CARLOS ALBERTO DE BRAGANÇA PEREIRA
Instrutor do Departamento de Estatística
do I.M.E. da U.S.P.

ESTIMATIVA DA PROBABILIDADE À PRIORI EM UM PROBLEMA DE CLASSIFICAÇÃO

Trabalho apresentado ao IME da USP para obtenção do grau de Mestre em Estatística Aplicada em 03-06-1971

JUNHO DE 1971

+ SÃO PAULO +

À

MEUS PAIS

MEUS IRMÃOS

E

CLÓVIS A. PERES

I - INTRODUÇÃO

Êste trabalho decorreu da necessidade de se resolver um problema concreto que surgiu no decorrer de análises dos resultados experimentais, obtidos no Laboratório de Genética de Populações do Prof. Dr. Luiz Edmundo Magalhães, no Departamento de Genética do Instituto de Biociências, da Universidade de São Paulo.

Em geral, os trabalhos experimentais de Genética de Populações são realizados tendo-se em vista um modêlo matemático que serve para o cálculo de estimativas.

O problema em causa no momento é o dos efeitos dos cruzamentos preferenciais na dinâmica dos fatôres genéticos envolvidos no comportamento preferencial. Já existem diversos modelos imaginados para essa situação entre os quais os de Karlin, 1968. Êsses modelos são genéricos, isto é, tratam da consequência do efeito preferencial na dinâmica do gene, e, por isso, nem sempre contêm informações para a estimativa da preferência nas condições experimentais com um organismo particular, que, exige também condições experimentais particulares.

Foi exatamente, para estimar a preferência sexual, em dado tipo de experiência que surgiu a necessidade da presente investigação.

A orientação teórica dêste trabalho, coube aos professôres Harold J. Larson (1) e Carlos Alberto Barbosa Dantas (2), a quem agradecemos a colaboração recebida.

---

(1) Dept.of Operations Analysis, Naval Postgraduate School, Monterey, California, U.S.A. - Deptº de Estatística do I.M.E. - U.S.P.

(2) Departamento de Estatística do I.M.E. - U.S.P.

II - O PROBLEMA GENÉTICO

Vamos supor que uma fêmea do tipo A está em presença de dois machos, um do tipo A e outro do tipo B. Vamos definir o valor preferencial da fêmea pelo macho do tipo A, ao valor da probabilidade de haver cruzamento do tipo (A×A) e pelo macho B, ao valor da probabilidade do cruzamento do tipo (A×B). Supondo que haverá sempre cruzamento, vamos representar a primeira probabilidade por g então, a segunda será 1-g. Se g = 1-g = 0,5, dizemos que não existe preferência. Se g < 1-g dizemos que existe preferência negativa e se g > 1-g existe preferência positiva. O organismo de interêsse é a Drosóphila Melanogaster.

Em nosso trabalho, vamos estudar a fêmea selvagem Heterozigota (e+) em presença dos machos: selvagem heterozigoto (e+) e ébony (ee). Então, os tipos de cruzamentos possíveis são:

$$\underset{\text{♀}\quad\text{♂}}{\text{I - (e+)×(e+)}} \qquad \text{e} \qquad \underset{\text{♀}\quad\text{♂}}{\text{II - (e+)×(ee)}}$$

A hipótese que o Prof. Edmundo levantou, foi de que, nêstes tipos de cruzamentos, existe preferência positiva, ou seja, a probabilidade do cruzamento ser do tipo I é maior que a de ser do tipo II.

Pelas Leis genéticas sabemos que:

i) O cruzamento do tipo I produz descendentes nas seguintes proporções:
   a) ébony (ee) na proporção de 1/4;
   b) selvagem na proporção de 3/4, dos quais temos 1/4 de homozigotos (++) e 1/2 de heterozigotos (e+).

ii) O cruzamento do tipo II produz descendentes da seguinte maneira:
   a) ébony (ee) na proporção de 1/2;
   b) selvagem na proporção de 1/2.

iii) Os descendentes de um cruzamento, são produzidos independentemente. Então, se temos n descendentes em um cruzamento, o número de selvagens, X, é uma variável aleatória binomial. Se fôr do tipo I, temos uma binomial com parâmetros n e 3/4 e se fôr do tipo II, temos n e 1/2 como parâmetros.

O nosso experimento foi construido da seguinte forma: dentro de uma caixa de população, foram colocadas N fêmeas (e+), N machos (e+) e N machos (ee) e aí permaneceram durante 3 (três) horas. Êste é o tempo estipulado em experiências dessa natureza, dando oportunidade para tôdas as fêmeas cruzarem. É importante acrescentar, que

uma fêmea não cruza mais de uma vez antes de pôr os ovos. Mas, o macho pode cruzar mais de uma vez,sendo que o tempo estipulado diminui bastante esta possibilidade.A relação de 2N machos para N fêmeas foi fixada,exatamente,para haver maior concorrência entre os machos e assim,diminuir a probabilidade de um macho cruzar mais de uma vez. Mesmo assim,não podemos garantir que houve no máximo um cruzamento por macho. Pretendemos,em estudos posteriores,verificar qual a relação ótima entre o número de fêmeas e de machos.

Se considerarmos um número N,bem grande,é razoável admitir que os cruzamentos são assintòticamente independentes. Então nestas condições, teremos u'a amostra aleatória simples de N cruzamentos. A imposição de tomarmos um N grande vem do fato de que se N fôr pequeno,as probabilidades em cada cruzamento serão,significantemente,diferentes. Por exemplo,consideremos N = 2,o que implica em 2 fêmeas (e+),2 machos (e+) e 2 machos (ee). Neste caso,se não existe preferência,a probabilidade de um primeiro cruzamento ser do tipo I é 1/2,mas,a probabilidade do segundo ser do tipo I,dado que o outro foi do tipo I,é 1/3.

É claro que se a amostra é grande,torna-se impossível a um observador contar,diretamente,o número de cruzamentos do tipo I e do tipo II nos N cruzamentos. Então,o primeiro problema que surge é o da classificação de uma fêmea,como tendo sido cruzada com um determinado tipo de macho. Mas,dado um cruzamento com n descendentes, conhecemos a distribuição da variável X (número de descendentes selvagens),nos dois tipos de cruzamentos. Então,podemos utilizar a teoria da classificação.

O segundo problema que vamos enfrentar, é o de encontrar um estimador do parâmetro g (que é a probabilidade a priori do problema da classificação),utilizando as observações classificadas.

Nos dois capítulos que seguem, fizemos uma generalização teórica,utilizando algumas particularizações do problema e em seguida fizemos a aplicação ao nosso problema.

III - O PROBLEMA DA CLASSIFICAÇÃO

Seja X uma variável aleatória definida em duas populações, π_1 e π_2, com densidades $f_1(x)$ e $f_2(x)$, respectivamente. Consideremos uma observação x de X. O nosso problema é classificar essa observação como sendo proveniente de π_1 ou π_2. Nestas condições, a variável X será denominada: *variável de classificação*.

Vamos supor que existe um conjunto S tal que, se $x \in S$ então, $f_1(x) > 0$ e $f_2(x) > 0$ e se $x \notin S$, então, $f_1(x) = 0$ e $f_2(x) = 0$ o que equivale a dizer que $f_1(x) > 0 \not\equiv f_2(x) > 0$ e então S é o conjunto seguinte:

$$S : \{x / f_1(x) > 0 \quad e \quad f_2(x) > 0\}$$

cujo complementar é:

$$\bar{S} : \{x / f_1(x) = f_2(x) = 0\}$$

Vamos considerar:

1) O espaço das *ações* A : $\{a_1, a_2\}$ onde:
   a_1 : classificar a observação x como proveniente de π_1 e
   a_2 : classificar a observação x como proveniente de π_2

2) O espaço dos *estados da natureza* θ $\{\theta_1, \theta_2\}$ onde:
   θ_1 : a observação x é proveniente de π_1
   θ_2 : a observação x é proveniente de π_2

3) A *função de perda* $L(a_i, \theta_j)$ que é a perda de classificarmos x como proveniente de π_i dado que o estado da natureza é θ_j. No caso particular do nosso problema é razoável admitirmos a função:

$$L(a_i, \theta_j) = \begin{cases} 0 & \text{se } i = j \\ C & \text{se } i \neq j \end{cases} \qquad (III\text{-}1)$$

onde i,j = 1,2 e C é uma constante positiva.

Representando em quadro, teremos:

	a_1	a_2
θ_1	0	C
θ_2	C	0

Uma *função de decisão* estará determinada por uma partição (S_1, S_2) do conjunto S e uma correspondência do tipo:

$$d(x) = \begin{cases} a_1 & \text{se } x \in S_1 \\ a_2 & \text{se } x \in S_2 \end{cases} \qquad (III\text{-}2)$$

Supondo que a observação é de π_1, temos:

a) Probabilidade de tomarmos a ação a_1 dado que o estado da natureza é θ_1, que será representada por:

$$Pr\{a_1/\theta_1\} = \int_{S_1} f_1(x)dx = p$$

b) Da mesma maneira se tomarmos a ação a_2, vem:

$$Pr\{a_2/\theta_1\} = \int_{S_2} f_1(x)dx = 1-p$$

O valor dessa probabilidade é denominado *êrro de 1a.espécie.*

Supondo que a observação é de π_2, temos:

a) Probabilidade de tomarmos a ação a_2 dado que o estado da natureza é θ_2, que será representada por:

$$Pr\{a_2/\theta_2\} = \int_{S_2} f_2(x)dx = 1-q$$

b) Da mesma maneira se tomarmos a ação a_1, vem:

$$Pr\{a_1/\theta_2\} = \int_{S_1} f_2(x)dx = q$$

O valor dessa probabilidade é denominado *êrro de 2a.espécie.*

Dada a função de perda *(III-1)* e a função de decisão *(III-2)* podemos escrever:

$$L(a; \theta) = L(d(X); \theta)$$

que é uma variável aleatória.

Podemos, então, definir a *função de risco*, como:

$$r(\theta_i; d) = E\{L(d(X); \theta_i)\} \qquad i = 1, 2$$

que representa a perda esperada quando o estado da natureza é θ_i. Então, a função de risco será:

$$(III\text{-}3) \qquad r(\theta_i; d) = \begin{cases} 0p + C(1-p) = C(1-p) & \text{se } i = 1 \\ Cp + 0(1-q) = Cq & \text{se } i = 2 \end{cases}$$

Se o estado da natureza é uma variável aleatória, vamos representar sua distribuição (também chamada distribuição a priori) por:

$$(III\text{-}4) \qquad Pr(\theta_i) = \begin{cases} g & \text{se } i = 1 \\ 1-g & \text{se } i = 2 \end{cases}$$

onde, $Pr(\theta_i)$ = probabilidade do estado da natureza ser θ_i.

Considerando que θ é uma variável aleatória com distribuição *(III-4)*, então, $r(\theta,d)$ passa a ser, também, uma variável aleató

ria e a função de *Bayes* para o problema, será:

$$(III\text{-}5) \qquad \beta(d) = E\{r(\theta,d)\} = gC(1-p) + (1-g)Cq$$

que representa o risco médio.

Somando e subtraindo, em (III-5), (1-g)(1-q), temos:

$$(III\text{-}6) \quad \beta(d)=gC(1-p)-(1-g)C(1-q)+(1-g)C=\int_{S_2}C\{gf_1(x)-(1-g)f_2(x)\}dx+(1-g)$$

A partição de Bayes será então, $B(S) = (S_1, S_2)$, onde

$$(III\text{-}7) \qquad S_1:\{x/\frac{f_1(x)}{f_2(x)} \geq K\}$$

$$S_2:\{x/\frac{f_1(x)}{f_2(x)} < K\} \quad \text{sendo } K = \frac{1-g}{g}$$

Vamos agora demonstrar um teorema que será útil, posteriormente.

TEOREMA I - Dada a função de decisão (III-2), cuja partição de S é (S_1, S_2), onde:

$$S_1:\{x/\frac{f_1(x)}{f_2(x)} \geq K\},$$

$$S_2:\{x/\frac{f_1(x)}{f_2(x)} < K\} \text{ e}$$

$S:\{x/f_1(x)>0,\ f_2(x)>0\}$ cujo complementar é $\bar{S}:\{x/f_1(x) = f_2(x) = 0\}$.

Então, $p = q$ se e sòmente se $p = 1$ ou $p = 0$.

PROVA

a) Se $p = 1$ temos, $\int_{S_1} f_1(x)dx = 1$.

Se $f_1(x)$ é uma função de densidade, então, $\int_S f_1(x)dx = 1$; pois, se $x \notin S \rightarrow f_1(x) = 0$. Logo, $\int_{S_1} f_1(x)dx = 1 \rightarrow f_1(x)=0$ qualquer que seja $x \notin S_1$. Então, $S = S_1$ e $S_2 = \phi$. Temos, então, $q = \int_{S_1} f_2(x)dx = \int_S f_2(x)dx = 1$, logo, $p = q$.

Da mesma forma provaríamos que: se $p = 0 \rightarrow q = 0$, pois. neste caso $S_1 = \phi$ e $S_2 = S$.

É fácil ver também, que se $p\neq 0 \rightarrow q\neq 0$ e se $p\neq 1 \rightarrow q\neq 1$.

b) Vamos supor por absurdo que $p = q$ onde, $p \neq 1$ e $p \neq 0$.

Se $p=q \rightarrow \int_{S_1} f_1(x)dx - \int_{S_1} f_2(x)dx = 0 \rightarrow \int_{S_1}\{f_1(x)-f_2(x)\}dx = 0$

Como em S_1, $f_1(x) \geq f_2(x)K \rightarrow$

$\rightarrow 0 = \int_{S_1}\{f_1(x) - f_2(x)\}dx \geq (K-1)\int_{S_1} f_2(x)dx = (K-1)q$

então $(K-1)q \leq 0$. Como estamos supondo $p \neq 0$, vem $K \leq 1$, pois, $p \neq 0 \rightarrow q \neq 0$.

Além disso $p=q \rightarrow (1-p) = (1-q) \rightarrow \int_{S_2} f_2(x)dx - \int_{S_2} f_1(x)dx=0$

$\rightarrow \int_{S_2}\{f_2(x) - f_1(x)\}dx = 0$. Como em S_2, $f_1(x) < Kf_2(x)$ vem:

$0 = \int_{S_2}\{f_2(x)- f_1(x)\}dx > (1-K)\int_{S_2} f_2(x)dx = (1-K)(1-q)$

então, $(1-K)(1-q) < 0$. Como $q \neq 0$ e $q \neq 1$, vem $K > 1$, o que é um absurdo, pois, tinhamos encontrado $K \leq 1$. Então $p \neq q$ quando $p \neq 1$ e $p \neq 0$.

Corolário

Se $p \neq 1$ e $p \neq 0$, então $p > q$.

Prova - Vamos provar por absurdo. Notando, neste caso, que $p \neq q$, suponhamos por absurdo que $p < q$.

$p < q \rightarrow \int_{S_1} f_1(x)dx - \int_{S_1} f_2(x)dx < 0 \rightarrow \int_{S_1}\{f_1(x)-f_2(x)\}dx < 0$, mas, em S_1 temos, $f_1(x) \geq Kf_2(x)$, então $0>\int_{S_1}\{f_1(x)- f_2(x)\}dx\geq(K-1)\int_{S_1} f_2(x)dx$ $= (K-1)q$. Logo, $(K-1)q < 0$ e como $q > 0$, temos $K < 1$.

Além disso, se $p < q$ temos $(1-p) > (1-q)$ então vem:

$\int_{S_2} f_1(x)dx > \int_{S_2} f_2(x)dx \rightarrow \int_{S_2}\{f_1(x) - f_2(x)\}dx > 0$.

Como em S_2, $f_1(x) < Kf_2(x)$, então:

$(K-1)\int_{S_2} f_2(x)dx > \int_{S_2}\{f_1(x) - f_2(x)\}dx > 0$

então, $(K-1)(1-q) > 0$ e como $(1-q) > 0$, temos $K > 1$, o que é um absurdo, pois, tinhamos encontrado que $K < 1$. Então, concluimos que $p > q$ quando $p \neq 0$ e $p \neq 1$, pois, demonstramos que $p \neq q$.

IV - ESTIMAÇÃO DE g

Suponhamos, agora, que temos uma amostra aleatória simples de tamanho N da variável de classificação X, especificada no capítulo III, e queremos estimar a probabilidade a priori, g, dada por (III-4). Como não conhecemos o valor g, não conhecemos também, o valor de K = (1-g)/g, que caracteriza a partição de Bayes (III-7).

Vamos, então, fixar o êrro de 1a. especie, (1-p), o que implica em fixar p. Utilizando a partição (S_1, S_2) do conjunto S especificado em III, onde:

$$S_1 : \{x / \frac{f_1(x)}{f_2(x)} \geq K\} \qquad \text{e} \qquad S_2 : \{x / \frac{f_1(x)}{f_2(x)} < K\} ,$$

podemos encontrar o valor de K fixando p. Ao encontrarmos o valor de K, estamos caracterizando uma partição do conjunto S e, consequentemente, a função de decisão, d(x), dada por (III-2). É claro, então, que podemos calcular o êrro de 2a. espécie, q, pois, a partição assim está fixada. Êste procedimento equivale a construção de um teste da razão de verossimilhança com alternativa simples para uma amostra de tamanho 1 (um).

Se temos uma amostra aleatória simples de tamanho N de X, podemos supor que temos N repetições de uma variável de Bernoulli, W*, com probabilidade g, ou seja:

$$W^* = \begin{cases} 1 \text{ se a observação x de X é proveniente de } \pi_1 \\ 0 \text{ se a observação x de X é proveniente de } \pi_2 \end{cases}$$

onde, $Pr\{W^* = 1\} = Pr\{\theta_1\} = g$.

Se a variável Z, representa o número de observações da população π_1 nas N repetições, temos que Z tem distribuição binomial com parâmetro N e g, ou seja:

$$(IV\text{-}1) \qquad Pr\{Z = z\} = \binom{N}{z} g^z (1-g)^{N-z}$$

Vamos, agora, considerar a variável V, onde:

$$V = \begin{cases} 1 \text{ se classificarmos a observação x de X em } \pi_1 \\ 0 \text{ se classificarmos a observação x de X em } \pi_2 \end{cases}$$

Temos, então, as condicionadas:

$$V/\theta_1 = \begin{cases} 1 \text{ com probabilidade } p \\ 0 \text{ com probabilidade } 1-p \end{cases}$$

$$V/\theta_2 = \begin{cases} 1 \text{ com probabilidade } q \\ 0 \text{ com probabilidade } 1-q \end{cases}$$

Consideremos a variável Y, que representa o número de observações de X classificadas em π_1, dentro de N.

A distribuição condicionada de Y/Z é dada por:

$$(IV\text{-}2) \quad Pr\{Y = y/Z = z\} = \sum_{j=0}^{y} \binom{z}{j} p^j (1-p)^{z-j} \binom{N-z}{y-j} q^{y-j} (1-q)^{(N-z)-(y-j)}$$

Podemos encontrar,então,através do produto de (IV-1) por (IV-2) a distribuição conjunta de Y e Z, ou seja:

$$Pr\{Y = y, Z = z\} = Pr\{Z = z\} Pr\{Y = y/Z = z\}$$

e temos a marginal de Y:

$$(IV\text{-}3) \quad Pr\{Y = y\} = \sum_{z=0}^{N} Pr\{Z = z ; Y = y\} =$$

$$= \sum_{z=0}^{N} \binom{N}{z} g^z (1-g)^{N-z} \sum_{j=0}^{y} \binom{z}{j} p^j (1-p)^{z-j} \binom{N-z}{y-j} q^{y-j} (1-q)^{(N-z)-(y-j)}$$

Após ter efetuado tôdas as transformações algébricas, necessárias (vide apêndice), vamos encontrar:

$$(IV\text{-}4) \quad Pr\{Y = y\} = \binom{N}{y} \{g(p-q) + q\}^y \{1-g(p-q) - q\}^{N-y}$$

que nada mais é que, uma binomial com parâmetros N e $\{g(p-q) + q\}$. Podemos verificar, também, que $g(p-q)+q = gp+(1-g)q$ é a probabilidade de classificarmos uma observação em π_1, ou seja:

$$Pr(V = 1) = gp + (1-g)q$$

Vamos encontrar,agora o estimador de máxima verossimilhança de g. Para isso,vamos considerar,que temos uma amostra aleatória simples de tamanho N,da variável V. Então,a função de verossimilhança será:

$$(IV\text{-}5) \quad L(g) = \{g(p-q) + g\}^y \{1 - g(P-q) - q\}^{N-y}$$

Tomando o logarítimo neperiano em (IV-5). temos:

$$\lg L(g) = y \lg\{g(p-q) + q\} + (N-y)\lg\{1-g(p-q) - q\}$$

Para encontrar o máximo de L(g), vamos derivar lgL(g) em relação a g e igualar a zero, isto é:

$$\frac{d\lg L(g)}{dg} = y\frac{p-q}{g(p-q)+q} - (N-y)\frac{p-q}{1-g(p-q)-q} = 0$$

teremos, então:

$$y - N\{g(p-q) + q\} = 0$$

logo, teremos a estimativa g* de g:

(IV-6) $$g^* = \{y/N - q\}\, 1/(p-q) ,$$

pois, sabemos que (p-q) > 0 (ver corolário). Então o estimador de máximo verossimilhança de g é:

(IV-7) $$\hat{g} = \{Y/N - q\}\, 1/(p-q)$$

Pelas propriedades da binomial, sabemos que:

$$E\{Y\} = N\{g(p-q) + q\} \quad e$$

$$V\{Y\} = N\{g(p-q) + q\}\{1-g(p-q) - q\}$$

onde, E{Y} é a média de Y e V{Y} é a variância de Y. Podemos, então, calcular a média e a variância do estimador $\hat{g}$, que são respectivamente:

(IV-8) $$E\{\hat{g}\} = (E\{Y\}/N - q)\, 1/(p-q) = \{g(p-q) + q - q\}\, 1/(p-q) = g$$

(IV-9) $$V\{\hat{g}\} = (V\{Y\}/N^2)1/(p-q)^2 = \frac{\{g(p-q)+q\}\{1-g(p-q) - q\}}{N(p-q)^2}$$

Verificamos, então, que $\hat{g}$ é um estimador não viciado de g.

Por definição, sabemos que $0 \leq g \leq 1$, então, só tem sentido utilizarmos $\hat{g}$ quando $0 \leq g^* \leq 1$ ou seja, quando $q \leq Y/N \leq p$ onde, g* é dado em (IV-6). Podemos verificar que:

a) Se $y < Nq$ temos que o número de classificados em π_1 é menor que o número esperado de classificações erradas em π_1.

b) Se $y > Np$ temos que o número de classificados em π_1 é maior que o número esperado de classificações certas em π_1.

No primeiro caso o valor estimado de g será g*=0 e no segundo caso será g*=1.

O problema maior que surge, quando da utilização do estimador dado por (IV-7), é a escolha do valor de p. Para cada valor fixado de p, vamos encontrar um estimador diferente, e por conseguinte, estimativas diferentes. Então, na verdade, o que encontramos foi uma classe de estimadores $\hat{g}$. Mas, estamos interessados, em apenas, um estimador de g. O mais razoável seria encontrar o valor p tal que a variância do estimador fôsse mínima, isto é, encontrar a partição ou a função de decisão que nos dá o estimador com mínima variância, den

tre a classe de estimadores $\hat{g}$. Mas, para isso seria necessário encontrar a função que liga p e q, isto é a função q(p). Assim, o problema se torna bem difícil, pois, para cada par de funções $\{f_1(x), f_2(x)\}$ teríamos uma particular q(p).

Para resolver êste problema, vamos utilizar uma solução aproximada. Isto é, encontrar uma função de decisão que nos dê um estimador, dentre a classe de estimadores $\hat{g}$, com variância próxima da mínima. Analizemos, então, a função *(IV-9)*:

Notemos que o numerador de $V\{\hat{g}\}$, no máximo, assume o valor 1/4, pois, temos $\{g(p-q)+q\}\{1-g(p-q)-q\}$. Então:

(IV-10) $$V\{\hat{g}\} \leq 1/4N(p-q)^2$$

Se o tamanho da amostra, N, fôr um número bem grande e se tomarmos o máximo valor de p-q, poderemos garantir que $1/4N(p-q)^2$ será um número razoàvelmente pequeno. Então, se fixamos p de tal forma que p-q seja máximo, o estimador $\hat{g}$ terá uma variância pequena, lembrando *(IV-10)*. Sendo que ao tomarmos o estimador $\hat{g}$ onde p-q é máximo, supomos intuitivamente que estamos utilizando um estimador, dentro da classe de estimadores $\hat{g}$, que possui a menor variância. Esta nossa suposição vem do fato de que quando trabalhávamos com o nosso exemplo de aplicação, calculamos as estimativas para vários p's e ao calcularmos as variâncias, verificamos que o menor valor foi encontrado quando p-q era máximo.

TEOREMA 2

Consideremos: 1) A variável de classificação X, nas condições do capítulo III, onde existe um conjunto S tal que, se $x \in S \rightarrow f_1(x) > 0$ e $f_2(x) > 0$ e se $x \notin S \rightarrow f_1(x) = f_2(x) = 0$;

2) Uma partição (S_1, S_2) de S onde:

$$S_1: \{x / \frac{f_1(x)}{f_2(x)} \geq K\} \qquad \text{e} \qquad S_2: \{x / \frac{f_1(x)}{f_2(x)} < K\}$$

Nestas condições, se K = 1, então, p-q é máximo.

PROVA

Seja a partição (S_1^*, S_2^*), onde:

$$S_1^*: \{x / \frac{f_1(x)}{f_2(x)} \geq 1\} \qquad \text{e} \qquad S_2^*: \{x / \frac{f_1(x)}{f_2(x)} < 1\}$$

e representemos:

$$p^* = \int_{S_1^*} f_1(x)dx \qquad e \qquad q^* = \int_{S_1^*} f_2(x)dx.$$

a) Se K < 1, temos que S_1^* está contido em S_1 e S_2 está contido em S_2^* Então,

$$p = \int_{S_1} f_1(x)dx = p^* + \int_{S_1 - S_1^*} f_1(x)dx \qquad e$$

$$q = \int_{S_1} f_2(x)dx = q^* + \int_{S_1 - S_1^*} f_2(x)dx$$

o que implica em:

$$p-q = p^*-q^* - \int_{S_1-S_1^*}\{f_2(x) - f_1(x)\}dx \qquad (IV\text{-}11)$$

Notemos, que $S_1 - S_1^*$ é o conjunto $\{x/K \leq \frac{f_1(x)}{f_2(x)} < 1\}$ então, $f_1(x) < f_2(x)$ nêste conjunto. Logo, admitindo que $S_1 - S_1^* \neq \phi$, temos

$$\int_{S_1-S_1^*}\{f_2(x) - f_1(x)\}dx > 0$$

Então, verificando (IV-11), temos:

$$p - q < p^* - q^* \quad \text{se} \quad S_1 - S_1^* \neq \phi \quad e$$

$$p - q = p^* - q^* \quad \text{se} \quad S_1 - S_1^* = \phi$$

b) Se K > 1, temos que S_1 está contido em S_1^* e S_2^* está contido em S_2; Então,

$$p = \int_{S_1} f_1(x)\,dx = p^* - \int_{S_1^*-S_1} f_1(x)dx \qquad e$$

$$q = \int_{S_1} f_2(x)\,dx = q^* - \int_{S_1^*-S_1} f_2(x)dx$$

o que implica em:

$$p - q = p^* - q^* - \int_{S_1^*-S_1}\{f_1(x) - f_2(x)\}dx \qquad (IV\text{-}12)$$

Notemos que $S_1^* - S_1$ é o conjunto $\{x/1 \leq \frac{f_1(x)}{f_2(x)} < K\}$ e podemos escrever: $S_1^* - S_1 = \{A \text{ ou } B\}$ onde:

$$A: \{x/\frac{f_1(x)}{f_2(x)} = 1\} \qquad e$$

$$B: \{x/1 < \frac{f_1(x)}{f_2(x)} < K\}$$

Então, vamos escrever *(IV-12)* da seguinte forma:

$$p - q = p^* - q^* - \int_A \{f_1(x) - f_2(x)\} - \int_B \{f_1(x) - f_2(x)\}dx$$

mas, $\int_A \{f_1(x) - f_2(x)\}dx = 0$ pois, em A $f_1(x) = f_2(x)$, então, podemos escrever:

(IV-13) $$p - q = p^* - q^* - \int_B \{f_1(x) - f_2(x)\}dx$$

Se o conjunto $B \neq \phi$, então, como em B $f_1(x) > f_2(x)$, vem:

$$\int_B \{f_1(x) - f_2(x)\}dx > 0$$

Logo, verificando *(IV-13)*, temos:

$$p - q < p^* - q^* \quad \text{se} \quad B \neq \phi \quad \text{e}$$

$$p - q = p^* - q^* \quad \text{se} \quad B = \phi$$

Através de a) e b), podemos verificar que p-q assume, no máximo o valor p*-q*. Como p*-q* é o valor de p-q quando K = 1, podemos concluir que: Se K = 1 então, p-q é máximo.

Em alguns casos particulares, p-q será máximo se e sòmente se K = 1. Por exemplo, no caso em que $Pr\{f_1(x)/f_2(x) = c\} = 0$, onde, c é uma constante qualquer.

Após todos êsse comentários, verificamos que um bom estimador de g, é:

(IV-14) $$\hat{g} = \{Y/N-q\}1/p-q) \quad \text{onde,}$$

$$p = \int_{S_1} f_1(x)dx, \qquad q = \int_{S_1} f_2(x)dx,$$

$$S_1: \{x/\frac{f_1(x)}{f_2(x)} \geq 1\} \qquad \text{e} \qquad S_2: \{x/\frac{f_1(x)}{f_2(x)} < 1\}$$

V - APLICAÇÃO AO PROBLEMA GENÉTICO

Como vimos no capítulo II, o tamanho de nossa amostra, é exatamente, o número N de cruzamentos e como tôdas as fêmeas cruzam, N é o número de fêmeas. O tamanho da amostra foi fixado em N = 250, foram perdidas 18 fêmeas e, então, ficamos reduzidos a uma amostra de N = 232 fêmeas.

A variável de classificação X, observada em cada cruzamento, foi o número de descendentes selvagens, que é uma variável binomial.

Vamos considerar, que do i-ésimo cruzamento, resultaram n_i descendentes dos quais x_i eram selvagens. Então, queremos, através da observação x_i, classificar êste cruzamento como sendo do tipo I (população π_1) ou do tipo II (população π_2). Podemos verificar que x_i assume valores de zero a n_i, com probabilidade positiva, em qualquer dos dois tipos de cruzamentos. Então, podemos dizer que $\{x/x\in\{0,n_i\}\}$ é o conjunto S especificado no capítulo III.

Sabemos que:

a) Se o cruzamento é do tipo I (estado da natureza θ_1),

$$Pr\{X = x_i\} = \binom{n_i}{x_i}(3/4)^{x_i}(1/4)^{n_i-x_i}$$

b) Se o cruzamento é do tipo II (estado da natureza θ_2),

$$Pr\{X = x_i\} = \binom{n_i}{x_i}(1/2)^{n_i}$$

A partição (S_1^i, S_2^i) do i-ésimo cruzamento, será:

$$S_1^i:\{x_i/\frac{(3/4)^{x_i}(1/4)^{n_i-x_i}}{(1/2)^{n_i}} \geq K\}$$

$$S_2^i:\{x_i/\frac{(3/4)^{x_i}(1/4)^{n_i-x_i}}{(1/2)^{n_i}} < K\}$$

Fazendo-se as transformações necessárias e tomando-se o logarítimo decimal, vamos encontrar:

$$S_1^i:\{x_i/x_i \geq \log K + (0,631)n_i\}$$

$$S_2^i:\{x_i/x_i < \log K + (0,631)n_i\}$$

Notemos que os cruzamentos produzem descendentes em números diferentes (ver TABELA I). Então, para um determinado cruzamento i, fixado K, vamos encontrar os erros $(1-p_i)$ e q_i, que podem ser diferentes em outro cruzamento. Para fazer com que os erros sejam constantes e a partição seja a mesma em todos os cruzamentos, tomamos apenas os n primeiros descendentes de cada cruzamento e observamos o número de selvagens, x, dentre os n. Como o número mínimo de descentes das N = 232 fêmeas foi de 40 (quarenta), fixamos n = 40. A variável de classificação X, terá, desta maneira, as seguintes distribuições:

a) Se o cruzamento fôr do tipo I, temos:

$$Pr\{X = x\} = \binom{40}{x}(3/4)^x(1/4)^{40-x}$$

b) Se o cruzamento fôr do tipo II, temos:

$$Pr\{X = x\} = \binom{40}{x}(1/2)^{40}$$

Logo, temos agora, uma partição (S_1, S_2) de S para um cruzamento qualquer, dentre os N, onde:

$$S_1:\{x/\frac{(3/4)^x(1/4)^{40-x}}{(1/2)^{40}} \geq K\},$$

$$S_2:\{x/\frac{(3/4)^x(1/4)^{40-x}}{(1/2)^{40}} < K\} \text{ e } S:\{x/x\in(0,40)\}$$

Da mesma forma que fizemos em (V-1), temos que:

$$S_1:\{x/x \geq \log K + (0,631)40\}$$

$$S_2:\{x/x < \log K + (0,631)40\}$$

No capítulo anterior verificamos que o valor de K fixado deverá ser um valor que torne p-q máximo. Pelo teorema 2, sabemos que se K=1, p-q é máximo. Então, a partição fixada será (S_1, S_2), onde:

$$S_1:\{x/x \geq 25,24\}$$

$$S_2:\{x/x < 25,24\}$$

pois, log 1 = 0.

Como X é binomial, temos que:

$$p = \sum_{x=26}^{40}\binom{40}{x}(3/4)^x(1/4)^{40}$$

$$q = \sum_{x=26}^{40}\binom{40}{x}(1/2)^{40}$$

A função de decisão fixada será, desta meneira, a seguinte:

Suponhamos que em um determinado cruzamento encontramos, x selvagens dentre os 40 primeiros descendentes. Se $x \geq 25,24$, classificamos o cruzamento como sendo do tipo I, se $x < 25,24$ classificamos o cruzamento como sendo do tipo II.

Como x é uma variável binomial, com parâmetros 40 e 3/4 ou 1/2, podemos utilizar a aproximação da binomial para a normal (Ver Feller, 1968), para os cálculos de p e q. Logo, vem:

a) Se o cruzamento é do tipo I, temos:

$X \sim N(30; 7,5)$ que equivale à $z = (X-30)/\sqrt{7,5} \sim N(0,1)$

então,

$$P \simeq \int_{25,24}^{\infty} f_1(x)dx = \int_{-1,74}^{\infty} \phi(z)dz = 0,95907$$

b) Se o cruzamento é do tipo II, temos:

$X \sim N(20; 10)$ que equivale à $z = (X-20)/\sqrt{10} \sim N(0,1)$

então,

$$q \simeq \int_{25,24}^{\infty} f_2(x)dx = \int_{1,66}^{\infty} \phi(z)dz = 0,04846$$

onde, $\phi(\cdot)$ representa a densidade da normal padrão e $f_1(\cdot)$ e $f_2(\cdot)$ representam as densidades quando os cruzamentos são do tipo I e do tipo II, respectivamente.

O valor máximo de p-q será, dêste modo:

$$p - q = 0,95907 - 0,04846 = 0,91061$$

Na tabela II, vamos encontrar os resultados com apenas os 40 primeiros descendentes de cada cruzamento e podemos verificar que o valor y assumido pela variável Y, que representa o número de classificados como cruzamentos do tipo I, foi de y = 140. A estimativa de g, utilizando o estimador *(IV-14)*, será:

$$g^* = \{140/232 - 0,04846\}1/0,91061 = 0,60947$$

Sabemos que a variável Y, tem distribuição binomial *(IV-4)* com parâmetros N = 232 e $g(p-q) + q = g(0,91061) + 0,04846$. Se fizermos aproximação da binomial para a normal, temos que:

$$Y \sim N(\mu;\sigma^2) \text{ onde, } \mu = 232\{g(0,91061) + 0,04846\} \text{ e}$$

$$\sigma^2 = \mu\{1-g(0,91061) - 0,04846\}$$

Como $\hat{g}$ é combinação linear de Y, vem:

$$\hat{g} \sim N(E\{\hat{g}\}; V\{\hat{g}\})$$

Mas, por *(IV-8)* e *(IV-9)* podemos escrever:

$$\hat{g} \sim N(g;\ V\{\hat{g}\}) \qquad \text{onde,}$$

$$V\{\hat{g}\} = \frac{\{g(p-q)+q\}\{1-g(p-q)-q\}}{N(p-q)^2} \simeq g(0,004275)-g^2(0,004310)+ 0.000240$$

Se quizermos, agora, testar a hipótese da não existência de preferência, contra a alternativa, da existência de preferência positiva basta tomar:

$$H_o : g = 0,5 \quad e \quad H_1 : g > 0,5$$

Se H_o é verdadeira, temos:

$$\hat{g} \sim N(0,5\ ;\ 0,0013)$$

A um nível de significância, α, a região crítica será:

$$C_\alpha : \{g^*/g^* > 0,5 + z_\alpha \sqrt{0,0013}\}$$

Se fixarmos $\alpha = 0,05$, vamos encontrar $z_\alpha = 1,64$. Logo, a região crítica do teste, com nível de significância de 5%, será:

$$C_{0,05} : \{g^*/g^* > 0,55904\}$$

Como o valor observado g^* de $\hat{g}$, foi $g^* = 0,60947$, rejeitamos H_o, pois, $g^* \in C_{0,05}$.

A conclusão a que chegamos, através do teste estatístico, é que existe preferência positiva da fêmea *(heterozigota)* selvagem, ou seja, a fêmea *(heterozigota)* selvagem, tem preferência do cruzamento pelo macho *(heterozigoto)* selvagem.

O valor estimado da preferência é $g^* = 0,60947$. Para a construção de um intervalo de confiança de g, vamos supor válida a aproximação:

$$V\{\hat{g}\} = g(0,004275)-g^2(0,00431)+ 0,00024 \simeq g^*(0,004275)-g^{*2}(0,00431)+ 0,00024 = 0,001244$$

Fixado um coeficiente de confiança de 90%, temos o intervalo de confiança: $g \in \{g^*-1,64\ \sqrt{V\{\hat{g}\}}\ ;\ g^*+1,64\ \sqrt{V\{\hat{g}\}}\}$ com 90% de confiança. Valendo a aproximação $V\{\hat{g}\} \simeq 0,001244$, temos $g \in \{0,55163;\ 0,66731\}$ com 90% de confiança.

VI-O ESTIMADOR DE g QUANDO OS ERROS SÃO DIFERENTES EM CADA UNIDADE AMOSTRAL

Considerando, tôdas as suposições do capítulo III, vamos admitir, agora, que o i-ésimo elemento amostral X_i, tenhamos a partição (S_1^i, S_2^i), o que nos dá:

$$p_i = \int_{S^i} f_1(x_i)dx_i \qquad \text{e} \qquad q_i = \int_{S^i} f_2(x_i)dx_i$$

onde, (S_1^i, S_2^i) é a partição que nos dá $p_i - q_i$ máximo, ou seja:

$$S_1^i : \{x_i / \frac{f_1(x_i)}{f_2(x_i)} \geq 1\}$$

$$S_2^i : \{x_i / \frac{f_1(x_i)}{f_2(x_i)} < 1\}$$

Imaginemos uma variável aleatória W_i, tal que:

$$W_i = \begin{cases} 1 & \text{se classificarmos a i-ésima observação } x_i \text{ em } \pi_1 \\ 0 & \text{se classificarmos a i-ésima observação } x_i \text{ em } \pi_2 \end{cases}$$

Então W_i é uma variável aleatória de Bernoulli com parâmetro $r_i = gp_i + (1-g)q_i$, ou seja:

$$Pr\{W_i = 1\} = r_i$$

$$Pr\{W_i = 0\} = 1-r_i$$

Consideremos uma sequência de variáveis aleatórias independentes W_1, W_2, ..., W_N e uma observação de cada uma das variáveis, w_1, w_2, ..., w_N. A variável aleatória $Y = \sum_{i=1}^{N} W_i$, representa o número de classificações em π_1, nas sequências das N variáveis. Podemos então escrever a distribuição de Y.

Seja $y = \sum_{i=1}^{N} w_i$, temos:

$$Pr\{Y = y\} = \sum_{i \in \Omega} \prod_{j=1}^{y} r_{i_j} \prod_{\substack{K=1 \\ K \neq i_j}}^{N} (1-r_K)$$

onde Ω é o conjunto de tôdas y-uplas possíveis formadas no conjunto das N observações.

Considerando, agora, as N observações, $w_1, w_2, \ldots, w_N$ se $\sum_{i=1}^{N} W_i = y$, podemos ordenar as W_i, de tal forma que as y primeiras sejam iguais a 1 e que as N-y restantes sejam iguais a zero. Daí, a função de verossimilhança, será:

$$L(g) = \prod_{i=1}^{y} r_i \prod_{i=y+1}^{N} (1-r_i)$$

Tomando-se o logarítimo neperiano, vem:

$$\lg L(g) = \sum_{i=1} \lg r_i + \sum_{i=y+1}^{N} \lg(1-r_i)$$

Para encontrar o máximo de L(g), vamos derivar lgL(g) e igualar a zero:

$$\frac{d\lg L(g)}{dg} = \sum_{i=1}^{y} \frac{p_i - q_i}{r_i} - \sum_{i=y+1}^{N} \frac{p_i - q_i}{1 - r_i} = 0$$

podemos ainda escrever:

$$\sum_{i=1}^{y} \frac{(p_i-q_i)(1-r_i)}{r_i(1-r_i)} - \sum_{i=y+1}^{N} \frac{(p_i-q_i)r_i}{r_i(1-r_i)} = 0$$

temos finalmente a função:

$$\sum_{i=1}^{y} \frac{p_i-q_i}{r_i(1-r_i)} - \sum_{i=1}^{N} \frac{p_i-q_i}{1-r_i} = 0$$

o que é o mesmo que:

$$\sum_{i=1}^{y} \frac{p_i - q_i}{\{g(p_i-q_i)+q_i\}\{1-g(p_i-q_i)-q_i\}} - \sum_{i=1}^{N} \frac{p_i-q_i}{1-g(p_i-q_i)-q_i} = 0$$

pois: $r_i = gp_i + (1-g)q_i = g(p_i-q_i) + q_i$.

Então, a estimativa de máxima verossimilhança de g, será o valor de g que satisfaz a igualdade acima.

Se considerarmos agora, nosso exemplo de aplicação, podemos ver que a partição para o i-ésimo elemento será: (S_1^i, S_2^i). onde:

$$S_1^i: \{x_i / x_i : n_i \geq 0{,}631\}$$

$$S_2^i: \{x_i / x_i : n_i < 0{,}631\}$$

Na coluna 4 da tabela I, vamos encontrar os valores observados das variáveis W_i. Podemos ver também que o valor observado y de Y foi $\sum w_i = 145$.

Então, ordenando-se as observações de modo que as 145 classificações de π_1 estejam em primeiro lugar, temos que, a estimativa de g, será o valor de g* que satisfaz a equação:

$$\sum_{i=1}^{145} \frac{p_i - q_i}{\{g^*(p_i-q_i)+q_i\}\{1-g^*(p_i-q_i)-q_i\}} - \sum_{i=1}^{232} \frac{p_i - q_i}{1-g^*(p_i-q_i)-q_i} = 0$$

Êste cálculo é muito trabalhoso, pois, teríamos que calcular os 232 valores dos p_i e dos q_i. Sòmente com o auxílio de um computador isto seria possível. Êste cálculo não nos interêssa muito, pois, não sabemos qual a distribuição do estimador $\hat{g}$. Assim sendo, não poderíamos construir um intervalo de confiança para g.

Ficará em aberto, o estudo das propriedades da estimativa g* de g, calculada dessa maneira.

TABELA I

(i) ordem do elemento amostral; (n_i) total de descendentes de cada cruzamento; (x_i) número de selvagens em cada cruzamento e (w_i) determinações da variável W_i.

i	n_i	x_i	w_i	i	n_i	x_i	w_i	i	n_i	x_i	w_i
1	141	77	0	51	118	67	0	101	74	50	1
2	75	57	1	52	123	63	0	102	103	81	1
3	121	62	0	53	91	70	1	103	62	44	1
4	93	76	1	54	144	111	1	104	123	66	0
5	65	42	1	55	103	76	1	105	89	56	0
6	75	52	1	56	115	79	1	106	75	51	1
7	98	54	0	57	122	62	0	107	54	44	1
8	88	49	0	58	110	55	0	108	72	38	0
9	74	57	1	59	137	103	1	109	101	51	0
10	67	36	1	60	109	87	1	110	105	41	0
11	64	53	1	61	102	70	1	111	78	60	1
12	116	62	0	62	138	74	0	112	108	67	0
13	94	44	0	63	115	84	1	113	84	47	0
14	88	63	1	64	135	69	0	114	112	63	0
15	107	58	0	65	110	93	1	115	84	43	0
16	74	55	1	66	114	88	1	116	116	93	1
17	97	49	0	67	99	74	1	117	41	33	1
18	137	107	1	68	90	69	1	118	109	77	1
19	76	45	0	69	93	41	0	119	92	63	1
20	97	55	0	70	118	63	0	120	64	39	0
21	111	51	0	71	121	61	0	121	87	64	1
22	81	64	1	72	91	63	1	122	110	88	1
23	118	81	1	73	122	89	1	123	113	84	1
24	79	52	1	74	133	68	0	124	112	48	0
25	43	28	1	75	84	64	1	125	84	34	0
26	58	50	1	76	90	64	1	126	64	34	0
27	122	92	1	77	100	71	1	127	57	33	0
28	58	32	0	78	70	53	1	128	91	67	1
29	127	92	1	79	163	84	0	129	73	57	1
30	122	65	0	80	40	31	1	130	82	62	1
31	146	62	0	81	100	81	1	131	60	50	1
32	135	61	0	82	115	64	0	132	80	60	1
33	87	76	1	83	73	33	0	133	112	79	1
34	105	74	1	84	103	60	0	134	101	57	0
35	95	60	1	85	93	66	1	135	104	54	0
36	98	56	0	86	96	59	0	136	98	76	1
37	75	48	1	87	135	104	1	137	111	85	1
38	86	69	1	88	104	79	1	138	110	57	0
39	109	55	0	89	62	46	1	139	108	86	1
40	133	60	0	90	98	78	1	140	110	74	1
41	108	57	0	91	74	60	1	141	103	70	1
42	106	54	0	92	104	74	1	142	110	52	0
43	126	63	0	93	99	76	1	143	119	90	1
44	107	62	0	94	71	52	1	144	118	88	1
45	105	85	1	95	82	61	1	145	72	41	0
46	102	80	1	96	106	56	0	146	107	74	1
47	90	70	1	97	144	68	0	147	111	84	1
48	79	60	1	98	83	63	1	148	80	61	1
49	127	76	0	99	60	46	1	149	126	91	1
50	91	48	0	100	107	79	1	150	49	33	1

TABELA I (continuação e conclusão)

i	n_i	x_i	w_i	i	n_i	x_i	w_i	i	n_i	x_i	w_i
151	95	51	0	181	81	58	1	211	78	65	1
152	83	66	1	182	107	63	0	212	75	35	0
153	93	44	0	183	96	69	1	213	94	49	0
154	99	72	1	184	106	79	1	214	90	68	1
155	115	85	1	185	96	73	1	215	122	91	1
156	84	45	0	186	96	46	0	216	113	69	0
157	80	63	1	187	117	91	1	217	108	86	1
158	106	81	1	188	63	52	1	218	77	52	1
159	94	68	1	189	99	76	1	219	75	61	1
160	97	70	1	190	69	46	1	220	129	80	0
161	43	30	1	191	87	67	1	221	83	59	1
162	113	61	0	192	114	57	0	222	96	72	1
163	102	68	1	193	97	76	1	223	100	48	0
164	118	92	1	194	93	73	1	224	70	48	1
165	124	88	1	195	105	76	1	225	118	61	0
166	108	77	1	196	86	42	0	226	142	75	0
167	77	58	1	197	77	61	1	227	121	90	1
168	95	53	0	198	116	47	0	228	74	50	1
169	129	69	0	199	96	72	1	229	81	60	1
170	115	82	1	200	84	64	1	230	118	66	1
171	46	27	0	201	135	61	0	231	90	41	0
172	92	78	1	202	106	88	1	232	119	100	1
173	124	100	1	203	90	67	1				
174	77	41	0	204	65	46	1				
175	79	41	0	205	98	40	0				
176	50	34	1	206	94	70	1				
177	69	35	0	207	90	63	1				
178	103	52	0	208	120	61	0				
179	113	76	1	209	48	32	1				
180	92	64	1	210	69	51	1				

$y = \Sigma w_i = 145$

TABELA II

(i) ordem do elemento amostral; (x_i) número de selvagens dentre os 40 primeiros descendentes de cada cruzamento; (v_i) determinação da variável V_i.

i	x_i	v_i	i	x_i	v_i	i	x_i	v_i	i	x_i	v_i	i	x_i	v_i
1	17	0	51	23	0	101	26	1	151	20	0	201	19	0
2	30	1	52	21	0	102	31	1	152	26	1	202	37	1
3	21	0	53	30	1	103	34	1	153	18	0	203	26	1
4	36	1	54	34	1	104	22	0	154	32	1	204	32	1
5	29	1	55	28	1	105	26	1	155	34	1	205	12	0
6	32	1	56	30	1	106	27	1	156	18	0	206	28	1
7	23	0	57	21	0	107	38	1	157	27	1	207	25	0
8	25	0	58	18	0	108	25	0	158	31	1	208	19	0
9	34	1	59	29	1	109	17	0	159	28	1	209	26	1
10	19	0	60	34	1	110	14	0	160	30	1	210	26	1
11	38	1	61	24	0	111	26	1	161	26	1	211	29	1
12	21	0	62	19	0	112	23	0	162	20	0	212	23	0
13	15	0	63	32	1	113	26	1	163	29	1	213	21	0
14	31	1	64	17	0	114	21	0	164	34	1	214	28	1
15	23	0	65	33	1	115	20	0	165	36	1	215	33	1
16	26	1	66	32	1	116	34	1	166	29	1	216	20	0
17	20	0	67	32	1	117	33	1	167	29	1	217	34	1
18	39	1	68	29	1	118	29	1	168	26	1	218	29	1
19	24	0	69	16	0	119	27	1	169	23	0	219	39	1
20	21	0	70	20	0	120	25	0	170	32	1	220	20	0
21	13	0	71	20	0	121	22	0	171	20	0	221	26	1
22	27	1	72	26	1	122	34	1	172	36	1	222	39	1
23	34	1	73	32	1	123	34	1	173	40	1	223	10	0
24	28	1	74	17	0	124	15	0	174	14	0	224	28	1
25	27	1	75	26	1	125	12	0	175	23	0	225	20	0
26	35	1	76	27	1	126	17	0	176	30	1	226	25	0
27	30	1	77	26	1	127	13	0	177	20	0	227	39	1
28	25	0	78	26	1	128	26	1	178	13	0	228	36	1
29	27	1	79	20	0	129	30	1	179	25	0	229	33	1
30	20	0	80	31	1	130	34	1	180	30	1	230	14	0
31	10	0	81	30	1	131	37	1	181	31	1	231	19	0
32	17	0	82	21	0	132	39	1	182	27	1	232	24	0
33	38	1	83	16	0	133	30	1	183	30	1			
34	27	1	84	19	0	134	18	0	184	33	1	$y = \Sigma v_i = 140$		
35	24	0	85	28	1	135	18	0	185	29	1			
36	21	0	86	23	0	136	27	1	186	11	0			
37	21	0	87	31	1	137	34	1	187	39	1			
38	22	1	88	30	1	138	18	0	188	31	1			
39	20	0	89	27	1	139	34	1	189	28	1			
40	11	0	90	36	1	140	26	1	190	27	1			
41	18	0	91	27	1	141	24	0	191	29	1			
42	23	0	92	26	1	142	18	0	192	13	0			
43	20	0	93	35	1	143	35	1	193	34	1			
44	25	0	94	27	1	144	34	1	194	37	1			
45	37	1	95	29	1	145	26	1	195	26	1			
46	39	1	96	18	0	146	26	1	196	25	0			
47	28	1	97	13	0	147	34	1	197	38	1			
48	33	1	98	29	1	148	31	1	198	15	0			
49	23	0	99	31	1	149	35	1	199	26	1			
50	15	0	100	30	1	150	27	1	200	27	1			

A P Ê N D I C E

LEMA

Seja Y uma variável aleatória com função de probabilidade dada por (IV-3). Então, Y tem distribuição binomial com parâmetros N e $\{g(p-q)+q\}$.

PROVA

$$\Pr\{Y=y\} = \sum_{z=0}^{N} \binom{N}{z} g^z (1-g)^{N-z} \sum_{j=0}^{y} \binom{z}{j} p^j (1-p)^{z-j} \binom{N-z}{y-j} q^{y-j} (1-q)^{(N-z)-(y-j)} =$$

$$= \sum_{j=0}^{y} p^j q^{y-j} \sum_{z=0}^{N} \binom{N}{z}\binom{z}{j}\binom{N-z}{y-j} g^z (1-g)^{N-z} (1-p)^{z-j} (1-q)^{(N-z)-(y-j)} = *$$

$$\binom{N}{z}\binom{z}{j}\binom{N-z}{y-j} = \frac{N!}{z!(N-z)!}\,\frac{z!}{j!(z-j)!}\,\frac{(N-z)!}{(y-j)!(N-z-y+j)!} =$$

$$= \frac{N!}{y!(N-y)!}\,\frac{y!}{j!(y-j)!}\,\frac{(N-y)!}{(z-j)!(N-y-z+j)!} = \binom{N}{y}\binom{y}{j}\binom{N-y}{z-j}$$

$$* = \sum_{j=0}^{y} \binom{N}{y}\binom{y}{j} p^j q^{y-j} \sum_{z=0}^{N} \binom{N-y}{z-j} g^z (1-g)^{N-z} (1-p)^{z-j} (1-q)^{(N-z)-(y-j)} =$$

$$= \binom{N}{y} \sum_{j=0}^{y} \binom{y}{j} p^j q^{y-j} \sum_{z=j}^{N} \binom{N-y}{z-j} g^z (1-g)^{N-z} (1-p)^{z-j} (1-q)^{(N-z)-(y-j)} =$$

$$= \binom{N}{y} \sum_{j=0}^{y} \binom{y}{j} (pg)^j (q\{1-g\})^{y-j} \sum_{z=j}^{N} \binom{N-y}{z-j} \{g(1-p)\}^{z-j} \{(1-g)(1-q)\}^{(N-z)-(y-j)}$$

$$= \binom{N}{y} \sum_{j=0}^{y} \binom{y}{j} (pg)^j \{(q(1-g)\}^{y-j} \sum_{k=0}^{N-j} \binom{N-y}{k} \{g(1-p)\}^k \{(1-g)(1-q)\}^{(N-y)-k} = *$$

Mas, $N-j \geq N-y$, então:

$$* = \binom{N}{y} \sum_{j=0}^{y} \binom{y}{j} (pg)^j \{q(1-g)\}^{y-j} \sum_{k=0}^{N-y} \binom{N-y}{k} \{g(1-p)\}^k \{(1-g)(1-q)\}^{(N-y)-k} =$$

$$= \binom{N}{y} \sum_{j=0}^{y} \binom{y}{j} (pg)^j \{q(1-g)\}^{y-j} \{g(1-p)+(1-g)(1-q)\}^{N-y} =$$

$$= \binom{N}{y} \{pg+q(1-g)\}^y \{g(1-p)+(1-g)(1-q)\}^{N-y} = \binom{N}{y} \{g(p-q)+q\}^y \{1-g(p-q)-q\}^N$$

então, $\Pr\{Y=y\} = \binom{N}{y} \{g(p-q)+q\}^y \{1-g(p-q)-q\}^{N-y}$

C.Q.D.

BIBLIOGRAFIA

1. ANDERSON,T.M. - AN INTRODUCTION TO MULTIVARIATE ANALYSIS - John Wiley, N. York, 1958.

2. CRAMÉR,H.-MATHEMATICAL METHODS OF STATISTICS - Princeton University Press, Princeton, N. Jersey, 1945.

3. DANTAS,C.A.B. - ESTATÍSTICA MATEMÁTICA - IMPA - 7º Colóquio Brasileiro de Matemática - Poços de Caldas, 1960.

4. FELLER,W. - AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS - Wiley International Edition 3nd Edition, 1968.

5. KARLIN,S. - PRELIMINARIES AND SPECIAL MATING SYSTEMS - JOURNAL OF APPLIED PROBABILITY - Vol. 5 (233 a 311), 1968.

6. KEMPTHORNE,O. - AN INTRODUCTION TO GENETIC STATISTICS - John Wiley, N. York, 1957

7. KRISHNAIAH,P.R.- MULTIVARIATE ANALYSIS - Academic Press, N.York, 1963.

8. LINDGREN,B.W. - STATISTICAL THEORY - McMillan, N. York, 1963.

9. MOOD,A.M. and GRAYBILL,F.A. - INTRODUCTION TO THE THEORY OF STATISTICS - McGraw Hill - Koga Kusha International Student Edition, 2nd Edition, 1963.

10. RAIFFA,H. and SCHLAIFER,R. - APPLIED STATISTICAL DECISION THEORY - Harvard University, Boston, 1961.

11. WALD,A. - STATISTICAL DECISION FUNCTIONS - John Wiley, N. York, 1950.

12. WILKS,S.S. - MATHEMATICAL STATISTICS - John Wiley - N. York, 1962.

.o0o.

ALGUMAS PALAVRAS

Desejamos dedicar êste trabalho a *Clôvis de Araujo Peres* que foi quem nos levou a realização do mesmo e de quem recebemos todo o apoio e incentivo para que isto fôsse possível.

Êste estudo só pôde se realizar com a colaboração e interêsse de inúmeras pessoas às quais desejamos expressar aqui, nossos profundos agradecimentos:

ao Prof. Dr. Carlos Alberto Barbosa Dantas, tôda a nossa formação teórica imprescindível a êste estudo e a outros que pretendemos desenvolver no futuro;

ao Prof. Dr. Harold J. Larson, nosso orientador desde sua chegada ao Departamento de Estatística do I.M.E.;

ao Prof. Dr. Luiz Edmundo de Magalhães, pela orientação na área da genética;

a Adolpho Walter P. Canton, os diálogos prolongados que tornaram possível o esclarecimento de dúvidas e falhas;

ao Prof. Dr. Lindo Fava, Chefe do Departamento de Estatística do I.M.E. e a todos os membros do Departamento, o apôio e estímulo;

a Maria Agusta Querubim e Carlos Ribeiro Vilela, a precisão e eficiência na coleta dos dados no laboratório e a todos os pesquisadores do Departamento de Genética do Instituto de Biociências, o interêsse demonstrado pelo estudo;

a Nancy Mitiko Yamazoe, pela revisão de todo o trabalho de redação;

a Fundação de Amparo à Pesquisa do Estado de São Paulo, sem cujo apôio material não teria sido possível a realização dêste trabalho:

a João Baptista Esteves de Oliveira, a paciência em ler e datilografar os manuscritos.

Í N D I C E

. - o ° o - .