$$

2) – Percentagem de não inoculados que não foram atacados:

$$\frac{(\tilde{\alpha}\beta)}{(\tilde{\alpha})} = \frac{473}{539} = 87{,}9\%$$

São, talvez, mais sugestivos (embora equivalentes) as proporções complementares:

1') – Percentagem de inoculados que foram atacados: 1,1

2') – Percentagem de não inoculados que foram atacados: 12,2

Em qualquer das formas é visível a associação positiva entre a *inoculação* e a *isenção do ataque* (ou associação negativa entre *inoculação e ataque*).

A comparação por colunas indica, igualmente, uma associação apreciável, mas é claro que não responde directamente à questão implícita no inquérito:

1'') – Percentagem de não atacados que foram inoculados: 36,8

2'') – Percentagem de atacados que foram inoculados: 4,3

14. Associação e independência de partições múltiplas. Tábuas de contingência

Consideremos, agora, mais geralmente, duas partições em atributos

$$\{\alpha_1, \alpha_2, \ldots, \alpha_p\} \{\beta_1, \beta_2, \ldots, \beta_q\}$$

num mesmo universo U com N indivíduos. Seja α a variável que toma por valores $\alpha_1, \alpha_2, \ldots, \alpha_p$ e seja β a variável que toma por valores $\beta_1, \beta_2, \ldots, \beta_q$. As duas partições, cruzadas sobre U, determinam uma nova partição, mais fina, constituída pelos $p.q$ atributos $\alpha_i\ \beta_k$ ($i = 1, 2, \ldots, p$; $k = 1, 2, \ldots, q$); em particular, alguns destes atributos, podem ser impossíveis, o que reduz o seu número efectivo. As frequências dos atributos α_i, β_k podem ser indicadas numa tabela de duas entradas do seguinte tipo:

α \ β	β_1	β_2	.....	β_q	Totais
α_1	$(\alpha_1\ \beta_1)$	$(\alpha_1\ \beta_2)$	.....	$(\alpha_1\ \beta_q)$	(α_1)
α_2	$(\alpha_2\ \beta_1)$	$(\alpha_2\ \beta_2)$	.....	$(\alpha_2\ \beta_q)$	(α_2)
.....	.....	.....	.....	.....	.....
α_p	$(\alpha_p\ \beta_1)$	$(\alpha_p\ \beta_2)$	.....	$(\alpha_p\ \beta_p)$	(α_p)
Totais	(β_1)	(β_2)	.....	(β_q)	N

Trata-se, manifestamente, duma generalização do tipo de tábuas consideradas no número precedente, para o caso de duas partições dicotómicas ($p = q = 0$). Dá-se-lhe o nome de *tábuas de contingência*. Como exemplo, indicaremos, a seguinte tabela apresentado por YULE e KENDALL (obra citada, p. 66), relativa às 2 variáveis "cor dos olhos" e "cor dos cabelos", numa população constituída por 6.800 habitantes masculinos de Baden.

TABELA N.º 6

Cor dos olhos \ Cor dos cabelos	Louro	Castanho	Preto	Ruivo	Total
Azul	1.768	807	189	47	2.811
Cinzento ou verde	946	1.387	746	53	3.132
Castanho	115	438	288	16	857
Total	2.829	2.632	1.223	116	6.800

Aqui, como se vê, a variável "cor dos olhos" é susceptível de três valores ("azul", "cinzento ou verde" e "castanho"), enquanto a variável "cor dos cabelos" é susceptível de quatro valores ("louro", "castanho, "preto" e "ruivo").

Dum modo geral, diremos que a variável α é *independente* da variável β (em U), quando cada um dos valores de α é independente de cada um dos valores de β, no sentido precisado no número precedente; isto é, quando se tiver

$$(\alpha_i\, \beta_k) = \frac{(\alpha_i)\,(\beta_k)}{N}, \text{ ou seja, } \mathrm{fr}(\alpha_i\, \beta_k) = \mathrm{fr}(\alpha_i)\, \mathrm{fr}(\beta_k)$$

para todos os valores possíveis de i e de k. É claro, que se α é independente de β, também β é independente de α: diremos, então, simplesmente, que *as variáveis*, α, β *são independentes* em U.

Se tal não acontecer, poremos, ainda,

$$(\alpha_i\, \beta_k)_0 = \frac{(\alpha_i)\,(\beta_k)}{N}, \quad \text{para } i = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, q,$$

e chamaremos aos números $(\alpha_i \beta_k)_0$ *valores de independência*. Por sua vez, chamaremos *discrepâncias* às diferenças entre os valores observados e os valores de independência, isto é, aos números

$$\delta_{ik} = (\alpha_i \beta_k) - (\alpha_i \beta_k)_0, \text{ para } i = 1, 2, \ldots, p; \; k = 1,2, \ldots, q.$$

Desde já, convém salientar que *as discrepâncias não são algebricamente independentes*. Tem-se, com efeito, para cada valor de i:

$$\delta_{i1} + \delta_{i2} + \cdots + \delta_{iq} = \sum_{k=1}^{q} \left[(\alpha_i \beta_k) - \frac{(\alpha_i)\,(\beta_k)}{N} \right]$$

$$= \sum_{k=1}^{q} (\alpha_i \beta_k) - \sum_{k=1}^{q} \frac{(\alpha_i)\,(\beta_k)}{N}.$$

Mas

$$\sum_{k=1}^{q} (\alpha_i \beta_k) = (\alpha_i \beta_1) + (\alpha_i \beta_2) + \cdots + (\alpha_i \beta_q) = (\alpha_i)$$

e

$$\sum_{k=1}^{q} \frac{(\alpha_i)\,(\beta_k)}{N} = \frac{(\alpha_i)}{N} \sum_{k=1}^{q} (\beta_k) = \frac{(\alpha_i)}{N} \cdot N = (\alpha_i).$$

Logo

$$\delta_{i1} + \delta_{i2} + \cdots + \delta_{iq} = 0, \quad \text{para} \quad i = 1, 2, \ldots, p;$$

isto é, *a soma das discrepâncias em cada linha é igual a zero*.

Analogamente se conclui que é *nula a soma das discrepâncias em cada coluna*, isto é, que:

$$\delta_{1k} + \delta_{2k} + \cdots + \delta_{pk} = 0, \quad \text{para} \quad k = 1, 2, \ldots, q.$$

O número total dos δ_{ik} é pq. Mas é claro que basta conhecer os valores das discrepâncias em $p-1$ linhas e em $q-1$ colunas, pois

os valores correspondentes à linha e à coluna restantes se determinam por meio das relações estabelecidas. É fácil ver, agora, que não existem outras relações obrigatórias entre as discrepâncias; o número total dos δ_{ik} *independentes* é, pois, $(p-1)\,(q-1)$. Dá-se-lhe o nome de *número de graus de liberdade* da tábua em questão.

O conjunto das discrepâncias dá uma ideia do grau de associação entre os diferentes valores de α e de β. Mas convém resumir esses dados num índice único, de maneira tão simples e elucidativa quanto possível. A soma dos δ_{ik} não serve, visto ser nula, como vimos. Um índice independente dos sinais dos δ_{ik} é o que se obtém *dividindo o quadrado de cada discrepância pelo correspondente valor de independência e somando os resultados obtidos*. Este índice que se representa por χ^2 (*qui quadrado*) e se chama *contingência quadrática*, apresenta, como veremos, um grande interesse em ciências experimentais. Será, pois, por definição,

$$\chi^2 = \sum \frac{\delta_{ik}^2}{(\alpha_i \beta_k)_0}$$

(Segundo um hábito corrente em Estatística, omitiremos as indicações relativas aos índices nos somatórios, quando não houver perigo de confusão. Note-se que muitos autores usam a letra *S* em vez de Σ como símbolo de somatório).

Na prática, procede-se do seguinte modo para a determinação dos δ_{ik} e do χ^2. Começa-se por deduzir da tábua inicial uma segunda tábua com os valores de independência nos lugares dos valores observados. Em seguida, constroi-se a tabela das discrepâncias por simples subtracção entre os valores da primeira e os correspondentes da segunda. Finalmente, calcula-se o χ^2 segundo a definição deste índice.

No cálculo dos valores de independência convém, ainda, observar certos preceitos práticos. Visto que se tem

$$(\alpha_i\, \beta_k)_0 = \frac{(\alpha_i)\,(\beta_k)}{N} = (\alpha_i)\,\mathrm{fr}(\beta_k) = (\beta_k)\,\mathrm{fr}(\alpha_i),$$

pode-se começar por calcular, por exemplo, a frequência relativa de cada β e multiplicá-la, depois, pelas frequências absolutas de todos os α (ou vice-versa); sem esquecer que a soma das frequências relativas de todos os β ou de todos os α é 1, isto é:

$$\mathrm{fr}(\alpha_1) + \mathrm{fr}(\alpha_2) + \cdots + \mathrm{fr}(\alpha_p) = 1,$$
$$\mathrm{fr}(\beta_1) + \mathrm{fr}(\beta_2) + \cdots + \mathrm{fr}(\beta_q) = 1;$$

portanto, uma vez calculadas todas as frequências relativas menos uma, a restante obtém-se por subtracção da unidade. Daremos exemplos de cálculo do χ^2 no n.° 16, em que abordaremos o estudo interpretativo deste índice.

15. Associações parciais de atributos. Independência de atributos no caso em que o seu número é superior a dois

Suponhamos agora, mais geralmente, que no universo U são dadas várias partições em atributos. Podemos limitar-nos ao caso em que as partições são dicotómicas. Consideremos, pois, as partições determinadas em U por vários atributos

$$\alpha, \beta, \gamma, \delta, \ldots$$

e pelos seus contrários (em número finito).

Além da frequência condicional de α a respeito de β, podemos, agora, considerar:

1) – A frequência condicional de γ a respeito de α e β:

$$\mathrm{fr}(\gamma|\alpha\beta) = \frac{(\alpha\beta\gamma)}{(\alpha\beta)} = \frac{\mathrm{fr}(\alpha\beta\gamma)}{\mathrm{fr}(\alpha\beta)};$$

2) – A frequência condicional de δ a respeito de α, β e γ:

$$\mathrm{fr}(\delta|\alpha\beta\gamma) = \frac{(\alpha\beta\gamma\delta)}{(\alpha\beta\gamma)} = \frac{\mathrm{fr}(\alpha\beta\gamma\delta)}{\mathrm{fr}(\alpha\beta\gamma)};$$

etc., etc.

Destas definições resulta, desde logo, que se tem, em geral, a *fórmula do produto*:

$$(15.1) \qquad \mathrm{fr}(\alpha\beta\gamma\delta\ldots) = \mathrm{fr}(\alpha)\,\mathrm{fr}(\beta|\alpha)\,\mathrm{fr}(\gamma|\alpha\beta)\,\mathrm{fr}(\delta|\alpha\beta\gamma)\ldots,$$

fórmula esta que continua a ser válida permutando entre si, de todos os modos possíveis, as letras α, β, γ, δ, ...

Os atributos β, γ dizem-se *parcialmente independentes a respeito de* α, se

$$\mathrm{fr}(\gamma|\alpha\beta) = \mathrm{fr}(\gamma|\alpha)$$

isto é, se

$$\frac{(\alpha\beta\gamma)}{(\alpha\beta)} = \frac{(\alpha\gamma)}{(\alpha)},$$

ou, ainda, o que é equivalente, se:

$$(\alpha\beta\gamma) = \frac{(\alpha\beta)\,(\alpha\gamma)}{(\alpha)}.$$

Caso contrário, os atributos β, γ dir-se-ão *parcialmente associados a respeito de* α *– positivamente ou negativamente*, conforme for[1]

$$(\alpha\beta\gamma) \gtrless \frac{(\alpha\beta)\,(\alpha\gamma)}{(\alpha)}.$$

O caso da independência parcial também se pode conceber como associação parcial nula.

Analogamente, se define associação parcial de 3 ou mais atributos a respeito de um ou mais atributos.

---

(1) – O conceito de associação parcial a respeito de α corresponde, assim, a substituir o universo inicial U pelo universo dos α_i, isto é, pelo conjunto dos indivíduos com o atributo α.

Note-se que, para 3 atributos α, β, γ, se têm ao todo 3 associações totais (de α com β, de α com γ e de β com γ) e 6 associações parciais (de α com β a respeito de γ, de α com γ a respeito de β, etc.).

Os atributos α, β, γ, δ, ..., dizem-se *independentes* em U, se são nulas todas as possíveis associações (totais e parciais) entre estes atributos e seus contrários. Da fórmula (15.1) deduz-se a seguinte generalização do

TEOREMA DO PRODUTO. *Se os atributos* α, β, γ, ..., *são independentes, tem-se*

$$\text{(15.2)} \qquad \mathrm{fr}(\alpha\beta\gamma\ldots) = \mathrm{fr}(\alpha)\,\mathrm{fr}(\beta)\,\mathrm{fr}(\gamma)\ldots$$

Todavia, pode reconhecer-se com exemplos que a recíproca não é, agora, verdadeira.

Também importa salientar que o facto de os atributos serem independentes dois a dois não implica a fórmula (15.2), ao contrário do que poderia pensar-se à primeira vista.

É, porém, verdadeiro o seguinte teorema, que contém o anterior como caso particular:

TEOREMA. *Para que os atributos* α, β, γ, ... *sejam independentes, é necessário e suficiente que sejam válidas, não só a fórmula*

$$\mathrm{fr}(\alpha\beta\gamma\ldots) = \mathrm{fr}(\alpha)\,\mathrm{fr}(\beta)\,\mathrm{fr}(\gamma)\ldots$$

como todas as que se deduzem desta, substituindo um ou mais dos atributos α, β, γ, ..., *pelos seus contrários.*

A condição é, evidentemente, necessária: se os atributos dados são independentes, a frequência condicional de cada atributo (ou do seu contrário) a respeito de um ou mais dos restantes atributos (ou dos seus contrários) é igual à frequência relativa do primeiro no universo. Então, a fórmula (15.1) simplifica-se tomando o aspecto (15.2), pois que

$$\mathrm{fr}(\beta|\alpha) = \mathrm{fr}(\beta), \quad \mathrm{fr}(\gamma|\alpha\beta) = \mathrm{fr}(\gamma|\alpha) = \mathrm{fr}(\gamma), \text{ etc.;}$$

e o mesmo se pode dizer da fórmula que se deduz de (15.1), substituindo um ou mais dos atributos pelos seus contrários.

Demonstremos agora que a condição é suficiente. Limitar-nos-emos ao caso de 3 atributos α, β, γ, pois que, no caso geral, a demonstração é análoga. Por se ter

$$\alpha\beta = \alpha\beta(\gamma + \tilde{\gamma}) = \alpha\beta\gamma + \alpha\beta\tilde{\gamma},$$

será

$$\mathrm{fr}(\alpha\beta) = \mathrm{fr}(\alpha\beta\gamma) + \mathrm{fr}(\alpha\beta\tilde{\gamma}).$$

Então, se forem verdadeiras as igualdades

$$\mathrm{fr}(\alpha\beta\gamma) = \mathrm{fr}(\alpha).\,\mathrm{fr}(\beta).\,\mathrm{fr}(\gamma), \qquad \mathrm{fr}(\alpha\beta\tilde{\gamma}) = \mathrm{fr}(\alpha).\,\mathrm{fr}(\beta).\,\mathrm{fr}(\tilde{\gamma}),$$

virá

$$\mathrm{fr}(\alpha\beta) = \mathrm{fr}(\alpha).\,\mathrm{fr}(\beta).\,\mathrm{fr}(\gamma) + \mathrm{fr}(\alpha).\,\mathrm{fr}(\beta).\,\mathrm{fr}(\tilde{\gamma}) = \mathrm{fr}(\alpha).\,\mathrm{fr}(\beta),$$

donde

$$\mathrm{fr}(\alpha|\beta) = \frac{\mathrm{fr}(\alpha\beta)}{\mathrm{fr}(\alpha)} = \mathrm{fr}(\beta),$$

$$\mathrm{fr}(\gamma|\alpha\beta) = \frac{\mathrm{fr}(\alpha\beta\gamma)}{\mathrm{fr}(\alpha\beta)} = \mathrm{fr}(\gamma),$$

e, analogamente, para os restantes casos.

Para complementos e exemplos sobre este ponto, veja-se a citada obra de YULE e KENDALL.

Note-se, ainda, que todas as considerações deste número e dos dois precedentes se estendem, mutatis mutandis, ao caso dos acontecimentos.

16. Interpretação duma tábua de contingência. Testes de significância

Vamos, agora, abordar o problema da interpretação duma tábua de contingência, aproximando-nos do conceito de probabilidade.

Como primeiro exemplo, consideremos o seguinte, extraído da obra de FINNEY citada na advertência prévia:

Trata-se duma experiência hipotética relativa à protecção de animais contra determinada doença. Supõe-se que o experimentador dispõe de 300 animais, e que decide submeter 100 destes a um certo tratamento, reservando os outros 200 para controle. Supõe-se, além disso, que são os seguintes os resultados obtidos:

TABELA N.º 7

	Sobreviventes	Mortos	Total	Mortos %
Não tratados	152	48	200	24
Tratados	88	12	100	12
Total	240	60	300	20

É claro que se pretende averiguar em que medida estes resultados autorizam a admitir uma real influência do tratamento na doença.

Para isso, começa-se por adoptar uma atitude psicológica, que faz lembrar o método de demonstração por redução ao absurdo usado em Matemática, e que consiste em admitir precisamente o contrário do que se pretende estabelecer: *suponhamos que o tratamento não influi na doença nem num sentido nem noutro*. Nisto consiste a chamada *hipótese nula*. Admitida esta hipótese, seria natural esperar que os resultados obtidos fossem aqueles a que, no n.º 13, chamámos valores de independência e que constam da seguinte tabela:

TABELA N.° 8

	Sobreviventes	Mortos	Total	Mortos %
Não tratados	160	40	200	20
Tratados	80	20	100	20
Total	240	60	300	20

Observe-se como foi construída esta tabela:

De acordo com as instruções práticas do n.° 14, calculou-se a frequência relativa dos mortos, ou seja,

$$\frac{60}{300} = 0{,}2 = 20\%;$$

multiplicou-se depois, este resultado, pelos totais de animais não tratados e de animais tratados, obtendo-se

$$0{,}2 \times 200 = 40, \quad 0{,}2 \times 100 = 20.$$

As frequências dos sobreviventes entre animais não tratados e animais tratados são, respectivamente,

$$160 = 200 - 40, \quad 80 = 100 - 20.$$

Posto isto, subtraindo os valores da tabela n.° 8 (*valores de independência* ou *valores esperados* de acordo com a hipótese nula) dos valores correspondentes da tabela n.° 7 (*valores observados*), obtém-se a tabela das *discrepâncias*.

TABELA N.° 9

	Sobreviventes	Mortos	Total
Não tratados	− 8	+ 8	0
Tratados	+ 8	− 8	0
Total	0	0	0

Verificam-se, é claro, as relações algébricas entre os δ_{ik} indicadas no n.º 14 (o número de graus de liberdade é aqui igual a 1).

Mas pergunta-se agora:

Estão estes resultados em contradição com a hipótese nula? A resposta exige reflexão.

É claro que o facto de o tratamento não influir na doença não obriga, de nenhum modo, a uma rígida confirmação dos valores de independência: *pode haver discrepâncias não nulas devidas ao acaso; só se as discrepâncias excederem um "certo limite" haverá motivo para refutar a hipótese nula.* Tudo está, portanto, em avaliar esse limite para além do qual a hipótese nula é insustentável.

Note-se que uma questão análoga se nos apresenta, quando se lança ao ar uma moeda, várias vezes seguidas, e se regista de cada vez o lado que fica para cima: *coroa* ou *face*. Se a moeda estiver bem *balançada* (isto é, se os dois lados forem sensivelmente iguais do ponto de vista da gravidade) é de *esperar* que, por exemplo, em 100 lançamentos sucessivos, se apresente coroa 50 vezes, isto é, que seja 1/2 a frequência relativa deste acontecimento. Mas é bem possível (e até mais *provável*) que, nos 100 lançamentos, se apresente coroa mais ou menos de 50 vezes: *simplesmente, a frequência absoluta do acontecimento será tanto menos de esperar* (*tanto menos provável*), *quanto mais se afastar do valor esperado, isto é, quanto maior for a discrepância.*

Este exemplo sugere o seguinte procedimento para averiguar em que medida as discrepâncias observadas no caso em estudo são atribuíveis ao acaso:

Tomem-se 300 bocados de cartão, sensivelmente iguais em substância, forma e dimensões, e distingam-se 100 desses cartões com a cor vermelha (representativos dos 100 animais tratados) e os 200 restantes com a cor branca (representativos dos animais não tratados). Execute-se, depois, *um grande número de vezes seguidas*, a prova $\mathcal{P}$ que consiste nas seguintes operações:

1.º – *Fechar todos os cartões numa mesma caixa.*

2.º – *Agitar esta várias vezes.*

3.º – *Tirar da caixa ao acaso 60 cartões* (representativos dos animais mortos).

4.º – *Registar o número de cartões vermelhos existentes nessa amostra de 60 cartões.*

É claro que a repetição $\mathcal{P}$ implica a reposição dos cartões retirados na prova anterior.

A cada realização da prova corresponderá, assim, uma tábua de contingência de valores observados e, portanto, uma outra de discrepâncias. O que está agora naturalmente indicado é *verificar a frequência relativa com que se registam discrepâncias iguais ou superiores em valor absoluto às da tabela n.º 9, isto é, iguais ou superiores a 8 (em valor absoluto).*

Ora, quem tiver a paciência de efectuar um número muito grande de tais provas (pelo menos 1.000), verificará que *só em cerca de 2 por cento das provas se apresentam discrepâncias iguais ou superiores a 8, em valor absoluto.* A raridade de tais discrepâncias, embora não autorize, em absoluto, a rejeitar a hipótese nula, torna já bastante plausível a hipótese da eficácia do tratamento.

Convém, desde já, registar este facto: a frequência relativa com que, numa série de realizações da prova $\mathcal{P}$, se apresentam discrepâncias cujo valor absoluto é igual ou superior a um dado limite L, tende a estabilizar-se num valor determinado, à medida que o número de realizações aumenta. Ao tratar do conceito de probabilidade, veremos, mais precisamente, em que consiste essa estabilização. Entretanto, podemos já dizer que o referido valor se chama *probabilidade duma discrepância igual ou superior a L (em valor absoluto)*[1].

Ora, como teremos ocasião de ver, tal valor pode ser calculado por dedução puramente matemática, que dispensa as longas e fastidiosas séries de provas com cartões a que fizemos referência. Por este simples exemplo se pode já fazer uma ideia da utilidade e do alcance do Cálculo das probabilidades.

O critério de interpretação atrás descrito pertence à categoria dos chamados *testes de significância.* Para averiguar se dado tratamento é eficaz, começa-se por formular a *hipótese nula*, que consiste em

---

(1) – Nesta antecipação, que nos parece útil do ponto de vista didáctico (fazendo preceder a teoria de exemplos concretos) seguimos a orientação da citada obra de FINNEY.

supor o contrário, isto é, que o tratamento não tem efeito nenhum. Da tabela dos *valores observados*, deduz-se, então, a dos *valores esperados* (de acordo com a hipótese nula) e, em seguida, a das discrepâncias. Posto isto, investiga-se a frequência relativa com a qual, sendo válida a hipótese nula, se apresentariam discrepâncias tão grandes ou maiores (em valor absoluto) do que os desvios observados. Tal investigação poderia ser feita directamente, pelo processo empírico, laborioso, de amostragem casual de cartões, segundo o esquema atrás descrito; mas a Matemática permite evitar esse trabalho, por meio do *Cálculo das probabilidades*. Se a frequência relativa (ou melhor, a *probabilidade*) de tais desvios for *muito pequena*, então, sim, haverá razão para rejeitar a hipótese nula, admitindo como real a influência do tratamento. Os desvios observados dizem-se, então, *estatisticamente significantes* ou apenas *significantes*. De contrário, continuará de pé a hipótese nula, enquanto novas experiências não vierem enriquecer o material de dados, para que o investigador se possa pronunciar num ou noutro sentido.

Mas o que se entende aqui por frequência relativa "muito pequena" e por frequência relativa "não muito pequena"? Haverá, necessariamente, uma extensa margem de subjectividade no emprego destes termos. O critério a adoptar no seu uso depende não só da experiência realizada, como, ainda, da intuição do experimentador, que neste campo dispõe de ampla liberdade.

Em investigações agronómicas é usual considerar já como "muito pequena", em casos tais, a frequência relativa de 1 por 20 (0,05), embora, por vezes, se prefira considerar como "muito pequena" a frequência de 1 por 100 (0,01), sem esquecer que se trata aqui das frequências respeitantes a um grande número de provas (ou melhor, de *probabilidades*). Para distinguir aqueles dois graus de significância (ou *níveis de significância*, como usa dizer-se nestes casos), convenciona-se chamar ao primeiro, "significante" (sem qualquer restritivo) e ao segundo, "altamente significante". Convém salientar que se trata aqui apenas duma convenção de linguagem, reconhecida como apropriada por estatísticos e experimentadores. De resto, não são estes os únicos níveis de significância empregados: casos há em que uma probabilidade de 1 por 10 é já considerada signifi-

cante (especialmente em investigações médicas, em que o tratamento consista numa ligeira alteração de regime, pouco dispendiosa e sem risco para o doente); e casos há, pelo contrário, em que se requer um nível de 1 por 1.000 (quando se trate de alterações de regime que sejam profundas e dispendiosas).

Em síntese, podemos dizer que a aplicação dum teste de significância comporta os seguintes passos:

1) – *Formular a hipótese nula, contraditória do facto experimental a estabelecer.*

2) – *Escolher o nível de probabilidade que pareça mais adequado à natureza da experiência e ao tipo de decisão a tomar.*

3) – *Empregando o cálculo das probabilidades ou efectuando uma longa série de provas com cartões no género das que foram descritas, avaliar a probabilidade de que, sendo válida a hipótese nula, as discrepâncias casuais sejam pelo menos tão grandes (em módulo) como aquelas observadas na experiência.*

4) – *Se a probabilidade obtida em 3) é inferior à probabilidade escolhida em 2), rejeitar a hipótese nula; de contrário, deixar suspensa a conclusão.*

É, agora, o momento de mostrar o partido que se pode tirar da função χ^2 dos desvios, como teste de significância aplicado a uma experiência, cujos resultados estejam inscritos numa tábua de contingência. Mostra a análise matemática o seguinte: a probabilidade de que, sendo válida a hipótese nula, o valor de χ^2 exceda um dado limite, pode, sob certas restrições, ser calculada com suficiente aproximação, sem fazer intervir os números de indivíduos observados, mas, unicamente, o número n de graus de liberdade da tábua em questão. Deste modo, será possível calcular, uma vez por todas, para cada valor de n, as probabilidades correspondentes a diversos valores de χ^2 e registar os resultados numa tábua de duas entradas, que pode ser usada pelo investigador, com enorme economia de esforço e de tempo[1].

---

(1) – Só ao tratar das distribuições de variáveis contínuas, podemos referir-nos aos fundamentos teóricos deste método.

Assim, por exemplo, para $n = 2$, sendo válida a hipótese nula, está calculado que se tem:

$\chi^2 \geq 0{,}5$ em cerca de 50 % dos casos,
$\chi^2 \geq 2{,}7$ em cerca de 10 % dos casos,
$\chi^2 \geq 3{,}8$ em cerca de 5 % dos casos,
$\chi^2 \geq 6{,}6$ em cerca de 1% dos casos,
$\chi^2 \geq 10{,}8$ em cerca de 1% dos casos.

Por outras palavras, a *probabilidade* de se ter $\chi^2 \geq 0{,}5$ é aproximadamente igual a 0,5, o que também se exprime escrevendo

$$\Pr(\chi^2 \geq 0{,}5) = 0{,}5$$

e, analogamente: $\Pr(\chi^2 \geq 2{,}7) = 0{,}1$; $\Pr(\chi^2 \geq 3{,}8) = 0{,}05$, etc.

Estes resultados podem ser confirmados, experimentalmente pelo processo das provas repetidas atrás descrito, calculando o valor de χ^2 correspondente aos resultados de cada prova e determinando, no final, a frequência relativa dos valores $\geq 0{,}5$, a dos valores $\geq 2{,}7$, etc.

Ora, tornando ao exemplo das tabelas nºs 7, 8 e 9, verifica-se que é:

$$\chi^2 = \frac{(-8)^2}{160} + \frac{8^2}{40} + \frac{8^2}{80} + \frac{(-8)^2}{20} = 6{,}00.$$

Segundo a doutrina exposta, a probabilidade de que seja $\chi^2 \geq 6{,}6$ deve ser um pouco superior a 1% (mais precisamente 0,0143), o que concorda com a frequência de cerca de 2% que se obteria numa longa série de provas com cartões.

Mas é preciso notar que *o uso do* χ^2 *e dos respectivos níveis de probabilidade envolve aproximações que só são aceitáveis quando os números de indivíduos são grandes*. Uma regra prática que costuma ser recomendada é a de que nenhum dos valores esperados (ver tabela nº 8) deve ser inferior a 5. Contudo, a aproximação pode ser melhorada consideravelmente, em qualquer caso, por um simples artifício, que consiste em diminuir 0,5 ao módulo de cada discrepância, antes de calcular o χ^2 (*correcção de* YATES). Assim, no caso em estudo, viria

$$\chi^2 = \frac{(-7,5)^2}{160} + \frac{(7,5)^2}{40} + \frac{(7,5)^2}{80} + \frac{(-7,5)^2}{20} = 5,27.$$

O valor da $\Pr(\chi^2 \geq 5,27)$, no caso ideal a que se refere a teoria matemática, é 0,0217, em vez de 0,0143 – bastante mais próximo do verdadeiro valor que é, no caso considerado, 0,0210 (a menos de 0,0001).

Da mesma obra extraímos um segundo exemplo, relativo a uma tábua de contingência com 3 graus de liberdade. Trata-se de colheitas de cevada em 260 campos, feitas em 1942. No inverno e primavera precedentes, tinha-se avaliado em cada campo a população da larva dum coleóptero, comum em Inglaterra, com o nome de "wireworm" (alfinete). Segundo determinado critério, classificou-se a infestação dos diferentes campos em "baixa", "moderada", "alta" e "muito alta". Por sua vez, o resultado das colheitas foi classificado em "satisfatório" e "não satisfatório". Os resultados são os que constam da seguinte tabela:

TABELA N.º 10

Resultados das colheitas	Infestação				Total
	Baixa	Moderada	Alta	Muito alta	
Satisfatórios	94	62	31	15	202
Não satisfatórios	15	15	17	11	58
Total	109	77	48	26	260

Daqui se deduziu uma tabela de valores de independência e uma outra de discrepâncias (tabelas que não reproduzimos), achando-se o seguinte valor:

$$\chi^2 = \frac{(9,3)^2}{84,7} + \frac{(2,2)^2}{59,9} + \frac{(-6,3)^2}{37,3} + \frac{(-5,2)^2}{20,2} +$$

$$+ \frac{(-9,3)^2}{24,3} + \frac{(-2,2)^2}{17,2} + \frac{(6,3)^2}{10,7} + \frac{(5,2)^2}{5,8} = 15,7.$$

A tabela n.° 11, em que são indicados alguns níveis de significância para χ^2, mostra-nos que o valor 15,7 calculado é sensivelmente superior ao valor 11,3, que, para 3 graus de liberdade, marca o nível "altamente significante". Há, pois, razão suficiente para excluir a hipótese nula.

TABELA N.° 11

Graus de liberdade	Frequências relativas (Probabilidades)		
	0,1	0,05	0,01
1	2,7	3,8	6,6
2	4,6	6,0	9,2
3	6,3	7,8	11,3
4	7,8	9,5	13,3
6	10,6	12,6	16,8
8	13,4	15,5	20,1
10	16,0	18,3	23,2
15	22,3	25,0	30,6
20	28,4	31,4	37,6
25	34,4	37,7	44,3
30	40,3	43,8	50,9

Convem notar que:

1) *Dum modo geral, o teste do* χ^2, *como foi descrito, não deve aplicar-se quando algum dos valores esperados for demasiado pequeno, exigindo-se, nesse caso, uma análise estatística especial.*

2) *A correcção de Yates é somente aplicável nas tábuas de* 2×2 (com 1 grau de liberdade).

Há, ainda, certos pormenores, relativos à realização da experiência, que precisam de ser observados, para que o teste de significância não resulte ilusório. Tornando ao primeiro exemplo citado neste número, é óbvio, que, se os animais tratados tiverem uma proveniência diferente da dos animais não tratados, as diferenças registadas podem muito bem não ser devidas ao tratamento. Para atenuar o mais possível os factores estranhos ao tratamento, o que há a fazer é destacar os 300 animais duma população tão *homogénea* quanto

possível, e, entre esses, escolher, completamente *ao acaso*, os 100 animais a serem tratados, dando a todos os 300 *igual probabilidade* de serem escolhidos. Esta operação, chamada *casualização*, tem de ser efectuada estritamente por um processo objectivo, em que não intervenha qualquer factor pessoal na escolha. Um tal processo pode consistir no seguinte: numerar os animais de 1 a 300, deitar numa caixa 300 cartões iguais, numerados de 1 a 300, e tirar 100 à sorte; os números saídos serão os dos animais a tratar.

Nunca é demais encarecer a importância da casualização, nestes e noutros tipos de experiência, como base duma investigação estatística bem orientada. Não podemos, contudo, indicar aqui os pormenores da técnica, a que por vezes tem de obedecer uma tal operação.

Note-se que o papel do estatístico não se limita à interpretação dos resultados experimentais: ele pode contribuir, com grande vantagem, para o *planeamento da experiência*. Não pertence ao âmbito do nosso curso o estudo deste assunto de alto interesse. Limitar-nos-emos, por isso, a breves indicações sobre a natureza da questão, no caso particular do tratamento hipotético atrás considerado. Suponhamos, por exemplo, que se tinham escolhido, apenas, 2 ou 3 animais para serem tratados, reservando os restantes para controle (ou vice-versa); é manifesto, mesmo *a priori*, que os resultados não autorizariam um juízo seguro. Deve, portanto, haver uma proporção *óptima* para revelar a eficácia do tratamento, caso este tenha, de facto, efeito. Porém, essa proporção não é, como poderia parecer à primeira vista, a de 50% para os dois grupos. Na Tabela n.º 12, estão indicados os valores do χ^2 para diferentes proporções de animais tratados, entre os 300, na hipótese de a percentagem de mortos entre animais tratados ser em todos os casos a mesma que se observa na Tabela n.º 7 (12%). Vê-se que o máximo valor do χ^2 é atingido na proporção de 175 animais tratados para 125 não tratados. Será, portanto, essa proporção ideal para a experiência.

É claro que, quanto maior for o número total de casos individuais estudados, melhor informação fornecem os resultados obtidos; mas esse número é forçosamente limitado por razões de ordem económica, e, por isso, mais necessário se torna tirar o maior rendimento possível do material de que se dispõe.

TABELA N° 12

N.° de animais tratados	χ^2 para 12% de mortalidade
25	1,2
50	2,8
100	5,3
125	6,1
150	6,5
175	6,6
200	6,3
250	4,0
275	2,0

Muito mais haveria a dizer sobre testes de significância, mas, embora o assunto seja de grande interesse para agrónomos, não pode ser desenvolvido nesta cadeira. E, se nos alongámos um pouco a respeito do teste do χ^2, foi sobretudo para dar uma ideia de como o Cálculo das Probabilidades pode intervir nas aplicações de carácter agronómico.

B – Probabilidades

1. Lógica indutiva

Suponhamos que, em n realizações, duma mesma prova $\mathcal{P}$, um dado acontecimento α se realizou n vezes: então, a frequência relativa do acontecimento α, na referida série de provas, será $f = n/n = 1$. Como já tivemos ocasião de observar, este facto não habilita a concluir que α seja um acontecimento certo. No entanto, se o número n de provas realizadas for *muito grande*, é-se levado a admitir que o acontecimento é *praticamente certo*, embora não se possa concluir que seja *certo* (*em absoluto*). Nisto consiste, esquematicamente, a *indução* ou *raciocínio indutivo*, que se encontra na base de toda a ciência experimental. É bem sabido que as *leis físicas* ou, melhor, as *leis da Natureza*, têm carácter *contingente*: não se pode garantir que sejam absolutamente infalíveis.

Seja, por exemplo, a seguinte experiência: aproximar uma chama dum frasco, com a boca para baixo, no qual se tenha introduzido

uma mistura de hidrogénio e oxigénio, em proporções convenientes. É, então, praticamente certo que se verifica o fenómeno "explosão".

Trata-se, aqui, duma lei química qualitativa, de cuja contingência nos apercebemos, sobretudo, se não tiverem sido definidas, com certa precisão, as condições em que a experiência deve ser realizada. Muitos outros exemplos poderíamos citar, de leis qualitativas, com o mesmo carácter contingente.

As leis quantitativas correspondem a um grau mais elevado de conhecimento, a uma fase mais avançada da ciência: mas nelas subsiste o carácter de contingência. Primeiro que tudo, devemos lembrar-nos de que as medidas das grandezas físicas *nunca* podem ser exactas (não faz, mesmo, sentido falar de medida exacta duma grandeza empírica – seja comprimento, seja velocidade, seja carga eléctrica, seja um pH, seja qualquer outra). Já daqui resulta uma certa margem de incerteza, isto é, de contingência.

Consideremos, por exemplo, esta experiência: aquecer um pedaço de chumbo a uma temperatura de cerca de 335 graus centígrados, em condições normais. É praticamente certo que se verifica o fenómeno *fusão*. Mas no próprio carácter aproximativo da temperatura indicada, bem como no significado da expressão "condições normais", há uma origem de incerteza, que pode ser reduzida (indicando, por exemplo, uma vizinhança de 335 na qual se prevê o acontecimento), mas nunca eliminada.

Recordemos, ainda, leis físicas, de nível mais elevado: a lei da gravitação, as leis da termodinâmica, as leis do electro-magnetismo, etc. Como é sabido, nenhuma destas leis se adapta exactamente à realidade: assentam todas em hipóteses simplificadoras, que só aproximadamente se podem realizar. Essas leis são, pois, apenas esquematizações duma realidade que é sempre demasiado complexa, demasiado mutável, para se deixar traduzir fielmente na simplicidade dos nossos símbolos.

Assim, a lei dos gases perfeitos deve o seu nome ao facto de se ter convencionado chamar "gases perfeitos" aos gases que a seguem rigorosamente, o que imprime a essa lei um certo carácter de definição ou de postulado. A verdade, porém, é que não há gases perfeitos: há, apenas, gases que se aproximam, mais ou menos, dessa forma

ideal. Uma lei que substitui aquela, dando uma melhor aproximação da realidade, é a de Van der Waals; mas nem mesmo essa poderá ser exacta. Porque é lei metafísica das leis físicas, o serem contingentes.

O método experimental, e com ele o método matemático, tem-se estendido progressivamente, do âmbito restrito das ciências físico-químicas, ao das ciências biológicas, ao das ciências sociais, etc. Nestes novos domínios, mais acentuado se torna o carácter contingente das leis naturais. No entanto, algumas se apresentam com visos de certeza inabalável, como aquela que tem servido de premissa a um exemplo clássico de silogismo:

"Todos os homens são mortais".

Há aqui, de certo modo, uma lei biológica qualitativa. Mas se tentarmos precisá-la quantitativamente, afirmando, por exemplo:

"Todos os homens morrem antes dos 300 anos de idade",

já não sentiremos o mesmo grau de segurança. Conhecemos nós, suficientemente, o passado da espécie humana? E que sabemos nós sobre o futuro?

O que pode dizer-se é que, nas condições actuais, *é extremamente improvável, praticamente impossível* que um ser humano atinja a idade de 300 anos. Um outro exemplo análogo é o que se refere a alturas: é praticamente impossível que um ser humano cresça até atingir a altura de 4 metros.

Se formos baixando estes limites, o grau de incerteza aumentará – e entraremos, abertamente, no campo das probabilidades. É de salientar que o Cálculo das Probabilidades e a Estatística se têm desenvolvido principalmente no sector das ciências biológicas e das ciências sociais. Mas certo é, também, que, por um movimento de retrocesso, acabaram por invadir o campo das ciências físicas, principalmente no que se refere ao estudo do átomo. A física moderna tem carácter probabilista.

NOTA. Uma fonte primária de contingência está na impossibilidade que há em definir exactamente os conceitos relativos ao mundo empírico, – em delimitar com precisão os atributos dos entes naturais. Esta impossibilidade tem sido origem constante de especulações filosóficas através dos séculos. Platão resolvia a dificuldade, distinguindo duas formas de existência: a *realidade sensível* ou *mundo dos fenómenos*, que conhecemos por meio dos sentidos, e a *realidade inteligível* ou *mundo das Ideias,* que conhecemos por meio da razão e da qual a primeira é, apenas, imitação grosseira. Recorde-se, ainda, a querela que, na Idade Média, dividiu os filósofos em *realistas* e *nominalistas*, os primeiros proclamando a *realidade* dos universais (isto é, das classes, dos entes abstractos) e a sua supremacia sobre o contingente e o transitório; os segundos afirmando que os conceitos abstractos são, apenas, *nomes*, ficções cómodas, e que só os indivíduos existem.

No fundo, trata-se de duas tendências complementares, inerentes ao nosso espírito, com predomínio duma ou doutra conforme as pessoas e as situações. Assim, o matemático assume geralmente a atitude platónica (realista), reportando ao mundo dos seres ideais, de maneira mais ou menos consciente, os conceitos abstractos que são objecto do seu pensamento. Por sua vez, o experimentador tende para a atitude nominalista (ou empirista). Do equilíbrio das duas tendências, segundo o bom-senso, é que pode resultar o êxito da actividade intelectual: toda a teoria deve ser controlada pela experiência, assim como a prática deve sempre ser guiada pela teoria.

2. Lógica dedutiva

Ao raciocínio indutivo das ciências experimentais contrapõe a Matemática o raciocínio dedutivo. À contingência *física* das leis físicas opõe-se a certeza *matemática* das deduções matemáticas. Considerem-se, por exemplo, os seguintes teoremas:

"O quadrado dum número ímpar é sempre um número ímpar".

"Toda a equação de coeficientes complexos admite pelo menos uma solução no campo complexo".

"Toda a série de potências de x é derivável, termo a termo (em ordem a x), no seu intervalo de convergência".

Encontramos aqui um carácter de certeza e precisão que contrasta vivamente com a natureza dúbia e aproximativa das leis naturais. Contudo, já se nota diferença, quando se passa da matemática pura para as matemáticas aplicadas. A própria Geometria é, sob certo aspecto, um ramo da Física. Consideremos, por exemplo, o bem conhecido teorema da geometria euclideana:

"A soma dos ângulos dum triângulo (plano) é sempre igual a um ângulo raso".

Dois métodos diferentes se podem seguir para estabelecer a veracidade desta proposição:

1.º – *Método indutivo ou experimental* – Medem-se os ângulos internos de vários triângulos e verifica-se que a soma dos ângulos de cada triângulo se *aproxima bastante* de 180°, tanto mais quanto mais perfeito for o traçado dos triângulos e mais precisa for a medição. É este o método seguido na primeira fase do ensino da Matemática no liceu.

2.º – *Método dedutivo ou racional* – Demonstra-se a proposição, *deduzindo-a logicamente* de proposições mais simples, cuja veracidade se considera *evidente*. Estas proposições evidentes são os chamados *axiomas* ou *postulados* da geometria euclideana. No teorema em questão, o postulado que intervém de maneira essencial é precisamente o postulado de Euclides:

"Dados um ponto e uma recta, existe sempre uma recta, e uma só, que passa pelo ponto dado e é paralela à recta dada".

Mas o declarar *evidentes* os postulados é uma atitude cómoda, que exige reflexão. Têm eles, de facto, carácter de certeza absoluta?

Um dos grandes progressos da história do pensamento, realizado no século passado, foi precisamente o de reconhecer que os postulados da geometria euclideana têm a natureza contingente e aproximativa das leis naturais. Nem leis se podem dizer, propriamente, mas sim hipóteses, cuja legitimidade se avalia indirectamente pelas suas consequências lógicas. Insinuam-se no nosso espírito, por um longo processo indutivo, dificilmente controlável, que se desenvolve na experiência quotidiana, no exercício constante da nossa actividade neuro-muscular. É esta uma forma de indução que se confunde com aquele processo directo de conhecimento a que se dá o nome de *intuição*.

Mas a nossa experiência quotidiana refere-se a uma região limitada do espaço. A teoria da relatividade veio mostrar, precisamente, que, nos espaços astronómicos, não é a geometria euclideana a que mais se aproxima da realidade: aí, a soma dos ângulos internos dum triângulo pode tornar-se sensivelmente superior a 180°.

O que se disse a respeito do teorema citado aplica-se a qualquer outro. A contingência dos postulados transmite-se a todos os teoremas. O que é matematicamente certo não é o teorema, mas, sim, o facto de o teorema ser consequência lógica dos postulados.

Entretanto registem-se as vantagens do método racional. Enquanto o método empírico exige um grande número de verificações fastidiosas, o método racional, com elegante simplicidade e perfeito rigor, reduz a veracidade do teorema à de proposições já reconhecidas como verdadeiras. Estas vantagens patenteiam-se, em particular, no Cálculo das Probabilidades.

3. Conceito natural de probabilidade

As considerações precedentes mostram que os termos "verdadeiro" e "falso", aplicados a proposições, se tornam insuficientes, quando, da realidade inteligível da Matemática, se passa à realidade sensível dos fenómenos. Os conceitos de "verdadeiro" e "falso" cedem, então, o lugar ao conceito de "probabilidade", correlativo do de "incerteza" ou "contingência": um facto dir-se-á tanto mais *provável* quanto menos contingente for.

O conceito de probabilidade, como todos os conceitos relativos ao mundo empírico, não é susceptível de definição lógica: gera-se no nosso espírito por um processo indutivo. Mais até, faz parte intrínseca do próprio mecanismo da indução. É o que vamos ver, tentando esclarecer este conceito.

Seja α um dos acontecimentos a prever numa certa experiência $\mathcal{P}$ e suponhamos que esta experiência foi efectuada em várias séries de provas, todas em *grande número*:

Série 1 – $\mathcal{P}_{1,1}, \mathcal{P}_{1,2}, \ldots\ldots\ldots, \mathcal{P}_{1,n_1}$ (n_1 provas)
Série 2 – $\mathcal{P}_{2,1}, \mathcal{P}_{2,2}, \ldots\ldots\ldots, \mathcal{P}_{2,n_2}$ (n_2 provas)
....
Série m – $\mathcal{P}_{m,1}, \mathcal{P}_{m,2}, \ldots\ldots\ldots, \mathcal{P}_{m,n_m}$ (n_m provas).

Se forem, respectivamente, $\nu_1, \nu_2, \ldots, \nu_m$ as frequências absolutas do acontecimento α nestas séries de provas, serão

$$\frac{\nu_1}{n_1}, \frac{\nu_2}{n_2}, \ldots, \frac{\nu_m}{n_m},$$

as correspondentes frequências relativas. Suponhamos, ainda, que se verificou uma sensível concordância entre estas frequências, isto é, que todas se localizaram num pequeno intervalo $]f_0-\varepsilon, f_0+\varepsilon[$:

$$f_0 - \varepsilon < \frac{\nu_i}{n_i} < f_0 + \varepsilon, \quad \text{para } i = 1, 2, \ldots, m.$$

Aplicando o raciocínio indutivo, exactamente como se faz ao estabelecer as leis naturais, seremos levados a admitir, como praticamente certo, que:

Em toda a série formada por muitas realizações de $\mathcal{P}$ *(em número não inferior ao maior dos números* $n_1, n_2, \ldots, n_m$*) a frequência relativa de* α *ficará situada entre* $f_0 - \varepsilon$ e $f_0 + \varepsilon$.

É, então, natural dizer que f_0 é um *valor aproximado da probabilidade de* α, *com erro inferior a* ε (*ou a menos de* ε).

Por exemplo, se em tais séries de provas se registou sempre uma frequência relativa de α entre 7% e 11% (ou seja, entre $0{,}09-0{,}02$ e $0{,}09+0{,}02$), dir-se-á que 0,09 é um valor aproximado, a menos de 0,02, da probabilidade de α.

Suponhamos, agora, que se efectuaram novas séries de realizações de $\mathcal{P}$, em *números bastante maiores* que os anteriores, e que passou a registar-se uma frequência relativa de α, entre

$$f_1 - \frac{\varepsilon}{10} \text{ e } f_1 + \frac{\varepsilon}{10};$$

dir-se-á, então, que f_1 é valor aproximado da probabilidade de α, a menos de $\varepsilon/10$. Analogamente, poderíamos imaginar valores $f_2, f_3, \ldots$, aproximados da probabilidade de α a menos de $\varepsilon/10^2$, de $\varepsilon/10^3$, etc.

A probabilidade aparece-nos, assim, como *grandeza que se mede*, tal como se fosse uma grandeza física (uma velocidade, um ponto

de fusão, uma densidade, etc.), com o carácter contingente e aproximativo que revestem todas as medições empíricas. Já as expressões "grandes números", "números bastante maiores", atrás usadas, envolvem a aplicação de critérios subjectivos, humanos, a que falta rigor matemático, mas que são inevitáveis.

Tal como sucede com as grandezas físicas, é-se levado a idealizar, para cada acontecimento α (em certos tipos de experiências), um valor *exacto*, p, da probabilidade de α, que seria, por assim dizer, o *limite* da sucessão dos valores aproximados $f_0, f_1, f_2, \ldots$ Mas é claro que um tal valor exacto *não existe nos casos concretos* – assim como não existe o valor exacto duma grandeza física: trata-se apenas de convenções cómodas, que o nosso espírito transporta para o mundo platónico dos entes ideais, para sobre as mesmas poder construir uma teoria racional – *uma teoria matemática*[1].

Note-se, ainda, que a unidade adoptada para medir a probabilidade do acontecimento α é a probabilidade do acontecimento certo – isto é, a *certeza*[2]. Por sua vez, ao acontecimento impossível é atribuída a probabilidade 0. Deste modo, a probabilidade p de α deverá ser um número tal que

$$0 \leq p \leq 1 .$$

Se p for aproximadamente igual a 1, dir-se-á que α é *praticamente certo*. Se p for aproximadamente igual a 0, dir-se-á que α é *praticamente impossível*. Mas o grau de aproximação a partir do qual se dirá uma ou outra coisa é questão manifestamente subjectiva. Registe-se, entretanto, que, para um acontecimento ser considerado praticamente certo, não é necessário que se realize em *todas* as provas que tenham sido efectuadas: admite-se a possibilidade de excepções, contanto que estas sejam extremamente raras.

E, analogamente, no caso oposto.

---

(1) – Esta concepção da probabilidade como constante física dos acontecimentos é devida ao matemático francês MAURICE FRÉCHET.

(2) – Recorde-se que, na medida dos arcos, se pode tomar unidade a circunferência. Uma das maneiras sugestivas de indicar graficamente as probabilidades (ou as frequências relativas) de vários acontecimentos incompatíveis, consiste em representá-los por meio de sectores dum círculo, de amplitudes proporcionais a essas probabilidades (ou frequências).

O termo "probabilidade", tal como acabamos de o precisar, é usado na prática com maior ou menor propriedade. Recordemos o exemplo das tábuas de mortalidade, sobre as quais os actuários das companhias de seguros baseiam o cálculo dos prémios a pagar em seguros de vida. Essas tábuas são, no fundo, estatísticas, isto é, tabelas de frequência, relativas a populações numerosas. Assim, por exemplo, a tábua de mortalidade alemã para homens, correspondente ao decénio 1901-1910, indica que, entre 10.000 crianças com menos de 1 ano, 3.608 chegaram à idade de 65 anos. Deste modo, a probabilidade de que um recém-nascido venha a atingir os 65 anos será, aproximadamente:

$$\frac{3.608}{10.000} \approx 36\%.$$

Por sua vez, a mesma tábua mostra que, entre 10.000 crianças com menos de 1 ano, 7.065 chegam aos 20 anos. Daqui se deduz que a probabilidade de um rapaz de 20 anos chegar aos 65 anos é

$$\frac{3.608}{7.065} \approx 51\%,$$

maior que a precedente.

A mesma tábua dá para crianças de menos de 1 ano, a probabilidade 0,68% de chegar aos 90 anos, a probabilidade 0,0038% de atingir os 100 anos, etc.

Mas note-se que as tábuas de mortalidade têm de ser revistas de tempos a tempos e não devem ser as mesmas para todas as regiões, para todas as raças, para os dois sexos, etc.

A probabilidade dum acontecimento pode, assim, estar sujeita a uma evolução imprevisível no tempo e no espaço. É este um facto que, pelo menos aparentemente, está em contradição com o conceito de probabilidade, tal como foi atrás explanado. O mesmo é dizer que as leis físicas evoluem de maneira imprevisível no tempo e no espaço – *deixando, portanto, de ser leis*. A explicação do facto está, ainda, no carácter contingente do raciocínio indutivo, que impõe sempre a seguinte norma prudencial ao experimentador: *as induções*

só merecem confiança em regiões limitadas do espaço e do tempo; não podem ser extrapoladas muito para além do domínio das experiências efectuadas.

Portanto, só quando a evolução for lenta será lícito falar de probabilidades.

Recordemos, ainda, o exemplo dos desastres de avião: é hoje menor do que há vinte anos a frequência relativa dos desastres de aviação em percursos iguais. Neste caso, embora seja mais rápida a evolução, ainda, de certo modo, é lícito falar de probabilidade no sentido atrás considerado(1).

Casos há, porém, em que o emprego desta palavra escapa, de todo, a critérios quantitativos, que lhe confiram interesse científico.

4. Axiomatização do conceito de probabilidade

Viu-se no número anterior que, para fundar uma teoria matemática das probabilidades, se idealiza para cada acontecimento, em certos tipos de experiências, um valor exacto da sua probabilidade. Contudo, seja esse valor uma convenção ou uma realidade, o que interessa para o desenvolvimento lógico da teoria não é, propriamente, o conceito em si, mas, antes, o conjunto das suas propriedades formais que intervêm nos cálculos e nos raciocínios. O que há a fazer, então, é escolher, entre essas propriedades, algumas mais simples, das quais se possam deduzir logicamente todas as outras. As propriedades escolhidas dir-se-ão *propriedades fundamentais* ou *axiomas* e o seu conjunto terá o nome de *axiomática*. Já em Matemáticas Gerais foi estudada uma axiomática dos grupos contínuos de grandezas. Trata-se, agora, de estabelecer uma axiomática do conceito de probabilidade, no caso dos corpos finitos de acontecimentos, considerando aquele conceito como *primitivo*, isto é, não definível logicamente à custa de outros.

Seja, pois, $\mathcal{R}$ um corpo finito de eventualidades a considerar numa dada prova $\mathcal{P}$, e suponhamos que a cada eventualidade $\alpha \in \mathcal{R}$, corresponde uma determinada probabilidade, que designaremos pelo

---

(1) – Nos seguros relativos a acidentes de aviação, as companhias americanas fazem hoje o cálculo dos prémios como se houvesse morte de um passageiro em 20.000 viagens de ida e volta.

símbolo $Pr(\alpha)$. Tomaremos como fundamentais as propriedades seguintes[1]:

AXIOMA 1. *A probabilidade de* α *é sempre um número real não negativo, isto é*:

$$Pr(\alpha) \geq 0, \textit{ qualquer que seja } \alpha \in \mathcal{R}.$$

AXIOMA 2. *Se* α *e* β *são acontecimentos incompatíveis de* $\mathcal{R}$, *tem-se*:

$$Pr(\alpha + \beta) = Pr(\alpha) + Pr(\beta).$$

(Convém aqui rever, cuidadosamente, as noções do n.° 2 e do n.° 3 relativas às álgebras de atributos e de acontecimentos).

AXIOMA 3. *A probabilidade do acontecimento certo é* 1; *isto é, tem-se*:

$$Pr(\alpha) = 1, \textit{ se } \alpha \textit{ é certo}.$$

Estas propriedades concordam com o conceito empírico de probabilidade, tal como foi atrás introduzido. Sendo as probabilidades, por assim dizer, limites de frequências relativas, é natural que se conservem as propriedades das frequências relativas que não mudam na passagem ao limite.

Os axiomas 1 e 2 dizem-nos, simplesmente, que a função $Pr(\alpha)$ é uma distribuição definida no corpo $\mathcal{R}$. O axioma 3 acrescenta que essa distribuição é relativa. Diremos, então, que se trata duma *distribuição de probabilidade* definida em $\mathcal{R}$. Mas só nas aplicações concretas será possível distinguir uma distribuição de probabilidade duma distribuição de frequência, pois que a anterior axiomática se limita a dar os caracteres duma distribuição abstracta.

---

(1) – Vários autores consideram, ainda, como axioma a propriedade relativa à probabilidade do produto. Mas essa propriedade pode considerar-se como definição ou como consequência de definição de probabilidade condicional, como veremos.

Da anterior axiomática deduzem-se, desde logo, as seguintes consequências (que se estendem a qualquer distribuição):

TEOREMA 1. *Se* α *implica* β, *então* $\Pr(\alpha) \leq \Pr(\beta)$.

Com efeito, se $\alpha \subset \beta$, tem-se $\beta = \alpha + \tilde{\alpha}\beta$ e, como α e $\tilde{\alpha}\beta$ são incompatíveis, vem, pelo axioma 2,

$$\Pr(\beta) = \Pr(\alpha) + \Pr(\tilde{\alpha}\beta)$$

donde se conclui, pelo axioma 1, que $\Pr(\beta) \geq \Pr(\alpha)$.

COROLÁRIO. *Qualquer que seja* $\alpha \in \mathcal{R}$, *tem-se:*

$$\Pr(\alpha) \leq 1.$$

Com efeito, todo o acontecimento α de $\mathcal{R}$ implica o acontecimento certo: isto é, em fórmula:

$$\alpha \subset I,$$

donde, pelo teorema 1 e pelo axioma 3:

$$\Pr(\alpha) \leq 1.$$

TEOREMA 2[1]. *Se* $\alpha_1, \alpha_2, \ldots, \alpha_n$ *são acontecimentos* (*de* $\mathcal{R}$) *incompatíveis dois a dois, a probabilidade de que se realize um dos acontecimentos* $\alpha_1, \alpha_2, \ldots, \alpha_n$ *é a soma das probabilidades destes acontecimentos, isto é, tem-se*:

$$\Pr(\sum \alpha_i) = \sum \Pr(\alpha_i), \text{ se } \alpha_i \alpha_k = \varnothing \text{ para } i \neq k.$$

Basta aplicar, repetidamente, o axioma 2, observando que é $\alpha_1 + \alpha_2 + \alpha_3 = (\alpha_1 + \alpha_2) + \alpha_3$, $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = (\alpha_1 + \alpha_2 + \alpha_3) + \alpha_4$, etc.,

---

(1) – Também conhecido por *princípio das probabilidades totais.*

e que, se $\alpha_1, \alpha_2, \ldots, \alpha_n$ são incompatíveis dois a dois, também α_3 é incompatível com $\alpha_1 + \alpha_2$, α_4 com $\alpha_1 + \alpha_2 + \alpha_3$, etc.

TEOREMA 3. *Quaisquer que sejam* $\alpha, \beta \in \mathcal{R}$, *tem-se*

$$\Pr(\alpha + \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha\beta).$$

Basta notar que $\alpha + \beta = \alpha\tilde{\beta} + \tilde{\alpha}\beta + \alpha\beta$ e aplicar o teorema 2, notando que $\Pr(\alpha) = \Pr(\alpha\tilde{\beta}) + \Pr(\alpha\beta)$, $\Pr(\beta) = \Pr(\tilde{\alpha}\beta) + \Pr(\alpha\beta)$. O teorema pode generalizar-se ao caso de vários acontecimentos, obtendo-se a fórmula de DANIEL DA SILVA (cf. A-10).

TEOREMA 4. *A probabilidade do acontecimento contrário de* α *é a diferença para* 1 *da probabilidade de* α; *isto é*:

$$\Pr(\tilde{\alpha}) = 1 - \Pr(\alpha).$$

Consequência imediata dos axiomas 2, 3 e da definição de contrário ($\alpha + \tilde{\alpha} = I$, $\alpha\tilde{\alpha} = \varnothing$), conforme já se viu para as distribuições em geral.

COROLÁRIO. *Se* α *é acontecimento impossível, tem-se*

$$\Pr(\alpha) = 0.$$

Basta lembrar que o acontecimento impossível é contrário do acontecimento certo e aplicar o teorema 4 com o axioma 3.

É claro que, em virtude do teorema 2, uma distribuição de probabilidade num corpo finito $\mathcal{R}$ fica determinada, logo que se conheçam as probabilidades das células de $\mathcal{R}$: a probabilidade dum acontecimento $\alpha \in \mathcal{R}$ será a soma das probabilidades dos acontecimentos celulares de que α é a soma. Tudo está, pois, em determinar as probabilidades das células.

5. Alguns exemplos de cálculo de probabilidades *a priori*

Por vezes, a distribuição de probabilidade num corpo finito pode ser determinada *a priori* (isto é, antes de qualquer experiência deliberada) por simples considerações ditadas pelo *senso-comum.*

Um dos exemplos clássicos, já atrás invocados, é o que se refere à extracção casual de bolas ou de cartões duma caixa. Consideremos o caso duma urna com N bolas, que designaremos por $x_1, x_2, \ldots, x_N$. Seja U o conjunto destas bolas e seja $\mathcal{P}$ a experiência que consiste em tirar *ao acaso* uma bola da urna.

Os acontecimentos elementares que se podem verificar nesta experiência são: "sair a bola x_1", "sair a bola x_2", ..., "sair a bola x_N"; ou, em notação proposicional:

$$x = x_1, x = x_2, \ldots, x = x_N.$$

(Como se disse atrás, x é neste caso uma *variável casual*, que toma, em cada prova, um e um só dos valores $x_1, x_2, \ldots, x_N$).

Estes acontecimentos elementares geram um corpo de que são as células. Os acontecimentos que formam o corpo serão, além das células e do acontecimento impossível, todos aqueles que se exprimem como somas lógicas de células[(1)], por exemplo:

$$(x = x_1) + (x = x_2) \text{ (sair } x_1 \text{ ou } x_2)$$

$$(x = x_1) + (x = x_3) + (x = x_5) \text{ (sair } x_1 \text{ ou } x_3 \text{ ou } x_5), \text{ etc.}$$

Entre estes está incluído o acontecimento certo:

$$\sum_{i=1}^{p} (x = x_i) \text{ (sair uma das bolas } x_1, x_2, \ldots, x_N).$$

É claro, agora, que se conhecermos as probabilidades dos acontecimentos elementares, estamos aptos a conhecer a probabilidade de qualquer outro acontecimento do corpo, por simples aplicação do teorema 2. Designemos por $p_1, p_2, \ldots, p_N$, respectivamente, as probabilidades de "sair x_1", de "sair x_2", ..., de "sair x_N", isto é, ponhamos

$$p_i = \Pr(x = x_i), \quad \text{para } i = 1, 2, \ldots, N.$$

---

(1) – É evidente que os acontecimentos considerados são incompatíveis dois a dois, uma vez estabelecido que se tira *uma só* bola de cada vez.

Então, a probabilidade de "sair x_1 ou x_2" será

$$\Pr[(x = x_1) + (x = x_2)] = p_1 + p_2,$$

a probabilidade de "sair x_1 ou x_3 ou x_5" será $p_1 + p_3 + p_5$.

Como a probabilidade do acontecimento certo é igual à soma das probabilidades de todas as células, deverá ter-se

$$p_1 + p_2 + \cdots + p_N = 1.$$

Suponhamos que as bolas $x_1, x_2, \ldots x_\nu$ são vermelhas e as restantes brancas. Designando por V o conjunto das bolas vermelhas e por B o conjunto das bolas brancas, a probabilidade de sair bola vermelha será

$$\Pr(x \in V) = p_1 + p_2 + \cdots + p_\nu$$

e a de sair bola branca

$$\Pr(x \in B) = p_{\nu+1} + p_{\nu+2} + \cdots + p_N = 1 - \Pr(x \in V)$$

Mas como determinar as probabilidades elementares $p_1, p_2, \ldots, p_N$, se é que existem?

Para isso, há que introduzir hipóteses suplementares. *Suponhamos, por exemplo, que as bolas são sensivelmente iguais, em substância, forma e dimensões*. Então, para dar a todas as bolas igual probabilidade de serem extraídas (*casualização*), ocorre a ideia de fechar a urna e agitá-la várias vezes antes de tirar a bola.

Esta ideia, ou melhor, esta intuição, é confirmada pela experiência: efectuando *muitas vezes* a prova nas condições indicadas (com reposição da bola retirada), serão sensivelmente iguais as frequências relativas com que aparecem as diferentes bolas[1].

---

(1) – Donde nos vem esta intuição? É, na verdade, anterior a qualquer esperiência? Não parece que o seja. Será, antes, o produto de inúmeras experiências que fazemos, sem dar por isso, no decurso da nossa existência. Qualquer coisa de semelhante à intuição que nos leva a admitir os postulados da geometria euclideana.

Uma vez admitido que todas as bolas têm a mesma probabilidade de saída, isto é, que:

$$p_1 = p_2 = \cdots = p_N,$$

como se tem $p_1 + p_2 + \cdots + p_N = 1$, será

$$p_i = \frac{1}{N}$$

a probabilidade de saída de qualquer bola. Então, a probabilidade de sair bola vermelha, será

$$\Pr(x \in V) = \frac{1}{N} + \frac{1}{N} + \cdots + \frac{1}{N} \quad (v \text{ vezes}) = \frac{v}{N}.$$

Este último resultado corresponde directamente à definição clássica de probabilidade:

"A probabilidade dum acontecimento é o quociente do número de casos favoráveis ao acontecimento [1] pelo número total de casos possíveis, supondo que estes são todos igualmente prováveis".

Segundo o ponto de vista adoptado nesta definição, considera-se como primitivo, não o conceito de probabilidade, mas, sim, o de "igualmente provável".

É claro que a hipótese de as bolas serem iguais (em substância, forma e dimensões) nunca se realiza exactamente na prática: podem ser, apenas, *aproximadamente iguais* e, então, as respectivas probabilidades, $p_1, p_2, \ldots, p_N$, serão aproximadamente iguais, com um grau de aproximação correspondente. Se as bolas diferem entre si de maneira sensível, as probabilidades elementares $p_1, p_2, \ldots, p_N$ serão, também, sensivelmente diferentes e só poderão ser avaliadas *a posteriori,* efectuando numerosas extracções com reposição e registando as frequências relativas dos acontecimentos $x = x_1$, $x = x_2$, ..., $x = x_N$. Por comodidade, em tudo o que segue, ao tratar de problemas de bolas numa urna, supomos verificada a hipótese da igualdade.

---

(1) – Aqui a palavra "caso" é empregada na acepção de acontecimento celular. Chamam-se "casos favoráveis ao acontecimento" os casos que *implicam* o acontecimento.

Recordemos que a extracção casual de bolas com números constitui a base de certos jogos, como o loto ou a lotaria. O cálculo das probabilidades nasceu precisamente das reflexões de certos matemáticos, nomeadamente PASCAL, sobre questões relativas a *jogos de azar* (1).

Um jogo de azar conhecido desde a antiguidade é o dos dados. Em princípio, um dado deve ser um cubo formado de substância homogénea (*dado perfeito*), mas é claro que, na prática, estas condições só aproximadamente se realizam. As faces do cubo estão numeradas de 1 a 6. Admite-se que, lançando o dado, depois de o agitar dentro dum copo (*casualização*), todas as faces têm igual probabilidade de se apresentar em cima e será, pois, 1/6 a probabilidade correspondente a cada face. Esta hipótese é, geralmente, confirmada pela experiência: quando não houver confirmação, a experiência vem, apenas, revelar imperfeições que tenham passado despercebidas.

Recordemos, ainda, os jogos de cartas, alguns dos quais são puramente de azar. Nestes jogos, a casualização consiste em *baralhar* as cartas.

É, ainda, de citar o jogo da roleta – círculo dividido em sectores numerados, que roda em torno dum eixo que passa pelo centro. A casualização consiste em imprimir à roleta um impulso que a faça dar várias voltas: se os sectores foram iguais, a probabilidade de parar num dado ponto é a mesma para todos. Se os sectores forem diferentes, as probabilidades serão proporcionais às respectivas amplitudes, de modo que , se foram $\Theta_1, \Theta_2, \ldots, \Theta_N$, essas amplitudes em radianos, serão

$$p_1 = \frac{\Theta_1}{2\pi}, p_2 = \frac{\Theta_2}{2\pi}, \ldots, p_N = \frac{\Theta_N}{2\pi},$$

as respectivas probabilidades, pois que se deve ter

$$p_1 + p_2 + \cdots + p_N = 1.$$

(Este exemplo conduz, naturalmente, à consideração de probabilidade no contínuo).

---

(1) – Aqui ''azar'' não é usado como sinónimo de ''pouca sorte'', mas, simplesmente, como sinónimo de ''acaso'' (tal como o francês ''hasard'').

Recordemos, por último, o mais simples de todos os jogos de azar: o lançamento duma moeda ao ar. Se a moeda for bem *balançada*, as duas eventualidades "sair coroa" e "sair face" terão igual probabilidade, ou seja, 1/2. Se não, haverá uma diferença de probabilidade que a experiência acusará.

Para ilustrar as considerações precedentes, convém apresentar aqui alguns exemplos numéricos, que, no fundo, se reduzem geralmente a problemas de Cálculo Combinatório.

Exemplo 1 – *Determinar a probabilidade de que, no lançamento dum dado perfeito, se obtenha um múltiplo de* 3.

São 2 os casos favoráveis ao acontecimento: $x = 3$, $x = 6$. Visto haver ao todo 6 casos possíveis (igualmente prováveis), a probabilidade pedida será

$$\frac{2}{6} = \frac{1}{3}.$$

Exemplo 2 – *Determinar a probabilidade de que, em dois lançamentos sucessivos dum dado perfeito, se obtenham* 2 *números pares.*

As células, neste caso, podem ser representadas pelos pares ordenados (x, y), em que x designa o número saído no 1.º lançamento e y o número saído no 2.º lançamento. Estes pares são em número de $6^2 = 36$ e todos igualmente prováveis, como é evidente. Os casos que implicam o acontecimento "saída de dois números pares" são: (2,2), (2,4), (2,6), (4,2), (4,4), (4,6), (6,2), (6,4), (6,6), em número de 9. A probabilidade pedida será, pois,

$$\frac{9}{36} = \frac{1}{4}.$$

NOTA. Como se disse em Matemáticas Gerais, chama-se produto cartesiano dum conjunto A por um conjunto B, e designa-se por $A \times B$, o conjunto de todos os pares ordenados (a, b) que se obtém tomando um elemento a em A e um elemento b em B; o número de elementos de $A \times B$ será, então, o produto dos números de elementos de A e de B. Analogamente, chama-se produto cartesiano de 3

conjuntos A, B, C, na ordem por que estão escritos, e designa-se por $A \times B \times C$, o conjunto de todos os ternos ordenados (a, b, c) com $a \in A$, $b \in B$, $c \in C$; o número destes ternos é igual ao produto dos números de elementos de A, B, e C. E assim por diante.

Também convencionámos escrever A^2 como abreviatura de $A \times A$, A^3 como abreviatura de $A \times A \times A$, etc.

Deste modo, os casos possíveis, no exemplo 2, serão os elementos do produto cartesiano

$$\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\} = \{1, 2, 3, 4, 5, 6\}^2$$

e os casos que implicam o acontecimento considerado serão os elementos do produto cartesiano $\{2, 4, 6\}^2$.

Note-se que os pares ordenados, os ternos ordenados, etc., também se chamam *arranjos com repetição* (*dois a dois, três a três*, etc.).

Exemplo 3 – *Determinar a probabilidade de que, em dois lançamentos sucessivos dum dado perfeito, saia pelo menos uma vez número ímpar.*

O acontecimento "sair pelo menos uma vez número ímpar" é o contrário do acontecimento "sair duas vezes número par" cuja probabilidade, há pouco calculada, é 1/4. A probabilidade pedida será, pois,

$$1 - \frac{1}{4} = \frac{3}{4}.$$

Exemplo 4 – *Determinar a probabilidade de que, em dois lançamentos sucessivos dum dado perfeito, saia primeiro um número par e depois um múltiplo de* 3.

Os casos possíveis (células) são os mesmos que no exemplo 2 e no exemplo 3. Os casos favoráveis ao acontecimento considerado obtêm-se desenvolvendo o produto cartesiano

$$\{2, 4, 6\} \times \{3, 6\},$$

cujos elementos são em número de $3 \times 2 = 6$. A probabilidade pedida será, pois, $6/36 = 1/6$.

Exemplo 5 – *Determinemos a probabilidade de que, em dois lançamentos sucessivos dum dado perfeito, saia uma vez um número par e outra vez um múltiplo de* 3.

O acontecimento considerado é a soma lógica dos dois seguintes: "sair primeiro número par e depois um múltiplo de 3" e "sair primeiro um múltiplo de 3 e depois um número par". Estes dois acontecimentos, que designaremos por α_1 e α_2, têm ambos a probabilidade 1/6 (exemplo 4), *mas não são incompatíveis*. O acontecimento $\alpha_1\alpha_2$ equivale a "sair 6 duas vezes"; como se trata de um só caso entre 36 possíveis, a sua probabilidade é de 1/36. A probabilidade pedida será, pois (teorema 3):

$$\Pr(\alpha_1) + \Pr(\alpha_2) - \Pr(\alpha_1\alpha_2) = 2 \cdot \frac{1}{6} - \frac{1}{36} = \frac{11}{36}.$$

NOTA. Nos exemplos anteriores, excepto o primeiro, as considerações ficam essencialmente as mesmas, se substituirmos a expressão "em dois lançamentos sucessivos dum dado perfeito" pela expressão "num lançamento simultâneo de dois dados perfeitos".

Exemplo 6 – *Duma urna que contem* 12 *bolas, das quais* 4 *são brancas e* 8 *pretas, tiram-se duas bolas à sorte, uma após a outra, sem reposição. Calcular a probabilidade de que*: a) *sejam ambas brancas*; b) *sejam ambas pretas*; c) *sejam da mesma cor*; d) *sejam de cores diferentes*; e) *sejam a 1.ª branca e a 2.ª preta*; f) *sejam a 2.ª preta e a 1.ª branca.*

Suponhamos as bolas numeradas de 1 a 12, sendo as 4 primeiras brancas; designemos por x o primeiro número saído e por y o segundo. Neste caso, os acontecimentos celulares são representados pelos pares ordenados, (x, y), com $x \neq y$; trata-se, portanto, de arranjos (sem repetição!) de 12 elementos 2 a 2; o número desses arranjos é $12 \times 11 = 132$.

O acontecimento considerado em a) corresponde aos pares (x, y), em que x e y variam de 1 a 4, mas sendo $x \neq y$. Trata-se, portanto, de arranjos de 4 elementos 2 a 2, cujo número é $4 \times 3 = 12$. A probabilidade pedida será, pois, 12/132.

Analogamente, se reconhece que a probabilidade pedida em b) é

$$\frac{8\times 7}{132}=\frac{56}{132}\,.$$

O acontecimento considerado em c) é a soma lógica dos acontecimentos anteriores, que são incompatíveis. Portanto, a probabilidade pedida em c) será

$$\frac{12}{132}+\frac{56}{132}=\frac{68}{132}\,.$$

O acontecimento considerado em d) é o contrário do anterior. Logo, a sua probabilidade será

$$1-\frac{68}{132}=\frac{64}{132}\,.$$

É fácil reconhecer, finalmente, que as probabilidades pedidas em e) e f) são ambas iguais a

$$\frac{8\times 4}{132}=\frac{32}{132}\,;$$

a sua soma dá a anterior, como era de esperar.

Exemplo 7 – *Duma urna que contem N bolas, sendo a brancas e b pretas, extrai-se, ao acaso, uma amostra de n bolas* (*tiradas, por exemplo, uma após outra, sem reposição*). *Achar a probabilidade de que, na amostra obtida, haja* ν *bolas brancas e* $n-\nu$ bolas pretas.

Comecemos por notar que a ordem pela qual são tiradas as bolas não interessa à questão. Os casos possíveis (todos igualmente prováveis) serão, pois, as combinações das N bolas tomadas n a n, cujo número é $\binom{N}{n}$. Por sua vez, as amostras com ν bolas brancas e $n-\nu$ bolas pretas podem obter-se, sem omissão nem repetição, do seguinte modo:

1) – Formando, por um lado, as combinações das a bolas brancas tomadas ν a ν_1 cujo número é $\binom{a}{\nu}$.

2) – Formando, por outro lado, as combinações das b bolas pretas tomadas $n-\nu$ a $n-\nu$, cujo número é $\binom{b}{n-\nu}$.

3) – Arranjando todos os possíveis pares $A_\nu\ B_{n-\nu}$ constituídos por uma combinação A_ν obtida em 1) e uma combinação $B_{n-\nu}$ obtida em 2). Como o número destes pares é $\binom{a}{\nu}\binom{b}{n-\nu}$, a probabilidade pedida será

$$\frac{\binom{a}{\nu}\binom{b}{n-\nu}}{\binom{N}{n}}.$$

Esta distribuição de variável casual ν é chamada *distribuição hipergeométrica.*

Exemplo 8 – *Duma urna que contem N bolas, sendo a brancas e b pretas, tiram-se, ao acaso, sucessivamente, n bolas, repondo a bola retirada após cada extracção. Achar a probabilidade de que saia* ν *vezes bola branca* e $n-\nu$ *vezes bola preta.*

Discorrendo como há pouco, seríamos levados a considerar, agora, os casos possíveis sob a forma de *combinações com repetição*. Porém, esses casos *já não são igualmente prováveis, como é fácil ver.* Suponhamos, por exemplo, $n=3$; neste caso, a combinação $x_1x_2x_3$ será mais provável que a combinação $x_1x_1x_3$, pois que a primeira se pode apresentar com as seguintes ordens de saída de bolas:

$$x_1x_2x_3,\ x_1x_3x_2,\ x_2x_1x_3,\ x_2x_3x_1,\ x_3x_1x_2,\ x_3x_2x_1,$$

ao passo que a segunda se pode apresentar só dos seguintes modos:

$$x_1x_1x_3,\ x_1x_3x_1,\ x_3x_1x_1.$$

Este exemplo mostra, bem, o cuidado que é necessário ter na escolha dos acontecimentos celulares, para que sejam igualmente prováveis. Neste caso, há que tomar para células os arranjos com repetição das N bolas n a n (isto é, sistema de n bolas).

Veremos, mais adiante, como se resolve este problema, considerado sob um aspecto mais geral (distribuição de BERNOULLI).

Exemplo 9 – *Nas condições do exemplo* 7, *achar a probabilidade de que, numa amostra, o número* ν de bolas *brancas:* a) *não seja inferior a um dado número* L_1 *nem superior a um dado número* L_2; b) *seja inferior a* L_1 *ou superior a* L_2.

Em a) trata-se de achar a probabilidade do acontecimento

$$L_1 \leq \nu \leq L_2$$

o qual é a soma lógica dos acontecimentos

$$\nu = L_1,\ \nu = L_1 + 1,\ \ldots,\ \nu = L_2 - 1,\ \nu = L_2,$$

incompatíveis dois a dois. Ter-se-á, portanto,

$$\Pr(L_1 \leq \nu \leq L_2) = \Pr(\nu = L_1) + \Pr(\nu = L_1 + 1) + \cdots + \Pr(\nu = L_2),$$

ou seja, atendendo ao resultado do exemplo 7:

$$\Pr(L_1 \leq \nu \leq L_2) = \sum_{\nu=L_1}^{L_2} \frac{\binom{a}{\nu}\binom{b}{n-\nu}}{\binom{N}{n}}.$$

Em b) pede-se a probabilidade do acontecimento

$$(\nu < L_1) + (L_2 < \nu) \quad (\nu \text{ inferior a } L_1 \text{ ou } L_2 \text{ inferior a } \nu)$$

que é o contrário do acontecimento $L_1 \leq \nu \leq L_2$. Será, pois,

$$\Pr[(\nu < L_1) + (\nu > L_2)] = 1 - \Pr(L_1 \leq \nu \leq L_2).$$

Eis como se poderiam calcular exactamente as probabilidades que fomos levados a considerar em A-16, a propósito da experiência hipotética sobre animais. Porém, estas fórmulas, embora simples na aparência, exigem cálculos muito laboriosos. É possível substituí-las por outras que, com muito menos trabalho, fornecem boas aproximações quando o número n excede um certo limite.

NOTA IMPORTANTE RELATIVA ÀS NOTAÇÕES E TERMINOLOGIA

Convencionámos no n.º 4 designar por $\Pr(\alpha)$ a probabilidade do acontecimento α. Nesta ordem de ideias, representámos, atrás, por $\Pr(x=x_1)$ a probabilidade de "sair a bola x_1", por $\Pr(x \in V)$ a probabilidade de "sair bola vermelha", etc. Mas poderíamos, igualmente, para abreviar, designar por $\Pr(x_1)$, $\Pr(V)$, etc., aquelas probabilidades, dizendo "a probabilidade do elemento x_1", a "probabilidade do conjunto V", etc. Assim, a distribuição de probabilidade passa a conceber-se como distribuição definida, não num corpo de acontecimentos, mas, sim, num corpo de conjuntos: todos os possíveis conjuntos de bolas da urna (cf. A-4). Poderíamos, também, conceber esta distribuição como função $\Pr(x)$ da variável casual x, tendo-se, na hipótese de as bolas serem iguais,

$$\Pr(x) = \frac{1}{N}, \text{ para todo o valor de } x.$$

Estas considerações estendem-se, *mutatis mutandis,* a qualquer distribuição de probabilidade.

6. Independência e associação de acontecimentos

Sendo as probabilidades valores ideais de frequências relativas, as definições de independência e associação que demos para o caso das frequências traduzem-se, imediatamente, em termos de probabilidade. Seja $\mathcal{R}$ um corpo de eventualidades a considerar numa dada prova $\mathcal{P}$, com determinadas probabilidades. Dados dois acontecimentos α, β, *chama-se probabilidade condicional de* β *em relação a* α e *designa-se por* $\Pr(\beta|\alpha)$ ao *número dado pela fórmula*

$$(6.1) \qquad \Pr(\beta|\alpha) = \frac{\Pr(\alpha\beta)}{\Pr(\alpha)}.$$

Para avaliar $\Pr(\beta|\alpha)$ empiricamente, o processo a seguir consistiria em efectuar a prova $\mathcal{P}$ um grande número de vezes e registar: 1) a frequência absoluta (α) de α; 2) a frequência absoluta $(\alpha\beta)$ de $\alpha\beta$. O quociente de $(\alpha\beta)$ por (α) dar-nos-ia, então, um valor aproximado de $\Pr(\beta|\alpha)$.

É claro que, da definição (6.1), resulta, logo, a *fórmula do produto*:

$$(6.2) \qquad \Pr(\alpha\beta) = \Pr(\alpha) \,.\, \Pr(\beta|\alpha) = \Pr(\beta) \,.\, \Pr(\alpha|\beta),$$

a qual nos diz que: *a probabilidade de se verificarem ao mesmo tempo os acontecimentos* α e β *é o produto da probabilidade de* α *pela probabilidade condicional de* β *na hipótese de se verificar* α (*ou vice-versa*).

Os acontecimentos α, β *dizem-se estocasticamente independentes* ou, apenas, *independentes*, quando $\Pr(\beta|\alpha) = \Pr(\beta)$ ou, o que é equivalente, quando $\Pr(\alpha|\beta) = \Pr(\alpha)$; no caso contrário, dizem-se *associados*. Desta definição e da fórmula do produto deduz-se, logo, o

TEOREMA DO PRODUTO. *Se os acontecimentos* α, β *são independentes* (*e só neste caso*), *a probabilidade de realização simultânea de* α *e* β *é igual ao produto das probabilidades de* α *e de* β.

Suponhamos, por exemplo, que uma urna contém 8 bolas brancas e 24 pretas, estando 6 bolas brancas marcadas com o sinal +, 18 bolas pretas com este mesmo sinal, e todas as restantes com o sinal –. Então, como existem na urna 32 bolas, ao todo, sendo 24 marcadas com o sinal +, a probabilidade (incondicional) de aparecer o sinal + é 24/32 = 3/4. Por sua vez, a probabilidade de aparecer o sinal + em bola branca é 6/8 = 3/4, igual à primeira. Os acontecimentos "sair bola branca" e "sair sinal +" são, pois, independentes; deste modo, a probabilidade de "sair bola branca e sinal +" é:

$$\frac{8}{32}\cdot\frac{3}{4}=\frac{1}{4}\cdot\frac{3}{4}=\frac{3}{16}$$

(probabilidade de sair bola branca vezes probabilidade de sair sinal +).

Um exemplo mais sugestivo, embora menos rigoroso, será o seguinte. Consideremos a prova que consiste em *semear uma certa quantidade de trigo num certo campo*. Seja p a probabilidade (incondicional) *de colher então uma quantidade de sementes superior a um dado limite L e* seja p' a probabilidade do mesmo acontecimento, *na hipótese de a altura pluviométrica, nos meses de outono e inverno, ser inferior a um dado limite* λ. Será, então, p' uma probabilidade condicional do primeiro acontecimento em relação ao segundo; e é natural que seja $p' \neq p$, isto é, que a eventualidade "colher trigo em quantidade superior a L" depende da eventualidade "chover em quantidade inferior a λ".

Mas, geralmente, dados vários acontecimentos α, β, γ, ... do corpo $\mathcal{R}$ tem-se, como é fácil ver,

$$\text{(6.3)} \qquad \Pr(\alpha\beta\gamma\ldots) = \Pr(\alpha)\Pr(\beta|\alpha)\Pr(\gamma|\alpha\beta)\ldots .$$

Os acontecimentos α, β, γ, ... *são independentes, se for verdadeira, não só a igualdade*

$$\Pr(\alpha\beta\gamma\ldots) = \Pr(\alpha)\Pr(\beta)\Pr(\gamma)\ldots$$

como todas as que resultam desta substituindo um ou mais dos acontecimentos α, β, γ, ... *pelos seus contrários* (cf. A-15)[1].

7. Sistema de duas experiências

Os conceitos de associação e independência de acontecimentos têm interesse, principalmente, quando se trata de várias experiências combinadas. Comecemos por considerar o caso das duas experiências $\mathcal{P}$, $\mathcal{P}$', iguais ou diferentes, e sejam $\mathcal{R}$ e $\mathcal{R}$' dois corpos de eventualidades a considerar, respectivamente, nas provas $\mathcal{P}$ e $\mathcal{P}$'. A realização destas duas provas, ao mesmo tempo ou uma após a outra, pode

(1) – Pode considerar-se esta proposição como definição de independência, no caso de vários acontecimentos.

conceber-se como *experiência* composta, única, que designaremos por $(\mathcal{P}, \mathcal{P}')$. Sejam $\alpha_1, \alpha_2, \ldots \alpha_m$ as células de $\mathcal{R}$ e $\beta_1, \beta_2, \ldots, \beta_n$ as células de $\mathcal{R}'$. Designaremos, então, por

$$(\alpha_i, \beta_k)$$

o acontecimento que consiste em realizar-se α_i na prova $\mathcal{P}$ e β_k na prova $\mathcal{P}'$ (*acontecimento composto de* α_i *e* β_k). Todos estes acontecimentos (α_i, β_k) são resultados possíveis da prova $(\mathcal{P}, \mathcal{P}')$, são as células dum novo corpo, que designaremos por $\mathcal{R} \times \mathcal{R}'$. Note-se que cada acontecimento α_i pode ser considerado como elemento do corpo $\mathcal{R} \times \mathcal{R}'$, pondo

$$\alpha_i = (\alpha_i, \beta_1) + (\alpha_i, \beta_2) + \cdots + (\alpha_i, \beta_n),$$

visto que, $\beta_1 + \beta_2 + \cdots + \beta_n = I$. Analogamente:

$$\beta_k = (\alpha_1, \beta_k) + (\alpha_2, \beta_k) + \cdots + (\alpha_m, \beta_k).$$

Nesta ordem de ideias, podemos identificar o acontecimento composto (α_i, β_k) com o produto lógico $\alpha_i \beta_k$, desde que não haja perigo de confusão[1].

Suponhamos, agora, definido no corpo $\mathcal{R} \times \mathcal{R}'$ uma distribuição de probabilidade. Esta poderá indicar-se numa tábua do seguinte tipo (cf. tabela-tipo de A-14):

	β_1	β_2	.....	β_n	Total
α_1	$\Pr(\alpha_1 \beta_1)$	$\Pr(\alpha_1 \beta_2)$	.....	$\Pr(\alpha_1 \beta_n)$	$\Pr(\alpha_1)$
α_2	$\Pr(\alpha_2 \beta_1)$	$\Pr(\alpha_2 \beta_2)$	.....	$\Pr(\alpha_2 \beta_n)$	$\Pr(\alpha_2)$
.....	.....	.....	.....	.....	.....
α_m	$\Pr(\alpha_m \beta_1)$	$\Pr(\alpha_m \beta_2)$	.....	$\Pr(\alpha_m \beta_n)$	$\Pr(\alpha_m)$
Total	$\Pr(\beta_1)$	$\Pr(\beta_2)$	.....	$\Pr(\beta_n)$	1

(1) – A confusão pode surgir quando $\mathcal{P}'$ é apenas a repetição de $\mathcal{P}$. Neste caso, um acontecimento α deverá ser designado por dois símbolos diferentes, conforme se realize em $\mathcal{P}$ ou em $\mathcal{P}'$. Um outro modo de evitar a confusão é chamar a (α_i, β_k) o *produto cartesiano de* α_i *por* β_k. Este produto, *que não é comutativo*, será designado por $\alpha_i \times \beta_k$ para o distinguir do produto lógico vulgar $\alpha_i \beta_k$.

Note-se que, para $i = 1, \ldots, m$, $k = 1, 2, \ldots, n$, se tem:

$$\Pr(\alpha_i) = \Pr(\alpha_i \beta_1) + \Pr(\alpha_i \beta_2) + \cdots + \Pr(\alpha_i \beta_n),$$

$$\Pr(\beta_k) = \Pr(\alpha_1 \beta_2) + \Pr(\alpha_2 \beta_k) + \cdots + \Pr(\alpha_m \beta_k).$$

Como estas probabilidades estão registadas à margem (sob a indicação "total"), dá-se-lhes, também, o nome de *probabilidades marginais* da distribuição considerada. Pode acontecer, em particular, que se tenha

$$\Pr(\alpha_i \beta_k) = \Pr(\alpha_i) \Pr(\beta_k),$$

quaisquer que sejam i, k. Então, é fácil ver que *a probabilidade de qualquer acontecimento* $\alpha\beta$, *composto dum acontecimento* α *de* $\mathcal{R}$ *e dum acontecimento* β *de* $\mathcal{R}'$ *será igual ao produto das probabilidades de* α *e de* β[1]. Os acontecimentos de $\mathcal{R}$ dir-se-ão, neste caso, *independentes* dos acontecimentos de $\mathcal{R}'$. De contrário, terá de usar-se a fórmula (6.2).

Estas considerações tornam-se mais intuitivas quando se fala em termos de variáveis casuais. Consideremos um par (x, y) de variáveis casuais com uma determinada distribuição de probabilidade, $\Pr(x, y)$. Chama-se *probabilidade condicional dum valor* x_i *de* x *a respeito dum valor* y_k de y, e representa-se por $\Pr(x_i | y_k)$, o número dado por

$$\Pr(x_i | y_k) = \frac{\Pr(x_i, y_k)}{\Pr(y_k)}.$$

As variáveis x, y dizem-se *independentes* (*estocasticamente*), se for $\Pr(x | y) = \Pr(x)$ para todos os valores de x e y. Tem-se, pois, nesta hipótese

$$\Pr(x, y) = \Pr(x) \Pr(y).$$

---

(1) – Esta regra, ou, mais geralmente, a fórmula (6.3), é conhecida tradicionalmente, como *princípio das probabilidades compostas.*

NOTA. Seria mais correcto designar por símbolos diferentes as distribuições de probabilidade de x e y: por exemplo, a primeira por $\Pr_1(x)$ e a segunda por $\Pr_2(y)$. Mas, para não sobrecarregar as notações, e não havendo perigo de confusão, usamos aqui, para ambas as distribuições, o mesmo símbolo.

Exemplos – Consideremos uma urna que contenha M bolas $x_1, x_2, \ldots, x_M$, sendo as v primeiras vermelhas e as b últimas brancas. E consideremos uma outra urna, que contenha N bolas $y_1, y_2, \ldots, y_N$, sendo as v' primeiras vermelhas e as b' últimas brancas. Sejam $\mathcal{P}$, $\mathcal{P}'$, respectivamente, as experiências que consistem em tirar à sorte uma bola da primeira urna e tirar à sorte uma bola da segunda urna. Neste caso, os resultados elementares da experiência composta $(\mathcal{P}, \mathcal{P}')$ serão expressos pelos diferentes valores (x_i, y_k) da variável casual (x, y). Como estes valores são em número de MN e *todos igualmente prováveis*, a probabilidade de cada par (x_i, y_k) será

$$\frac{1}{MN} = \frac{1}{M} \cdot \frac{1}{N} = \Pr(x_i)\Pr(y_k),$$

o que significa que *as duas variáveis casuais x, y são independentes* (*estocasticamente*). Os resultados da prova $\mathcal{P}'$ são pois independentes dos resultados da prova $\mathcal{P}$. Por exemplo, a probabilidade do acontecimento *sair bola branca em* $\mathcal{P}$ *e bola vermelha em* $\mathcal{P}'$, será

$$\frac{b}{M} \cdot \frac{v'}{N} = \frac{bv'}{MN}.$$

Casos análogos ao anterior serão, ainda, todos aqueles em que $\mathcal{P}'$ é simplesmente a repetição de $\mathcal{P}$. Suponhamos, por exemplo, que se trata de duas extracções sucessivas duma bola da 1.ª urna, *com reposição*. A probabilidade de sair, em primeiro lugar, bola vermelha e, depois, bola branca será

$$\frac{v}{M} \cdot \frac{b}{M} = \frac{bv}{M^2}.$$

igual à probabilidade de sair, primeiro, branca e, depois, vermelha (será, portanto, $2bv/M^2$ a probabilidade do acontecimento "sairem cores diferentes").

Mas suponhamos, agora, que $\mathcal{P}$ consiste em tirar uma bola da 1.ª urna, e $\mathcal{P}'$ em tirar uma segunda bola da mesma urna *sem repor a primeira bola tirada*. Os resultados elementares da prova composta $(\mathcal{P}, \mathcal{P}')$ serão, ainda, os valores da variável casual (x, y), *mas com a condição suplementar* $x \neq y$. Isto basta para ver que, neste caso, as variáveis casuais x, y não são independentes. Por exemplo, a probabilidade (condicional) de sair bola vermelha na 2.ª extracção, *tendo saído branca na* 1.ª, é

$$\frac{v}{M-1},$$

ao passo que a probabilidade (condicional) de sair vermelha na 2.ª, *tendo saído vermelha* na 1.ª é

$$\frac{v-1}{M-1}.$$

Assim, a probabilidade de *sair vermelha duas vezes* é

$$\frac{v}{M} \cdot \frac{v-1}{M-1} = \frac{v(v-1)}{M(M-1)}$$

e a de *sair primeiro branca e depois vermelha*:

$$\frac{b}{M} \cdot \frac{v}{M-1} = \frac{bv}{M(M-1)}.$$

Estes resultados podem recolher-se na seguinte tabela:

TABELA N.° 13

1.ª \ 2.ª	Vermelha	Branca	Total
Vermelha	$\frac{v(v-1)}{M(M-1)}$	$\frac{vb}{M(M-1)}$	$\frac{v}{M}$
Branca	$\frac{bv}{M(M-1)}$	$\frac{b(b-1)}{M(M-1)}$	$\frac{b}{M}$
Total	$\frac{v}{M}$	$\frac{b}{M}$	1

Na margem direita estão indicadas as probabilidades de sair bola vermelha e a de sair bola branca na 2.ª extracção, respectivamente, iguais às probabilidades de sair vermelha e de sair branca na 1.ª extracção (registadas na margem inferior). A tabela patenteia, assim, a associação dos acontecimentos considerados.

8. Sistema de várias experiências

Podemos, agora, conceber, mais geralmente, uma experiência $(\mathcal{P}, \mathcal{P}', \mathcal{P}'', \ldots)$, composta de várias experiências $\mathcal{P}$, $\mathcal{P}'$, $\mathcal{P}''$, ..., (em número finito). Sendo α, β, γ, ... eventualidades a considerar em cada uma destas experiências, designaremos por $(\alpha, \beta, \gamma, \ldots)$ a eventualidade que consiste em acontecer α em $\mathcal{P}$, β em $\mathcal{P}'$, γ em $\mathcal{P}''$, etc., e diremos que $(\alpha, \beta, \gamma, \ldots)$ é o *acontecimento composto* de α, β, γ, ... Porém, segundo o ponto de vista explanado no número anterior, podemos considerar este novo acontecimento como o produto lógico dos acontecimentos α, β, γ, ..., desde que se tomem as devidas precauções[1]. Suponhamos que, a todo o acontecimento a considerar na prova $(\mathcal{P}, \mathcal{P}', \mathcal{P}'', \ldots)$, corresponde uma determinada probabilidade. Então, o que se disse em B-6 é aqui aplicável: os acontecimentos α, β, γ, ... dizem-se *independentes* (*estocasticamente*), se for verdadeira não só a igualdade

$$\Pr(\alpha\beta\gamma \ldots) = \Pr(\alpha)\Pr(\beta)\Pr(\gamma)\ldots, \tag{8.1}$$

como todas as que se deduzem desta substituindo um ou mais dos acontecimentos α, β, γ, ... pelos seus contrários. Não sendo assim, terá de usar-se a fórmula geral do produto, com as probabilidades condicionais.

Poderíamos, de novo, dar aqui exemplos de tiragens de bolas de uma ou várias urnas, com ou sem reposição, mas as considerações do número seguinte bastam para esclarecer as precedentes.

---

(1) – Isto é, se $\mathcal{P}'$, $\mathcal{P}''$, ... consistem apenas na repetição de $\mathcal{P}$, um mesmo acontecimento deverá ser designado de modos diversos, conforme as provas em que se realiza. É o que faremos mais adiante.

9. Distribuição binomial ou de BERNOULLI

Seja α um acontecimento a esperar com determinada probabilidade numa certa prova $\mathcal{P}$ e consideremos n realizações $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n$, da prova genérica $\mathcal{P}$ [1]. Estas provas particulares constituem uma experiência composta $(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n)$. Tendo em vista o conceito empírico de probabilidade, é fácil reconhecer que a probabilidade do acontecimento α deverá ser a mesma em cada uma das provas, quaisquer que sejam os resultados das restantes provas. Por outras palavras:

Os acontecimentos a esperar na repetição duma dada prova são independentes dos resultados das provas já efectuadas.

É claro que este facto não se deduz da anterior axiomática das probabilidades. Poderia, sim, assumir-se como novo axioma, porém com carácter bem diverso do dos primeiros.

Um exemplo típico é o das sucessivas extracções casuais de bolas de uma urna, com reposição da bola após cada extracção.

Pode, pois, aplicar-se nestes casos a fórmula (8.1).

Proponhamo-nos, então, resolver o seguinte problema:

Determinar a probabilidade de que, em n realizações da prova $\mathcal{P}$, *um acontecimento* α *de probabilidade p se realize x vezes.*

(Para concretizar, podemos supor que α consiste em tirar à sorte uma bola duma urna com bolas brancas e pretas, sendo α o acontecimento "sair bola branca". Porém, as considerações que se seguem são de todo gerais).

Representemos por q a probabilidade de $\tilde{\alpha}$, isto é, ponhamos $q = 1 - p$. Para evitar confusões, designaremos por α_i o acontecimento particular que consiste em realizar-se α na prova $\mathcal{P}_i$ $(i = 1, 2, \ldots, n)$; é claro que será sempre

$$\Pr(\alpha_i) = p, \Pr(\tilde{\alpha}_i) = q.$$

Suponhamos, por exemplo, $n = 3$ e $x = 2$. Três são os modos de α se realizar 2 vezes numa série de 3 provas: na 1.ª e na 2.ª, na 1.ª e na 3.ª ou na 2.ª e na 3.ª; isto é, em símbolos:

$$\alpha_1\alpha_2\tilde{\alpha}_3, \qquad \alpha_1\tilde{\alpha}_2\alpha_3, \qquad \tilde{\alpha}_1\alpha_2\alpha_3.$$

---

(1) – Dizendo que $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n$ são realizações da mesma prova $\mathcal{P}$, fica implícito que $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n$ se realizam em condições idênticas.

O acontecimento que consiste em α *se realizar* 2 *vezes nas* 3 *provas*, será, pois, a soma lógica

$$\alpha_1\alpha_2\tilde{\alpha}_3 + \alpha_1\tilde{\alpha}_2\alpha_3 + \tilde{\alpha}_1\alpha_2\alpha_3.$$

Ora, as três modalidades consideradas são incompatíveis duas a duas e têm todas as mesmas probabilidades; por exemplo:

$$\Pr(\alpha_1\tilde{\alpha}_2\alpha_3) = \Pr(\alpha_1)\Pr(\tilde{\alpha}_2)\Pr(\alpha_3) = pqp = p^2q.$$

A probabilidade pedida será, pois, neste caso, $3p^2q$.

Passemos, agora, ao caso geral. Um dos modos de α se realizar x vezes nas n provas será

$$\alpha_1\alpha_2 \ldots \alpha_x\tilde{\alpha}_{x+1} \ldots \tilde{\alpha}_n.$$

Qualquer outra modalidade se obtém escolhendo, entre as n provas, as x provas em que se realize α. As modalidades possíveis correspondem, pois, às combinações das n provas x a x. Como o número total destas combinações é $\binom{n}{x}$, o acontecimento que consiste em α *se realizar* x *vezes nas* n *provas* é a soma lógica de $\binom{n}{x}$ eventualidades, incompatíveis duas a duas e todas com a mesma probabilidade, que é

$$\Pr(\alpha_1)\Pr(\alpha_2) \ldots \Pr(\alpha_x)\Pr(\tilde{\alpha}_{x+1}) \ldots \Pr(\tilde{\alpha}_n) = p^x q^{n-x}.$$

A probabilidade pedida será, pois,

$$\Pr(x) = \binom{n}{x} p^x q^{n-x}$$

fórmula esta muito importante. Note-se que a variável casual x *representa, aqui, a frequência absoluta de* α *numa série de n provas.* A distribuição de probabilidade desta variável, dada pela fórmula anterior, tem o nome de *distribuição de BERNOULLI.* Também se

lhe chama distribuição *binominal*, atendendo a que os diferentes valores de $\Pr(x)$ são os termos do desenvolvimento da potência n do binómio $p + q$:

$$(p+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}.$$

É claro que $(p + q)^n = 1$. Procuremos, agora, a probabilidade de que a frequência absoluta de α seja inferior ou igual a um dado limite L (com $L \leq n$). Como o acontecimento $x \leq L$ é a soma lógica dos acontecimentos incompatíveis $x = 0$, $x = 1, \ldots, x = L$, virá:

$$\Pr(x \leq \mathrm{L}) = \Pr(x=0) + \Pr(x=1) + \cdots + \Pr(x=L) = \sum_{x=0}^{L} \binom{n}{x} p^x q^{n-x}.$$

Analogamente, se reconhece que

$$\Pr(L_1 \leq x \leq L_2) = \sum_{x=L_1}^{L_2} \binom{n}{x} p^x q^{n-x}.$$

(Confrontar com os exemplos 7, 8 e 9 do n.° 5).

Se pusermos $\xi = x/n$ (frequência relativa de α nas n provas), a distribuição da variável ξ será, manifestamente:

$$\Pr(\xi) = \binom{n}{n\xi} p^{n\xi} q^{n(1-\xi)}.$$

A tabela n.° 14 representa a distribuição binominal para $p = 1/2$, $n = 10$. Por exemplo, a probabilidade de que em 10 lançamentos sucessivos duma moeda "correcta" se apresente 3 vezes coroa, é

$$\Pr(3) = \binom{10}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 = \left(\frac{10.9.8}{3.2.1}\right) \cdot \left(\frac{1}{2^{10}}\right) = \frac{15}{128},$$

valor que concorda com o correspondente da tabela (a menos de 0,001).

TABELA N.° 14

x	$\Pr(x)$
0	0,001
1	0,010
2	0,044
3	0,117
4	0,205
5	0,246

x	$\Pr(x)$
6	0,205
7	0,117
8	0,044
9	0,010
10	0,001

Na Fig. 3 é dado o histograma desta distribuição.

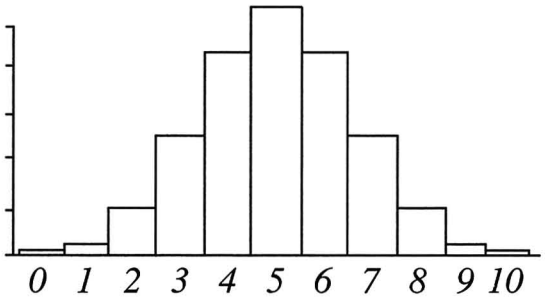


Fig. 3

Note-se que o histograma é *simétrico* a respeito da recta $x = 5$ e que a função $\Pr(x)$ atinge um máximo no ponto 5.

Podemos ver um exemplo curioso desta distribuição no campo da genética. Sabe-se que, no cruzamento dum touro avermelhado do Shorthorn com uma vaca malhada da mesma raça, há a probabilidade de 1/2 de se obter um vitelo avermelhado sem malhas[1]. Então, a probabilidade de que, em 10 destes cruzamentos, se obtenham x vitelos avermelhados (sem malhas), é dada pela tabela n.° 14.

(1) – Traduz-se por ''touro avermelhado'' a expressão inglesa ''red bull'' e por ''vaca malhada'' a expressão ''roan cow''. Neste caso, ''malhado'' significa ''avermelhado, com malhas brancas ou cinzentas dispersas''.

Por sua vez, sabe-se que a probabilidade de que, no cruzamento dum touro malhado Shorthorn com uma vaca malhada da mesma raça se obtenha um vitelo avermelhado (sem malhas) é 1/4. Então, a probabilidade de que, em 10 destes cruzamentos, se obtenham x vitelos avermelhados (sem malhas) será:

$$\Pr(x) = \binom{10}{x}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{10-x}.$$

Na Fig. 4 é dado o histograma desta distribuição, que já não apresenta simetria.

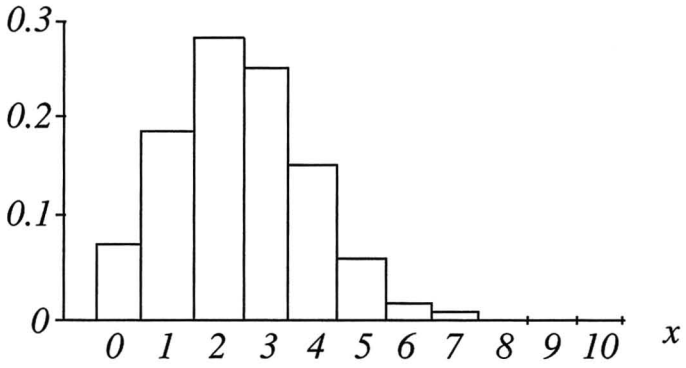


Fig. 4

Os dois casos extremos na distribuição binominal são os seguintes:

1) $x = n$ (o acontecimento α realiza-se nas n provas)

2) $x = 0$ (o acontecimento α não se realiza em nenhuma das provas).

A probabilidade do primeiro caso é (como se poderia reconhecer mesmo directamente):

$$\Pr(n) = p^n.$$

A probabilidade do segundo caso é:

$$\Pr(0) = q^n.$$

Note-se, porém, que a negação do acontecimento $x = 0$ não é o acontecimentos $x = n$, mas sim o acontecimento que consiste em α *se realizar pelo menos uma vez em n provas.*

A probabilidade deste acontecimento será, pois,

$$\Pr(x \neq 0) = 1 - q^n = 1 - (1-p)^n = \sum_{x=1}^{n} (-1)^{x-1} \binom{n}{x} p^x.$$

Exemplo – *Calcular a probabilidade de que, jogando* 1.000 *vezes num número duma lotaria com* 30.000 *números, se tenha pelo menos uma vez a sorte grande.*

A probabilidade de, em cada extracção, se ter a sorte grande, é, manifestamente, 1/30.000. Então, a probabilidade pedida será

$$1 - \left(1 - \frac{1}{30.000}\right)^{1.000} = \frac{1.000}{30.000} - \binom{1.000}{2} \frac{1}{(30.000)^2} + \cdots \approx 0{,}0328$$

probabilidade esta ainda muito fraca, apesar do grande número de tentativas.

10. Conceito de moda. Caso da distribuição normal

Chama-se *moda* duma distribuição de probabilidade duma variável x (com um número finito de valores) todo o valor de x, cuja probabilidade seja superior ou igual à de qualquer outro valor de x. Há distribuições com uma só moda, distribuições com mais de uma moda e distribuições sem moda.

Vamos ver que a distribuição binominal

$$\Pr(x) = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x} \quad (\text{com } q = 1-p)$$

apresenta uma ou, quando muito, duas modas.

Considerando três valores consecutivos, $X-1$, X, $X+1$ da variável x, tem-se

$$\Pr(X-1) = \frac{n!}{(X-1)!\,(n-X+1)!}\, p^{X-1} q^{n-X+1}$$

$$\Pr(X) = \frac{n!}{X!\,(n-X)!}\, p^X q^{n-X}$$

$$\Pr(X+1) = \frac{n!}{(X+1)!\,(n-X-1)!}\, p^{X+1} q^{n-X-1}.$$

Então, virá, como é fácil verificar,

$$\frac{\Pr(X)}{\Pr(X-1)} = \frac{(n-X+1)p}{Xq},$$

$$\frac{\Pr(X)}{\Pr(X+1)} = \frac{(X+1)q}{(n-X)p}.$$

Por conseguinte, para que X seja uma moda, deve ser, simultaneamente,

$$(n-X+1)p \geq Xq,$$
$$(X+1)q \geq (n-X)p.$$

Ora, esta dupla condição equivale à seguinte

$$np - q \leq X \leq np + p.$$

Como a diferença entre $np+p$ e $np-q$ é igual a $p+q=1$, a anterior condição será verificada por um só valor inteiro de X (a moda), a não ser que $np+p$ e $np-q$ sejam números inteiros, que serão, nesse caso, as duas modas existentes.

Assim, a moda ou as modas da distribuição binominal são sempre números inteiros que diferem de np menos de uma unidade.

No primeiro exemplo atrás considerado ($p=1/2$, $n=10$), a moda é precisamente $np=10.1/2=5$. No segundo exemplo ($p=1/4$, $n=10$), a moda X deve verificar a condição

$$1{,}75 = 10 \cdot \frac{1}{4} - \frac{3}{4} \leq X \leq 10 \cdot \frac{1}{4} + \frac{1}{4} = 2{,}75;$$

só poderá, então, ser $X=2$. Mas, se em vez de $n=10$, tivéssemos tomado $n=15$, com $p=1/4$, já teríamos duas modas: $X_1=3$, $X_2=4$.

11. Distribuição polinomial. Amostras casuais

Consideremos agora, mais geralmente, uma partição formada por r acontecimentos $\alpha_1, \alpha_2, \ldots, \alpha_r$, a prever numa certa prova $\mathcal{P}$, com probabilidades $p_1, p_2, \ldots, p_r$ respectivamente[1]. Pergunta-se:

Qual a probabilidade de que, em n realizações da prova $\mathcal{P}$, as frequências absolutas dos acontecimentos $\alpha_1, \alpha_2, \ldots, \alpha_r$, sejam, respectivamente, $x_1, x_2, \ldots, x_r$?

É claro que, sendo as eventualidades $\alpha_1, \alpha_2, \ldots, \alpha_r$, por hipótese, incompatíveis duas a duas, e sendo a sua soma lógica o acontecimento certo, deverá ter-se

$$p_1 + p_2 + \cdots + p_r = 1, \qquad x_1 + x_2 + \cdots + x_r = n.$$

Suponhamos que os acontecimentos $\alpha_1, \alpha_2, \ldots, \alpha_r$ se verificam numa determinada ordem, por exemplo:

$$\begin{aligned}
&\alpha_1 \text{ nas provas } \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{x_1} \\
&\alpha_2 \text{ nas provas } \mathcal{P}_{x_1+1}, \mathcal{P}_{x_1+2}, \ldots, \mathcal{P}_{x_1+x_2} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\alpha_r \text{ nas provas } \mathcal{P}_{n-x_r}; \mathcal{P}_{n-x_r+1}, \ldots, \mathcal{P}_n \quad .
\end{aligned}$$

Obtém-se, então, um acontecimento composto, cuja probabilidade é $p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}$. É claro que, qualquer que seja a ordem de realização, a probabilidade será esta; a probabilidade pedida será, pois, o produto de $p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}$ pelo número total de ordens possíveis. Mas este número não é $n!$ (permutações das n provas), como poderia parecer, visto que, permutando entre si as x_1 provas em que se verifica α_1, as x_2 provas em que se verifica α_2, ..., as x_r provas em que se verifica α_r, se obtém sempre o mesmo acontecimento. Como se vê facilmente, o número das ordens possíveis será $n!$ dividido por $x_1!\, x_2! \ldots x_r!$, e a probabilidade pedida será, então,

$$\Pr(x_1, x_2, \ldots x_r) = \frac{n!}{x_1!\, x_2! \ldots x_r!}\, p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}.$$

---

(1) – As eventualidades $\alpha_1, \alpha_2, \ldots, \alpha_r$ serão pois incompatíveis duas a duas e a sua soma lógica será o acontecimento certo.

Temos aqui, pois, um exemplo duma distribuição de r variáveis casuais $x_1, x_2, \ldots, x_r$. É claro que, por ser $x_1 + x_2 + \ldots + x_r = n$, estas variáveis não são independentes, nem sequer algebricamente.

A referida distribuição diz-se *polinominal*, atendendo a que os valores de $\Pr(x_1, x_2, \ldots, x_r)$ são os termos do desenvolvimento da potência n do polinómio $p_1 + p_2 + \ldots + p_r$. Tem-se, com efeito, segundo a fórmula de LEIBNIZ:

$$(p_1 + p_2 + \cdots + p_r)^n = \sum_{x_1 + \ldots + x_r = n} \frac{n!}{x_1!\, x_2! \ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}.$$

Note-se que a distribuição binominal é um caso particular desta, correspondente a $r = 2$ (partição dicotómica). Tem-se, então, $x_2 = n - x_1$, donde

$$\frac{n!}{x_1!\, x_2!} = \frac{n!}{x_1!\, (n - x_1)!} = \binom{n}{x_1}.$$

Este resultado aplica-se em questões de *amostragem casual*. Chama-se *amostra casual* a todo o sistema $(u_1, u_2, \ldots, u_n)$ de n indivíduos tirados *ao acaso* duma dada população U.

Como já tivemos ocasião de verificar em exemplos, a amostragem casual pode fazer-se de dois modos: *com reposição ou sem reposição*. No primeiro caso, cada um dos indivíduos $u_1, u_2, \ldots, u_n$ é tirado e, em seguida, reposto na população, antes de se tirar o seguinte: *pode assim acontecer que um mesmo indivíduo apareça repetido na amostra*. No segundo caso, os indivíduos são tirados sucessivamente, sem regressarem à população após as tiragens.

Suponhamos dada no universo U uma partição em r atributos $\alpha_1, \alpha_2, \ldots, \alpha_r$ e sejam $a_1, a_2, \ldots, a_r$, respectivamente, as frequências absolutas de $\alpha_1, \alpha_2, \ldots, \alpha_r$ em U. Suponhamos, ainda, que todos os indivíduos têm igual probabilidade de ser tirados. Então, a probabilidade de aparecer um indivíduo com o atributo α_i será, manifestamente,

$$p_i = \frac{a_i}{N}, \; i = 1, 2, \ldots, r,$$

em que $N = a_1 + a_2 + \ldots + a_r$ (número de elementos de U).

Qual é, neste caso, a probabilidade de que, numa amostragem casual de n elementos, com reposição, as frequências absolutas de $\alpha_1, \alpha_2, \ldots, \alpha_r$ *sejam, respectivamente,* $x_1, x_2, \ldots, x_r$?

A resposta é, segundo o que vimos,

$$\Pr(x_1, x_2, \ldots, x_r) = \frac{n!}{x_1!\, x_2! \ldots x_r!} \left(\frac{a_1}{N}\right)^{x_1} \left(\frac{a_2}{N}\right)^{x_2} \cdots \left(\frac{a_r}{N}\right)^{x_r}.$$

Mas seja, agora, este outro problema:

Qual é a probabilidade de que, numa amostragem casual de n elementos, sem reposição, as frequências absolutas de $\alpha_1, \alpha_2, \ldots, \alpha_r$, *sejam, respectivamente,* $x_1, x_2, \ldots, x_r$?

Raciocinando como no exemplo 7 do n.° 21, chega-se, agora, ao resultado

$$\Pr(x_1, x_2, \ldots, x_r) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2} \cdots \binom{a_r}{x_r}}{\binom{N}{n}}.$$

Eis aqui um novo exemplo de distribuição de r variáveis casuais, que, no caso particular $r = 2$, tem o nome de *distribuição hipergeométrica.*

É fácil ver que, se n é bastante pequeno em relação a N (isto é, se a razão n/N é bastante pequena), esta distribuição coincide, sensivelmente, com a anterior.

Note-se, ainda, que, na prática, as fórmulas obtidas exigem cálculos muito laboriosos quando n é grande. Têm de ser então substituídas por outras que, embora não sejam exactas, se adaptam muito melhor ao cálculo numérico, dando uma boa aproximação, para valores de n elevados.

BIBLIOGRAFIA

Além das obras indicadas na advertência prévia, recomendamos, ainda, as seguintes:

A. C. AITKEN – *Statistical Mathematics*. University Mathematical Texts. Oliver and Boyd; Edimburg and London, 1944.

A. HALD – *Statistical Theory with Engineering Application*, New York (John Wiley & Sons Inc.) and London (Chapman & Hall, Ld.), 1952.

P. LEVY – *Calcul des probabilités*, Gauthiers – Villars, Paris.

I.4.2

APONTAMENTOS DE CÁLCULO DAS PROBABILIDADES

A – Distribuições de uma variável contínua (real)

Consideremos na recta real, **R**, um intervalo U de extremos a, b, que, para fixar ideias, vamos supor limitado e fechado: $U = [a, b]$. Seja x uma variável casual (comprimento, densidade, percentagem ou qualquer outra espécie de grandeza) que toma todos os seus valores no intervalo $[a, b]$. Trata-se, pois, agora, duma *variável casual contínua* (variável real).

Suponhamos que, para cada intervalo J contido em U, existe uma determinada *probabilidade de que o valor de x esteja em J*. Essa probabilidade representa-se por qualquer dos símbolos

$$\Pr(x \in J) \text{ ou } \Pr(J).$$

Em particular, se for $J = [x_1, x_2]$, poderá, ainda, escrever-se

$$\Pr(J) = \Pr(x_1 \leq x \leq x_2);$$

se for $J = [x_1, x_2[$, poderá escrever-se

$$\Pr(J) = \Pr(x_1 \leq x < x_2), \text{ etc.}$$

É claro que $\Pr(J)$, função numérica do intervalo variável $J \subset U$, deve satisfazer às seguintes condições (ou axiomas):

I. $\Pr(J) \geq 0$, *qualquer que seja J.*

II. $\Pr(J) = 1$, se $J = U$.

III. $\Pr(J_1 + J_2) = \Pr(J_1) + \Pr(J_2)$, *se* J_1 e J_2 *forem intervalos contíguos mas disjuntos.*

Note-se que esta última condição pode apresentar-se com vários aspectos. Assim, se forem x_1, x_2, x_3 três pontos de U, tais que $x_1 < x_2 < x_3$, pode ter-se:

$$\Pr(x_1 \leq x \leq x_3) = \Pr(x_1 \leq x < x_2) + \Pr(x_2 \leq x \leq x_3)$$

$$= \Pr(x_1 \leq x \leq x_2) + \Pr(x_2 < x \leq x_3)$$

$$\Pr(x_1 \leq x < x_3) = \Pr(x_1 \leq x < x_2) + \Pr(x_2 \leq x < x_3), \text{ etc.}$$

Dum modo geral, quando $x_1 = x_2$, identificaremos o intervalo $[x_1, x_2]$ com o ponto x_1 e escreveremos

$$\Pr([x_1, x_1]) = \Pr(x = x_1) = \Pr(x_1).$$

Verificadas as referidas condições, diremos que a função $\Pr(J)$ é uma *distribuição de x definida sobre o intervalo U* (universo infinito, visto ser formado por uma infinidade de pontos).

Das condições I, II e III, resulta que é sempre

$$0 \leq \Pr(J) \leq 1.$$

Também é fácil ver que, por exemplo:

$$\Pr(x_1 \leq x \leq x_2) = \Pr(x_1 \leq x < x_2) + \Pr(x_2).$$

Outras consequências poder-se-iam deduzir, ainda, de tais axiomas.

NOTA. Às condições anteriores pode juntar-se, para fins teóricos, uma outra (*axioma de continuidade*), que enunciaremos nos seguintes termos:

IV. *Dada uma sucessão infinita* (J_n) *de intervalos tais que*

$$J_1 \subset J_2 \subset \ldots \subset J_n \subset \ldots,$$

sendo J a reunião de todos estes intervalos, ter-se-á

$$\Pr(J) = \lim_{n \to \infty} \Pr(J_n).$$

Em estudos de nível mais elevado, os axiomas III e IV são substituidos por um outro, mais geral, relativo à soma de sucessões infinitas de conjuntos C_n disjuntos dois a dois. O axioma traduz-se, simplesmente, pela fórmula

$$\Pr\left(\sum_{n=1}^{\infty} C_n\right) = \sum_{n=1}^{\infty} \Pr(C_n).$$

Os conjuntos C_n podem ser intervalos ou quaisquer outros conjuntos que se possam construir a partir dos intervalos pela aplicação sucessiva ou alternada do símbolo

$$\sum_{1}^{\infty}$$

e de passagens ao complementar: dá-se-lhes o nome de *conjuntos borelianos ou conjuntos de* BOREL.

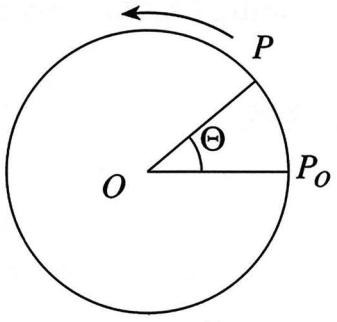


Fig. 1

Exemplo – Numa roleta "perfeita", cada ponto P da sua circunferência é determinado por um número real Θ, igual ou superior a 0 e

menor que 2π, que é o arco $\widehat{P_0P}$, medido em radianos num sentido prefixo, tomando como origem um ponto P_0. É claro que a correspondência $P \mapsto \Theta$ assim estabelecida, entre os pontos da circunferência e os pontos do intervalo $[0, 2\pi[$, é biunívoca. As probabilidades correspondentes a dois arcos iguais serão também iguais (na hipótese de a roleta ser perfeita). Por sua vez, a probabilidade correspondente à circunferência será

$$\Pr(0 \leq \Theta \leq 2\pi) = 1.$$

Então, atendendo à condição III, é fácil ver que

$$\Pr(\Theta_1 \leq \Theta \leq \Theta_2) = \Pr(\Theta_1 < \Theta < \Theta_2)$$

$$= \frac{\Theta_2 - \Theta_1}{2\pi}.$$

Em particular, tem-se

$$\Pr(\Theta = \Theta_1) = \frac{\Theta_1 - \Theta_1}{2\pi} = 0,$$

isto é, a probabilidade de cada valor particular Θ_1 de Θ é *nula; não quer isto dizer, porém, que qualquer dos acontecimentos individuais* $\Theta = \Theta_1$ *seja impossível, pois que um deles há-de realizar-se, necessariamente, em cada prova.* O que podemos dizer, ainda aqui, é que se trata de acontecimentos *praticamente impossíveis.*

Dada uma distribuição de probabilidade, $\Pr(J)$, sobre o intervalo $U = [a, b]$, o número

$$\Pr(x \leq u) = \Pr([a, u])$$

é, manifestamente, uma função de u, a que chamaremos *cumulant da distribuição.* Representamo-la por $\Phi(u)$ ou, mesmo, por $\Phi(x)$, tomando x para variável independente, em vez de u. É claro que, aumentando u, $\Phi(u)$ não pode diminuir, em virtude das condições I, II, III: *a função* $\Phi(x)$ *é, pois, crescente em sentido lato no intervalo* U.

NOTA. Do axioma IV deduz-se que

$$\Pr(x < u) = \lim_{x \to u^-} \Phi(x) = \Phi(u^-).$$

Com efeito, dada uma sucessão *crescente* de pontos x_n, convergente para u, o intervalo $[a, u[$ será a reunião de todos os intervalos $[a, x_n]$ e, portanto, a probabilidade de x estar em $[a, u[$, ou seja, $\Pr(x < u)$, será o limite, quando $n \to \infty$, da probabilidade de x estar em $[a, x_n]$. Tem-se, pois, $\lim_{n \to \infty} \Pr(x < x_n) = \lim_{x \to u^-} \Pr(x \leq u) = \Phi(u^-)$.

Então, será

$$\Pr(x_1 \leq x \leq x_2) = \Phi(x_2) - \Phi(x_1^-),$$

e, analogamente,

$$\Pr(x_1 < x < x_2) = \Phi(x_2^-) - \Phi(x_1),$$

$$\Pr(x_1 < x \leq x_2) = \Phi(x_2) - \Phi(x_1),$$

$$\Pr(x_1 \leq x < x_2) = \Phi(x_2^-) - \Phi(x_1^-).$$

Assim, *toda a distribuição* $\Pr(J)$ *é determinada pela sua função cumulante*, $\Phi(x)$.

É claro que será

$$\Pr(x) = \Phi(x) - \Phi(x^-), \text{ para todo o } x \in U.$$

$$\Phi(b) = \Pr(U) = 1.$$

A função $\Phi(x)$ será contínua à direita em todos os pontos, isto é: $\Phi(x^+) = \Phi(x)$, qualquer que seja x.

Em particular, pode suceder que, num ponto x_0, se tenha $\Phi(x_0^-) = \Phi(x_0)$. *Então, a função* $\Phi(x)$ *é contínua em* x_0 *e a probabilidade deste ponto é nula, visto que* $\Phi(x_0) - \Phi(x_0^-) = 0$.

Um caso particular importante é aquele em que a função $\Pr(x)$ *do ponto* x é nula, excepto num número finito de pontos $x_1, x_2, \ldots, x_n$ do intervalo U: recaímos, então, no caso já estudado das variáveis casuais descontínuas, com um número finito de valores. Neste caso, a

probabilidade $\Pr(J)$, correspondente a um dado intervalo J, será a soma $\sum \Pr(x_i)$ das probabilidades dos valores x_i situados em J. A função cumulante $\Phi(x)$ será, pois, neste caso,

$$\Phi(x) = \sum_{x_i \leq x} \Pr(x_i),$$

sendo fácil ver que uma tal função $\Phi(x)$ apresenta uma descontinuidade de 1ª espécie em cada um dos pontos x_i, com um salto positivo igual a $\Pr(x_i)$:

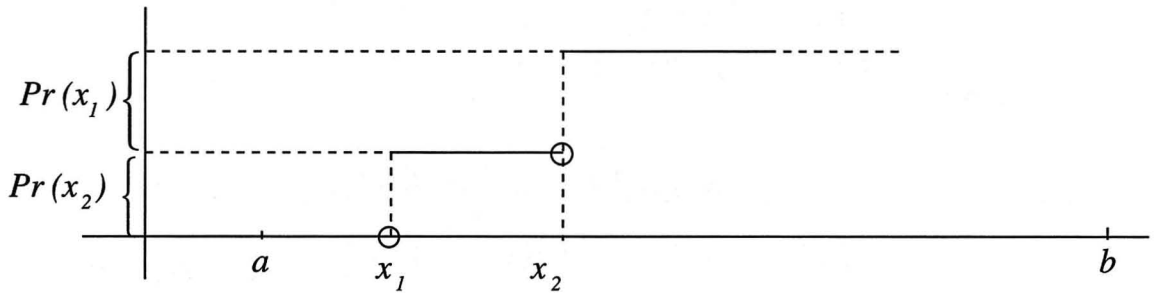


Fig. 2

Fique, pois, assente que o caso das variáveis descontínuas, com um número finito de valores $x_1, x_2, \ldots, x_r$, se pode sempre englobar no caso das variáveis contínuas, desde que se atribua probabilidade nula a todo o valor de x diferente daqueles.

Um outro caso particular importante, mas que já não se reduz ao caso dos universos finitos, é aquele em que *a cumulante $\Phi(x)$ admite derivada contínua no intervalo U*. Procuremos o significado desta derivada.

Comecemos por notar que, sendo neste caso $\Phi(x)$ uma função contínua, não apresenta saltos. *Então, a probabilidade de cada valor de $x \neq a$ é nula,*[1] *tendo-se, pois, sempre*:

$$\Pr(x_1 < x < x_2) = \Pr(x_1 \leq x \leq x_2).$$

---

(1) – Veja-se nota precedente.

Suponhamos que é também $\Pr(a) = 0$. Nestas condições, tem-se, para todo o ponto x_0 de U:

$$\Phi(x_0 + h) - \Phi(x_0) = \begin{cases} \Pr(x_0 \leq x \leq x_0 + h), \text{ para } h > 0 \\ \Pr(x_0 + h \leq x \leq x_0), \text{ para } h < 0 \end{cases}$$

À razão incremental

$$\frac{\Phi(x_0 + h) - \Phi(x_0)}{h}$$

podemos, então, chamar *densidade média de probabilidade no intervalo de extremos* x_0, $x_0 + h$. Por sua vez, à derivada $\Phi'(x_0)$, *limite da razão incremental quando* $h \to 0$ (que existe, por hipótese), será natural chamar *a densidade de probabilidade no ponto* x_0.

Representemos por $\varphi(x)$ a derivada de $\Phi(x)$ no intervalo U. Então, segundo o teorema fundamental do cálculo integral, será

$$\Phi(u) = \Phi(a) + \int_a^u \varphi(x)dx = \int_a^u \varphi(x)dx.$$

É claro que o diferencial $d\Phi = \varphi(x)dx$ (*probabilidade elementar*), representa, a menos de um infinitésimo de ordem superior à de dx, a probabilidade correspondente ao intervalo infinitésimo $[x, x + dx]$.

Será, ainda, evidentemente:

$$\int_a^b \varphi(x)dx = \Phi(b) = 1.$$

Estas considerações tornam-se mais intuitivas, se imaginarmos a função $\Pr(J)$ como indicando uma distribuição de matéria, de massa total 1, sobre o intervalo $[a, b]$: então, $\varphi(x)$ representará a densidade (ou melhor, a massa específica) no ponto variável x.

Exemplos – 1) No caso duma roleta perfeita, tem-se

$$\Phi(u) = Pr(0 \leq \Theta \leq u) = \frac{u}{2\pi}$$

e, portanto, $\varphi(u) = \Phi'(u) = \frac{1}{2\pi}$. A densidade de probabilidade é, portanto, a mesma em todos os pontos. Mas basta que a roleta não esteja bem centrada ou que não seja homogénea, para que a densidade varie de ponto para ponto.

2) Vejamos, agora, um outro exemplo sugestivo que se apresenta na teoria dos seguros. A probabilidade de que uma criança recém-nascida venha a falecer antes duma certa idade x pode considerar-se como função da variável contínua x, definida num intervalo $[0, L]$, em que L representa um majorante da duração possível da vida humana. Supondo que esta função admite derivada contínua, a probabilidade de que a criança venha a viver até uma idade compreendida entre x_1 e x_2 será

$$\Phi(x_2) - \Phi(x_1) = \int_{x_1}^{x_2} \varphi(x)dx;$$

mas, neste caso, a densidade $\varphi(x)$ é função decrescente de x. Por sua vez, a probabilidade de que uma pessoa de idade a venha a viver até uma idade x entre x_1 e x_2 (com $x_1 > a$) será:

$$\frac{\int_{x_1}^{x_2} \varphi(x)dx}{\int_{a}^{L} \varphi(x)dx}$$

probabilidade condicional do acontecimento $x_1 < x < x_2$ *a respeito do acontecimento* $x \geq a$ (substituição do universo $[0, L]$, pelo universo $[a, L]$).

As precedentes considerações generalizam-se imediatamente ao caso dum intervalo não limitado, por exemplo, o intervalo $]-\infty, +\infty[$. Será, então, $U=\mathbf{R}$. Neste caso, se a cumulante $\Phi(u) = \Pr(x \leq u)$ admite derivada contínua $\varphi(u)$ em todos os pontos, ter-se-á, ainda,

$$\Pr(v \leq x \leq u) = \Phi(u) - \Phi(v) = \int_v^u \varphi(x)dx .$$

Como se deve ter, além disso[1],

$$\Pr(x \leq u) = \lim_{v \to -\infty} \Pr(v \leq x \leq u),$$

será

$$\Phi(u) = \int_{-\infty}^u \varphi(x)dx .$$

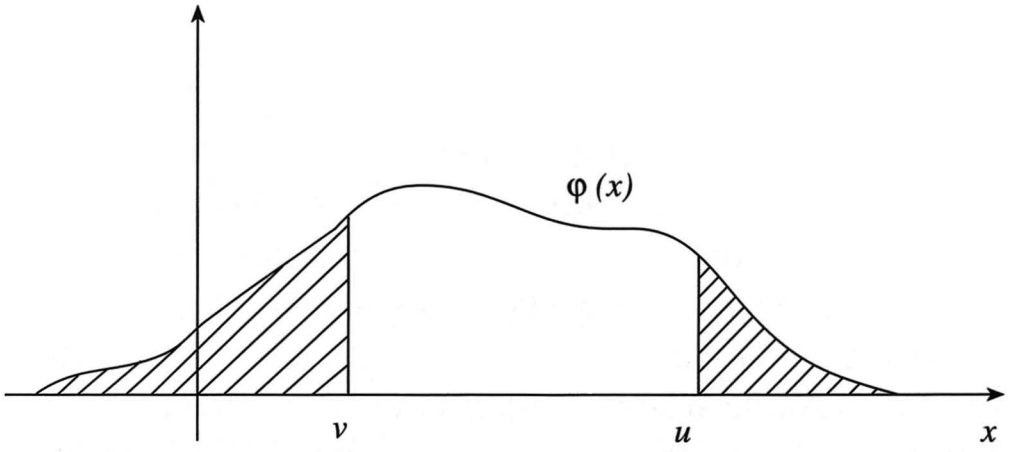


Fig. 3

Se for C a curva representativa da função $\varphi(x)$, o valor de $\Phi(u)$ será a área do domínio limitado por C e pelo eixo dos xx, à esquerda da recta $x = u$. A região a tracejado indica na figura a probabilidade de x estar *fora do intervalo* $[v, u]$, probabilidade esta igual a $1 - \Pr(v \leq x \leq u)$, sendo, por sua vez,

$$\Pr(v \leq x \leq u) = \int_v^u \varphi(x)dx \quad \text{(região a branco)}.$$

---

(1) – Em virtude do axioma IV, enunciado numa nota precedente.

É claro que[2]

$$\int_{-\infty}^{+\infty} \varphi(x)dx = \Pr(U) = 1 \quad \text{(domínio total)}.$$

Note-se, ainda, que toda a distribuição de probabilidade sobre um intervalo $[a, b]$ *se pode conceber como distribuição sobre a recta inteira, considerando como nula a probabilidade correspondente a qualquer intervalo contido no complementar de* $[a, b]$.

NOTA. Dum modo geral, sendo $\Pr(J)$ uma distribuição sobre um intervalo U, ter-se-á, naturalmente:

$$(1) \qquad \Pr(J_1 + J_2 + \cdots + J_n) = \Pr(J_1) + \Pr(J_2) + \cdots + \Pr(J_n)$$

para todo o sistema finito de intervalos $J_1, J_2, \ldots, J_n$, disjuntos dois a dois, *mesmo que estes não sejam contíguos.*

Se representarmos por $\mathcal{H}$ a totalidade dos conjuntos C contidos em U, que são somas de intervalos, em número finito, disjuntos dois a dois, e se incluirmos em $\mathcal{H}$ o conjunto vazio, é fácil ver que:

1) – *A soma lógica (ou produto lógico) de dois quaisquer conjuntos da família* $\mathcal{H}$ *ainda é um conjunto desta família.*

2) – O complementar de qualquer conjunto de $\mathcal{H}$ a respeito de U *ainda é um conjunto de* $\mathcal{H}$ (*por exemplo, o complementar dum intervalo é, geralmente, a soma de dois intervalos disjuntos*).

Exprimiremos este facto dizendo que a família $\mathcal{H}$ é *um corpo de conjuntos* (o que não sucede com a família dos intervalos, pois que a soma de dois intervalos ou o complementar dum intervalo pode não ser um intervalo).

Chamaremos *distribuição em* $\mathcal{H}$ a toda a função real não negativa $\mu(C)$, definida em $\mathcal{H}$, de modo que se tenha:

$$\mu(C_1 + C_2) = \mu(C_1) + \mu(C_2)$$

quando C_1, C_2 são conjuntos disjuntos da família $\mathcal{H}$.

---

(2) – Este facto é, ainda, uma consequência do axioma IV.

Deste modo, a função $\Pr(C)$ será uma distribuição em $\mathcal{H}$ a qual verifica a condição suplementar seguinte: $\Pr(U)=1$ (*distribuição relativa*).

É claro que, para definir uma distribuição em $\mathcal{H}$, basta defini-la na família dos intervalos contidos em U, atendendo à fórmula (1). Mas ao contrário do que sucede com os corpos finitos, não é agora suficiente, em geral, conhecer a distribuição nas células do corpo, que são, neste caso, os pontos de U. Como vimos, pode até acontecer que as probabilidades dos pontos sejam todas nulas sem que o sejam a dos intervalos não nulos: é o que sucede nos casos em que existe em cada ponto uma densidade de probabilidade finita e diferente de 0.

Note-se, ainda, que a toda a função não negativa $\Phi(x)$, definida em U, crescente em sentido lato e contínua à direita, corresponde uma distribuição $\mu(C)$ sobre U de que $\Phi(x)$ é a cumulante, isto é, tal que

$$\Phi(u) = \mu(x \leq u).$$

De resto, além do corpo $\mathcal{H}$, existem outros corpos de conjuntos aos quais se pode prolongar qualquer distribuição definida na família dos intervalos; por exemplo, o corpo dos conjuntos borelianos, citado numa nota precedente.

B – Valores médios para distribuições duma variável real

Comecemos por um exemplo: suponhamos que se trata de calcular a média μ das classificações de todos os alunos num exame. A fórmula a usar será, então,

$$\mu = \frac{\sum x\, \nu(x)}{N},$$

em que N representa o número total dos alunos, x cada uma das classificações 0, 1,..., 20 e $\nu(x)$ o número de alunos que tiveram a classificação x. É claro que será, também,

$$\mu = \sum x \frac{\nu(x)}{N} = \sum x\, \mathrm{fr}(x),$$

em que $\mathrm{fr}(x)$ designa a frequência relativa de x.

Consideremos, agora, em geral, uma qualquer variável numérica x, susceptível dum número finito de valores $x_1, x_2, \ldots, x_r$, com uma dada distribuição de frequência $\mathrm{fr}(x)$. Chama-se *valor médio* da variável x ao número μ dado pela fórmula

$$\mu = \sum_{i=1}^{r} x_i \, \mathrm{fr}(x_i).$$

O valor médio de x também se designa por $M\{x\}$ ou, até, abreviadamente, por $\bar{x}$. Importa salientar que $M\{x\}$ não é propriamente uma *função da variável* x, mas sim uma *função da distribuição* $\mathrm{fr}(x)$[1]. Por isso mesmo se diz, também, (com mais propriedade) que μ é o *valor médio da distribuição*. Ainda com o mesmo significado se usa a expressão *centro da distribuição*, por analogia com o conceito mecânico de centro da gravidade; com efeito, se assimilarmos cada valor x_i de x a um ponto material de abcissa x_i e massa $\mathrm{fr}(x_i)$, o conjunto de tais valores será um sistema material que tem por centro de gravidade o ponto μ.

O conceito do valor médio traduz-se, naturalmente, em termos de probabilidade. Sendo x uma variável numérica de valores $x_1, x_2, \ldots, x_r$, com uma distribuição da probabilidade $\mathrm{Pr}(x)$, chamaremos valor médio de x ao número

$$\mu = \sum_{i=1}^{r} x_i \, \mathrm{Pr}(x_i).$$

Alguns autores usam, neste caso, a expressão *esperança matemática*, em vez de *valor médio*, e a notação $E\{x\}$ em vez de $M\{x\}$.

Por exemplo, se for x a variável casual cujos valores são os números que se obtêm nos lançamentos dum dado, a esperança matemática de x será

$$E\{x\} = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = \frac{21}{6}.$$

É claro que tudo o que dissermos sobre valores médios para distribuições de frequência se aplica, *mutatis mutandis*, a distribuições

---

(1) – É aquilo a que, modernamente, se chama *funcional*.

de probabilidade. Também deve, desde já, ficar assente que todas as variáveis a que faremos agora referência são sempre variáveis *numéricas*, com um número finito ou infinito de valores reais.

Seja x uma variável casual contínua definida num intervalo $[a, b]$, com uma função de probabilidade $\varphi(x)$ (densidade). É natural chamar, neste caso, valor médio ou esperança matemática de x ao número μ dado pela fórmula

$$\mu = \int_a^b x\,\varphi(x)\,dx,$$

em que o somatório foi substituído por um integral.

O valor deste integral é, ainda, designado por $M\{x\}$ (ou por $E\{x\}$).

Exemplo – Sendo $\varphi(x)\,dx$ a probabilidade de um recém-nascido viver até uma idade compreendida entre x e $x+dx$, o valor médio de x (que neste caso se chama *esperança de vida média*) será:

$$M\{x\} = \int_0^L x\,\varphi(x)\,dx,$$

em que L é um majorante da duração possível da vida. Para uma pessoa de idade a, a esperança de vida média será o integral de $x\,\varphi(x)$ entre 0 e L dividido pelo integral de $\varphi(x)$ entre a e L.

Na prática usam-se, apenas, valores aproximados destes integrais, obtidos pela regra do trapézio. Por exemplo, a tábua de mortalidade alemã, já citada, dá, para esperança de vida média dum recém-nascido, o valor 44,9 (anos); para um rapaz de 20 anos, o valor 42,6; para um homem de 50 anos, o valor 19,4, etc.

A anterior definição estende-se ao caso dum intervalo infinito, substituindo o integral próprio por um integral impróprio de 2.ª espécie.

Todas as demonstrações que faremos mais adiante acerca de valores médios pressupõem que a variável toma só um número finito de valores. Mas podem facilmente estender-se ao caso das variáveis casuais contínuas, com uma dada densidade de probabilidade $\varphi(x)$. Basta atender às propriedades dos integrais que generalizam as dos somatórios.

Cálculo prático do valor médio – Na prática, quando é muito grande o número de valores possíveis de x, o cálculo dos valores médios

deve subordinar-se a certas normas. Suponhamos que é dada uma tabela de frequências com intervalos-classes de comprimento h. *Neste caso, obtém-se um valor aproximado de* $M\{x\}$, *com erro inferior a h, multiplicando o ponto médio de cada classe pela frequência relativa dessa classe e somando os resultados obtidos.*

Para facilitar os cálculos, procede-se do modo seguinte:

1) – Toma-se para a unidade o comprimento h das classes, para que os valores das variáveis sejam inteiros.

2) – Escolhe-se, arbitrariamente, um valor c próximo do centro do intervalo em que varia x e calcula-se o valor médio da variável $x - c$. Como se tem, por definição,

$$M\{x - c\} = \sum (x_i - c)\,\mathrm{fr}(x_i),$$

será

$$M\{x - c\} = \sum x_i\,\mathrm{fr}(x_i) - c \sum \mathrm{fr}(x_i) = M\{x\} - c,$$

donde

$$M\{x\} = c + M\{x - c\}.$$

O valor c diz-se *média arbitrária*; o valor $\gamma = M\{x - c\}$ diz-se correcção da *média arbitrária.*

A vantagem do processo está em que os números $x_i - c$ são, geralmente, mais pequenos do que os correspondentes valores x_i.

3) – Exprime-se, finalmente, a média $M\{x\}$ na primitiva unidade.

Veremos adiante um exemplo de aplicação.

Valores médios de funções de x – Seja, ainda, x uma variável susceptível dum número finito de valores $x_1, x_2, \ldots, x_r$, com uma dada distribuição de frequência, $\mathrm{fr}(x)$. Qualquer variável y, que seja função unívoca de x, terá também uma distribuição de frequência que se deduz facilmente da primeira: *a frequência dum dado valor de y será, evidentemente, a soma das frequências dos valores de x a que corresponde esse valor de y.* Seja $y = g(x)$; então, o valor médio de y será dado pela fórmula

$$M\{y\} = \sum g(x_i)\,\mathrm{fr}(x_i).$$

Com efeito, se tivermos, por exemplo, $y_1 = g(x_1) = g(x_2) = \cdots = g(x_{n_1})$, a frequência relativa de y_1 será $\mathrm{fr}^*(y_1) = \mathrm{fr}(x_1) + \mathrm{fr}(x_2) + \cdots + \mathrm{fr}(x_{n_1})$, donde

$$y_1\,\mathrm{fr}^*(y_1) = y_1\,[\mathrm{fr}(x_1) + \mathrm{fr}(x_2) + \cdots + \mathrm{fr}(x_{n_1})]$$

$$= g(x_1)\,\mathrm{fr}(x_1) + \cdots + g(x_{n_1})\,\mathrm{fr}(x_{n_1})$$

e, analogamente, para os outros valores de y, o que justifica a fórmula precedente.

Desta definição resultam, desde logo, as seguintes proposições:

PROPOSIÇÃO 1. *O valor médio da soma de duas variáveis u, v funções de x é igual à soma dos valores médios dessas variáveis:*

$$M\{u + v\} = M\{u\} + M\{v\}.$$

Com efeito, se for $u = g(x)$, $v = h(x)$, será

$$M\{u + v\} = \sum [g(x_i) + h(x_i)]\,\mathrm{fr}(x_i) =$$

$$= \sum g(x_i)\,\mathrm{fr}(x_i) + \sum h(x_i)\,\mathrm{fr}(x_i) = M\{u\} + M\{v\}.$$

PROPOSIÇÃO 2. *O valor médio duma constante* (1) *k é essa mesma constante k.*

Com efeito, tem-se

$$M\{k\} = \sum k\,\mathrm{fr}(x_i) = k \sum \mathrm{fr}(x_i) = k.$$

PROPOSIÇÃO 3. *O valor médio do produto duma constante k por uma função u de x é igual ao produto da constante pelo valor médio da função.*

---

(1) – Isto é, duma função $g(x)$ cujos valores $g(x_1), \ldots, g(x_r)$ sejam todos iguais a k.

Com efeito, se for $u = g(x)$, será

$$M\{ku\} = \sum k\, g(x_i)\, \mathrm{fr}(x_i) = k \sum g(x_i)\, \mathrm{fr}(x_i) = kM\{u\}.$$

As proposições 1 e 3 exprimem-se dizendo que o símbolo M representa um *operador linear*. As três propriedades anteriores combinadas entre si conduzem à proposição mais geral seguinte:

PROPOSIÇÃO 4. *Dadas n variáveis* $y_1, \ldots, y_n$, *funções da variável* x *e* $n + 1$ *constantes* $a_0, a_1, \ldots, a_n$, *tem-se*:

$$M\{a_0 + a_1 y_1 + \cdots + a_n y_n\} = a_0 + a_1 M\{y_1\} + \cdots + a_n M\{y_n\}.$$

Como já se disse previamente, estas considerações generalizam-se, imediatamente, ao caso das distribuições de probabilidade duma variável discreta ou duma variável contínua. Seja, por exemplo, x uma variável casual contínua definida no intervalo $[a, b]$ com uma densidade de probabilidade $\varphi(x)$, e seja $y = f(x)$ uma função de x integrável em $[a, b]$; então, o valor médio de y será dado pela fórmula

$$M\{y\} = \int_a^b f(x)\, \varphi(x)\, dx.$$

Como se tem $\int_a^b \varphi(x)\, dx = 1$, o teorema da média permite-nos afirmar que o valor médio de $f(x)$, ou seja, $M\{y\}$, é um número k compreendido entre os extremos inferior e superior de $f(x)$ em $[a, b]$. As propriedades elementares do integral de RIEMANN habilitam-nos a demonstrar que se tem, ainda,

$$M\{a_0 + a_1 y_1 + \cdots + a_n y_n\} = a_0 + a_1 M\{y_1\} + \cdots + a_n M\{y_n\},$$

sendo $a_0, \ldots, a_n$ constantes, e $y_1, \ldots, y_n$ funções de x.

Momentos – Entre as funções da variável x apresentam-se-nos, desde logo, as potências de expoente inteiro ≥ 0. Chamam-se *momentos* de x os valores médios das funções x^0, x^1, x^2, ..., x^n, ... Põe-se, habitualmente:

$$\mu_n = M\{x^n\} \quad (\textit{momento de ordem } n).$$

É claro que $\mu_0 = M\{1\} = 1$, $\mu_1 = M\{x\} = \mu$.

Em vez de "momentos da variável x" também se diz (e até com mais propriedade) "momentos da distribuição de x".

Dado um número c qualquer, chama-se *desvio de x a respeito* de c à variável $x - c$. Os desvios a respeito da média dizem-se, simplesmente, *desvios* ou *discrepâncias*, sem qualquer outra referência.

Os momentos da variável $x - c$ (*chamados momentos a respeito de c*) são comparáveis aos momentos dum sistema material a respeito dum ponto ou dum eixo. Interessam, especialmente, os *momentos a respeito do centro* (isto é, a respeito do valor médio). São estes:

$$(1)\ M\{x-\mu\} = M\{x\} - M\{\mu\} = \mu - \mu = 0,$$

$$(2)\ M\{(x-\mu)^2\} = M\{x^2 - 2\mu x + \mu^2\} = M\{x^2\} - 2\mu M\{x\} + \mu^2 = \\ = M\{x^2\} - 2\mu^2 + \mu^2 = \mu_2 - \mu^2,$$

$$(3)\ M\{(x-\mu)^3\} = M\{x^3 - 3x^2\mu + 3x\mu^2 - \mu^3\} = \\ = \mu_3 - 3\mu_2\mu + 3\mu\mu^2 - \mu^3 = \mu_3 + 2\mu^3 - 3\mu\mu_2,$$

etc.

É particularmente importante o segundo momento a respeito do centro, análogo ao momento de inércia dum sistema material: dá-se-lhe o nome de *variância* de x (ou da distribuição considerada) e representa-se por $V\{x\}$. Tem-se, pois, por definição,

$$V\{x\} = M\{(x - M\{x\})^2\}.$$

Como já se viu em (2), é

$$V\{x\} = \mu_2 - \mu^2 = M\{x^2\} - (M\{x\})^2,$$

isto é:

PROPOSIÇÃO 5. *A variância duma variável x é igual à diferença entre o valor médio do seu quadrado e o quadrado do seu valor médio.*

Daqui e das proposições 2 e 3 resultam logo, as seguintes consequências:

PROPOSIÇÃO 6. *A variância duma constante é nula.*

PROPOSIÇÃO 7. *A variância do produto de x por uma constante k é igual ao produto k^2 pela variância de x*:

$$V\{kx\} = k^2 V\{x\}.$$

Uma outra propriedade fundamental da variância é a seguinte:

PROPOSIÇÃO 8. *Qualquer que seja a constante a, a variância de $x + a$ é igual à variância de x*:

$$V\{x + a\} = V\{x\}.$$

Com efeito, $V\{x + a\}$ é, por definição, o valor médio do quadrado de $x + a - M\{x + a\}$; mas, como $M\{x + a\} = M\{x\} + a$, tem-se:

$$x + a - M\{x + a\} = x + a - (M\{x\} + a) = x - M\{x\},$$

donde,

$$V\{x + a\} = M\{(x - M\{x\})^2\} = V\{x\}.$$

Cálculo prático da variância – As proposições 5 e 8 são úteis na prática para o cálculo da variância. Com efeito, uma vez escolhida a "média arbitrária" c, tem-se, em virtude daquelas proposições,

$$V\{x\} = V\{x - c\} = M\{(x - c)^2\} - (M\{x - c\})^2,$$

ou seja:

$$V\{x\} = M\{(x - c)^2\} - \gamma^2,$$

pondo $\gamma = M\{x - c\}$ (correcção da média arbitrária).

Basta, portanto, calcular o valor médio de $(x - c)^2$ e subtrair-lhe o quadrado de γ.

Suponhamos, por exemplo, que se trata de achar o centro e a variância da distribuição de frequência dada pela tabela n.° 4 do capítulo anterior. Tomando para média arbitrária c o valor 22,5 (ponto médio do intervalo [22, 23]), podemos dispor os cálculos preliminares no seguinte quadro, em que $\nu(x)$ designa a *frequência absoluta* de x:

x	$\nu(x)$	$x - c$	$(x - c)\,\nu(x)$	$(x - c)^2\,\nu(x)$
14,5	1	– 8	– 8	64
15,5	2	– 7	– 14	98
16,5	2	– 6	– 12	72
17,5	9	– 5	– 45	225
18,5	11	– 4	– 44	176
19,5	20	– 3	– 60	180
20,5	75	– 2	– 150	300
21,5	84	– 1	– 84	84
22,5	95	0	0	0
23,5	60	1	60	60
24,5	20	2	40	80
25,5	14	3	42	126
26,5	3	4	12	48
27,5	2	5	10	50
28,5	1	6	6	36
29,5	1	7	7	49
	$\Sigma\nu(x) = 400$	—	$\Sigma(x-c)\,\nu(x) = -240$	$\Sigma(x-c)^2\,\nu(x) = 1648$

Será, então:

$$\gamma = \frac{\sum (x - c)\,\nu(x)}{N} = -\frac{240}{400} = -0{,}60$$

e, portanto, o valor médio de x será

$$\bar{x} = c + \gamma = 22{,}5 + (-0{,}60) = 21{,}90.$$

Por sua vez, o segundo momento de $x - c$ é

$$M\{(x-c)^2\} = \frac{\sum (x-c)^2 \nu(x)}{N} = \frac{1.648}{400} = 4{,}12\,,$$

donde,

$$V\{x\} = M\{(x-c)^2\} - \gamma^2 = 4{,}12 - 0{,}36 = 3{,}76.$$

Dum modo geral, no cálculo da variância por este método, o erro proveniente de se agruparem os valores de x por classes pode ser reduzido subtraindo ao valor calculado a quantidade $h^2/12$, sendo h o comprimento das classes. Nisto consiste a chamada *correcção de* SHEPPARD.

Uma distribuição diz-se mais ou menos *concentrada*, conforme os valores da variável se acumulam mais ou menos à volta do valor médio. O oposto de concentração é dispersão. Para dar uma ideia do grau de dispersão, poderia utilizar-se a *média dos módulos dos desvios*, isto é, o valor $M\{|x - \mu|\}$ (a média dos desvios não nos diz nada, visto ser nula). Uma medida de dispersão de grande interesse, teórico e prático, é a raiz quadrada da variância: dá-se-lhe o nome de *desvio padrão de x* ("standard deviation", em inglês), também chamado *desvio quadrático médio*, e representa-se por $\sigma\{x\}$, e por σ_x ou, simplesmente, por σ, quando estiver subentendida a variável de que se trata. Ter-se-á, pois, por definição:

$$\sigma_x = \sqrt{V\{x\}}\,.$$

É claro que, assim como σ se pode tomar para índice de dispersão, assim, também, o seu inverso $1/\sigma$ se pode tomar para *índice de concentração*.

O valor máximo de $1/\sigma$ só é atingido, evidentemente, quando x assume um único valor, que será, então, a média μ: diz-se, neste caso, que toda a *massa* de distribuição está concentrada em μ. Neste caso será $V\{x\} = 0$ e, portanto, $1/\sigma = \infty$.

Dá-se o nome de *desvio reduzido de x* ao desvio de x dividido pelo desvio padrão. O desvio reduzido de x será, pois, a variável

$$h = \frac{x - \mu}{\sigma},$$

que *dá a medida do desvio de x, tomando para unidade o desvio padrão.*

É claro que, por sua vez:

$$M\{h\} = \frac{1}{\sigma} M\{x - \mu\} = \frac{1}{\sigma} (M\{x\} - \mu) = 0,$$

$$\sigma\{h\} = \sqrt{V\{h\}} = \sqrt{\frac{1}{\sigma^2} V\{x - \mu\}} = \frac{\sqrt{V\{x\}}}{\sigma} = 1,$$

isto é:

PROPOSIÇÃO 9. *O desvio reduzido duma variável x tem sempre o valor médio igual a* 0 *e o desvio padrão igual a* 1.

Dum modo geral, dada uma distribuição de frequência ou de probabilidade, podemos sempre substituí-la pela distribuição do desvio reduzido, tomando para nova origem dos eixos o valor médio μ e para unidade dos valores da variável o desvio padrão.

Diremos, então, que a distribuição foi *estandardizada ou reduzida.* A estandardização consiste, pois, na mudança de variável

$$x \to h = \frac{x - \mu}{\sigma}.$$

Convém registar a seguinte

PROPOSIÇÃO 10. *Se for* $\Phi(x)$ *a cumulante da distribuição dada, será*

$$\Psi(h) = \Phi(\mu + \sigma h)$$

a cumulante da distribuição estandardizada. Reciprocamente, se for $\Psi(h)$ *a cumulante da distribuição estandardizada, será*

$$\Phi(x) = \Psi\left(\frac{x - \mu}{\sigma}\right)$$

a cumulante da distribuição dada.

Assim, em resumo, o valor médio e o desvio padrão constituem duas características fundamentais duma distribuição: o primeiro indica sumariamente a *posição* ou *localização* da distribuição; o segundo quantifica a *dispersão* dos valores da variável à volta do valor médio.

Teorema de TCHEBICHEFF – O poder representativo do desvio padrão é posto em evidência por um teorema de TCHEBICHEFF, que podemos enunciar do seguinte modo:

Qualquer que seja a distribuição de probabilidade de x, a probabilidade de que o módulo do desvio $x - \mu$ *seja igual ou superior a k vezes o desvio padrão (sendo k um número qualquer) é sempre igual ou inferior a* $1/k^2$. *Isto é, simbolicamente*:

$$\Pr(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2} .$$

Faremos a demonstração apenas para o caso em que x toma um número finito de valores $x_1, x_2, \ldots, x_r$, com probabilidades que designaremos, respectivamente, por $p_1, p_2, \ldots, p_r$. Pondo $\varepsilon_i = x_i - \mu$, tem-se, por definição,

$$\sigma^2 = V\{x\} = p_1\varepsilon_1^2 + p_2\varepsilon_2^2 + \cdots + p_r\varepsilon_r^2 .$$

Sejam $\varepsilon_a, \varepsilon_b, \ldots$ os desvios de módulo inferior a $k\sigma$ e sejam $\varepsilon_m, \varepsilon_n, \ldots$ os desvios de módulo superior ou igual a $k\sigma$. É claro que, substituindo na soma $\Sigma p_i\varepsilon_i^2$ os números $\varepsilon_a, \varepsilon_b, \ldots$ por 0 e os números $\varepsilon_m, \varepsilon_n, \ldots$ por $k\sigma$, se obtém o resultado

$$\sigma^2 k^2 (p_m + p_n + \ldots) \leq \sum p_i\varepsilon_i^2 = \sigma^2 .$$

Mas a soma $p_m + p_n + \ldots$ é a probabilidade dum desvio de módulo igual ou superior a $k\sigma$. Designando essa probabilidade por p, virá

$$\sigma^2 k^2 p \leq \sigma^2 ,$$

ou seja,

$$p \leq \frac{1}{k^2} ,$$

como queríamos demonstrar.

Este teorema costuma também ser apresentado com o seguinte aspecto:

A probabilidade dum desvio de módulo inferior a $k\sigma$ *é igual ou superior a* $1 - 1/k^2$. *Simbolicamente:*

$$\Pr(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Para reconhecer a equivalência dos dois enunciados, basta notar que o acontecimento $|x - \mu| < k\sigma$ é o contrário do acontecimento $|x - \mu| \geq k\sigma$. A sua probabilidade é, pois, complemento para 1 da probabilidade p deste último. Como $p \leq 1/k^2$, será $1 - p \geq 1 - 1/k^2$.

NOTAS IMPORTANTES A RESPEITO DA TERMINOLOGIA E DAS NOTAÇÕES

Dum modo geral, chama-se *parâmetro* duma distribuição de x toda a constante numérica associada a essa distribuição. Assim, serão parâmetros da distribuição os valores médios, não só de x, como de qualquer função de x; e, ainda, as funções desses valores médios (como, por exemplo, o desvio padrão).

Muitas vezes, para estudar a distribuição duma variável (atributo quantitativo) numa determinada população, é-se obrigado a substituir a população por uma amostra, tão representativa quanto possível da população, e a considerar os parâmetros da distribuição dessa variável *na amostra*, como valores aproximados dos parâmetros da distribuição da mesma variável *na população total*. (É o que se faria, por exemplo, para estudar a distribuição da variável "diâmetro do tronco" numa extensa mata de eucaliptos).

Nesta ordem de ideias, é costume designar os parâmetros da distribuição, na amostra, pelas letras latinas correspondentes às letras gregas com as quais se representam os mesmos parâmetros na população considerada: por exemplo, a média por m, o segundo

momento por m_2, o desvio padrão por s, etc., etc. É, ainda, nas amostras que, de preferência, se usa a notação $\bar{x}$ para designar o valor médio de x.

Aos parâmetros da distribuição na amostra daremos, ainda, o nome de *constantes estatísticas* da mesma (em inglês, *statistics*).

Importa, finalmente, observar que a distribuição da variável considerada na população tem, muitas vezes, de ser concebida como distribuição de probabilidade. Sucede isto, primeiro que tudo, quando se trata duma população que esteja constantemente a ser acrescida de novos indivíduos, podendo, assim, considerar-se praticamente infinita (por exemplo, uma espécie, uma raça, uma variedade, etc.). Nestes casos, é corrente atribuir à distribuição de probabilidade uma determinada expressão analítica, com base em considerações de carácter teórico-experimental. A distribuição de frequência relativa da variável na amostra deverá, então, *tender* para a distribuição de probabilidade pressuposta na população quando o número de elementos da amostra aumenta indefinidamente.

Este ponto visto é, ainda, aplicado a populações que, embora circunscritas no tempo e no espaço, sejam muito numerosas.

Todas estas considerações se aplicam, *mutatis mutandis*, ao caso das distribuições de duas ou mais variáveis numéricas, que passamos a estudar.

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

Comecemos por considerar um sistema (x, y) de duas variáveis reais, susceptível dum número finito de valores,

$$(x_i, y_k), \quad i = 1, 2, \ldots, r, \quad k = 1, 2, \ldots, s,$$

e suponhamos dada uma distribuição de frequências, $\mathrm{fr}(x, y)$, sobre o universo destes valores. Nestas condições, qualquer variável z que seja função unívoca de (x, y) terá, também, uma distribuição de frequência que se deduz da primeira do seguinte modo: *a frequência dum dado valor de z será a soma das frequências dos valores de (x, y) a que corresponde esse valor de z.* Seja $z = g(x, y)$; então, é

fácil reconhecer, como para as distribuições duma só variável, que o valor médio de z é dado pela fórmula

$$M\{z\} = \sum_{i,\,k} g(x_i, y_k)\,\mathrm{fr}(x_i, y_k).$$

Entre as possíveis funções de (x, y) aparecem-nos, primeiro que tudo, as próprias variáveis x, y. Será, então,

$$M\{x\} = \sum_{i,\,k} x_i \mathrm{fr}(x_i, y_k) = \sum_i \sum_k x_i \mathrm{fr}(x_i, y_k)$$

$$= \sum_i x_i \sum_k \mathrm{fr}(x_i, y_k) = \sum_i x_i\,\mathrm{fr}(x_i),$$

visto que $\mathrm{fr}(x_i) = \sum_k \mathrm{fr}(x_i, y_k)$ (1ª *frequência marginal*).

Analogamente,

$$M\{y\} = \sum_k y_k \mathrm{fr}(y_k),$$

com $\mathrm{fr}(y_k) = \sum_i \mathrm{fr}(x_i, y_k)$ (2ª *frequência marginal*)[1].

Uma outra função simples de (x, y) é a sua soma $x + y$. A proposição 1 de B, pode agora generalizar-se do seguinte modo:

PROPOSIÇÃO 1. *O valor médio da soma das duas variáveis x, y é igual à soma dos valores médios de x e de y.*

Com efeito, tem-se, por definição:

$$M\{x + y\} = \sum_{i,\,k} (x_i + y_k)\,\mathrm{fr}(x_i, y_k)$$

$$= \sum_{i,\,k} x_i\,\mathrm{fr}(x_i, y_k) + \sum_{i,\,k} y_k \mathrm{fr}(x_i, y_k)$$

$$= M\{x\} + M\{y\},$$

em virtude do que se disse, há pouco, sobre $M\{x\}$ e $M\{y\}$.

---

(1) – As duas funções $\mathrm{fr}(x)$, $\mathrm{fr}(y)$ tomam, geralmente, valores diferentes para valores iguais das variáveis x, y. Seria, por isso, mais correcto designá-las por símbolos diferentes, por exemplo, $\mathrm{fr}_1(x)$, $\mathrm{fr}_2(y)$. Não o fazemos para não sobrecarregar as notações.

Chamam-se *momentos de distribuição* $\mathrm{fr}(x, y)$ os valores médios das funções $x^m y^n$, sendo m, n números inteiros não negativos. Dum modo geral, põe-se

$$\mu_{m,n} = M\{x^m y^n\}.$$

Em particular, $\mu_{1,0} = M\{x\}$, $\mu_{0,1} = M\{y\}$, $\mu_{2,0} = M\{x^2\}$, $\mu_{0,2} = M\{y^2\}$ (primeiros e segundos momentos das variáveis x, y, isoladas). Ao par $(\mu_{1,0}, \mu_{0,1})$ dá-se o nome de *centro da distribuição,* ainda por analogia com o centro da gravidade dum sistema de pontos materiais (x_i, y_k) que tivessem massas iguais a $\mathrm{fr}(x_i, y_k)$, respectivamente.

O primeiro momento misto é $\mu_{1,1} = M\{xy\}$. Algumas vezes, para simplificar as notações, representaremos por $\bar{x}$ o valor médio de x e por $\bar{y}$ o valor médio de y (isto é, pomos $\bar{x} = \mu_{1,0}$, $\bar{y} = \mu_{0,1}$, embora as notações $\bar{x}$, $\bar{y}$ se devam usar, de preferência, para as amostras).

PROPOSIÇÃO 2. *Se as variáveis x, y são independentes (a respeito da distribuição considerada), tem-se*

$$M\{xy\} = M\{x\}\, M\{y\}, \text{ ou seja, } \mu_{1,1} = \mu_{1,0}\, \mu_{0,1}.$$

Com efeito, dizer que x, y são independentes equivale a dizer que $\mathrm{fr}(xy) = \mathrm{fr}(x) \cdot \mathrm{fr}(y)$. Então, será

$$M\{xy\} = \sum_{i,k} x_i y_k \,\mathrm{fr}(x_i, y_k) = \sum_{i,k} x_i y_k \,\mathrm{fr}(x_i)\,\mathrm{fr}(y_k)$$

$$= \sum_i x_i \,\mathrm{fr}(x_i) \cdot \sum_k y_k \,\mathrm{fr}(y_k) = M\{x\}\, M\{y\}.$$

São particularmente importantes os *momentos a respeito do centro,*

$$M\{(x-\bar{x})^m \cdot (y-\bar{y})^n\}, \quad m, n = 0, 1, 2, \ldots$$

Entre estes, destacaremos o valor médio do produto dos desvios $x - \bar{x}$, $y - \bar{y}$. Dá-se-lhe o nome de *covariância* das variáveis x, y e designa-se por $C\{x, y\}$. Portanto:

$$C\{x, y\} = M\{(x-\bar{x})\,(y-\bar{y})\}.$$

Será, então:

$$\begin{aligned} C\{x, y\} &= M\{xy - x\bar{y} - y\bar{x} + \bar{y}\bar{x}\} \\ &= M\{xy\} - \bar{y}M\{x\} - \bar{x}M\{y\} + \bar{x}\bar{y} = M\{xy\} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} \\ &= M\{xy\} - M\{x\}\,M\{y\} = \mu_{1,1} - \mu_{1,0} \cdot \mu_{0,1}, \end{aligned}$$

isto é:

PROPOSIÇÃO 3. *A covariância das variáveis x, y é igual à diferença entre o valor médio do produto dessas variáveis e o produto dos valores médios das mesmas.*

Daqui e da proposição 2 deduz-se, logo, o seguinte

COROLÁRIO. *Se as variáveis x, y são independentes, a sua covariância é nula* (a recíproca, porém, não é verdadeira).

Por sua vez, tem-se

PROPOSIÇÃO 4. *A variância da soma das duas variáveis x, y é igual à soma das respectivas variâncias mais o dobro da sua covariância, isto é*:

$$V\{x + y\} = V\{x\} + V\{y\} + 2C\{x, y\}.$$

Comecemos por notar que o desvio de $x + y$ é

$$x + y - M\{x + y\} = x + y - (\bar{x} + \bar{y}) = (x - \bar{x}) + (y - \bar{y}),$$

o que se pode exprimir dizendo que o *desvio da soma é igual à soma dos desvios das parcelas*. O quadrado do desvio de $x + y$ será, pois,

$$(x - \bar{x})^2 + (y - \bar{y})^2 + 2(x - \bar{x}) \cdot (y - \bar{y}),$$

donde, pela definição de variância e pela proposição 1:

$$\begin{aligned} V\{x + y\} &= M\{(x - \bar{x})^2 + (y - \bar{y})^2 + 2(x - \bar{x}) \cdot (y - \bar{y})\} \\ &= M\{(x - \bar{x})^2\} + M\{(y - \bar{y})^2\} + 2M\{(x - \bar{x}) \cdot (y - \bar{y})\} \\ &= V\{x\} + V\{y\} + 2C\{x, y\}. \end{aligned}$$

Daqui e do corolário anterior vem logo este outro

COROLÁRIO. *Se as variáveis x, y são independentes, tem-se* $V\{x+y\} = V\{x\} + V\{y\}$.

Registem-se, ainda, as seguintes propriedades, cuja demonstração é imediata:

$$C\{x, x\} = V\{x\}, \quad C\{ax, by\} = ab\,C\{x, y\}$$

sendo *a, b* constantes quaisquer.

Será, então, em virtude das propriedades já demonstradas da variância

$$(1) \qquad \boxed{V\{ax + by + c\} = a^2V\{x\} + b^2V\{y\} + 2ab\,C\{x, y\}}$$

em que *a, b, c* são constantes quaisquer.

Correlação e regressão – Já vimos que se tem $C\{x, y\} = 0$ quando *x, y* são independentes. Para avaliar o *grau de dependência* ou *associação* das variáveis *x, y,* usa-se o seguinte índice:

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

chamado *coeficiente de correlação* ou, apenas, *correlação de x* e *y*.

Vamos ver que é sempre

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

e que se tem $\rho = 1$ ou $\rho = -1$, se, e só se, a variável *y* é função linear de *x*, estando, então, os pontos (x_i, y_k) sobre uma recta. Chega-se a esta conclusão mediante as seguintes considerações:

Se as variáveis *x, y,* não são independentes, pode presumir-se a existência duma relação funcional entre elas, isto é, duma *lei natural*, que, em primeira aproximação, se procurará exprimir por meio duma função linear,

$$y = a + bx.$$

É claro que, só num caso excepcional, teórico, se poderá ter $y_k = a + bx_i$, ou seja, $y_k - a - bx_i = 0$, para todos os pontos (x_i, y_k) de *frequência não nula*. Em geral, estes pontos não estão em linha recta. O que se procura, então, é determinar a, b de modo que os desvios $y_k - a - bx_i$ (distâncias verticais dos pontos (x_i, y_k) à recta $y = a + bx$) sejam mínimos; mais precisamente, procura-se tornar mínima a média

$$E = M\{(y - a - bx)^2\} = \sum_{i,k} (y_k - a - bx_i)^2 \, \mathrm{fr}(x_i, y_k),$$

dos quadrados dos referidos desvios (*método dos mínimos quadrados*). Ora, tem-se, desenvolvendo $[(y - bx) - a]^2$ e atendendo à linearidade do operador M:

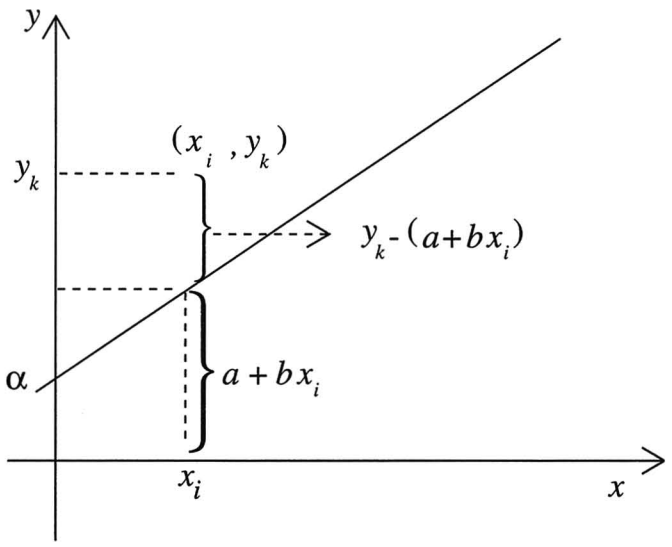


Fig. 4

$$\begin{aligned} E &= M\{(y - bx)^2\} - 2M\{a(y - bx)\} + M\{a^2\} \\ &= M\{y^2\} - 2bM\{xy\} + b^2M\{x^2\} - 2aM\{y\} + 2abM\{x\} + a^2 \\ &= \mu_{0,2} - 2b\mu_{1,1} + b^2\mu_{2,0} - 2a\bar{y} + 2ab\bar{x} + a^2, \end{aligned}$$

onde, para simplificar as notações, pusemos $\bar{x}$ em vez de $\mu_{1,0}$ e $\bar{y}$ em vez de $\mu_{0,1}$. Trata-se, pois, de minimizar a função E de a, b.

Virá, então,

$$\frac{\partial E}{\partial a} = 2a + 2(b\bar{x} - \bar{y}), \quad \frac{\partial E}{\partial b} = 2b\mu_{2,0} + 2(a\bar{x} - \mu_{1,1}),$$

donde, o sistema de equações nas incógnitas a, b:

$$\begin{cases} a + \bar{x}b = \bar{y} \\ \bar{x}a + \mu_{2,0}\, b = \mu_{1,1} \end{cases}$$

que, resolvido, dá o ponto de estacionaridade (a_1, b_1) definido pelas fórmulas:

$$b_1 = \frac{\mu_{1,1} - \bar{x}\bar{y}}{\mu_{2,0} - \bar{x}^2}, \quad a_1 = \bar{y} - b_1 \bar{x}.$$

Notemos, ainda, que

$$\mu_{1,1} - \bar{x}\bar{y} = M\{xy\} - M\{x\}\, M\{y\} = C\{x, y\}$$

e

$$\mu_{2,0} - \bar{x}^2 = M\{x^2\} - (M\{x\})^2 = V\{x\},$$

o que permite escrever b_1 sob a forma

$$b_1 = \frac{C\{x, y\}}{V\{x\}} = \rho \frac{\sigma_y}{\sigma_x},$$

visto que $C\{x, y\} = \rho \sqrt{V\{x\} V\{y\}} = \rho \sigma_x \sigma_y$ (pondo $\sigma_x = \sqrt{V\{x\}}$, $\sigma_y = \sqrt{V\{y\}}$).

É fácil verificar que se trata, efectivamente, dum mínimo (absoluto). A recta pedida será, pois,

$$\boxed{y = \bar{y} + b_1 (x - \bar{x}), \quad \text{com } b_1 = \frac{C\{x, y\}}{V\{x\}} = \rho \frac{\sigma_y}{\sigma_x}}$$

Dá-se-lhe o nome de *recta de regressão de y sobre x*. Analogamente, a recta

$$\boxed{x = \bar{x} + b_2 (y - \bar{y}), \quad \text{com } b_2 = \frac{C\{x, y\}}{V\{y\}} = \rho \frac{\sigma_x}{\sigma_y}}$$

é chamada *recta de regressão de x sobre y*. Como se vê, estas duas rectas passam pelo ponto $(\bar{x}, \bar{y})$, centro da distribuição. Os coeficientes b_1, b_2 dizem-se *coeficientes de regressão*, ou, apenas, *regressões*.

Note-se que, para $a = a_1$, $b = b_1$, vem

$$\begin{aligned} E &= M\{[y - \bar{y} - b_1(x - \bar{x})]^2\} \\ &= M\{(y - \bar{y})^2\} - 2b_1 M\{(x - \bar{x})(y - \bar{y})\} + b_1^2 M\{(x - \bar{x})^2\} \\ &= V\{y\} - 2b_1 C\{x, y\} + b_1^2 V\{x\} = V\{y\} - 2\rho^2 V\{y\} + \rho^2 V\{y\} \\ &= V\{y\}(1 - \rho^2), \end{aligned}$$

visto que $C\{x, y\} = \rho\sigma_x\sigma_y$ e $b_1 = \rho\sigma_y/\sigma_x$. Portanto, o valor mínimo de E será $V\{y\}(1 - \rho^2)$. *Ora, é evidente que este mínimo (valor médio dos quadrados dos desvios verticais a respeito da 1ª recta) só será nulo, se os pontos* (x_i, y_k) *de frequência não nula estiverem sobre aquela recta.* Vê-se, pois, que, como tínhamos afirmado, os referidos pontos *estão em linha recta, se, e só se,* $1 - \rho^2 = 0$, *ou seja,* $\rho^2 = 1$.

Neste caso, as duas rectas de regressão coincidem, pois que será $b_2 = 1/b_1$. As variáveis *x, y* dizem-se, então, *perfeitamente correlacionadas* (*positivamente, se* $\rho = +1$, *negativamente, se* $\rho = -1$). Mas este é um caso ideal, que nunca se verifica exactamente na prática. *Vê-se, entretanto, que, se o valor de* ρ^2 *for bastante próximo de* 1, *as duas rectas se aproximam bastante uma da outra, ajustando-se ambas, com boa aproximação, ao conjunto dos pontos* (x_i, y_k) *considerados* (*regressão linear*).

Casos há, todavia, em que a linearidade está longe de traduzir uma possível relação funcional entre as variáveis *x, y* em questão. Recorre-se, então, a funções mais complicadas – polinómios, exponenciais, etc. – para tentar traduzir a dita relação. Trata-se, portanto, de *ajustar* o mais possível, ao conjunto dos pontos (x_i, y_k), uma curva de dado tipo, usando, por exemplo, o método dos mínimos quadrados. Assim, a *regressão linear* cede o lugar à *regressão curvilínea*, cuja teoria não podemos expor aqui.

O exemplo que damos a seguir, extraído da citada obra de FINNEY[1], refere-se ao estudo da correlação entre a densidade de produção de trigo (variável x) e o teor do trigo em proteína (variável y), sendo os dados relativos a 10 talhões. A densidade de produção é medida em cwt por acre (o cwt, abreviatura de "hundred weight", equivale a 50,802 kg, e o acre equivale, aproximadamente, a 0,40467 ha).

TABELA N.° 1

N.° do talhão	x = Produção em cwt por acre	y = Proteína %
1	14,3	10,8
2	12,8	11,4
3	12,7	13,0
4	10,6	14,6
5	10,7	13,8
6	13,0	12,2
7	14,4	10,7
8	12,5	12,8
9	8,7	16,2
10	12,2	11,8

O coeficiente de correlação é, neste caso:

$$\rho = -\frac{26,33}{\sqrt{27,65 \times 27,52}} = -0,955.$$

Por sua vez, tem-se

$$\bar{x} = 12,19,\ \bar{y} = 12,73,\ b_1 = -\frac{26,33}{27,65} = -0,952,$$

(1) – Importa salientar que, na referida obra, este exemplo é, por sua vez, uma adaptação de resultados expostos pelo Engenheiro Augusto José de Oliveira, num trabalho publicado em "Agronomia Lusitana", vol. 8 (1946), pp. 147–159 (Estação Agronómica Nacional).

donde, a equação de regressão

$$y = 12{,}73 - 0{,}952\ (x - 12{,}19),$$

ou seja,

$$y = 24{,}3 - 0{,}95\ x.$$

Esta recta está representada na figura junta, em que os pontos marcados indicam os pares (x_i, y_k) observados. *É visível que o teor das sementes em proteínas diminui quando a densidade de produção aumenta, seguindo esta variação uma lei sensivelmente linear.*

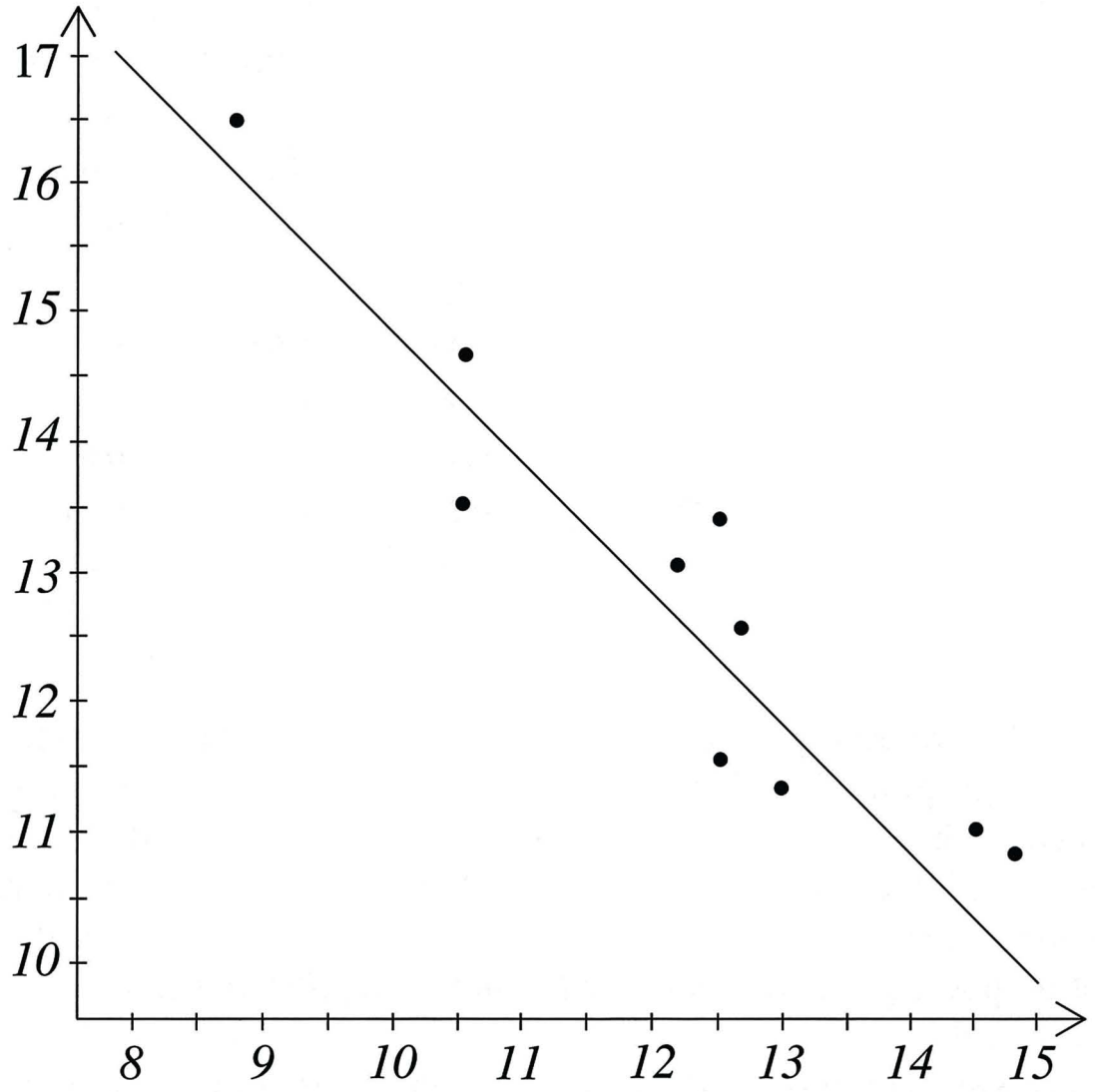


Fig. 5

Para ter uma ideia mais precisa do significado dos valores de ρ, institui-se sobre este índice um teste de significância, que dependerá, naturalmente, do número de pontos (x_i, y_k) observados: é a esse número, diminuído de 2 unidades, que, neste caso, se dá o nome de *número de graus de liberdade.* A *hipótese nula* consiste, agora, em supor ρ = 0; para averiguar em que medida o valor de ρ observado é ou não atribuível ao acaso, recorre-se às tábuas que dão valores de ρ correspondentes a diversos níveis de significância. Da mesma obra extraímos a seguinte tabela:

TABELA N.° 2

Nº de graus de liberdade	Nº de pares	Probabilidade		
		0,1	0,05	0,01
1	3	0.988	0.9969	0.9999
2	4	0.90	0.95	0.99
3	5	0.81	0.88	0.96
4	6	0.73	0.81	0.92
6	8	0.62	0.71	0.83
8	10	0.55	0.63	0.76
10	12	0.50	0.58	0.71
15	17	0.41	0.48	0.61
20	22	0.36	0.42	0.54
30	32	0.30	0.35	0.45
60	62	0.21	0.25	0.32

Não esquecer que a tabela dá valores de $|\rho|$. No caso anterior, o número de graus de liberdade é 8, sendo $|\rho| = 0{,}955$. Ora, este valor excede o limite 0,76 que marca o nível 0,01: quer isto dizer que, sendo válida a hipótese nula, a probabilidade dum ρ igual ou superior a 0,955 (em módulo) é bastante inferior a 1%. O valor de ρ obtido pode, pois, considerar-se *altamente significante contra a hipótese nula.*

Neste exemplo, o número de pares observados é pequeno. Quando esse número for grande, torna-se necessário repartir os valores de *x, y* por intervalos de classe, como foi indicado para as

tábuas de frequências, e organizar uma tábua de contingência que, neste caso (atributos quantitativos) se dirá uma *tábua de correlação*. Do próprio exame da tábua se pode já inferir se os dados tendem ou não a acumular-se à volta duma recta.

Caso das distribuições de probabilidade de duas variáveis reais – É claro que todas as considerações precedentes, relativas a distribuições de frequência de duas variáveis *x, y,* se podem estender, *mutatis mutandis*, ao caso das probabilidades. Poderão, ainda, considerar-se variáveis contínuas em vez de variáveis discretas. Para isso, haverá que estender os conceitos definidos em A) ao caso de duas variáveis contínuas *x, y,* substituindo os intervalos *J* por rectângulos Δ de lados paralelos aos eixos coordenados. Assimilando a distribuição de probabilidade a uma distribuição de massa, chega-se, intuitivamente, ao conceito de *densidade* (*superficial*) *de probabilidade* $\varphi(x, y)$[1]. Então, a probabilidade correspondente a um dado rectângulo Δ será dada pelo integral duplo

$$\Pr(\Delta) = \iint_{\Delta} \varphi(x, y)\, dx\, dy.$$

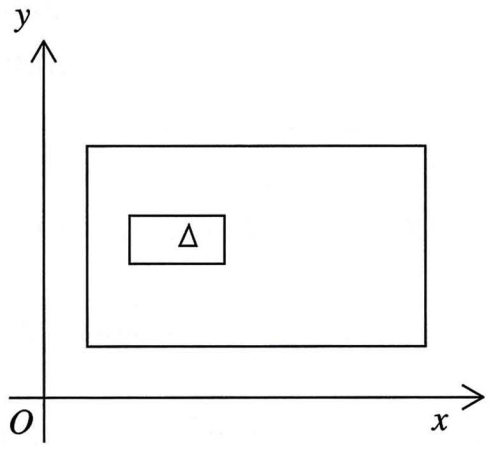


Fig. 6

(1) – Note-se que o conhecido conceito de *densidade de população se refere a uma densidade superficial de frequência absoluta* (número de habitantes por unidade de área). O próprio conceito de massa específica se confunde, na escala atómica, com o de densidade de população (de partículas materiais).

É claro que, se for U o domínio da distribuição, será

$$\Pr(U) = \iint_U \varphi(x, y)\, dx\, dy = 1.$$

Por sua vez, o valor médio duma função $z = f(x, y)$, integrável em U, será

$$M\{z\} = \iint_U f(x, y)\, \varphi(x, y)\, dx\, dy,$$

e, pelo teorema da média, tem-se

$$L_1 \leq M\{z\} \leq L_2,$$

sendo L_1 e L_2 os extremos inferior e superior de $f(x, y)$ em U.

Todas as anteriores proposições se podem, então, generalizar a este caso, atendendo às propriedades elementares do integral duplo.

Distribuição de n variáveis reais – Somos, finalmente, conduzidos, na mesma ordem de ideias, a considerar distribuições de n variáveis $x_1, x_2, \ldots, x_n$, discretas ou contínuas. É claro que o caso das variáveis contínuas obriga a considerar domínios do espaço $\mathbf{R}^n$ e a utilizar integrais múltiplos (duplos, triplos, quádruplos, etc., conforme for $n = 2, 3, 4, \ldots$).

Quanto a valores médios em $\mathbf{R}^n$, a sua teoria é perfeitamente análoga à precedente.

Suponhamos, por exemplo, que o sistema de variáveis $(x_1, x_2, \ldots, x_n)$ só pode tomar um número finito de valores. Então, dada uma distribuição de frequência, $\mathrm{fr}(x_1, x_2, \ldots, x_n)$, destas variáveis, e uma função $y = g(x_1, x_2, \ldots, x_n)$ das mesmas, o valor médio de y, será, por definição,

$$M\{y\} = \sum g(x_1, \ldots, x_n)\, \mathrm{fr}(x_1, \ldots, x_n).$$

Todas as anteriores proposições se generalizam a este caso. Mas bastará registar as seguintes fórmulas:

$$M\{x_1 + x_2 + \cdots + x_n\} = M\{x_1\} + M\{x_2\} + \cdots + M\{x_n\}$$

$$V\{x_1 + x_2 + \cdots + x_n\} = \sum_i V\{x_i\} + 2\sum_{i<k} C\{x_i, x_k\}$$

Em particular, *se* $x_1, \ldots, x_n$ *são independentes*, será $C\{x_i, x_k\} = 0$ para $i \neq k$ e a anterior fórmula simplifica-se:

$$V\{x_1 + x_2 + \cdots + x_n\} = V\{x_1\} + V\{x_2\} + \cdots + V\{x_n\}$$

Não esquecer, ainda, que se tem

$$V\{x_i\} = M\{x_i^2\} - (M\{x_i\})^2,$$
$$C\{x_i, x_k\} = M\{x_i x_k\} - M\{x_i\} M\{x_k\},$$
$$M\{ay\} = a M\{y\}, \quad V\{ay + b\} = a^2 V\{y\},$$

sendo y uma função qualquer de $x_1, \ldots, x_n$, e a, b constantes arbitrárias.

É claro que estas fórmulas se aplicam, igualmente, às distribuições de probabilidade, podendo, nesse caso, substituir-se a notação $M\{x\}$ por $E\{x\}$ (esperança matemática de x).

D – Aplicação à distribuição binomial. Teorema de BERNOULLI

Como se sabe, a distribuição de BERNOULLI

$$\Pr(x) = \binom{n}{x} p^x q^{n-x}$$

dá a probabilidade de que um acontecimento α, de probabilidade p, se realize x vezes em n provas

$$\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n.$$

Para determinar o centro e o desvio padrão desta distribuição, vamos recorrer a um artifício, em que se utilizam as propriedades precedentes, e que consiste no seguinte:

Designemos por $x^{(i)}$ a variável casual assim definida:

$$x^{(i)} \begin{cases} = 1, \text{ se } \alpha \text{ se realiza na prova } \mathcal{P}_i, \\ = 0, \text{ se } \alpha \text{ não se realiza na prova } \mathcal{P}_i, \end{cases}$$

(para $i = 1, 2, \ldots, n$). Então, qualquer que seja i, será p a probabilidade de ser $x^{(i)} = 1$ e será $1 - p$ a probabilidade de ser $x^{(i)} = 0$. Pondo $q = 1 - p$, virá, portanto,

$$M\{x^{(i)}\} = 1 \cdot p + 0 \cdot q = p. \tag{1}$$

O desvio de $x^{(i)}$ será, pois, $x^{(i)} - p$, com os valores $1 - p$, $0 - p$ de probabilidades p, q, respectivamente. Portanto:

$$V\{x^{(i)}\} = M\{(x^{(i)} - p)^2\} = (1-p)^2 p + (-p)^2 q = \tag{2}$$
$$= q^2 p + p^2 q = pq(p+q) = pq.$$

Notemos, agora, que a soma $x^{(1)} + x^{(2)} + \cdots + x^{(n)}$ tem tantas parcelas iguais a 1 quantas as vezes que α se realiza, sendo as restantes parcelas nulas; a soma será, pois, igual ao *número de realizações de* α *nas n provas, ou seja*, x. Tem-se, pois,

$$x = x^{(1)} + x^{(2)} + \cdots + x^{(n)}.$$

Além disso, já sabemos que as variáveis casuais $x^{(i)}$ *são independentes*. Logo, atendendo a (1), vem:

$$M\{x\} = M\{x^{(1)}\} + \cdots + M\{x^{(n)}\} = np,$$

e atendendo a (2):

$$V\{x\} = V\{x^{(1)}\} + \cdots + V\{x^{(n)}\} = npq.$$

Serão, pois,

$$\mu = np, \quad \sigma = \sqrt{npq},$$

o valor médio e o desvio padrão da frequência absoluta x de α.

A frequência relativa de α nas n provas é

$$f = \frac{x}{n}.$$

Aplicando os resultados anteriores, tem-se

$$M\{f\} = \frac{1}{n} M\{x\} = p, \quad V\{f\} = \frac{1}{n^2} V\{x\} = \frac{pq}{n}.$$

O valor médio e o desvio padrão da frequência relativa de α são, pois, respectivamente,

$$p \quad \text{e} \quad \sqrt{\frac{pq}{n}}.$$

Diz-se que uma sucessão

$$u_1, u_2, \ldots, u_n, \ldots$$

de números reais *converge estocasticamente* (ou *converge em probabilidade*) para um número real a, quando, dado um número $\delta > 0$, por menor que ele seja, a probabilidade de

$$|u_n - a| < \delta$$

tende para 1 quando n tende para ∞.

Podemos, agora, enunciar o

TEOREMA DE BERNOULLI. *Seja α um acontecimento de probabilidade p. A frequência relativa de α em n provas converge estocasticamente para p, quando $n \to \infty$.*

Demonstração. Designemos, agora, mais precisamente por f_n a frequência relativa de α nas n provas. Como se viu atrás, tem-se $M\{f_n\} = p$, $\sigma\{f_n\} = \sqrt{pq/n}$. Então, segundo o teorema de Tchebicheff atrás demonstrado, a probabilidade P_n de que se terá

$$|f_n - p| < \delta$$

satisfaz à condição

$$P_n \geq 1 - \frac{1}{k^2}, \text{ em que } k = \frac{\delta}{\sigma} = \delta \sqrt{\frac{n}{pq}}.$$

Ora, quando $n \to \infty$, também $k \to \infty$ e, portanto, $1 - 1/k^2$ tende para 1. Como se tem, por outro lado, $P_n \leq 1$, virá pois,

$$\lim_{n \to \infty} P_n = 1,$$

o que, segundo a definição anterior, é a tese do teorema.

Este teorema vem lançar nova luz sobre as relações entre os conceitos de frequência relativa e de probabilidade. Como vemos, a *frequência relativa, f_n, converge estocasticamente para a probabilidade p, quando $n \to \infty$*. Daqui resulta que, dado um número positivo δ, tão pequeno quanto quisermos, existe sempre uma ordem a partir da qual é *praticamente certo que* $f_n - p < \delta$. É praticamente certo, mas não certo! *Não será, portanto, lícito escrever*

$$\lim_{n \to \infty} f_n = p,$$

conforme o conceito de limite da análise matemática, embora, na prática, as coisas se passem, de certo modo, como tal. A variável f_n converge para p de *maneira casual, irregular*, não se podendo, em absoluto, negar a existência de desvios grandes para valores de n elevados.

A probabilidade P_n que intervem na demonstração diz-se de *segunda ordem* a respeito de p. É nas probabilidades de segunda ordem que se baseia a técnica dos chamados *resseguros*, usada entre companhias de seguros, para garantir um maior grau de segurança.

E – Distribuição normal

Uma grande parte das distribuições que se encontram na prática aproxima-se mais ou menos duma distribuição que, em cada ponto x do intervalo $]-\infty, +\infty[$, tem uma densidade $\phi(x)$ tal que

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

sendo μ, σ constantes.

A uma tal distribuição dá-se o nome de *distribuição normal, de* GAUSS *ou* LAPLACE-GAUSS, *de parâmetros* μ, σ; ou, abreviadamente, distribuição (N; μ, σ).

Demonstra-se que é:

(1) $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 1,$$

(2) $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \mu,$$

(3) $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\mu)^2\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \sigma^2.$$

Para a demonstração, veja-se, por exemplo, CASTELNUOVO, obra citada, pp. 253-255 (feita a mudança de variáveis $h=(x-\mu)/\sigma$).

A segunda fórmula diz-nos que o centro da distribuição é μ. A terceira diz-nos que a variância da distribuição é σ^2, sendo, portanto, σ o desvio padrão. Fica, assim, justificado o uso das letras μ, σ, como parâmetros da distribuição. Convém registar, desde já, este facto: *a distribuição normal é completamente determinada pelo centro e pelo desvio padrão.*

Substituindo x pelo desvio reduzido

$$h = \frac{x - \mu}{\sigma},$$

obtém-se a *distribuição normal estandardizada ou distribuição* (N; 0,1), cuja função de densidade é

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

(É claro que podemos passar a usar aqui x no papel de h).

Interpretemos esta distribuição como distribuição de probabilidade. A probabilidade dum desvio reduzido compreendido entre dois limites, λ_1, λ_2, será dada pelo integral

$$\frac{1}{\sqrt{2\pi}} \int_{\lambda_1}^{\lambda_2} e^{-\frac{x^2}{2}} dx.$$

Se pusermos

$$\Phi(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-\frac{x^2}{2}} dx,$$

será $\Phi(\lambda)$ a cumulante da distribuição e o valor do anterior integral será $\Phi(\lambda_2) - \Phi(\lambda_1)$. Note-se que existem tabelas com os valores de $\Phi(\lambda)$ para diferentes valores de λ.

Estudemos a função

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

Comecemos por notar que esta função é definida e positiva em todo o intervalo $]-\infty, +\infty[$ e que se tem $\varphi(-x) = \varphi(x)$ (curva simétrica a respeito do eixo dos yy).

Por outro lado, tem-se

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

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

Como se tem sempre $e^{-x^2/2} > 0$, a função $\varphi'(x)$ é nula para $x = 0$, positiva para $x < 0$ e negativa para $x > 0$. Por sua vez, a função $\varphi''(x)$ tem o sinal de $x^2 - 1$, sendo, portanto, negativa para $-1 < x < 1$, positiva para $x < -1$ ou $x > 1$ e nula para $x = \pm 1$.

Temos, então, os seguintes esquemas:

	$-\infty$		0		$+\infty$
$\varphi'(x)$		$+$		$-$	
$\varphi(x)$		↗		↘	

Máximo

	$-\infty$		-1		$+1$		$+\infty$
$\varphi''(x)$		$+$		$-$		$+$	
$\varphi(x)$		$\cup$		$\cap$		$\cup$	

inflexão inflexão

Note-se, finalmente, que

$$\lim_{x \to \infty} \varphi(x) = 0,$$

o que significa que o eixo dos xx é assimptota do gráfico de φ. É fácil ver que não há outras assimptotas.

A curva representativa da função $\varphi(x)$ considerada, terá, pois, o aspecto indicado na figura.

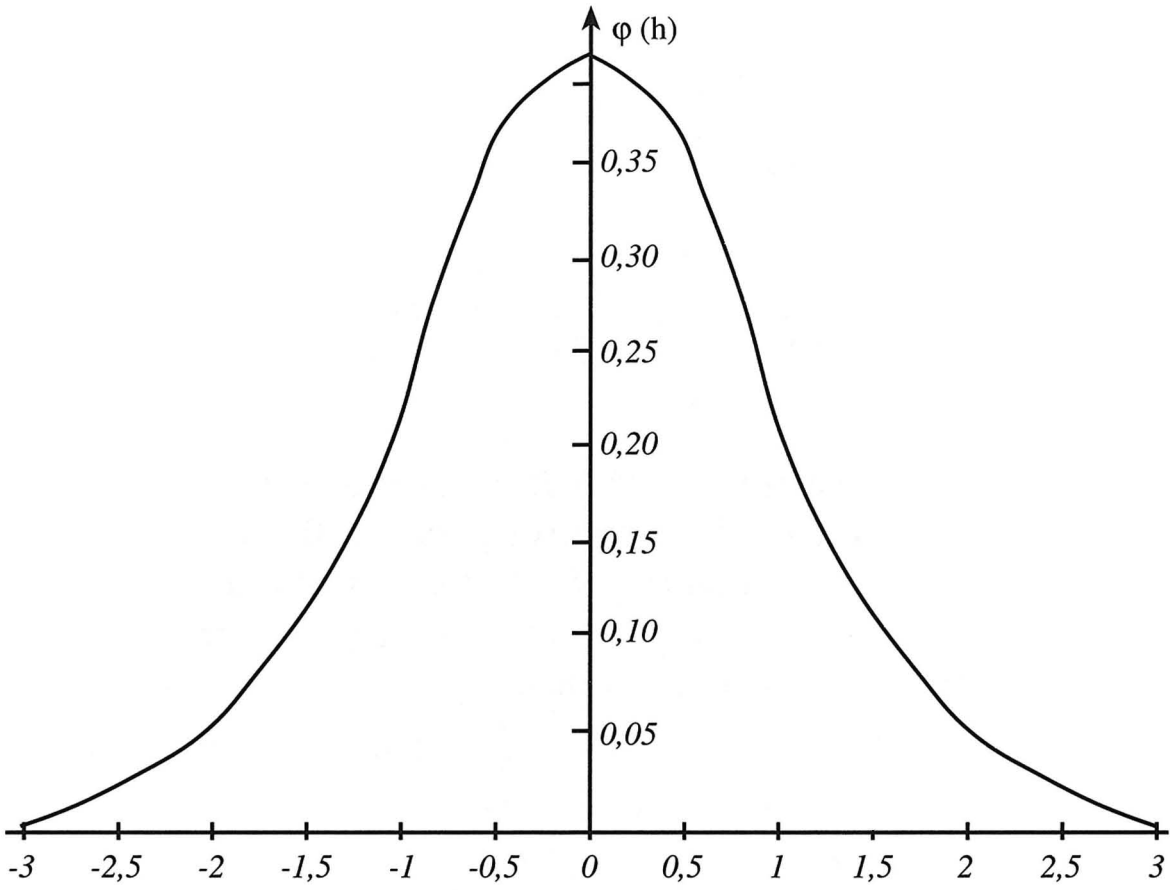


Fig. 7

Dá-se-lhe o nome de *curva de* GAUSS ou *curva em sino*.

Note-se que os pontos de inflexão têm as abcissas –1, 1 correspondentes ao desvio padrão (unitário neste caso).

Daqui se deduz, logo, que o gráfico da função mais geral

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

acusará um máximo no ponto da abcissa μ, será simétrico a respeito da recta $x = \mu$, terá inflexões nos pontos $\mu - \sigma$, $\mu + \sigma$, e admitirá o eixo dos xx como assimptota.

Por sua vez, a função $\Phi(x)$, cumulante da distribuição normal estandardizada, será crescente em todo o intervalo $]-\infty, +\infty[$, o seu

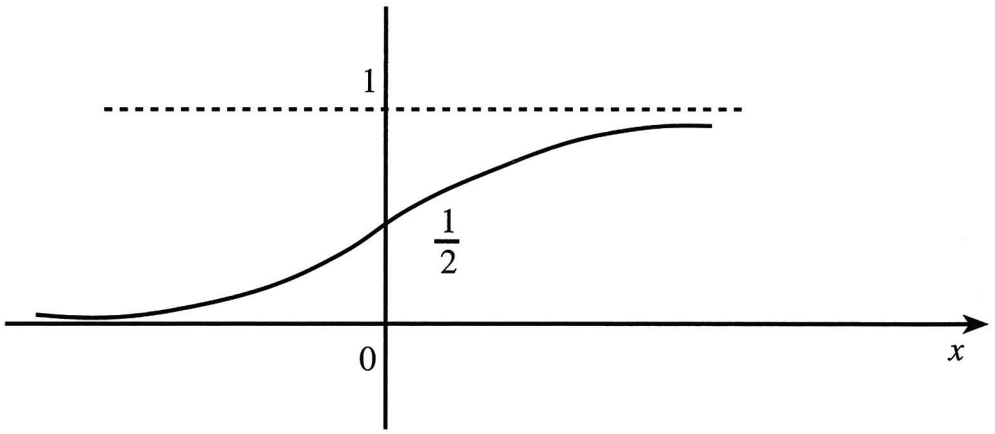


Fig. 8

gráfico terá uma inflexão no ponto de abcissa 0, a respeito do qual é simétrico, e admitirá como assimptotas as rectas $y = 0$, $y = 1$, visto que $\lim_{x \to -\infty} \Phi(x) = 0$, $\lim_{x \to +\infty} \Phi(x) = 1$[1].

Costuma, ainda, escrever-se

$$\Theta(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{x^2}{2}} dx.$$

A função $\Theta(u)$ também se encontra tabelada. É claro que para cada valor γ de u, $\Theta(\gamma)$ é a área do trapezóide determinado pela curva de GAUSS no intervalo $[0 , \lambda]$, igual à área do trapezóide correspondente ao intervalo $[0, -\lambda]$, visto a curva ser simétrica a respeito do eixo dos *yy*. Deste modo, a probabilidade dum desvio reduzido h compreendido entre $-\lambda$ e λ será

$$\Pr(-\lambda < h < \lambda) = 2\Theta(\lambda)$$

e a probabilidade dum desvio reduzido superior a λ, em valor absoluto, será $1 - 2\Theta(\lambda)$.

Por exemplo, a tabela n.° 3 reproduzida em YULE and KENDALL (loc. cit., pág. 533), dá os valores de $1 - 2\Theta(\lambda)$ a menos de 0,00001,

---

(1) – É claro que o facto de as rectas $y = 0$, $y = 1$ serem assimptotas do gráfico se verifica para a cumulante de *qualquer* distribuição definida no intervalo $]-\infty, +\infty[$.

com primeiras e segundas diferenças. Para $\lambda = 3$, o valor registado é 0,00270; quer isto dizer que:

Em qualquer distribuição normal, a probabilidade dum desvio de módulo superior ao triplo do desvio padrão é um pouco inferior a 3%.

(O teorema de TCHEBICHEFF, que, como vimos, é aplicável a qualquer distribuição, dá para um tal desvio uma probabilidade inferior ou igual a $1/3^2 \approx 11\%$).

Para $\lambda = 4$, a mesma tabela dá o valor 0,00006 e, para $\lambda = 4{,}5$, o valor de 0,00001.

Exemplos – Numerosos são os exemplos concretos de variáveis casuais que seguem aproximadamente a lei normal. Assim, é-se geralmente induzido a considerar como *normalmente distribuídas* as variáveis:

altura, numa espécie, raça ou variedade de animais ou plantas (dentro de certos limites de idade);

diâmetro do tronco, numa conveniente população de árvores;

comprimento duma espiga, teor em proteínas, produtividade, etc., numa dada variedade de trigo;

teores em gordura, em proteínas e em hidratos de carbono do leite produzido por vacas duma determinada raça;

etc., etc.

Note-se que os agrupamentos biológicos, tais como as espécies, as raças e as variedades, constituem populações praticamente infinitas, que vêm do passado e se prolongam no futuro, com uma certa constância de caracteres. (Muitas vezes, em vez de populações praticamente infinitas como estas, consideram-se população que, embora numerosas e homogéneas, estão bem delimitadas no tempo e no espaço: por exemplo, uma mata constituída por indivíduos duma mesma variedade). Entre as constantes biométricas duma tal população, figuram, precisamente, os valores médios e os desvios padrões de atributos quantitativos como os precedentes. À semelhança do que sucede para as grandezas físicas, idealizam-se para esses parâmetros valores exactos, *dos quais se calculam valores aproximados em amostras casuais extraídas da população: a aproximação*

obtida considera-se tanto melhor quanto mais representativa, e portanto, mais numerosa for a amostra[1]. (Convém recordar aqui as considerações que fizemos a propósito do conceito de probabilidade).

Nessas constantes biométricas e nos seus valores aproximados baseiam-se novos testes de significância, relativos, por exemplo, à comparação de duas variedades de trigo do ponto de vista de produtividade, ao efeito dum adubo ou dum insecticida, etc., etc. Trata-se, como se pode imaginar, de assuntos de maior interesse para o agrónomo e para o silvicultor, mas não é possível desenvolver neste curso.

Erros de observação – Invocámos há pouco o exemplo das grandezas físicas. Como se sabe, efectuando várias medições duma mesma grandeza, os resultados não concordam geralmente entre si. Admitida a existência duma medida exacta da grandeza, as diferenças $x - x_0$ entre os valores obtidos, x, e o valor exacto, x_0, chamam-se *erros de observação*. Estes classificam-se em *sistemáticos* (devidos a uma causa bem determinada, que pode ser um defeito do instrumento de medida ou do observador) e *acidentais, fortuitos* ou *casuais* (dependentes do "acaso"). Suponhamos eliminados os erros sistemáticos. Então, admitidos certos axiomas, demonstra-se que os resultados de medição duma grandeza se distribuem segundo a lei normal, com centro no valor exacto da grandeza. Esta conclusão é confirmada pela experiência, de maneira satisfatória.

É claro que, sendo assim, os erros de observação também se distribuem normalmente, sendo 0 o seu valor médio. O desvio padrão σ chama-se, agora, *erro padrão* ou *erro quadrático médio*. Dá-se o nome de *parâmetro de precisão* ao número $1/(\sigma\sqrt{2})$, que permite avaliar o grau de concentração da distribuição considerada.

É claro que, na prática, se trabalha apenas com um número finito de valores aproximados,

$$x_1, x_2, \ldots, x_N,$$

da grandeza medida, alguns dos quais podem aparecer repetidos.

---

(1) – A teoria da avaliação dos parâmetros duma distribuição, a partir de amostra da população, constitui um dos mais importantes capítulos da Estatística.

Estes valores constituem, por assim dizer, uma amostra dos possíveis resultados x da medição. Os parâmetros da distribuição de x na amostra, nomeadamente, o valor médio,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N},$$

e o desvio padrão,

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N}},$$

constituem valores aproximados, respectivamente, do valor exacto x_0 da grandeza, e do desvio padrão σ da distribuição normal considerada.

Propriedade reprodutiva da distribuição normal – Uma importante propriedade das distribuições normais é a seguinte, chamada PROPRIEDADE REPRODUTIVA ou TEOREMA DA ESTABILIDADE, que não demonstraremos:

Dadas n variáveis casuais $x_1, x_2, \ldots, x_n$, *independentes e normalmente distribuídas, toda a função linear*

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

daquelas variáveis segue ainda a lei normal.

As proposições estabelecidas em C permitem-nos determinar o valor médio e o desvio padrão da variável y a que se refere este enunciado, em função dos valores médios e dos desvios padrões de $x_1, x_2, \ldots, x_n$. Deverá ter-se, com efeito,

$$M\{y\} = a_0 + a_1 M\{x_1\} + \cdots + a_n M\{x_n\},$$

$$V\{y\} = a_1^2 V\{x_1\} + \cdots + a_n^2 V\{x_n\},$$

donde, designando por σ o desvio padrão de y e por $\sigma_1, \sigma_2, \ldots, \sigma_n$, os desvios padrões de $x_1, x_2, \ldots, x_n$:

$$\sigma = \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2}\,.$$

Como exemplo de aplicação, consideremos o caso duma amostra casual constituída por N valores independentes, $x_1, x_2, \ldots, x_N$, duma mesma variável casual x, normalmente distribuída, com o valor médio μ e o desvio padrão σ. É claro que a média daqueles valores

$$\bar{x} = \frac{\sum x_i}{N} = \frac{1}{N}x_1 + \frac{1}{N}x_2 + \cdots + \frac{1}{N}x_N$$

é uma função linear das variáveis casuais $x_1, x_2, \ldots, x_N$, *independentes e normalmente distribuídas com os parâmetros* μ, σ. Então, segundo o teorema anterior, a média $\bar{x}$ será, também, uma variável normalmente distribuída, tendo por valor médio

$$M\{\bar{x}\} = \frac{1}{N}\sum M\{x_i\} = \frac{N\mu}{N} = \mu\,,$$

isto é,

$$M\{\bar{x}\} = M\{x\}\,,$$

e por desvio padrão

$$\sigma\{\bar{x}\} = \sqrt{\sum \frac{1}{N^2}\sigma^2} = \sqrt{\frac{N\sigma^2}{N^2}} = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}\,,$$

isto é,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}\,.$$

O desvio padrão da média $\bar{x}$ obtém-se, pois, a partir do desvio padrão de x, dividindo este por $\sqrt{N}$.

Estes resultados são de grande importância, não só em estatística agronómica como, ainda, na teoria dos erros.

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

Consideremos uma sucessão infinita de distribuições sobre a recta, cujas funções cumulantes sejam

$$\Phi_1(x), \Phi_2(x), \ldots, \Phi_n(x), \ldots .$$

Diz-se que esta sucessão de distribuições *converge para* uma determinada distribuição, cuja cumulante seja $\psi(x)$, se a sucessão de funções $\Phi_n(x)$ converge uniformemente para $\psi(x)$ em todo o intervalo limitado J da recta.

Consideremos, em particular, o caso da distribuição binomial de x, que dá a probabilidade de que um acontecimento de probabilidade p se verifique x vezes em n provas. Supondo a probabilidade p fixa, e atribuindo a n os valores 1, 2, 3, ..., obtém-se uma sucessão de distribuições cujo termo geral é

$$\Pr_n(x) = \binom{n}{x} p^x q^{n-x}, \quad \text{com } q = 1 - p.$$

É claro que cada uma destas distribuições é definida apenas para os valores inteiros de x tais que $0 \leq x \leq n$; mas podemos supô-la prolongada a toda a recta, atribuindo a probabilidade 0 a cada valor de x que não seja inteiro ou não verifique a condição $0 \leq x \leq n$. *A cumulante da distribuição* $\Pr_n(x)$ *acusará, então, saltos positivos nos pontos*

$$0, 1, 2, \ldots, n-1, n,$$

sendo constante no interior dos intervalos determinados por estes pontos: em particular, será nula no intervalo $]-\infty, 0[$ *e igual a* 1 *no intervalo* $[n, +\infty[$.

O valor médio e o desvio padrão de $\Pr_n(x)$ são, respectivamente, como vimos,

$$\mu_n = np, \quad \sigma_n = \sqrt{npq}\,.$$

Efectuando em $\mathrm{Pr}_n(x)$ a mudança de variável

$$x \to h = \frac{x - \mu_n}{\sigma_n} = \frac{x - np}{\sqrt{npq}},$$

obtém-se a *distribuição binomial estandardizada*. Pois bem, demonstra-se o seguinte teorema, de importância fundamental em Cálculo das Probabilidades e Estatística:

TEOREMA. *A distribuição binomial estandardizada converge para a distribuição normal estandardizada quando* $n \to \infty$.

Para a demonstração, deveras delicada, pode ver-se, por exemplo, a obra de CRAMÉR já citada.

Na prática, o anterior resultado interpreta-se do seguinte modo:

Para valores de n elevados, a distribuição binomial aproxima-se da normal. A aproximação aumenta quando n cresce, mas diminui quando o desvio reduzido aumenta.

G – A distribuição de χ^2 de PEARSON

Vimos que toda a função linear de variáveis independentes normalmente distribuídas também é normalmente distribuída. Porém, uma função *não linear* de variáveis independentes normalmente distribuídas não tem, geralmente, distribuição normal. Com efeito, demonstra-se o seguinte teorema:

Dadas n variáveis casuais $x_1, x_2, \ldots, x_n$, *independentes e com distribuição normal estandardizada, a raiz quadrada da soma dos seus quadrados, que costuma ser designada por* χ:

$$\chi = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

tem uma distribuição cuja densidade é uma função de χ *da forma*

$$\varphi(\chi) = C\ e^{-\frac{\chi^2}{2}}\ \chi^{n-1},$$

sendo C uma constante.

Para a demonstração, veja-se, por exemplo, P. de VARENNES E MENDONÇA, obra citada.

Note-se que, sendo a raiz aritmética de Σx_i^2, o valor de χ será sempre não negativo: o domínio de distribuição é, pois, o intervalo $[0, +\infty[$. Deverá ter-se, então,

$$\int_0^{+\infty} \varphi(\chi)\, d\chi = 1,$$

o que permite determinar a constante C:

$$C = 1 : \int_0^{+\infty} e^{-\frac{\chi^2}{2}}\, \chi^{n-1}\, d\chi .$$

Deste modo, a probabilidade de que χ^2 seja superior a um dado valor χ_0^2 (igual à probabilidade de que seja $\chi \geq \chi_0$), será dada pela fórmula

$$\Pr(\chi^2 \geq \chi_0^2) = C \int_{\chi_0}^{+\infty} e^{-\frac{\chi^2}{2}}\, \chi^{n-1}\, d\chi ,$$

com o valor de C dado pela fórmula anterior.

O que interessa, na prática, não é propriamente a distribuição de χ, mas sim a distribuição de χ^2, que, evidentemente, se reduz imediatamente à anterior como mostra a última fórmula.

Como se vê, a anterior distribuição de χ^2 (chamada distribuição de PEARSON) tem como único parâmetro a variável n, que recebe aqui o nome de *número de graus de liberdade* da distribuição. Para indicar a distribuição de χ^2 com n graus de liberdade, usa-se a notação $(\chi^2; n)$.

O teorema atrás enunciado admite a seguinte generalização:

TEOREMA. *Dadas M variáveis* $x_1, x_2, \ldots, x_M$, *todas com distribuição normal estandardizada, sendo n dessas variáveis independentes (estocasticamente) e as restantes funções lineares homogéneas das primeiras, a variável*

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

tem a distribuição de PEARSON com n graus de liberdade.

Para a demonstração, veja-se o trabalho do Prof. P. de VARENNES E MENDONÇA, "Das distribuições estatísticas mais usadas em provas de significação", publicado na Revista Agronómica XXVIII (1940).

Este importante resultado permite ver como a distribuição considerada pode ser usada em testes de significância (ou *provas de significação*) aplicáveis a tábuas de contingência, ajustamento de curvas ou superfícies, etc.

Consideremos, por exemplo, uma partição em M atributos $\alpha_1, \alpha_2, \ldots, \alpha_M$, num universo infinito U, e seja p_i a probabilidade de que um indivíduo escolhido ao acaso tenha o atributo α_i ($i = 1, 2, \ldots, M$). Consideremos, por outro lado, uma amostra casual constituída por N indivíduos. Será, então, sensivelmente,

$$p_i N$$

o valor *esperado* da frequência absoluta do atributo α_i na amostra (supondo N bastante grande). Designemos por m_i este valor e por o_i o valor observado (da referida frequência absoluta). A discrepância

$$\delta_i = o_i - m_i$$

terá, então, o valor médio $M\{\delta_i\} = 0$. Além disso, *supondo N bastante grande e nenhum dos* p_i *demasiado pequeno*, demonstra-se que *cada uma das variáveis* δ_i *se distribui normalmente em torno de* 0, *com um desvio padrão* σ_i *tal que*

$$\sigma_i^2 = m_i .$$

Então, supondo que n dos desvios δ_i são independentes estocasticamente, sendo os restantes função linear homogénea dos primeiros (como sucede nas tábuas de contingência), o anterior teorema habilita-nos a afirmar que a variável

$$\chi^2 = \sum \frac{\delta_i^2}{m_i} = \sum \left(\frac{\delta_i}{\sigma_i} \right)^2$$

tem aproximadamente a distribuição de PEARSON para n graus de liberdade, visto que cada uma das variáveis δ_i/σ_i tem valor médio nulo e desvio padrão unitário.

E assim ficam esboçados os fundamentos teóricos do teste do χ^2. A falta de tempo impede-nos de aprofundar este assunto e de abordar o estudo de outras distribuições de grande interesse em Estatística agronómica. Para complementos, lembramos as obras de Estatística já citadas, bem como o referido trabalho do Prof. Varennes e Mendonça.

A obra de R. A. FISHER e F. YATES, "*Statistical Tables for Biological, Agricultural and Medical Research*" (Oliver and Boyd, Edinburgh and London) contém tábuas referentes a várias distribuições importantes; a tábua IV desta obra fornece, para diferentes valores de n e de p, o valor de χ_0^2 tal que $\Pr(\chi^2 \geq \chi_0^2) = p$, na distribuição de χ^2 de PEARSON com n graus de liberdade.

NOTA SOBRE A AVALIAÇÃO DA VARIÂNCIA

Embora não possamos aqui abordar o estudo da teoria da avaliação, convém, desde já, esclarecer um ponto.

É fácil ver que o valor esperado da variância, S^2, numa amostra com N elementos, está relacionada com a variância, σ^2, na população, por meio da fórmula

$$M\{S^2\} = \frac{N-1}{N}\sigma^2.$$

Daqui, o considerar o valor $(N/N-1)S^2$ como *a melhor avaliação da variância* σ^2, a partir da amostra. Este valor é representado por $\hat{\sigma}^2$. Será, portanto,

$$\hat{\sigma} = S\sqrt{\frac{N}{N-1}}$$

a melhor avaliação do desvio padrão, σ.

Dum modo geral, o acento circunflexo sobre um símbolo é usado para indicar a melhor avaliação do parâmetro designado por esse símbolo.

Na fórmula anterior, o coeficiente $N/N-1$ é chamado *correcção de* BESSEL. É claro que, para valores de N elevados, aquele coeficiente é sensivelmente igual a 1; a correcção só é, portanto, necessária para pequenas amostras.

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

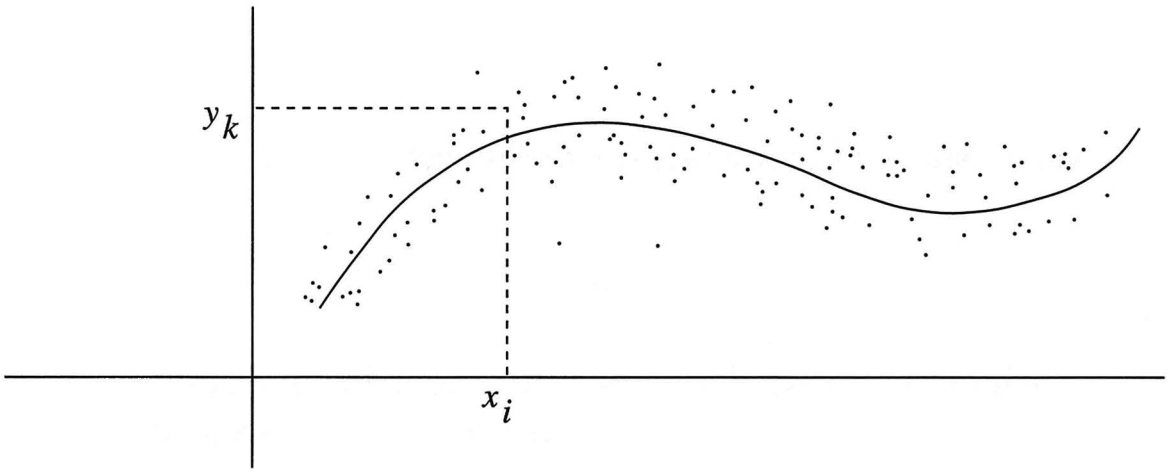


Fig. 1

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*.

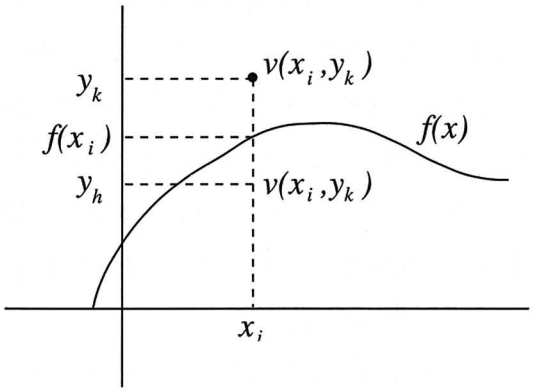


Fig. 2

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

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

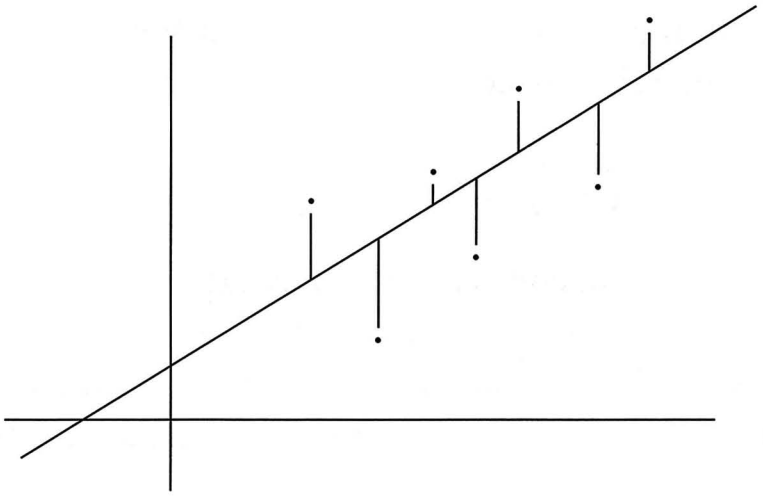
Para evitar confusões, o coeficiente de regressão de y sobre x passará a ser designado por b_{yx}, tendo-se, como vimos,

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

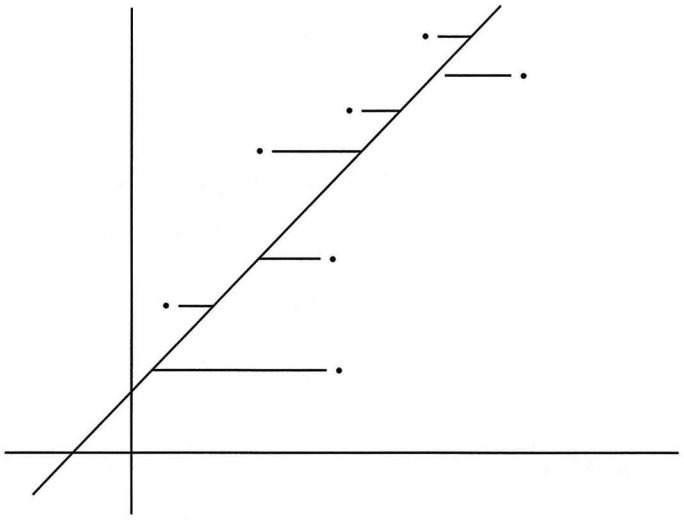
Mostram estas fórmulas que, se for $\rho \geq 0$, será também $b_{yx} \geq 0$ e $b_{xy} \geq 0$: y cresce com x e x cresce com y (*correlação directa ou positiva*); se for $\rho \leq 0$, será $b_{yx} \leq 0$ e $b_{xy} \leq 0$ (*correlação inversa ou negativa*).

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

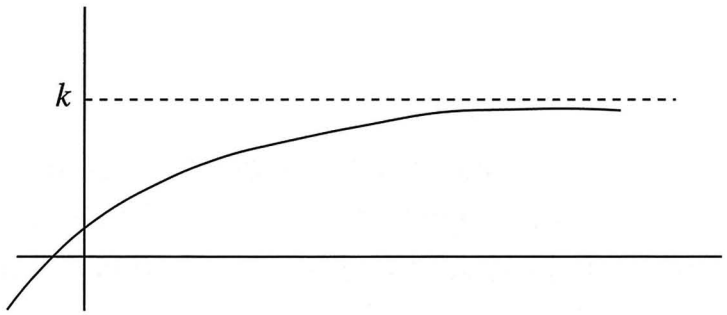
$$\frac{1}{y} = a + bx,$$

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 $\lim_{x \to +\infty} y = k$,

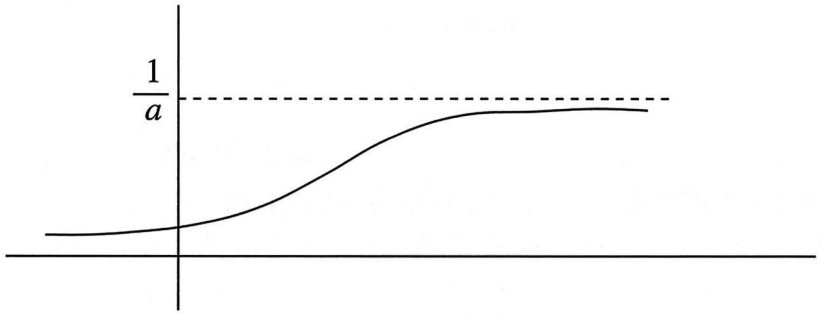


a recta $y=k$ é uma assimptota da curva geralmente conhecida *a priori* (este tipo de função é comum em fenómenos biológicos de crescimento). Ora, por ser

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

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.

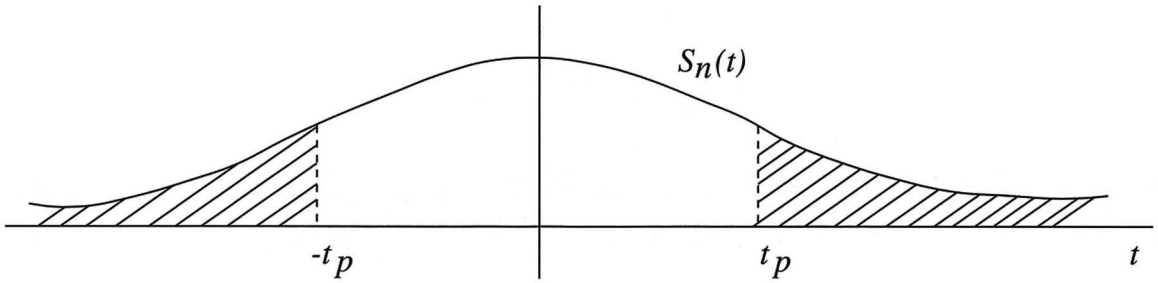
No que se segue, designaremos genericamente por P uma probabilidade e por p o número $100P$. Será, assim,

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

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

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

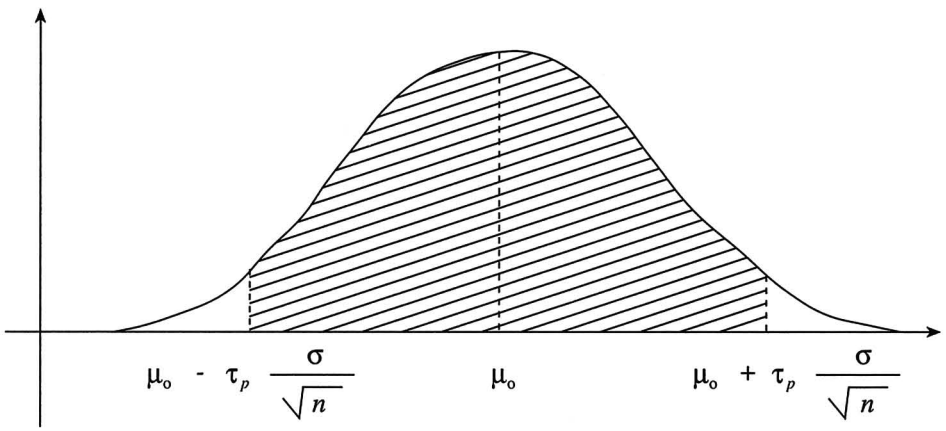
podemos concluir de (4) que os valores de $\bar{x}$ *tolerados* pela hipótese nula são os que verificam a condição

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

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