Rev. Saúde públ., S. Paulo
10:285-9, 1976

CLASSIFICAÇÃO DE PESSOAS NA PROVA TUBERCULÍNICA: APLICAÇÃO DE UM MODELO ESTATÍSTICO QUANDO A DISTRIBUIÇÃO DE FREQÜÊNCIAS DA INDURAÇÃO É UMA MISTURA DE COMPONENTES NORMAIS

Odécio Sanches *

RSPU-B/324

SANCHES, O. — *Classificação de pessoas na prova tuberculínica: aplicação de um modelo estatístico quando a distribuição de freqüências da indução é uma mistura de componentes normais.* **Rev. Saúde públ.**, *S. Paulo*, **10**: 285-9, 1976.

RESUMO: *O problema de classificar pessoas de acordo com o tamanho da indução, na prova tuberculínica segundo a técnica de Mantoux, é resolvido, para um conjunto de dados obtidos em uma população genérica, utilizando-se o critério estatístico de "melhores regiões possíveis de classificação". São obtidas estimativas das probabilidades de classificação errada.*

UNITERMOS: *Teste tuberculínico. Indução. Mistura de distribuições normais. Regiões de classificação. Probabilidade de classificação errada.*

1. INTRODUÇÃO

Em publicação anterior, Sanches [4] (1975), utilizando um método gráfico aproximado, discutiu-se a possibilidade de se decompor, em componentes normais, uma distribuição de freqüências de medidas de indução ou de eritema, na prova tuberculínica segundo a técnica de Mantoux, obtidas em uma população genérica, objetivando a posterior solução do problema de classificação segundo o tamanho da reação.

O objetivo do presente trabalho é uma complementação do anterior, isto é, utilizando os resultados anteriormente obtidos, relativamente à variável indução, mostrar que é possível estabelecer estimativas das regiões de classificação para cada um dos grupos componentes, segundo um modelo estatístico estabelecido na literatura, assim como determinar as estimativas das probabilidades de classificação errada.

2. CONSIDERAÇÕES SOBRE O PROBLEMA DE CLASSIFICAÇÃO EM UM ENTRE k GRUPOS COMPONENTES DE UMA MISTURA DE DISTRIBUIÇÕES NORMAIS

O que segue tem a finalidade de colocar o leitor interessado, não especialista em Estatística, em contacto com o modelo de classificação a ser utilizado. O assunto, no entanto, é discutido de forma ampla e exaustiva em quaisquer textos de análise multivariada.

* Da Escola de Enfermagem de Ribeirão Preto da USP — Campus de Ribeirão Preto, SP — Brasil.

SANCHES, O. — Classificação de pessoas na prova tuberculínica: aplicação de um modelo estatístico quando a distribuição de freqüências da indução é uma mistura de componentes normais. *Rev. Saúde públ.*, S. Paulo, **10**:285-9, 1976.

Consideremos uma mistura de k distribuições normais, univariadas, de médias μ_i e variâncias σ_i^2 , sendo p_i as proporções da mistura, i = 1, 2, ., k.

O problema geral consiste em classificar um elemento desta mistura, retirado ao acaso, como pertencente a um dos k grupos componentes. Nas condições consideradas, cada elemento é representado por um ponto de um espaço unidimensional. Assim, o problema de classificação é equivalente ao de dividir o espaço unidimensional referido em k regiões R_1, R_2, ., R_k, mutuamente exclusivas, de tal modo que permita colocar no i^{mo} grupo componente, um elemento representado por um ponto de R_i, i = 1, 2, ., k.

Se um elemento do i^{mo} grupo tem uma probabilidade β_i de pertencer a R_i, então o valor esperado da proporção de classificações erradas é:

$$\alpha = 1 - \sum_{i=1}^{k} p_i \beta_i$$

A questão é escolher as regiões de tal modo que α seja mínima, isto é,

$\sum_{i=1}^{k} p_i \beta_i$ seja máxima.

Tais regiões, quando existem, são denominadas "melhores regiões possíveis de classificação". Rao [2] (1952) e Anderson [1] (1958) demonstram que regiões definidas por:

$$R_i = \{x \,/\, p_i f_i(x) \geq p_j f_j(x);\ i \neq j;\ i, j = 1, 2, ., k\} \quad (2.1)$$

satisfazem o critério de melhores regiões possíveis ($f_i(x)$ é a função densidade de probabilidade da i^{ma} componente da mistura).

Para o caso em consideração as desigualdades contidas em (2.1) podem ser escritas, explicitamente:

$$(\sigma_j^2 - \sigma_i^2)\, x^2 - 2(\sigma_j^2 \mu_i - \sigma_i^2 \mu_j)\, x + [\sigma_j^2 \mu_i^2 - \sigma_i^2 \mu_j^2 - \sigma_i^2 \sigma_j^2 \log \left(\frac{p_i \sigma_j}{p_j \sigma_i}\right)^2] \leq 0 \quad (2.2)$$

$$i \neq j;\ i, j = 1, 2, ., k.$$

Se uma ou mais componentes são truncadas, é suficiente introduzir, em (2.1), a correção para o truncamento. Assim, apenas para fixar idéias, suponhamos a i^{ma} componente truncada, à esquerda, no ponto ξ_{1i}, suposto conhecido. Nestas condições a desigualdade (2.2) se escreve:

$$(\sigma_j^2 - \sigma_i^2)\, x^2 - 2(\sigma_j^2 \mu_i - \sigma_i^2 \mu_j)\, x + [\sigma_j^2 \mu_i^2 - \sigma_i^2 \mu_j^2 - \sigma_i^2 \sigma_j^2 \log \left(\frac{K_{ie}\, p_i \sigma_j}{p_j \sigma_i}\right)^2] \leq 0 \quad (2.2a)$$

onde K_{ie} é o fator de correção para o truncamento,

$$i \neq j;\ i, j = 1, 2, ., k.$$

Se μ_i^* , σ_i^{*2} e p_i^* são estimativas de μ_i, σ_i^2 e p_i, a sua substituição em (2.2) ou (2.2a) fornece estimativas R_i^* de R_i; i = 1, 2, ., k.

As probabilidades de classificar erradamente como pertencentes ao i^{mo} grupo elementos pertencentes aos (k−1) grupos restantes são dadas por:

$$\alpha_{ij} = \int_{R_i} p_j f_j(x)\, dx;$$

$$i \neq j;\ i, j = 1, 2, ., k \quad (2.3)$$

as quais são estimadas por:

$$\alpha_{ij}^* = \int_{R_i^*} p_j^*\, f_j^*(x)\, dx \quad (2.3a)$$

sendo $f_j^*(x)$ uma estimativa de $f_j(x)$.

3. CLASSIFICAÇÃO DE PESSOAS, EM UMA POPULAÇÃO GENÉRICA, SEGUNDO DIAMETROS DA INDURAÇÃO NA PROVA TUBERCULÍNICA

Utilizando-nos de dados obtidos por Ruffino Netto [3] (1970), Sanches [4] (1975) mostra que a distribuição de freqüências para as medidas de indução podiam ser decompostas em três grupos componentes, com distribuições normais, sendo a primeira componente, por decisão do autor, truncada à esquerda no ponto correspondente à medida de 2 mm.

A Tabela 1 apresenta a distribuição de freqüências observadas e as distribuições componentes normais esperadas, referentes às medidas de diâmetros de indução, estudadas no trabalho acima citado, enquanto que a Tabela 2 apresenta as estimativas dos parâmetros obtidas para tais componentes.

TABELA 1

Medidas de indução na prova tuberculínica, segundo a técnica de Mantoux: distribuição de freqüências observada e distribuições componentes esperadas.

Indução (mm)	Freqüências observadas	Freqüências esperadas		
		1.ª comp.	2.ª comp.	3.ª comp.
2 |— 4	253	251	1	2
4 |— 6	108	93	11	4
6 |— 8	37	1	26	10
8 |— 10	39		25	20
10 |— 12	30		5	27
12 |— 14	24			26
14 |— 16	16			18
16 |— 18	8			9
18 |— 20	2			3
Total	517	345	68	119

Fonte: Sanches [4] (1975).

TABELA 2

Estimativas dos parâmetros das três componentes esperadas, referidas na tabela 1

Componente	Estimativas		
	μ^* (mm)	σ^* (mm)	p* (%)
Primeira	3,53	0,94	65,34
Segunda	7,60	1,79	12,11
Terceira	11,80	3,43	22,54

De (2.2) e (2.2a), utilizando os dados da Tabela 2, obtemos as estimativas para as regiões de classificação dos três grupos componentes:

$$R_1^* = \{x \,/\, x < 5{,}27 \text{ mm}\}$$

$$R_2^* = \{x \,/\, 5{,}27 \text{ mm} \leq x < 8{,}98 \text{ mm}\}$$

$$R_3^* = \{x \,/\, 8{,}98 \text{ mm} < x\}$$

Com tais estimativas, a partir de (2.3a), utilizando-nos de uma tabela da N (0;1) obtemos as estimativas das probabilidades de classificação errada:

$$\alpha_{12}^* = 0{,}0116 \text{ ou } 1{,}16\%$$

$$\alpha_{13}^* = 0{,}0060 \text{ ou } 0{,}60\%$$

$$\alpha_{21}^* = 0{,}0222 \text{ ou } 2{,}22\%$$

$$\alpha_{23}^* = 0{,}0400 \text{ ou } 4{,}00\%$$

$$\alpha_{31}^* = 0{,}0000$$

$$\alpha_{32}^* = 0{,}0267 \text{ ou } 2{,}67\%$$

4. DISCUSSÃO E CONCLUSÃO

Se uma distribuição de freqüências observada, de medidas de indução na

prova tuberculínica, mostra-se decomponível em componentes normais, é possível a aplicação do critério estatístico de "melhores regiões possíveis" para estabelecer estimativas das regiões de classificação das pessoas pertencentes aos distintos grupos componentes, assim como estimativas das probabilidades de classificação errada.

Isto é interessante pois que o citado critério leva em consideração as proporções com que cada grupo entra na mistura. Ora, se estas proporções podem variar em função de alguns fatores como área geográfica considerada, grupos etários considerados, raça etc., há interesse buscar, em função de tais fatores, as regiões de classificação e estudá-las, em diferentes épocas, para se aquilatar de possíveis modificações nos seus valores.

A possibilidade do cálculo das estimativas das probabilidades de classificação errada permite ao pesquisador decidir se o critério estatístico de "melhores regiões possíveis" atendam ou não aos objetivos médicos.

Assim, com os dados utilizados, as maiores probabilidades de classificação errada foram:

i) $\alpha^*_{23} = 0{,}0400$, isto é, a probabilidade de classificar, erradamente, como pertencente ao segundo grupo, um elemento do terceiro grupo é de 4 em 100.

ii) $\alpha^*_{32} = 0.0267$, isto é, a probabilidade de classificar, erradamente, como pertencente ao terceiro grupo, um elemento pertencente ao segundo grupo é de, aproximadamente, 3 em 100.

iii) $\alpha^*_{21} = 0.0222$, isto é, a probabilidade de classificar, erradamente, como pertencente ao segundo grupo, um elemento do primeiro grupo é de, aproximadamente, 2 em 100.

Se tais estimativas forem consideradas elevadas, segundo os objetivos do pesquisador, é possível fixar uma probabilidade de classificação errada tão pequena quanto se deseje e, em função desta probabilidade pré-fixada, determinar novas regiões de classificação tentando diminuir, é claro, o custo da classificação errada. Procedimentos deste tipo são discutidos, por exemplo, em Rao [2] (1952), os quais, no entanto, não mais satisfazem o critério estatístico de "melhores regiões possíveis".

Finalmente, é interessante confrotar os resultados por nós obtidos com aqueles estabelecidos pela Comissão Nacional contra a Tuberculose. Esta, em sua 2.ª Recomendação (1968), considerando a dose de 2 T.U de R_t 23, estabelece que "até que seja possível fixar o exato tamanho em milímetros, acima e abaixo do qual devam as reações ser consideradas específicas (positivas) e inespecíficas (negativas) deve-se adotar o seguinte critério de interpretação dos resultados da prova Tuberculínica Padronizada:

Não reatores: 0 — 4 mm

Reatores fracos: 5 — 9 mm

Reatores fortes: 10 mm ou mais

Utilizando a nomenclatura acima, e arredondando-se para o número inteiro de milímetros mais próximo, o critério de classificação por nós utilizado, para os dados considerados, permite a seguinte classificação:

Não reatores: menores que 5 mm

Reatores fracos: de 5 a 8 mm inclusives

Reatores fortes: 9 mm ou mais

SANCHES, O. — Classificação de pessoas na prova tuberculínica: aplicação de um modelo estatístico quando a distribuição de freqüências da indução é uma mistura de componentes normais. *Rev. Saúde públ.*, S. Paulo, **10**:285-9, 1976.

RSPU-B/324

SANCHES, O. — [*Classification of subjects in the tuberculin test: application of a statistical model when the frequency distribution of the induration is a mixture of normal components*]. **Rev. Saúde públ.**, *S. Paulo*, **10**:285-9.

SUMMARY: *The classification problem in the tuberculin test was studied. A sample of induration data from a generic population was used and the "best possible regions" criterion was applied. Estimates of classification regions and estimates of probabilities of misclassification were obtained.*

UNITERMS: *Tuberculin test. Induration. Mixture of normal distributions. Classification regions. Misclassification probability.*

REFERÊNCIAS BIBLIOGRÁFICAS

1. ANDERSON, T. W — *An introduction to multivariate statistical analysis.* New York, John Wiley & Sons, 1958.

2. RAO, C. R. — Advanced statistical methods in biometric research. New York, John Wiley & Sons, 1952.

3. RUFFINO NETTO, A. — *Epidemiologia da tuberculose: estudo de alguns aspectos mensuráveis na prova tuberculínica.* Ribeirão Preto, 1970. [Tese de Doutoramento — Faculdade de Medicina de Ribeirão Preto da USP].

4. SANCHES, O. — Distribuição das medidas de indução e eritema na prova tuberculínica: aplicação de um método gráfico de decomposição de uma distribuição de freqüências em componentes normais. *Rev. Saúde públ.*, S. Paulo, 9:15-24, 1975.

Recebido para publicação em 22/03/1976
Aprovado para publicação em 14/06/1976