TESTE DE HIPÓTESES DEFINIDAS EM ESPAÇOS DE DIFERENTES DIMENSÕES: VISÃO BAYESIANA E INTERPRETAÇÃO CLÁSSICA

CARLOS ALBERTO DE BRAGANÇA PEREIRA

Tese apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo, para o Concurso de Professor Livre-Docente junto ao Departamento de Estatística.

SÃO PAULO
MARÇO
1985

À Lilian, minha esposa, aos meus filhos, André e Bruno, e à memória do Professor Prisco Bezerra, querido amigo, pelo exemplo como pessoa e como acadêmico.

PRÓLOGO

Este trabalho teve a sua origem em 1981, quando o autor percebeu a insatisfação dos geneticistas com o teste qui-quadrado de falta de ajustamento ao equilíbrio de Hardy-Weinberg. Esta insatisfação decorre de ser este teste muito conservativo. De fato, são raras as vezes em que o teste rejeita a hipótese de equilíbrio, mesmo quando esta é uma hipótese falsa. Com a nossa formação Bayesiana e o desejo de poder participar, com estas idéias, das atividades científicas de nossa comunidade, entendemos que na solução adequada do problema de testar equilíbrio estaria a oportunidade de demonstrarmos o poder do argumento Bayesiano. Acreditamos que o presente trabalho inclui tal solução a qual, embora Bayesiana, admite interpretações clássicas.

Rogatko[1] em sua tese de doutorado apresentou uma análise Bayesiana completa (teste e estimação; por ponto e intervalo) do problema do equilíbrio em sistemas simples (autossômico monogênico dialélico e codominante). O teste ali desenvolvido não permite extensões naturais para sistemas mais complexos (genes ligados ao sexo, por exemplo). Ao tentar adaptar aquele teste ao pro

---

(1) *Rogatko, A. (1983) Solução Bayesiana para dois problemas clássicos em Genética: Penetrância e equilíbrio de Hardy-Weinberg. Tese de doutorado, Instituto de Biociências - USP 115 p.*

blema da comparação de duas binomiais, Irony[2] em sua dissertação de mestrado redefiniu o teste através de integrais de linha. Com o uso das integrais de variedades (linhas e superfícies no R^3), esta nova formulação admite a generalidade desejada. O objetivo deste trabalho é desenvolver em detalhes o teste generalizado, suas propriedades e algumas aplicações relevantes.

Por ser um teste exato, o teste de Bayes que introduzimos apresenta dificuldades de cálculo no caso de grandes amostras. Contudo, as facilidades computacionais de nossa Universidade permitem que o teste seja aplicado mesmo em casos onde outros métodos exigem aproximações assintóticas. Assim, todos os resultados numéricos, descritos ao longo deste trabalho e decorrentes da aplicação do teste de Bayes, são resultados exatos, isto é, não se utilizam de métodos assintóticos de aproximação. Na seção 2 do capítulo III, sugerimos uma forma de, possivelmente, se obter tais métodos. Por outro lado, o uso da fórmula de Stirling facilitaria bastante os cálculos das tabelas de contingência unidimensionais, descritos no capítulo II. Como nos diversos casos de equilíbrio, as interações em tabelas multidimensionais são hipóteses não lineares que exigem métodos especiais

---

(2) *Irony, T.Z. (1984) Testes exatos para tabelas 2×2: Bayes × Fisher. Dissertação de mestrado. Instituto de Matemática e Estatística-USP 133p.*

de aproximação. Acreditamos que da procura por métodos adequados de aproximação possa surgir uma interessante linha de pesquisas.

Ao contrário de muitos trabalhos Bayesianos, este não teve como objetivo a crítica aos métodos clássicos em uso. O nosso objetivo é a obtenção de uma alternativa Bayesiana com as qualidades necessárias para substituir àqueles métodos. Formalmente, no capítulo I o teste é desenvolvido e os critérios de qualidade são caracterizados. Aplicações aos diversos problemas envolvendo tabelas de contingência são apresentadas no capítulo II. Os diversos casos de equilíbrio populacional são detalhadamente discutidos no capítulo III. Por problemas de espaço, apenas as tabelas de decisões 9,12,13,..., 21 e 22 foram incluídas. Contudo, tabelas para outros casos podem ser facilmente obtidas pois, a infra-estrutura computacional já está implantada.

Os critérios de qualidade aqui considerados são baseados nas idéias de Dawid[3]. O seguinte exemplo resume essas idéias: Um indivíduo é considerado um bom previsor de chuva (para o dia seguinte) se de todas as vezes que afirma "a chance de chover amanhã é 100p%", em 100p% das vezes chove (isto valendo para $\forall p \in [0,1]$). A de

---

(3) *Dawid, A.P. (1982). The well-calibrated Bayesian.* JASA, 77 (379):605-613.

finição 7 se utiliza dessas idéias para a caracterização da qualidade de um teste. Para que este critério seja avaliado de forma prática, efetuamos uma comparação (seção 5, capítulo III) entre os testes de Bayes e do qui-quadrado no problema do equilíbrio populacional com amostras de tamanho n=50. Considerou-se alí um grande número de amostras simuladas nas quais os dois testes foram aplicados. A seguir a proporção dos erros cometidos são comparados com os erros estabelecidos teoricamente pelos testes. Os resultados obtidos, neste caso particular, favorecem o teste de Bayes em concordância com os resultados obtidos por Irony[2] no teste de igualdade de proporções. Note que o critério de qualidade da definição 7 é um critério prático, independente da perspectiva (Bayesiana ou clássica) que se está considerando. A interpretação intuitiva da definição 7 é a seguinte: um bom teste é aquele que, além de errar menos, informa corretamente a proporção das vezes que erraria em situações idênticas.

Ao terminar esta apresentação gostaria de expressar meus sinceros agradecimentos a todos aqueles que direta ou indiretamente favoreceram a elaboração deste trabalho. O meu colega e amigo André Rogatko além de ter enunciado o problema original, elaborou e implantou todos os programas que permitiram os cálculos apresentados no capítulo III. A minha aluna e amiga Telba Zalkind

Irony aceitou minha orientação na sua dissertação de mestrado, de onde surgiram as idéias sobre a utilidade das integrais de variedades. Minhas dúvidas, sobre os efeitos do uso de tais integrais, foram eliminadas (acredito) em uma aula particular da Professora Elza Gomide. O ambiente agradável de nosso departamento é a maior contribuição de meus colegas a quem se dispõe a elaborar um trabalho acadêmico. O apoio do IME, do CNPq, da FAPESP e do projeto FINEP permitiram a visita de pesquisadores como D. Basu, D. Lindley e S. Zacks de quem recebi inúmeras contribuições durante o nosso convívio diário. A Tamico datilografou os manuscritos mesmo tendo de atender a inúmeros compromissos importantes. As ilustrações gráficas foram preparadas pelas mãos competentes do Yossio. O apoio e o incentivo da Lilian foram fundamentais para minha concentração no trabalho. A alegria do Bruno e o carinho do André foram os únicos calmantes dos meus momentos de angústia.

S U M Á R I O

CAPÍTULO I

DEFINIÇÃO DO PROBLEMA E DESCRIÇÃO DA SOLUÇÃO PROPOSTA

1 - PRELIMINARES

Grande parte da literatura estatística moderna é voltada à divulgação do ponto de vista Bayesiano. A maioria dos trabalhos, no entanto, procuram dar justificativas e interpretações Bayesianas aos métodos estatísticos clássicos[1] (Lindley (1965) e Box & Tiao (1973) são exemplos relevantes). Visando economia de esforços, os simpatizantes da metodologia Bayesiana procuram, assim, aproveitar toda a infra-estrutura já desenvolvida pela estatística clássica. O presente trabalho tem por objetivo o desenvolvimento de uma técnica Bayesiana para soluções de problemas de teste de hipótese, que, de modo recíproco, poderá ser útil aos estatísticos clássicos, principalmente quando a eliminação de parâmetros excedentes[2] se faz necessária.

No problema de teste de hipótese, a possibili-

---

(1) *Por métodos clássicos entende-se àqueles métodos construídos por meio das propriedades das distribuições amostrais. Por métodos Bayesianos entende-se àqueles construídos a partir de distribuições definidas nos espaços paramétricos.*

(2) *Usa-se excedente aqui como tradução de "nuisance".*

dade de um compromisso entre os dois pontos de vista (clássico e Bayesiano) é sugerida ao analisarmos o caso particular de confronto de hipóteses simples. Uma simples modificação no lema de Neyman-Pearson (N-P) elimina as contradições geradas pelo fato do nível de significância ser fixado independentemente do tamanho da amostra ou do valor do poder (De Groot, 1975).

A seguir reescrevemos este resultado e apresentamos uma extensão para o caso de três hipóteses.

Representemos por d os dados obtidos na realização de um experimento aleatório e por $f_\theta(d)$ a função (de densidade) de probabilidade associada ao experimento, onde θ é um parâmetro desconhecido que identifica essa função. Suponha que existe interesse em se testar a hipótese $H_0:\theta=0$ contra a alternativa $H_1:\theta=1$. Um teste de hipótese é uma função binária $\delta(.)$ dos dados e tal que, se $\delta(d)=1$ rejeita-se H_0 em favor de H_1 e se $\delta(d)=0$ não se rejeita H_0 em prejuízo de H_1. As probabilidades dos erros de primeira e segunda espécies são representadas, respectivamente, por $\alpha(\delta)$ e $\beta(\delta)$. Neste trabalho, o critério de seleção de um teste é estabelecido pelo seguinte resultado:

TEOREMA 1 - Considere duas constantes a e b. Se δ^* é um teste definido como

$$\delta^*(d)=0 \quad \text{se} \quad af_0(d) \underset{(>)}{\gtrsim} bf_1(d) \qquad (1)$$

e

$$\delta^*(d)=1 \quad \text{se} \quad af_0(d) \underset{(\leq)}{<} bf_1(d),$$

então qualquer que seja o teste δ, tem-se:

$$a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta).$$

A demonstração deste resultado, embora seja simples (De Groot, 1975), envolve aspectos interessantes do espaço amostral. Em um outro contexto, Pereira (1971) demonstrou o mesmo resultado para provar que $1-\alpha-\beta$ é máximo quando $a=b$. Note-se que embora o Lema N-P seja um Corolário deste resultado, o teste construído aqui tem uma atitude prática bem distinta do teste N-P. Tanto o valor de β como o valor de α variam com o tamanho da amostra. Os valores de a e b determinam o grau de importância dos erros. Por exemplo se $\frac{b}{a}=1$ diz-se que α e β são igualmente importantes enquanto que, se $\frac{b}{a} < 1$ então α sofre uma minimização mais intensa do que β. Os papéis se invertem quando $\frac{b}{a} > 1$. Note que minimizar $a\alpha + b\beta$ é equivalente a minimizar $\alpha + K\beta$ quando $K=\frac{b}{a}$.

O seguinte resultado estabelece a "robustez filosófica[3]" do teste estabelecido pelo Teorema 1.

TEOREMA 2 - O teste δ^* definido em (1) é uma solução Bayesiana para o problema do teste de H_0

---

(3) Um critério estatístico é dito ser robusto filosoficamente se for ótimo sob ambos os pontos de vista: clássico e Bayesiano.

contra H_1.

Demonstração

Provar este resultado exige a transferência para o contexto Bayesiano. Para o teste δ, a função de perda $L(\theta,\delta)$ é representada pela seguinte tabela

	$\delta=0$	$\delta=1$
$\theta=0$	0	ℓ_0
$\theta=1$	ℓ_1	0

e a probabilidade a priori de H_0: $\theta=0$ (H_1: $\theta=1$) é representada por ξ_0 ($\xi_1 = 1-\xi_0$). Assim, a perda esperada $r(\delta)$ do teste δ será

$r(\delta) = \xi_0 E\{L(\theta,\delta)|\theta=0\} + \xi_1 E\{L(\theta,\delta)|\theta=1\}$. Note que

$E\{L(\theta,\delta)|\ \theta=0\} = \ell_0\alpha(\delta)$

$E\{L(\theta,\delta)|\ \theta=1\} = \ell_1\beta(\delta)$.

Assim, $r(\delta) = \xi_0\ell_0\alpha(\delta)+\xi_1\ell_1\beta(\delta)$ e um teste que minimiza $r(\delta)$ é denominado um teste de Bayes. Isto é, se δ^* é um teste tal que, para qualquer outro teste δ

$\alpha(\delta^*)+K\beta(\delta^*) \leq \alpha(\delta) + K\beta(\delta)$ onde $K = \dfrac{\xi_1\ \ell_1}{\xi_0\ \ell_0}$, então

δ^* é um teste de Bayes. Assim, o teorema está demonstrado pois, qualquer que seja o valor de $\frac{b}{a}$ no teste do teorema 1, existirão valores de ξ_0 ℓ_0 e ℓ_1 tais que

$$\frac{\xi_1 \ell_1}{\xi_0 \ell_0} = \frac{b}{a} .$$

É razoável imaginarmos o que deverá acontecer quando no lugar de hipóteses simples encontrarmos hipóteses compostas. A próxima seção é dedicada à descrição da técnica proposta para construção de testes com hipóteses compostas, o objeto principal deste trabalho. Contudo, finalizamos esta seção apresentando uma extensão natural dos teoremas 1 e 2.

Suponha que além de H_0: $\theta=0$ e H_1:$\theta=1$ outra hipótese H_2:$\theta=2$ possa também ser escolhida. A função teste, δ, pode agora assumir três valores possíveis: 0, 1 e 2. Neste problema diversos tipos de erros podem ser cometidos. A tabela abaixo descreve esses erros e também a notação usada para as respectivas probabilidades quando um teste δ é usado.

Erros	Probabilidades
Aceitar H_1 quando H_0 é verdadeira	$\alpha_1(\delta)$
Aceitar H_2 quando H_0 é verdadeira	$\alpha_2(\delta)$
Aceitar H_0 quando H_1 é verdadeira	$\beta_0(\delta)$
Aceitar H_2 quando H_1 é verdadeira	$\beta_2(\delta)$
Aceitar H_0 quando H_2 é verdadeira	$\gamma_0(\delta)$
Aceitar H_1 quando H_2 é verdadeira	$\gamma_1(\delta)$

TEOREMA 1a. - Considere três constantes positivas a, b e c. Se δ^* é um teste definido como

$$\delta^*(d) = 0 \quad \text{se} \quad af_0(d) \underset{(>)}{\geq} bf_1(d) \quad \text{e} \quad af_0(d) \underset{(>)}{\geq} cf_2(d)$$

$$\delta^*(d) = 1 \quad \text{se} \quad af_0(d) \underset{(\leq)}{<} bf_1(d) \quad \text{e} \quad bf_1(d) \underset{(>)}{\geq} cf_2(d)$$

$$\delta^*(d) = 2 \quad \text{se} \quad af_0(d) \underset{(\leq)}{<} cf_2(d) \quad \text{e} \quad bf_1(d) \underset{(\leq)}{<} cf_2(d),$$

então qualquer que seja o teste δ, tem-se:

$$a[\alpha_1(\delta^*)+\alpha_2(\delta^*)] + b[\beta_0(\delta^*)+\beta_2(\delta^*)] + c[\gamma_0(\delta^*)+\gamma_1(\delta^*)] \leq$$

$$\leq a[\alpha_1(\delta)+\alpha_2(\delta)] + b[\beta_0(\delta)+\beta_2(\delta)]+c[\gamma_0(\delta)+ \gamma_1(\delta)].$$

Demonstração

Note que se δ^* é um teste que minimiza

$$a[\alpha_1(\delta)+\alpha_2(\delta)]+b[\beta_0(\delta)+ \beta_2(\delta)]+c[\gamma_0(\delta)+ \gamma_1(\delta)]$$

então, δ^* é o teste que minimiza

$$a[1- \textstyle\sum_0 f_0(d)] + b[1-\sum_1 f_1(d)] + c[1- \sum_2 f_2(d)]$$

onde $\sum_0$ $\sum_1$ $\sum_2$ são as somas sobre os conjuntos D_0, D_1 e D_2, respectivamente, os quais formam a partição definida pela inversa de δ. Assim, deseja-se encontrar o δ^* que maximiza

$$a \sum_0 f_0(d) + b \sum_1 f_1(d) + c \sum_2 f_2(d)$$

e isto é obtido se a inversa de δ^* definir uma partição (D_0^*, D_1^*, D_2^*) tal que

$$D_0^* = \{d;\ af_0(d) \geq bf_1(d) \quad e \quad af_0(d) \geq cf_2(d)\} ,$$
$$D_1^* = \{d;\ af_0(d) < bf_1(d) \quad e \quad bf_1(d) \geq cf_2(d)\} \quad e$$
$$D_2^* = \{d;\ af_0(d) < cf_2(d) \quad e \quad bf_1(d) < cf_2(d)\}$$

Note que o sinal de igualdade pode ser incluído, consistentemente, em qualquer das desigualdades.

TEOREMA 2a. - O teste δ^* definido pelo Teorema 1a é uma solução Bayesiana para o problema da escolha de uma hipótese dentre $H_0:\theta=0$, $H_1:\theta=1$ e $H_2:\theta=2$.

A demonstração é análoga a do teorema 2.

2 - HIPÓTESES COMPOSTAS

A seção anterior mostra que, quando apenas hipóteses simples são consideradas, uma solução clássica é também uma solução Bayesiana. Contudo, se nem todas as hipóteses envolvidas são simples, as soluções descritas na literatura não satisfazem esta propriedade. A razão para isto decorre do fato de não existir, dentro da vi-

são clássica da estatística, um método consistente de eliminação de parâmetro excedente. Diferentes situações utilizam diferentes métodos de eliminação de parâmetros excedentes. Basu (1976) e Mariotto (1983) descrevem uma série desses métodos analisando seus usos e inconsistências evidenciando o caráter pragmático de tais métodos.

Por outro lado, se as hipóteses confrontadas envolvem espaços de diferentes dimensões (por exemplo hipótese simples contra hipótese composta), o método Bayesiano apresenta dificuldades técnicas que são também contornáveis através de soluções ad hoc. Nesta seção tentaremos desenvolver um método para solucionar grande parte desses problemas. O objetivo é se obter um método que possa receber interpretação tanto clássica quanto Bayesiana e além disso não possuir restrição de aplicabilidade. Embora sejam estes objetivos pretenciosos, acreditamos que seja razoável perseguí-los.

Aqui, apresentaremos uma solução Bayesiana para o problema do teste de hipótese e mostraremos sua interpretação sob o ponto de vista clássico.

Vamos representar o parâmetro relacionado aos dados (através de verossimilhança) por ω e o espaço paramétrico, o conjunto dos possíveis valores de ω, por Ω. Uma distribuição de probabilidade, $P(.)$, definida em Ω descreverá as preferências de um estatístico pelos diver

sos pontos de Ω. As hipóteses a serem confrontadas, no caso binário, são H_0: $\omega \in \Omega_0$ e H_1: $\omega \in \Omega_1$ onde Ω_0 e Ω_1 formam uma partição de Ω. [No caso de três hipóteses teríamos uma partição formada por três subconjuntos de Ω.]. A preferência sobre a veracidade de H_0 (H_1) em prejuízo de H_1 (H_0) é descrito pela probabilidade $\xi_0(\xi_1)$, onde $\xi_0 + \xi_1 = 1$, no caso binário. [Analogamente teríamos (ξ_0, ξ_1, ξ_2) descrevendo as preferências entre H_0, H_1 e H_2.]

Com o intuito de simplificar, a linguagem da teoria das decisões será abandonada em benefício da intuição. É natural que para decidir-se em favor de H_0 deva-se relacionar os valores de ξ_0 e ξ_1. Assim, consideremos a razão de Bayes $R_{01} = \frac{\xi_0}{\xi_1}$ (alternativamente pode-se considerar $R_{10} = \frac{\xi_1}{\xi_0} = R_{01}^{-1}$) que indica a vantagem de $H_0(H_1)$ em relação a $H_1(H_0)$. [No caso de três hipóteses deve-se estudar conjuntamente os valores de R_{01} e $R_{02} = \frac{\xi_0}{\xi_2}$.]. Note que os valores de ξ_0 e $\xi_1 = 1-\xi_0$ oscilam de acordo com a oscilação da experiência do estatístico. Assim, antes de observar os dados, d, a preferência do estatístico é representada por (ξ_0,ξ_1) e após a observação de d o estatístico descreve a nova situação por $\xi_0(d)$ e $\xi_1(d)$. [Analogamente teríamos $\xi_0(d)$, $\xi_1(d)$ e $\xi_2(d)$.] A razão de Bayes então é representada por

$R_{01}(d)$. A seguinte definição indica o critério de decisão que deve ser utilizado.

Definição 1 - (Caso binário). Seja c uma constante positiva. Um teste δ_C definido por

$$\delta_C(d) = 0 \qquad \text{se} \qquad R_{01}(d) \geq c$$

e

$$\delta_C(d) = 1 \qquad \text{se} \qquad R_{01}(d) < c$$

denomina-se teste de Bayes. A constante c é escolhida levando-se em consideração o grau de importância dos erros.

Definição 1.a - (Caso de três hipóteses). Seja $C = (c_{01}, c_{02}, c_{12})$ um vetor real de componentes positivas. Um teste δ_C definido por

$$\delta_C(d) = 0 \qquad \text{se} \qquad R_{01}(d) \geq c_{01} \text{ e } R_{02}(d) \geq c_{02},$$

$$\delta_C(d) = 1 \qquad \text{se} \qquad R_{01}(d) < c_{01} \text{ e } R_{12}(d) \geq c_{12}$$

e

$$\delta_C(d) = 2 \qquad \text{se} \qquad R_{02}(d) < c_{02} \text{ e } R_{12}(d) < c_{12}$$

denomina-se teste de Bayes. O vetor C é definido levando-se em consideração o grau de importância dos erros envolvidos.

Até este ponto, nenhum problema técnico foi ressaltado. Contudo, para que seja possível calcular-se os valores de R_{ij}, necessita-se de alguns entes que ainda não foram definidos. É importante lembrarmos que apenas foi considerado uma distribuição, P, sobre Ω. Contudo, para se calcular os valores de $\xi_i(d)$, necessita-se de distribuições definidas nos conjuntos Ω_i. A primeira vista, poderíamos pensar em definir essas distribuições de ω, como distribuições condicionais, com respeito a P, de ω dado $\omega \in \Omega_i$. Porém, devido ao paradoxo de Borel (Lindley, 1983), esse método não é consistente no caso de algum Ω_i ter probabilidade (na medida P) nula. Para eliminar este problema definiremos medidas obtidas de P onde as integrais envolvidas são integrais de variedades (de superfícies em R^3 ou de linha em R^2 e R^3).

A figura 1 tenta dar uma interpretação intuitiva da definição que estamos considerando. Note que a distribuição P define ordem de preferência tanto em Ω quanto em Ω_i. A visão lateral (fig. 2) do contorno definido por P sobre o conjunto Ω_0 sugere de forma natural uma distribuição em Ω_0.

É importante lembrar que quando todos os Ω_i são de probabilidade (P) positiva, a definição aqui apresentada equivale a considerarmos a probabilidade condicional de ω dado Ω_i.

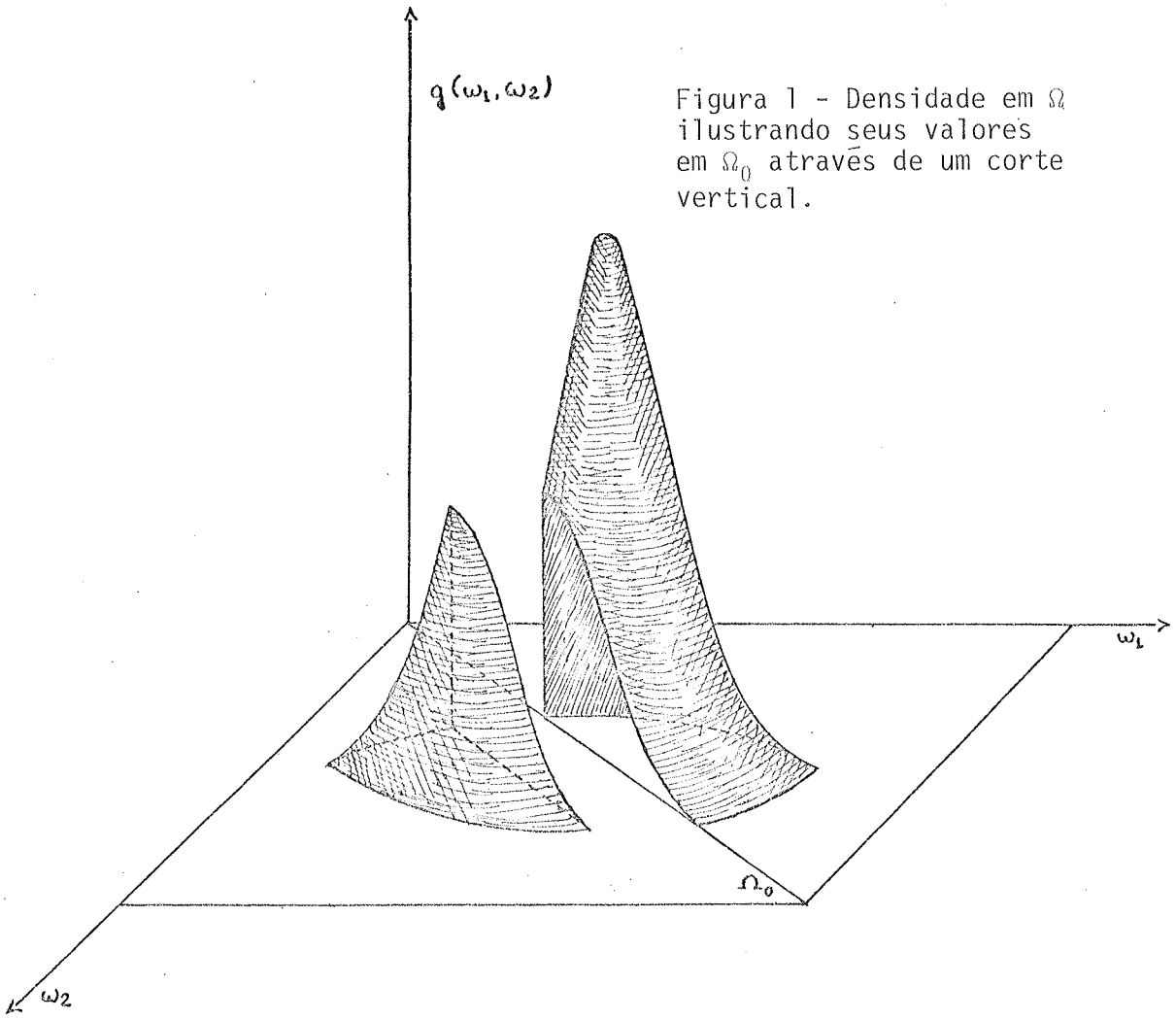
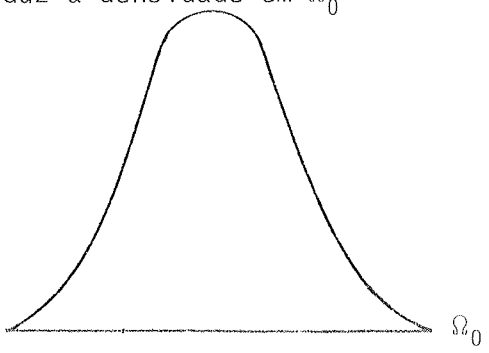


Figura 1 - Densidade em Ω ilustrando seus valores em Ω_0 através de um corte vertical.

Figura 2 - Valores da densidade $g(\omega_1, \omega_2)$ no conjunto Ω_0 que devidamente normali - zada produz a densidade em Ω_0



3 - O TESTE DE BAYES

O objetivo desta seção é descrever formalmente o teste proposto. Como nas seções anteriores, d representa os dados e ω o parâmetro relacionado a d pela verossimilhança. Na maioria dos exemplos aqui estudados os dados, d, admitem redução por suficiência porém, não será feita distinção de notação entre os dados originais e os reduzidos; d será usado em ambos os casos. Para manter a analogia com o caso elementar de hipóteses simples, introduziremos aqui um parâmetro θ,"indicador de hipótese". Isto é, $\theta=i$ se $\omega\in\Omega_i$ (ou se H_i é verdadeira). Assim, $\xi_0=P\{\theta=0\}$, $\xi_1=P\{\theta=1\}$ e $\xi_2=P\{\theta=2\}$, no caso mais geral de três hipóteses.

A função de verossimilhança será representada por $v(\omega|d)$. Note que embora para cada ω, $v(\omega|d)$ seja igual ao valor da função (de densidade) de probabilidade amostral $f(.|\omega)$ calculada nos dados (efetivamente observados) d, usamos $v(.|d)$ para indicar que a verossimilhança é uma função de ω e não de d, como é o caso de f.

À distribuição a priori, P, está associada uma função de densidade (de probabilidade) $g(\omega)$ definida no espaço paramétrico Ω. Nos exemplos discutidos neste trabalho, com o objetivo de obter-se expressões analíticas, toda distribuição a priori considerada pertence à classe conjugada de distribuições. Note que se um estatístico

ganha experiência apenas por resultados experimentais, necessariamente deve trabalhar com classes conjugadas. Veja Pereira e Viana (1982) para uma discussão mais completa sobre sua utilização. Com o uso dessas classes, na maioria dos casos, apenas o valor do parâmetro que indexa a classe é que caracteriza a diferença entre priori e posteriori.

Com a introdução do parâmetro θ, passamos a definir todos os entes envolvidos no processo de teste das hipóteses $H_i: \omega \in \Omega_i$ (equivalentemente $H_i: \theta = i$).

O volume sob uma função g na superfície Ω é calculado pela integral de variedades que recebe aqui a notação $\int g d\Omega$.

Definição 2 - Sob a hipótese $H_i: \theta = i$ (equivalentemente $H_i: \omega \in \Omega_i$) a função de densidade a priori é definida por

$$g_i(\omega) = \frac{g(\omega)\ C_i(\omega)}{\int g(\omega) d\Omega_i}$$ onde $C_i(\omega)$ é a função indicadora de $\omega \in \Omega_i$.

Ao considerarmos estas funções, na verdade estamos modificando a densidade a priori de ω, $g(\omega)$, de forma a garantir probabilidade positiva para os conjuntos Ω_i. Por outro lado, caso Ω_i seja um conjunto de probabilidade, P, positiva, $g_i(\omega)$ é a densidade condicional de

ω dado $\omega \in \Omega_i$. Devido ao paradoxo de Borel, no caso de Ω_i ter probabilidade zero, não podemos garantir que um cálculo da distribuição condicional de ω dado $\omega \in \Omega_i$ nos leve a $g_i(\omega)$. Porém, acreditamos que $g_i(\omega)$ seja uma forma de calcular a densidade condicional.

Definição 3 - Sob a hipótese $H_i: \theta = i$, a função preditiva (a função de probabilidade dos dados) é definida por

$$f_i(d) = \frac{\int g(\omega) v(\omega|d) d\Omega_i}{\int g(\omega) d\Omega_i}$$

Após os dados serem observados, os valores $f_i(d)$ definem uma função $V(.|d)$ de θ, onde $V(\theta|d) = f_\theta(d)$, a qual é denominada de verossimilhança de θ, o parâmetro de interesse.

O lema seguinte relaciona os resultados técnicos que se obtêm com o uso dos entes descritos acima.

Lema 1 - a) a probabilidade a posteriori, $\xi_i(d)$, do evento $\theta = i$ (H_i é verdadeira) é dada por

$$\xi_i(d) = \frac{\xi_i f_i(d)}{\xi_0 f_0(d) + \xi_1 f_1(d) + \xi_2 f_2(d)}$$

b) As razões de probabilidades a posteriori são obtidas como

$$R_{ij}(d) = \frac{\xi_i}{\xi_j} \quad \frac{f_i(d)}{f_j(d)}$$

c) A distribuição a priori, definida através das funções $g_i(\omega)$ e ξ_i, pertence a uma classe conjugada caso a densidade original, $g(\omega)$, seja conjugada em relação ao modelo $f(d|\omega)$.

Demonstração

a) Após entendermos $f_i(d)$ como a probabilidade condicional de d dado $\theta=i$, o resultado é uma aplicação direta do teorema de Bayes.

b) O resultado é obtido ao utilizarmos a) na razão $R_{ij}(d) = \dfrac{\xi_i(d)}{\xi_j(d)}$.

c) A densidade a posteriori de ω sob a hipótese H_i é

$$g_i(\omega|d) = \frac{g(\omega)v(\omega|d)\ c_i(\omega)}{\int g(\omega)v(\omega|d)d\Omega_i}$$

que corresponde a aplicar a definição 2 na densidade de ω obtida pela operação Bayesiana entre $g(\omega)$ e $v(\omega|d)$, isto é, na densidade a posteriori de ω quando $g(\omega)$ é a densidade a priori. Como $g(\omega)$ é conjugada, a mistura de $g_i(\omega)$ por ξ_i também será. Note-se que a densidade a priori modificada de ω é

$$g_*(\omega) = \sum_i \xi_i g_i(\omega)$$

a qual produz

$$g_*(\omega|d) = \sum_i \xi_i(d)\, g_i(\omega|d)$$

como densidade a posteriori modificada de ω.

O teste de Bayes recomendado é aquele incluído nas definições 1 e 1a. cujos elementos envolvidos são caracterizados acima. Para facilitar o relacionamento com os resultados da seção 1, o teste de Bayes pode ser reescrito como:

Teste de Bayes

i) Duas hipóteses: $H_0 \times H_1$.

A função amostral δ_C é um teste de Bayes quando

$$\delta_C(d) = 0 \quad \text{se} \quad d \in \{d;\ \frac{f_0(d)}{f_1(d)} \geq \frac{\xi_1}{\xi_0}\, c\ \} = D_0$$

e

$$\delta_C(d) = 1 \quad \text{se} \quad d \in D_1 = D - D_0$$

onde D é o espaço amostral e c é uma constante positiva fixada.

ii) Três hipóteses: $H_0 \times H_1 \times H_2$

A função amostral δ_C é um teste de Bayes quando

$$\delta_C(d)=0 \quad \text{se} \quad d \in \{d;\ \frac{f_0(d)}{f_1(d)} \geq \frac{\xi_1}{\xi_0}\, c_{01} \ \text{e} \ \frac{f_0(d)}{f_2(d)} \geq$$

$$\geq \frac{\xi_2}{\xi_0}\, c_{02}\ \} = D_0,$$

$$\delta_C(d) = 1 \quad \text{se } d \in \{d;\ \frac{f_0(d)}{f_1(d)} < \frac{\xi_1}{\xi_0}\ c_{01} \quad \text{e} \quad \frac{f_1(d)}{f_2(d)} \geq$$

$$\geq \frac{\xi_2}{\xi_1}\ c_{12}\} = D_1$$

e

$$\delta_C(d) = 2 \quad \text{se} \quad d \in D_2 = D - (D_0 \cup D_1)$$

onde $C=(c_{01}, c_{02}, c_{12})$ é um vetor de constantes positivas fixadas.

A qualidade do teste de Bayes assim definido deve estar relacionada aos tipos de erros que se pode cometer ao utilizá-lo e os quais dependem do valor de ω. Isto produz um impasse devido a infinidade de valores que ω pode assumir. Contudo, ao considerarmos a função $V(\theta|d) = f_\theta(d)$ como a verossimilhança de θ, o parâmetro ω fica eliminado e o impasse contornado.

Considera-se assim os erros modificados que relacionam as funções $f_\theta(d)$ e os conjuntos da partição obtida com o teste de Bayes. Seguindo a notação da seção 1, tem-se a seguinte lista de erros modificados.

a) Caso binário.

$$\alpha(\delta_C) = \sum_1 f_0(d)$$

$$\beta(\delta_C) = \sum_0 f_1(d)$$

onde $\sum_0$ e $\sum_1$ indicam somas sobre D_0 e D_1 respectivamente.

b) Caso de três hipóteses

$$\alpha_1(\delta_C) = \sum_1 f_0(d)$$

$$\alpha_2(\delta_C) = \sum_2 f_0(d)$$

$$\beta_0(\delta_C) = \sum_0 f_1(d)$$

$$\beta_2(\delta_C) = \sum_2 f_1(d)$$

$$\gamma_0(\delta_C) = \sum_0 f_2(d)$$

$$\gamma_1(\delta_C) = \sum_1 f_2(d)$$

onde $\sum_0$, $\sum_1$ e $\sum_2$ indicam soma sobre os conjuntos D_0, D_1 D_2, respectivamente.

Com esta notação, a qualidade do teste de Bayes fica expressada nos seguintes resultados que são as versões modificadas dos teoremas 1, 1a, 2 e 2a.

Teorema 3 - Caso binário

O teste de Bayes, δ_C, descrito nesta seção satisfaz a seguinte propriedade: Se δ é qualquer teste então

$$\alpha(\delta_C) + K\beta(\delta_C) \leq \alpha(\delta) + K\beta(\delta)$$

onde $K = \dfrac{\xi_1}{\xi_0} c$.

Teorema 3a - Caso de três hipóteses

Para $C=(c_{01}, c_{02}, c_{12})$, o teste de Bayes, δ_C,

descrito nesta seção, satisfaz a seguinte propriedade: Se δ é qualquer teste para $H_0 \times H_1 \times H_2$ e $c_{02} = c_{01}c_{12}$ então

$$\alpha_1(\delta_C) + \alpha_2(\delta_C) + K_0[\beta_0(\delta_C) + \beta_2(\delta_C)] + K_1[\gamma_0(\delta_C)+\gamma_1(\delta_C)]$$

$$\leq \alpha_1(\delta)+\alpha_2(\delta)+K_0[\beta_0(\delta)+\beta_2(\delta)]+K_1[\gamma_0(\delta)+\gamma_1(\delta)]$$

onde $K_0 = \frac{\xi_1}{\xi_0} c_{01}$ e $K_1 = \frac{\xi_2}{\xi_0} c_{02}$.

Demonstração

Reescrevendo os conjuntos D_0 e D_1 do teste de Bayes como

$$D_0 = \{d;\ f_0 \geq \frac{\xi_1}{\xi_0} c_{01}f_1 \text{ e } f_0 \geq \frac{\xi_2}{\xi_0} c_{02}f_2\}$$

$$D_1 = \{d;\ f_0 < \frac{\xi_1}{\xi_0} c_{01}f_1 \text{ e } \frac{\xi_1}{\xi_0} c_{01}f_1 \geq \frac{\xi_2}{\xi_0} c_{01}c_{12}f_2\}$$

e substituindo $c_{01}c_{12}$ por c_{02} obtemos um teste do tipo do *Teorema 1a* onde a=1, $b = \frac{\xi_1}{\xi_0} c_{01}$ e $c = \frac{\xi_2}{\xi_0} c_{02}$. Isto conclui a demonstração.

Embora o resultado acima exija uma particularização do teste de Bayes (ao considerar-se $c_{02}=c_{01}c_{12}$), introduz a "robustez filosófica" do teste aqui apresentado, caso a eliminação do parâmetro excedente produza as funções f_0, f_1 e f_2. A discussão desse ponto irá apare-

cer na seqüência. Evidentemente, propriedades mais gerais do teste de Bayes podem ser estabelecidas.

4 - EXEMPLOS

Os exemplos padrões apresentados nesta seção irão ilustrar o método descrito neste trabalho. Para simplificar a apresentação dos problemas utilizamos a seguinte notação:

i) $X \perp\!\!\!\perp Y$ indica que X e Y são variáveis independentes;

ii) $X \perp\!\!\!\perp Y|Z$ indica que X e Y são condicionalmente independentes dado Z;

iii) o símbolo $\sim$ substitui a expressão "tem distribuição".

Por exemplo, para indicar que X tem distribuição Beta com parâmetro (a,b) fixado, escrevemos $X \sim B(a,b)$. Se (a;b) for variável então escrevemos $X|(a,b) \sim B(a,b)$. No caso da distribuição Gama teríamos $X \sim G(a,b)$.

Exemplo 1

Suponha que x e y são observações de duas variáveis binomiais independentes, X e Y, cujos parâmetros são respectivamente (m;p) e (n;q). As hipóteses a serem confrontadas são H_0: p=q e H_1: $p \neq q$.

Como a verossimilhança é representada por $v(p,q|x,y) \propto p^x(1-p)^{m-x}\, q^y(1-q)^{n-y}$, onde $0 \leq p \leq 1$, $0 \leq q \leq 1$, então a classe conjugada de distribuições para (p,q) é formada pela conjunta de Betas independentes. Isto é,

se a priori $p \perp\!\!\!\perp q$, $p \sim B(a,b)$ e $q \sim B(c,d)$ então, a posteriori, $p \perp\!\!\!\perp q|(x,y)$, $p|x \sim B(a+x,\ b+m-x)$ e $q|y \sim B(c+y,\ d+n-y)$.

Ao admitirmos que H_0 tem chance de ser verdadeira, devemos considerar a priori modificada da definição 2. As verossimilhanças modificadas da definição 3 recebem aqui as expressões:

$$H_0 : f_0(x,y) = \frac{\binom{m}{x}\binom{n}{y} B[A+C-1;\ B+D-1]}{B[a+c-1;\ b+d-1]}$$

onde $B[a;b]$ é a função beta no ponto (a,b), $A=a+x$, $B=b+y$, $C=c+m-x$ e $D=d+n-y$.

$$H_1 - f_1(x,y) = \frac{\binom{m}{x}\binom{n}{y} B[A;B]B[C;D]}{B[a;b]B[c;d]}$$

A razão das verossimilhanças modificadas, na qual o teste é baseado, se expressa como:

$$\frac{f_0(x,y)}{f_1(x,y)} = \frac{B[A+C-1;B+D-1]}{B[A;B]B[C;D]} \quad \frac{B[a;b]\,B[c;d]}{B[a+c-1;b+d-1]}$$

No caso da priori uniforme, $a=b=c=d=1$, teríamos

$$\frac{f_0}{f_1} = \frac{\binom{x+y}{x}\binom{m+n-x-y}{m-x}}{\binom{m+n}{m}} \frac{(m+1)(n+1)}{(m+n+1)}$$

$$= \frac{\binom{m}{x}\binom{n}{y}}{\binom{m+n}{x+y}} \frac{(m+1)(n+1)}{(m+n+1)}$$

Note que o primeiro termo do produto é uma probabilidade hipergeométrica que é usada no teste de Fisher.

Exemplo 2:

Suponha que x e y representam os totais amostrais de duas amostras aleatórias simples de duas distribuições independentes de Poisson com médias λ e μ respectivamente. Se m e n são os tamanhos amostrais respectivos e X e Y as duas estatísticas suficientes que produziram x e y, então, $X \perp\!\!\!\perp Y$, $X \sim Poi(m\lambda)$ e $Y \sim Poi(n\mu)$. A verossimilhança é representada por

$$v(\lambda,\mu|x,y) = \frac{(m\lambda)^x (n\mu)^y}{x!\ y!} e^{-m\lambda} e^{-n\mu}$$

onde $\lambda>0$ e $\mu>0$. A classe de distribuições conjugadas para (λ,μ) é constituída por conjuntas de duas Gamas independentes. Isto se resume na seguinte propriedade:

Se, a priori, $\lambda \perp\!\!\!\perp \mu$, $\lambda \sim G(a,b)$ e $\mu \sim G(c,d)$, então, a posteriori, $\lambda \perp\!\!\!\perp \mu|(x,y)$, $\lambda|x \sim G(a+x,b+m)$ e

$\mu|y \sim G(c+y,\ d+n)$.

No caso de admitirmos que H_0: $\lambda = \mu$ tem chance de ser verdadeira, a distribuição a priori é modificada no sentido da definição 2 e, com a alternativa $H_1: \lambda \neq \mu$, utilizando a definição 3 calculamos as verossimilhancas modificadas que produzem a seguinte razão:

$$\frac{f_0(x,y)}{f_1(x,y)} = \frac{\Gamma(A+C-1)}{\Gamma(A)\Gamma(C)} \frac{\Gamma(a)\Gamma(c)}{\Gamma(a+c-1)} \frac{B^A D^C}{(B+D)^{A+C-1}} \frac{(b+d)^{a+c-1}}{b^a d^c}$$

No caso de a=b=c=d=1, que corresponde a considerarmos, como distribuições a priori para (λ,μ), duas exponências independentes com médias iguais a unidade, temos a seguinte expressão:

$$\frac{f_0}{f_1} = \binom{x+y}{x} \left(\frac{m+1}{m+n+2}\right)^{x+1} \left(\frac{n+1}{m+n+2}\right)^{y+1} (m+n+2)2$$

Finalmente, no caso de m=n teríamos

$$\frac{f_0}{f_1} = \binom{x+y}{x} \left(\frac{1}{2}\right)^{x+y} (m+1)$$

cujo produto dos dois primeiros fatores é uma probabilidade binomial com parâmetro $\frac{1}{2}$, que é usada em um teste clássico condicional.

O exemplo que apresentamos a seguir é o mais conhecido da comunidade estatística o que facilita o julgamento da metodologia descrita.

Exemplo 3

As médias λ e μ de duas distribuições normais, com variâncias iguais e conhecidas, devem ser comparadas por um teste de $H_0: \lambda = \mu$ contra $H_1: \lambda \neq \mu$. Das duas distribuições coletou-se amostras de tamanhos m e n, respectivamente, as quais produziram as médias amostrais x e y. Representando o inverso da variância por I, a verossimilhança se expressa como:

$$v(\lambda,\mu|x,y) = \frac{\sqrt{nm}\, I}{2\pi} \exp\{-2I[m(x-\lambda)^2 + n(y-\mu)^2]\}$$

Como distribuição a priori, consideramos aqui que λ e μ são independentes e identicamente distribuídas segundo uma normal com média a e inverso da variancia i. Assim, a densidade a priori é a função

$$g(\lambda,\mu) = \frac{i}{2\pi} \exp\{-2i[(\lambda-a)^2 + (\mu-a)^2]\}$$

a qual multiplicada pela verossimilhança produz a densidade conjunta entre os dados $d=(x,y)$ e o parâmetro $\omega=(\lambda,\mu)$, que reduz-se a seguinte expressão

$$gv = \frac{\sqrt{mn}\, Ii}{(2\pi)^2} \exp\{-2(mI+i)(\lambda-\bar{\lambda})^2 - 2(nI+i)(\mu-\bar{\mu})^2\} \times$$

$$\exp\{-2\frac{mIi}{mI+i}(x-a)^2 - 2\frac{nIi}{nI+i}(y-a)^2\}$$

onde

$$\bar{\lambda} = \frac{mIx + ia}{mI + i} \quad \text{e} \quad \bar{\mu} = \frac{nIy + ia}{nI + i} .$$

Após calcularmos as verossimilhanças modificadas, encontramos a razão destas que se expressa por:

$$\frac{f_0}{f_1} = \sqrt{2}\, \frac{\sqrt{mI+i}\ \sqrt{nI+i}}{\sqrt{i(mI+nI+2i)}} \exp\{ -2\, V_1(x-y)^2 - 2V_2(x-a)^2 - 2\, V_3(y-a)^2 \}$$

onde $V_1 = \frac{mnI^2}{mI+nI+2i}$, $V_2 = \frac{mI^2 i(m-n)}{(mI+nI+2i)(mI+i)}$

e $V_3 = \frac{nI^2 i(n-m)}{(mI+nI+2i)(nI+i)}$.

É interessante ressaltarmos que esta expressão dependente da distância das medias amostrais, x-y, e se m≠n também depende de x-a e y-a que indicam as distância entre a média da distribuição a priori e as médias amostrais.

5 - INTERPRETAÇÃO CLÁSSICA

Um dos testes clássicos que recebe maior divulgação na literatura é o teste da razão de verossimilhanças que consiste em se analisar uma razão cujo numerador

é o valor máximo da verossimilhança no conjunto Ω_0 e o denominador é o valor máximo da verossimilhança no conjunto Ω_1 (ou equivalentemente no conjunto Ω). Este método, entretanto, é heurístico. No lugar de utilizarmos a razão dos máximos, poderíamos pensar na razão das médias das verossimilhanças. Isto é, o numerador seria a média dos valores que $v(.|d)$ assume no conjunto Ω_0 e o denominador a média no conjunto Ω_1.

O método de Bayes desenvolvido nas seções precedentes é baseado na razão de médias ponderadas de $v(.|d)$. A distribuição a priori, na verdade, define a ponderação utilizada. Contudo, quando os conjuntos Ω_0 e Ω_1 são conjuntos limitados, distribuições uniformes são bem definidas e assim as médias ponderadas se reduzem à médias simples. Neste caso a interpretação clássica é natural. No exemplo 1, a razão entre as médias das verossimilhanças é a expressão que apresentamos ao final do exemplo. Assim, a solução ali apresentada possui ambas interpretações: clássica e Bayesiana.

Quando os conjuntos envolvidos não são limitados, a interpretação clássica seria possivel se considerassemos distribuições impróprias o que produz fortes inconsistências já bem descritas na literatura. No caso do exemplo 3, se tomarmos $i \to 0$, nos defrontaremos com o Paradoxo de Lindley (Berger, 1980).

O exemplo 2, no entanto, pode ser avaliado ao

considerarmos o princípio da condicionalidade (Basu, 1975). Cox (1958) sugere que testar H_0: $\lambda=\mu$ contra $H_1:\lambda\neq\mu$ é equivalente a testar H_0: $\Psi = \frac{\lambda}{\lambda+\mu} = \frac{1}{2}$ contra $H_1:\psi \neq \frac{1}{2}$ e no lugar de se utilizar a distribuição de d=(x,y), deve-se utilizar a distribuição condicional de x dado x+y. Isto é, para m=n, a verossimilhança a ser considerada é $\binom{x+y}{x} \psi^x(1-\psi)^y$. A razão das médias desta verossimilhança será

$$\frac{\binom{x+y}{x}\left(\frac{1}{2}\right)^{x+y}}{\binom{x+y}{x} B[x+1;y+1]} = \binom{x+y}{x}\left(\frac{1}{2}\right)^{x+y} (x+y+1)$$

o que difere da solução de Bayes apenas por considerar (x+y+1) no lugar de m+1.

Esta diferença, na realidade, se deve ao fato de a solução Bayesiana utilizar a verossimilhança original e a solução clássica, aqui considerada, a verossimilhança reduzida por condicionamento.

Nos próximos capítulos, discutimos problemas cujos espaços envolvidos são limitados e assim as soluções de Bayes possuem interpretação clássica.

6 - OBSERVAÇÕES

i) Os exemplos discutidos neste capítulo envolvem hipóteses lineares as quais produzem restrições lineares no espaço original Ω. Isto simplifica muito o

cálculo de integrais de variedades. Esta é a razão de não nos preocuparmos em descrever seus cálculos. Contudo os próximos capítulos incluem casos de hipóteses não lineares.

ii) O confronto entre três hipóteses foi aqui descrito para ser utilizado nos problemas estudados na seqüência.

iii) Recentemente, estão sendo publicados trabalhos que utilizam métodos Bayesianos de eliminação de parâmetros excedentes, mesmo quando métodos clássicos de testes de hipóteses estejam sendo considerados. Ver por exemplo Krewski, Brennan & Bickis (1984) e Lecoutre (1985). Contudo, somente parte dos parâmetros envolvidos recebem, nestes trabalhos, distribuições a priori. A mistura de métodos clássicos com os Bayesianos produz dificuldades técnicas ainda maiores do que as que enfrentamos neste trabalho.

iv) O teste de Bayes aqui descrito é, na verdade, uma generalização do método de teste introduzido por Jeffreys (1961). O teste de Jeffreys é conhecido como "Bayesian test of sharp hypothesis".

CAPÍTULO II

TABELAS DE CONTINGÊNCIA

1 - INTRODUÇÃO

O objetivo deste capítulo é mostrar como as técnicas desenvolvidas no capítulo precedente podem ser utilizadas na solução de problemas que surgem com a análise de tabelas de contingência. O contexto que produz tais tipos de tabelas é descrito a seguir.

Os membros de uma população — um termo genérico para designar um conjunto bem definido — podem ser classificados de diversas formas. Pessoas podem ser classificadas como; crianças, jovens ou adultas; homem ou mulher; solteira ou casada; etc. O nosso objetivo é o estudo de populações que são classificadas em um número finito de categorias "exaustivas" e "mutuamente exclusivas". As categorias de uma classificação são exaustivas se qualquer elemento da população pertence a pelo menos uma das categorias. A exaustão é necessária para garantir que todo membro da população seja classificado. As categorias de uma classificação são mutuamente exclusivas se qualquer elemento da população pertence a no máximo uma categoria. A exclusividade é necessária para impedir que algum elemento seja contado mais de uma vez. A seguir, o termo classificação substitui "classificação

por categorias exaustivas e mutuamente exclusivas".

Quando diversos tipos de classificações são combinadas, obtém-se uma nova classificação cujo total de categorias é o produto dos totais de categorias das classificações originais. As análises desenvolvidas na seqüência envolvem dados que são obtidos ao considerarmos uma amostra e registrarmos as freqüências amostrais de cada categoria de uma classificação combinada. Em geral, a forma de se descrever tais dados é através de tabelas cruzadas, *tabelas de contingência*, cujas entradas representam as classificações originais e em cujo corpo são apresentadas as freqüências das categorias da classificação combinada. Neste capítulo serão construídos testes para hipóteses que definem relações especiais entre as diversas classificações que foram combinadas (ou cruzadas).

Alguns resultados básicos são introduzidos na próxima seção.

2 - PRELIMINARES

A distribuição multinomial é o modelo normalmente considerado como gerador de freqüências amostrais. Isto se deve à necessidade que o estatístico clássico tem de descrever completamente o espaço amostral. Contudo, muitas vezes o tamanho da amostra não é conhecido antes de se observar a amostra. Em determinados experimen

tos, a decisão de se parar de observar no n-ésimo elemento é função do que foi observado nos n-1 iniciais. Existem ainda casos em que não é possível caracterizarmos razões probabilísticas que determinaram o tamanho da amostra. Por exemplo, os 17 clientes estudados por um clínico foram os que ele recebeu no seu consultório e logicamente não deve ter fixado esse número previamente nem deve ter decidido, sobre 17, quando possuia apenas 16. Veja Lindley & Phillips (1976), Pereira (1983) e Pereira & Lindley (1984) para discussões sobre o efeito de regras de parada na análise de modelos discretos. A seguir descrevemos um modelo gerador de freqüências mais geral que o multinomial. Consideramos aqui que a classificação combinada possui k categorias.

Seja U um vetor aleatório assumindo valores no conjunto $\{e_1, \ldots, e_k\}$ onde $e_1=(1,0,\ldots,0),\ldots, e_k=(0,\ldots,0,1)$ formam a base ortonormal padrão de R^k. Suponha ainda que $Pr\{U=e_i\} = p_i$, $i=1,\ldots, k$, onde $p_i \geq 0$ e $p_1+p_2+\ldots+ p_k=1$. Um processo de Bernoulli multivariado com parâmetro $\omega = (p_1, \ldots, p_k)$ é uma seqüência, $\{U_j\}_{j\geq 1}$, de vetores aleatórios independentes e identicamente distribuídos segundo U. Isto é, $\coprod_{j=1}^{\infty} U_j$ e $\forall j\geq 1$, $U_j \sim U$. Se os n primeiros elementos da seqüência são observados, podemos escrever

$$Pr\{U_1= e_{i_1}, U_2= e_{i_2},\ldots, U_n= e_{i_n} | \omega\} = \prod_{i=1}^{k} p_i^{x_i}$$

onde $x = (x_1, x_2, \ldots, x_k) = \sum_{j=1}^{n} e_{i_j}$.

Suponha que a decisão de parar de observar em n é função de x. Neste caso, se $X = \sum_{j=1}^{n} U_j$ então existe uma função h(x) tal que

$$Pr\{X=x|\omega\} = h(x) \prod_{i=1}^{k} p_i^{x_i} . \qquad (1)$$

No caso em que n é uma constante (fixada previamente), obtemos a distribuição multinomial,

$$Pr\{X=x|\omega\} = n! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!} ,$$

ou seja $h(x) = \frac{n!}{x_1! \ldots x_k!}$.

Independentemente da escolha de h, a classe de distribuições conjugadas naturais é formada por distribuições de Dirichlet. A seguir descrevemos os elementos desta classe conjugada e apresentamos algumas de suas propriedades importantes.

O parâmetro $\omega=(p_1, \ldots, p_k)$ tem distribuição de Dirichlet de ordem k com parâmetro $a=(a_1, \ldots, a_k)$, $a_i \geq 0$, se a densidade de ω é dada pela expressão

$$g(\omega) = \Gamma\left(\sum_{i=1}^{k} a_i\right) \prod_{i=1}^{k} \frac{p_i^{a_i}}{\Gamma(a_i)} ,$$

onde ω pertence ao simplex $\{(p_1, \ldots, p_k);\ p_i \geq 0,\ \sum_{i}^{k} p_i = 1\}$.

Esta definição é denotada por

$$\omega | a \sim D_k(a).$$

Note que a Distribuição de Dirichlet (representada no R^k) é uma representação singular da distribuição Beta generalizada a qual é definida no R^{k-1} pela mesma densidade acima onde p_k é substituído por $1-p_1- \ldots - p_{k-1}$. Esta representação, embora seja no R^{k-1}, possui um parâmetro, a, definido em R^k. Ao utilizarmos integrais de variedades, a representação no R^k não traz problemas de singularidade, apenas o fator de normalização $\frac{\Gamma(\sum a_i)}{\prod \Gamma(a_i)}$ é dividido por $\sqrt{k}$. Lembremos que o volume do simplex (no R^k) é maior do que o volume do conjunto $\{(p_1, \ldots, p_{k-1});\ p_i \geq 0,\ \sum p_i \leq 0\}$, contido no R^{k-1}. De fato, este volume (área ou comprimento) é $\frac{1}{(k-1)!}$ (Wilks, 1962) e o volume (área) do simplex é $\frac{\sqrt{k}}{(k-1)!}$ (Courant & John, 1974). As figuras 3 e 4 ilustram esta diferença.

As propriedades listadas abaixo são básicas e possuem demonstrações simples (Basu & Pereira, 1982 e Pereira & Viana 1982).

Propriedade 1

Sejam $Z_1, Z_2, \ldots, Z_k$ variáveis Gama indepen-

Figura 3 - Domínio da Dirichlet de ordem k=3 na forma não singular. Área = 1/2.

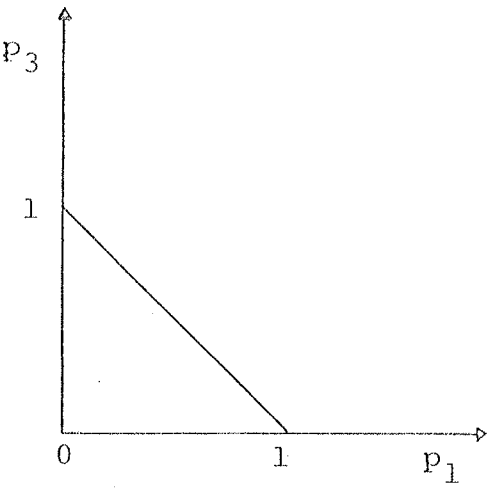
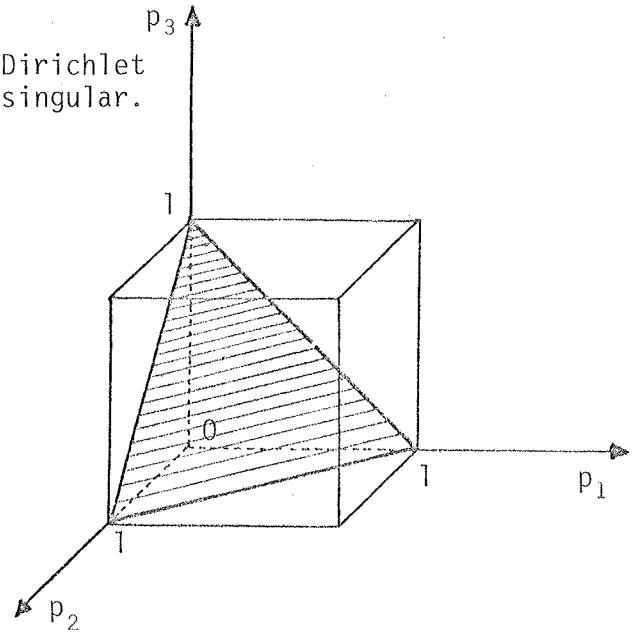


Figura 4 - Domínio da Dirichlet de ordem k=3 na forma singular. Área=$\frac{\sqrt{3}}{2}$.



dentes com parâmetros (a_1,b), (a_2,b), ..., (a_k,b) respectivamente. Se $Z=Z_1+ \ldots + Z_k$, então

a) $\frac{1}{Z}(Z_1, \ldots, Z_k) \sim D_k(a)$, $a=(a_1, \ldots, a_k)$ e

b) $\frac{1}{Z}(Z_1, \ldots, Z_k) \perp\!\!\!\perp Z$.

Corolário

Sejam $1<j<k$ e $P_j = p_1 + \ldots + p_j$.

Se $(p_1, \ldots, p_k) \sim D_k (a_1, \ldots, a_k)$, então

a) $\frac{1}{P_j}(p_1, \ldots, p_j) \perp\!\!\!\perp \frac{1}{1-P_j} (p_{j+1}, \ldots, p_k)$ e

b) $\frac{1}{P_j} (p_1, \ldots, p_j) \sim D_j(a_1, \ldots, a_j)$.

Suponha que o vetor de freqüências

$X = \sum_{i=1}^{n} U_i$ é distribuído como em (1).

Propriedade 2

Se a priori $\omega|a \sim D_k(a)$, então;

a) a posteriori $\omega|a,x \sim D_k(a+x)$,

b) a probabilidade marginal de X (preditiva de X) é dada por

$$f(x) = h(x) \frac{\Gamma(\sum a_i)}{\prod \Gamma(a_i)} \frac{\prod \Gamma (a_i+x_i)}{\Gamma (\sum a_i+n)} \quad \text{e}$$

c) se n é prefixado, f(x) é a função de probabilidade Dirichlet-Multinomial (Basu & Pereira, 1982).

Na próxima seção o modelo para tabelas de contingência é descrito com base nas propriedades aqui listadas.

3 - TABELA DE CONTINGÊNCIA $r \times s$

A forma geral de uma tabela de contingência é apresentada a seguir cuja notação é padrão e os elementos do vetor de freqüências recebem dois índices, identificando linhas e colunas.

TABELA 1

		Colunas (classificação 2) 1	2	.	.	.	s	Total
	1	n_{11}	n_{12}	.	.	.	n_{1s}	$n_{1.}$
Linhas	2	n_{21}	n_{22}	.	.	.	n_{2s}	$n_{2.}$
(classificação 1)	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	r	n_{r1}	n_{r2}	.	.	.	n_{rs}	$n_{r.}$
Total		$n_{.1}$	$n_{.2}$	.	.	.	$n_{.s}$	$n_{..} = n$

A notação análoga para a estrutura paramétrica é dada por

TABELA 2

	1	2	.	.	.	s	
1	p_{11}	p_{12}	.	.	.	p_{1s}	$p_{1.}$
2	p_{21}	p_{22}	.	.	.	p_{2s}	$p_{2.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
r	p_{r1}	p_{r2}	.	.	.	p_{rs}	$p_{r.}$
	$p_{.1}$	$p_{.2}$	.	.	.	$p_{.s}$	1

Uma outra reparametrização relevante é apresentada a seguir onde $q_{ij} = \frac{p_{ij}}{p_{i.}}$.

TABELA 3

	1	2	.	.	.	s	
1	q_{11}	q_{12}	.	.	.	q_{1s}	1
2	q_{21}	q_{22}	.	.	.	q_{2s}	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
r	q_{r1}	q_{r2}	.	.	.	q_{rs}	1

Com o vetor de freqüências $x=(n_{11}, \ldots, n_{rs})$ e o vetor de parâmetros $\omega = (p_{11}, \ldots, p_{rs})$, a veros-

similhança é expressada por

$$v(\omega|x) = h(x) \prod_{i,j} p_{ij}^{n_{ij}} .$$

Se a reparametrização (q,P), onde $q=(q_{11}, \ldots, q_{rs})$ e $P = (p_{1.}, \ldots, p_{r.})$, é utilizada, podemos escrever

$$V(q,P|x) = h(x) \left[\prod_{i,j} q_{ij}^{n_{ij}} \right] \prod_{i} p_{i.}^{n_{i.}} \qquad (2)$$

No caso de n ser fixado previamente $h(x) = \frac{n!}{\prod n_{ij}!}$. No caso de $(n_{1.}, \ldots, n_{r.})$ ser fixado previamente a verossimilhança se reduz a

$$V(q|x) = \left[\prod_{i} n_{i.}! \right] \prod_{i,j} \frac{q_{ij}^{n_{ij}}}{n_{ij}!} . \qquad (3)$$

No caso do modelo (2), a hipótese de interesse é a independência entre as duas classificações. Uma forma de se expressar independência entre dois eventos A e B é através das igualdades

$$P(A|B) = P(A) \quad e \quad P(A|B^C) = P(A).$$

No nosso caso, a independência entre as duas classificações corresponde a ser satisfeito o seguinte sistema de igualdades:

$$\begin{cases} q_{11} = q_{21} = \ldots = q_{r1} \\ q_{12} = q_{22} = \ldots = q_{r2} \\ \qquad\qquad \vdots \\ q_{1s} = q_{2s} = \ldots = q_{rs} \end{cases} .$$

Este mesmo sistema no modelo (3), corresponde a hipótese de homogeneidade.

Podemos assim concluir que a diferença entre um teste de homogeneidade e um teste de independência esta nas verossimilhanças e não nas hipóteses, como se poderia imaginar.

Para o desenvolvimento dos testes, a seguinte notação é conveniente:

(i) Parâmetros

$$\omega = (p_{11}, \ldots, p_{1s}, p_{21}, \ldots, p_{2s}, \ldots, p_{r1}, \ldots, p_{rs})$$

$$Q_i = (q_{i1}, \ldots, q_{is}) \quad \text{e} \quad P = (p_{1.}, \ldots, p_{r.})$$

(ii) Priori

$$a = (a_{11}, \ldots, a_{1s}, a_{21}, \ldots, a_{2s}, \ldots, a_{r1}, \ldots, a_{rs})$$

$$\tilde{a}_i = (a_{i1}, \ldots, a_{is}) \ , \ a_{i.} = \sum_j a_{ij}, \ a_{.j} = \sum_i a_{ij} \ ,$$

$\tilde{a} = (a_{1.}, \ldots, a_{r.})$ e $a_{..} = \sum_i a_{i.} = \sum_j a_{.j}$

(iii) Amostra

$x = (n_{11}, \ldots, n_{1s}, n_{21}, \ldots, n_{2s}, \ldots, n_{r1}, \ldots, n_{rs})$

$\tilde{x}_i = (n_{i1}, \ldots, n_{is})$ e $\tilde{x} = (n_{1.}, \ldots, n_{r.})$.

(iv) Posteriori

$A_{ij} = a_{ij} + n_{ij}$, $A = a+x$, $A_{i.} = a_{i.} + n_{i.}$, $A_{.j} = a_{.j} + n_{.j}$,

$\tilde{A}_i = \tilde{a}_i + \tilde{x}_i$, $\tilde{A} = \tilde{a}+\tilde{x}$ e $A_{..} = n+a_{..}$

Lema 2

Se $\omega|a \sim D_{sr}(a)$, então

(i) $Q_1 \perp\!\!\!\perp Q_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp Q_r \perp\!\!\!\perp P$

(ii) $Q_i|a \sim D_s(\tilde{a}_i)$, $\forall i = 1, \ldots, r$

(iii) $P|a \sim D_r(\tilde{a})$

Demonstração

Usando a propriedade 1 das distribuições de Dirichlet e adaptando a notação para o nosso caso temos que:

Se $Z_{ij} \sim G(a_{ij},1)$ e todos os Z_{ij} são mutuamente independentes, temos que

$$\frac{1}{Z}\,(Z_{\quad}, \ldots, Z_{rs}) \sim \omega$$

$$Q_i \sim \frac{1}{Z_i}\,(Z_{i1}, \ldots, Z_{is}) \quad \text{e} \quad P \sim \frac{1}{Z}\,(Z_{1.}, \ldots, Z_{r.})$$

Usando a propriedade 1 outra vez, concluímos os resultados desejados.

Evidentemente, os resultados acima são válidos quando A_{ij}, $A_{i.}$, $\tilde{A}_i$, A e $\tilde{A}$ substituem a_{ij}, $a_{i.}$, $\tilde{a}_i$, a e $\tilde{a}$, respectivamente. Isto é, estes resultados se aplicam tanto a priori quanto a posteriori visto que, estamos considerando uma classe conjugada de distribuições.

Inicialmente, vamos supor que a verossimilhança (2) esta sendo considerada e que a hipótese de independência deva ser testada. Assim, H_0 é representada pelo sistema

$$\begin{cases} q_{11} = \ldots = q_{r1} \\ \quad . \qquad\qquad . \\ \quad . \qquad\qquad . \\ \quad . \qquad\qquad . \\ q_{1s} = \ldots = q_{rs} \end{cases}$$

que define um subspaço Ω_0 do espaço original Ω. Assim, a hipótese alternativa H_1 é definida pelo espaço $\Omega-\Omega_0$. Para testar a hipótese H_0 contra a alternativa H_1, calculamos as verossimilhanças modificadas que estão incluídas no resultado abaixo.

Resultado 1

Para o teste de independência, com os entes considerados acima, a razão das verossimilhanças modificadas é dada por

$$\frac{\Gamma(a_{..}-sr+s)}{\Gamma(A_{..}-sr+s)}\left[\prod_{i,j}\frac{\Gamma(a_{ij})}{\Gamma(A_{ij})}\right]\left[\prod_{i}\frac{\Gamma(A_{i.})}{\Gamma(a_{i.})}\right]\left[\prod_{j}\frac{\Gamma(A_{.j}-r+1)}{\Gamma(a_{.j}-r+1)}\right]$$

Demonstração

$\frac{1-\xi}{\xi}R_{01} = \frac{f_0(d)}{f_1(d)}$ onde f_0 e f_1 são as verossimilhanças modificadas. Ao utilizarmos como priori a representação do Lema 2 e como verossimilhança o modelo 2, temos a conjunta de (ω,d) dada por

$$h(x)\frac{\Gamma(a_{..})}{\prod_{ij}\Gamma(a_{ij})}\left[\prod_{i,j}q_{ij}^{A_{ij}-1}\right]\left[\prod_{i}p_{i.}^{A_{i.}-1}\right]$$

cuja integral em Ω produz:

$$f_1 = h(x)\frac{\Gamma(a_{..})}{\Gamma(A_{..})}\left[\prod_{i,j}\frac{\Gamma(A_{ij})}{\Gamma(a_{ij})}\right]$$

Com a restrição definida por H_0, a conjunta de (ω,d) sob H_0 é dada por

$$h(x)\ \frac{\left[\prod_{i,j} q_{ij}^{A_{ij}-1}\right]\left[\prod_{i} p_{i.}^{A_{i.}-1}\right]}{\left[\int \prod_{ij} q_{ij}^{a_{ij}-1} d\Omega_0\right]\left[\frac{1}{\Gamma(a_{..})}\prod_{i}\Gamma(a_{i.})\right]}$$

onde

$$\int \prod_{i,j} q_{ij}^{a_{ij}-1}\, d\Omega_0 = C \int \prod_{j} q_{j}^{a_{.j}-r}\, dS,$$

$S = \{(q_1, \ldots, q_S)\ ;\ q_j \geq 0,\ \sum q_j = 1\}$ e C = tamanho de Ω_0 .

A integral em Ω_0 da conjunta será então

$$f_0 = h(x)\ \frac{C \int \prod_{j} q_{j}^{A_{.j}-r}\, dS}{C \int \prod_{j} g_{j}^{a_{.j}-r}\, dS}\left[\prod_{i}\frac{\Gamma(A_{i.})}{\Gamma(a_{i.})}\right]\left[\frac{\Gamma(a_{..})}{\Gamma(A_{..})}\right]$$

e finalmente

$$f_0 = h(x)\frac{\Gamma(a_{..})}{\Gamma(A_{..})}\ \frac{\Gamma(a_{..}-rs+s)}{\Gamma(A_{..}-rs+s)}\left[\prod_{j}\frac{\Gamma(A_{.j}-r+1)}{\Gamma(a_{.j}-r+1)}\right]\left[\prod_{i}\frac{\Gamma(A_{i.})}{(a_{i.})}\right]$$

que dividido por f_1 prova o resultado.

Resultado 2

No caso de considerarmos a priori uma distribuição uniforme, isto é $a_{ij} = 1\ \forall i,j$,

então, se $\xi=1-\xi=\frac{1}{2}$

$$R_{01} = \frac{\prod_i \binom{n_{i.}+s-1}{n_{i.}}}{\binom{n+s-1}{n}} \; \frac{\left(\prod_i n_{i.}!\right)\left(\prod_j n_{.j}!\right)}{n! \prod_{i,j} n_{ij}!} .$$

Este resultado é obtido ao considerarmos, no resultado 1, $a_{ij} = 1$, $a_{..} = rs$, $A_{..} = n+rs$, $a_{i.}=s$, $A_{i.} = n_{i.}+s$, $a_{.j}=r$ e $A_{.j} = n_{.j}+r$. Note que se $r=s=2$, obtemos

$$\frac{(n_{1.}+1)(n_{2.}+1)}{(n+1)} \; \frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{22}}}{\binom{n}{n_{1.}}}$$

que se identifica a fórmula do exemplo 1 (capítulo 1), após uma adaptação de notação.

Para concluir esta seção, desenvolvemos a seguir o teste de homogeneidade. No lugar do modelo (3), vamos considerar um caso mais geral onde a regra de parada, não necessariamente considera a marginal $(n_{1.},\ldots,n_{r.})$ como prefixada.

Aqui o modelo considerado é

$$V(q|x) = h(x) \prod_{i,j} g_{ij}^{n_{ij}}$$

onde $\sum_j q_{ij} = 1, \quad \forall i = 1, \ldots, r.$

Para um exemplo distinto de (3), suponha que 3 urnas com bolas pretas e brancas estão sendo consideradas. Da primeira urna retiramos 4 bolas com reposição ($n_{1.}$ fixado),

da segunda retiramos bolas (com reposição) até encontrarmos 2 pretas (n_{21} fixado) e da terceira retiramos bolas (com reposição) até encontrarmos 3 brancas (n_{32} fixado).

A distribuição conjugada para $V(q|x)$ é a seguinte:

a) $Q_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp Q_r$

b) $Q_i|\tilde{a}_i \sim D_s(\tilde{a}_i)$.

Como esperado, o resultado abaixo relaciona os dois testes.

Resultado 3

Com os entes considerados, a razão R_{01} para o teste de homogeneidade é idêntica a do teste de independência (resultado 1).

Demonstração

A conjunta de (q,d) é dada por

$$h(x)\ \frac{\prod_i \Gamma(a_{i.})}{\prod_{i,j}\Gamma(a_{ij})}\ \prod_{i,j} q_{ij}^{A_{ij}-1},$$

onde $q_{ij} \geq 0$ e $\sum_j q_{ij}=1$. Integrando-se em q obtemos

$$f_1 = h(x)\left[\prod_i \frac{\Gamma(a_{i.})}{\Gamma(A_{i.})}\right]\left[\prod_{i,j}\frac{\Gamma(A_{ij})}{\Gamma(a_{ij})}\right]$$

Sob a hipótese H_0, de igualdade entre as linhas da tabela paramétrica dos q_{ij}, obtemos a verossimilhança

$$f_0 = h(x) \frac{\int \prod_{i,j} q_{ij}^{A_{ij}-1} \, d\Omega_0}{\int \prod_{i,j} g_{ij}^{a_{ij}-1} \, d\Omega_0} =$$

$$= h(x) \left[\prod_j \frac{\Gamma(A_{.j}-r+1)}{\Gamma(a_{.j}-r+1)} \right] \frac{\Gamma(a_{..} - rs+s)}{\Gamma(A_{..}-rs+s)}$$

Tomando-se $\frac{f_0}{f_1}$ obtemos a fórmula do resultado 1.

O resultado 2 apresenta a razão R_{01}, no caso de distribuições uniformes a priori, tanto no caso de independência quanto no caso de homogeneidade. Neste caso, a interpretação clássica é possível. Na verdade, ao considerarmos $a_{ij} = 1$, R_{01} é a razão entre a média das verossimilhanças em Ω_0 sobre a média das verossimilhanças em Ω_1. É claro que para se conhecer o espaço amostral do experimento, h(x) deve ser especificada. A qualidade de um teste clássico está no valor de α (o nível de significância) e, algumas vezes, no valor de $1-\beta$ (o poder do teste). Isto faz com que h(x) determine a constante k com a qual R_{01} será comparada. Este fato mostra como o conhecimento do espaço amostral é fundamental na teoria clássica.

O teste de Bayes por sua vez é definido a partir de uma perda, ℓ_0 e ℓ_1, e de uma priori, (ξ_0,ξ_1), para (H_0,H_1). Estes dois entes são definidos basicamente no espaço paramétrico. Ao compararmos R_{01} com $k = \frac{\ell_1}{\ell_0}$ o conhecimento de h(x) é irrelevante. Contudo, qualquer que seja h(x), sabemos estar minimizando $\alpha + \frac{\xi_1}{\xi_0} k\beta$.

4 - EXEMPLOS

Para ilustrar os testes descritos na seção anterior, apresentamos nesta seção um problema de independência e um de homogeneidade. Aqui iremos considerar que as perdas são iguais e as probabilidades a priori das hipóteses também são iguais. Desta forma, estamos considerando prioridades iguais para as hipóteses em confronto. Para uma possível comparação com testes clássicos, distribuições uniformes em Ω são consideradas a priori.

Exemplo 1 (Everitt, 1977)

A tabela abaixo apresenta os dados de 141 pacientes com tumor cerebral que foram classificados com respeito ao tipo e ao local do tumor. Os três tipos são: A) Benigno, B) maligno e C) outros. Os três locais são:

I - lóbulo frontal, II - lóbulo temporal e III - outros locais.

	A	B	C	
I	23	9	6	38
II	21	4	3	28
III	34	24	17	75
	78	37	26	141

TABELA 4

A regra de parada que poderia indicar o porque da observação de 141 pacientes, ou mesmo das marginais observadas, não foi especificada. Uma análise clássica, necessariamente, exigiria uma suposição sobre a regra de parada; $n = 141$ é fixado previamente, $(n_{1.}, n_{2.}, n_{3.}) = (38, 28, 75)$ é fixado previamente ou $(n_{.1}, n_{.2}, n_{.3}) = (78,37,26)$ é fixado previamente. Por outro lado a análise Bayesiana é indiferente ao tipo de regra usada e a razão R_{01} é independente da escolha de $h(x)$. No caso específico do exemplo, usando-se uma uniforme em Ω, temos:

$$R_{01} = \frac{\binom{40}{2}\binom{30}{2}\binom{77}{2}}{\binom{143}{2}} \quad \frac{26!\ 28!\ 37!\ 38!\ 75!\ 78!}{141!\ 3!\ 4!\ 6!\ 9!\ 17!\ 21!23!\ 24!\ 34!} =$$

$$\cong 3,98.$$

No caso de considerarmos perdas iguais e a priori $Pr\{H_0\} = \frac{1}{2}$, devemos comparar este resultado com 1. Como $3,98 > 1$ aceitamos a hipótese H_0. Somente nos casos onde $k>3,98$ é que rejeitaríamos H_0 e nestes casos estaría

mos afirmando que o erro de 2a. espécie é pelo menos 4 vezes mais importantes que o de 1a. espécie. Isto devido ao fato deste teste minimizar $\alpha + 3{,}98\beta$.

Exemplo 2

Três amostras de 30 pacientes com depressão foram consideradas. Uma das amostras usou placebo, uma a droga A e a restante a droga B. A seguinte tabela apresenta os dados da pesquisa.

TABELA 5

	Melhora	Ñ melhora	
Placebo	8	28	30
A	21	9	30
B	19	11	30

Para este exemplo, no lugar de considerarmos duas hipóteses vamos considerar as seguintes hipóteses hierarquizadas.

H_0: $q_{11} = q_{21} = q_{31}$

H_1: $q_{21} = q_{31}$ (veja figuras 5, 6 e 7)

H_2: $q_{21} \neq q_{31}$

A verossimilhança neste caso é representada por $\binom{30}{x} q_1^x (1-q_1)^{30-x} \binom{30}{y} q_2^y (1-q_2)^{30-y} \binom{30}{z} q_3^z (1-q_3)^{30-z}$ onde $q_i = q_{i1}$,

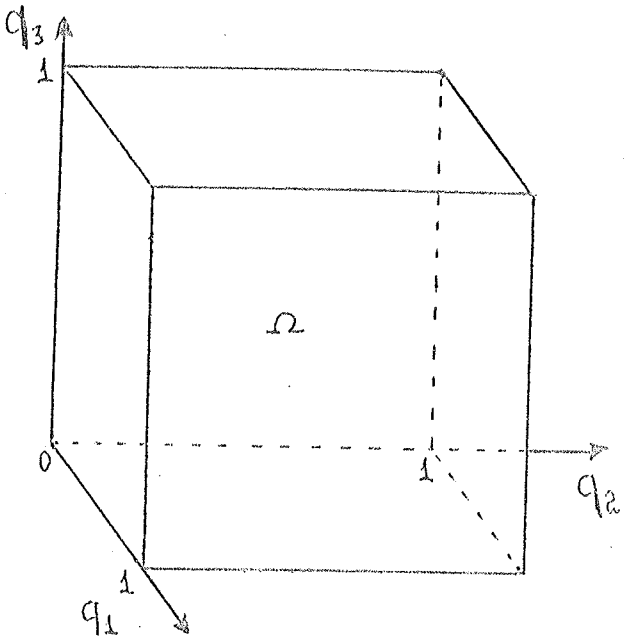


Figura 5 - Espaço Ω do exemplo 2. Representação não singular no R^3.

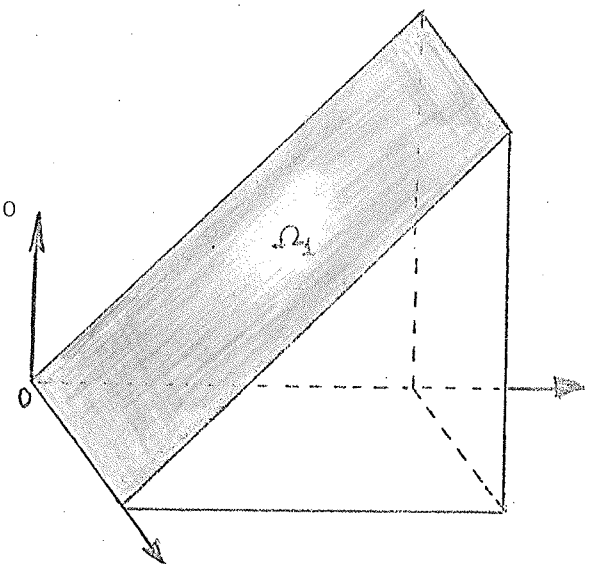


Figura 6 - Superfície Ω_1 do exemplo 2. Representação singular no R^3.

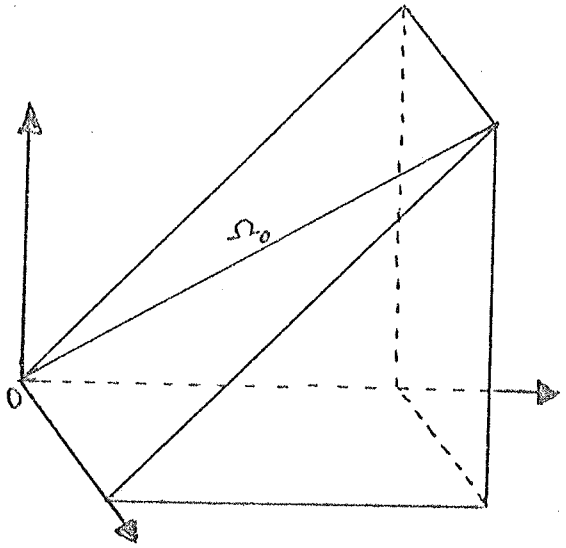


Figura 7 - Corda Ω_0 do exemplo 2. Representação singular no R^3.

$x=n_{11}$, $y=n_{21}$ e $z=n_{31}$. Ao considerarmos uma uniforme no cubo

$$\{(q_1, q_2, q_3);\ 0\leq q_1\leq 1,\ 0\leq q_2\leq 1 \text{ e } 0\leq q_3\leq 1\}\ ,$$

temos as seguintes verossimilhanças modificadas:

$$f_0 = \frac{\binom{30}{x}\binom{30}{y}\binom{30}{z}}{91\ \binom{90}{x+y+z}}$$

$$f_1 = \frac{\binom{30}{x}\binom{30}{y}\binom{30}{z}}{31\binom{30}{x}61\binom{60}{y+z}} = \frac{\binom{30}{y}\binom{30}{z}}{31.61\binom{60}{y+z}}$$

$$f_2 = \left(\frac{1}{31}\right)^3$$

As razões de verossimilhanças modificadas são:

$$R_{01} = \frac{\binom{30}{x}\binom{60}{y+z}}{\binom{90}{x+y+z}}\ \frac{31.61}{91}$$

$$R_{12} = \frac{\binom{30}{y}\binom{30}{z}}{\binom{60}{y+z}}\ \frac{31^2}{61}$$

$$R_{02} = R_{01}\ R_{12} = \frac{\binom{30}{x}\binom{30}{y}\binom{30}{z}}{\binom{90}{x+y+z}}$$

Substituindo os valores de x,y e z vamos obter:

$R_{01} \cong 0,006$, $R_{02} \cong 0,018$

$R_{12} \cong 2,944$ e $R_{10} \cong 163,556$

Estes resultados favorecem claramente a hipótese H_1. Evidentemente, vão existir valores de k_1 e k_2 que irão favorecer H_0.

Uma primeira tentativa de provar a superioridade do teste aqui desenvolvido, foi feita por Irony (1984) que, no caso r=s=2, comparou-o com o teste exato de Fisher. Por meio de amostras simuladas provou que a combinação linear dos erros $\alpha+k\beta$ era menor no teste de Bayes. Além disto, mostrou que o valor efetivo do nível de significância era bem superior ao nível indicado pelo teste de Fisher. Este grave defeito do teste de Fisher, seria corrigido se uma distribuição uniforme (discreta) fosse considerada para a freqüência da marginal. Uma análise complementar deste trabalho esta sendo preparada e em breve enviada para divulgação.

5 - TABELAS 2x2x2

Com o objetivo de não introduzir mais notação, esta seção é restrita ao caso particular, 2x2x2, de tabelas multidimensionais. A análise aqui descrita toma o ca

so onde apenas duas das 3 classificações são respostas. Isto é descrito a seguir.

Suponha 3 classificações binárias A,B e C que produzem a tabela abaixo,

TABELA 6

	A_1	A_2	TOTAL
B_1C_1	x_1	n_1-x_1	n_1
B_2C_1	x_2	n_2-x_2	n_2
B_1C_2	x_3	n_3-x_3	n_3
B_2C_2	x_4	n_4-x_4	n_4

onde $N_1 = n_1 + n_2$ e $N_2 = n_3 + n_4$ foram fixados previamente. Isto implica que o modelo contenha apenas 6 parâmetros, listados a seguir:

$$q_1 = P(A_1|B_1C_1)\ , \quad q_2 = P(A_1|B_2C_1), \quad q_3 = P(A_1|B_1C_2),$$

$$q_4 = P(A_1|B_2C_2)\ , \quad p_1 = P(B_1|C_1) \quad e \quad p_2 = P(B_1|C_2).$$

A distribuição conjugada para o modelo é dado por

$$q_i \sim B(a_i, b_i) \qquad \forall i = 1,2,3,4$$

$$p_j \sim B(a_j, b_j) \qquad \forall j = 1,2$$

e $(q_1, q_2, q_3, q_4, p_1, p_2)$ são mutuamente independentes.

Com estas suposições, a maioria das hipóteses de interesse são casos particulares das tabelas bidimensionais. É interessante descrevermos algumas dessas hipóteses.

A hipótese mais restritiva é a falta total de associação entre A, B e C. Essa hipótese é definida pelo sistema

$$H_0: \quad p_1 = p_2$$
$$q_1 = q_2 = q_3 = q_4$$

O cálculo de f_0 neste caso é obtido usando as fórmulas da seção 3. O mesmo acontecerá com a hipótese de falta de associação entre A e (B,C). Neste caso

$$H_1: q_1 = q_2 = q_3 = q_4.$$

Uma outra hipótese semelhante é a falta de associação marginal entre B e C, isto é

$$H_2: p_1 = p_2 .$$

A falta de associação condicional entre A e B dado C é representada pelo sistema

$$H_3: \quad q_1 = q_2$$
$$q_3 = q_4.$$

O cálculo de f_3 é semelhante ao cálculo de f_1 no exemplo 2.

Ressalte-se aqui que a simplicidade dos cálculos esta relacionada com o fato de todas as hipóteses serem relações lineares entre os parâmetros. Contudo, a hipótese menos restritiva é a hipótese de falta de interação de segunda ordem que representa igualdade de associação condicional entre A e B dado C. Isto é, a associação entre A e B dentro de C_1 é idêntica a associação entre A e B dentro de C_2. Este fato é representado pela hipótese

$$H_4 : \quad q_1 = q$$
$$q_2 = tq$$
$$q_3 = p$$
$$q_4 = tp$$

onde $0 \leq q, p \leq 1$ e $0 < t < \min\left(\frac{1}{p}, \frac{1}{q}\right) = t_0$.

Note que a dimensão de H_4 é 3, pois (q,p,t) substitui (q_1, q_2, q_3, q_4).

Para encontrarmos o tamanho da superfície definida por H_4, temos a seguinte matriz de derivadas

$$\begin{array}{c} \\ q_1 \\ q_2 \\ q_3 \\ q_4 \end{array} \begin{array}{c} \begin{array}{ccc} q & p & t \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ t & 0 & q \\ 0 & 1 & 0 \\ 0 & t & p \end{bmatrix} \end{array}$$

O tamanho da superfície é a integral:

$$\int_0^1 \int_0^1 \int_0^{t_0} \sqrt{q^2+q^2t^2+p^2+p^2t^2}\ dt\ dp\ dq =$$

$$\int_0^1 \int_0^1 \int_0^{t_0} \sqrt{(1+t^2)(p^2+q^2)}\ dt\ dp\ dq = I$$

Considerando a priori uma distribuição uniforme, em Ω, o cálculo de f_4 irá envolver uma integral do tipo

$$\frac{1}{I} \int_0^1 \int_0^1 \int_0^{t_0} \Delta(p,q,t)\ B(p,q,t)\ dt\ dq\ dp$$

onde $\Delta(p,q,t) = \sqrt{(1+t^2)(p^2+q^2)}$

e $B(p,q,t) = p^{A_1}(1-p)^{A_2} q^{A_3}(1-q)^{A_4} t^{A_5}(1-pt)^{A_6}(1-qt)^{A_7}$.

O cálculo desta integral exige métodos numéri-

cos especiais, visto que na maioria das vezes os expoentes são grandes. Esses métodos já estão implantados e seu cálculo é mais simples do que aqueles exigidos pelo teste clássico da hipótese H_4.

É interessante notarmos a hierarquia entre as hipóteses H_0, H_1, H_2, H_3 e H_4.

$$H_0 \quad \begin{matrix} H_1 \rightarrow H_3 \rightarrow H_4 \rightarrow H_5 \\ H_2 \end{matrix}$$

Aqui, a hipótese H_5 é a alternativa geral, onde não existem restrições.

No próximo capítulo, serão desenvolvidos testes para o equilíbrio populacional de Hardy-Weinberg, em diferentes situações genéticas as quais não permitem soluções clássicas adequadas. As hipóteses envolvidas nestes problemas não são lineares.

CAPÍTULO III

EQUILÍBRIO POPULACIONAL

1 - INTRODUÇÃO

A metodologia desenvolvida no capítulo I foi motivada pelos problemas discutidos neste capítulo. Pereira e Rogatko (1984) apresentam uma versão aproximada do teste de equilíbrio na situação genética mais elementar (sistema monogênico, dialélico e codominante). Contudo, o método de construção daquele teste não permite uma extensão adequada para situações mais complexas. O método desenvolvido no presente trabalho, sugerido por Irony (1984), é bastante geral e inclui situações genéticas que desmotivam a utilização de métodos clássicos. Por exemplo, o equilíbrio genético com genes ligados ao sexo. Para facilitar a leitura, revisitaremos alguns conceitos utilizados na seqüência.

Todas as características observáveis nos seres vivos são fruto da interação de componentes genéticos com o ambiente. Aqui, somente os componentes genéticos serão objeto de análise. A informação genética cuja unidade é denominada *gene*, está localizada em materiais genéticos denominados *cromossomos*. As espécies *diplóides*, das quais iremos nos referir, são aquelas onde os

cromossomos aparecem aos pares, *homólogos*. Um dos elementos do par é proveniente do pai e o outro da mãe.

Um gene ocupa um lugar físico específico, *loco*, no cromossomo. Um loco pode ser ocupado por genes que codificam diferentes tipos de informação, *alelos*. Nos cromossomos não relacionados com sexo, *autossônicos*, a característica genética final do indivíduo, *genótipo*, é determinada pelo par de genes presentes no par de homólogos, nos locos correspondentes. O número de diferentes alelos que podem ocupar um determinado loco pode variar com a característica estudada. O caso onde apenas um tipo de alelo ocupa o loco é denominado *monomórfico*. No caso de dois tipos *dialélico* e se existirem mais de dois tipos é denominado *polialélico*.

A característica genética final expressa (externamente) no indivíduo, *fenótipo*, é determinada pelo genótipo de um ou mais locos. No caso de apenas um loco, o sistema é denominado *monogênico*.

Suponha um sistema monogênico e dialético onde G_1 e G_2 são os dois tipos de genes. Os genótipos possíveis são, então, G_1G_1, G_1G_2 e G_2G_2. Se as três classes são fenotipicamente distintas, o sistema é *codominante*. Se G_1G_1 e G_1G_2 produzem o mesmo fenótipo, distinto de G_2G_2, então G_1 é um gene *dominante* e G_2 é *recessivo*. Evidentemente, para um sistema polialélico, poderemos

ter mistura das relações de dominância e codominância entre as várias combinações genotípicas. O sistema sanguíneo ABO é um exemplo relevante de tal mistura.

Nos cromossomos ligados ao sexo, o mecanismo descrito acima sofre uma modificação. Nos elementos do sexo feminino o mecanismo permanece o mesmo, contudo, nos do sexo masculino apenas os cromossomos provenientes da mãe carregam informação genética do tipo descrito acima. O cromossomo do par proveniente do pai, carrega, provavelmente, como única informação a determinação do sexo masculino do indivíduo. Assim, se os genes G_1 e G_2 são ligados ao sexo, produzirão, como antes, os genótipos G_1G_1, G_1G_2 e G_2G_2 na população feminina e apenas duas classes genotípicas, G_1 e G_2, na população masculina.

O conceito de equilíbrio que usaremos na seqüência está incluso na seguinte definição.

Definição 4

Uma população está em *equilíbrio*, segundo uma característica genética, se as freqüências das classes genotípicas, desta característica, não se alteram de uma geração para outra.

A próxima definição inclui um conceito com estreita ligação com o de equilíbrio. Desta ligação originou-se o teorema de Hardy-Weinberg, assim conhecido pe-

los trabalhos independentes de Hardy (1908) e Weinberg (1908).

Definição 5

Uma população está em *pan-mixia* se todos os cruzamentos ocorrem completamente ao acaso. Isto é, não existe estratificação social, nem a tendência de indivíduos do mesmo genótipo casarem preferencialmente entre si, nem a preferência por casamentos consanguíneos e nem a tendência de indivíduos fazerem a escolha de seus cônjuges com base em características físicas como estatura, cor de cabelos, etc.

Além de pan-míticas, as populações a que nos referimos neste capítulo devem satisfazer as seguintes restrições:

1) Possam ser consideradas infinitamente grandes.
2) Tenham a mesma freqüência de homens e mulheres.
3) A característica estudada seja independente da fertilidade dos casais.
4) Não sofram migrações relacionadas com a característica estudada.
5) Os genes não sofram mutação.
6) Os genes não sofram pressão seletiva, isto é, a característica em estudo não seja relacionada com

a sobrevivência dos indivíduos.

Para uma análise mais cuidadosa dessas restrições bem como para um estudo mais aprofundado dos conceitos descritos nesta seção, veja Beiguelman (1977) e Elandt - Johnson (1971).

Nas seções seguintes, os problemas tratados envolvem hipóteses não lineares e parâmetros excedentes que devem ser eliminados. Os cálculos apresentados foram executados no Centro de Computação Eletrônica da USP e os programas foram desenvolvidos pelo Dr. André Rogatko que para os cálculos das integrais envolvidas utilizou o método de interpolação adaptativa de Roemberg.

2 - SISTEMA AUTOSSÔMICO, MONOGÊNICO E DIALÉLICO

Considere como G_1G_1, G_1G_2 e G_2G_2 os genótipos relacionados ao par de alelos G_1 e G_2, não relacionados a sexo. Respectivamente a estas classes genotípicas, p_1, p_2 e p_3 representam as proporções destas na população. Vamos supor que o sistema é codominante, isto é, as classes genotípicas definem classes fenotípicas distintas. Assim, em uma amostra de tamanho n, a freqüência de elementos em cada classe, n_1, n_2 e n_3 onde $n_1+n_2+n_3 = n$, podem ser identificadas. Evidentemente, estamos admitindo que os elementos amostrais são obtidos in-

dependentemente, isto é, segundo um processo de Bernoulli (bivariado). Isto equivale a dizer que a verossimilhança é proporcional a

$$p_1^{n_1} p_2^{n_2} p_3^{n_3},$$

onde $p_1 \geq 0$, $p_2 \geq 0$, $p_3 \geq 0$ e $p_1 + p_2 + p_3 = 1$.

É comum o interesse de geneticistas em verificar se a população em estudo é pan-mítica e não foge das leis mendelianas por fatores adversos como seleção e outros já descritos. Isto equivale a verificar se a população está em equilíbrio. Esta equivalência é conhecida como a lei de Hardy-Weinberg (HW) e está incluída no teorema abaixo.

Teorema 4 - (Hardy-Weinberg)

Em uma população pan-mítica que não se afasta das leis mendelianas, o equilíbrio é obtido em uma etapa de reprodução e as proporções genotípicas satisfazem a seguinte propriedade:

$$\exists p \in (0,1) \text{ t.q. } \quad p_1 = p^2,\ p_2 = 2p(1-p) \text{ e } p_3 = (1-p)^2.$$

Demonstração.

Vamos representar por p_1, p_2 e p_3 as proporções genotípicas de uma determinada geração e por q_1, q_2

e q_3 as respectivas proporções na geração subsequente.

O seguinte quadro descreve o processo de geração de descendentes em uma população com as características desejadas.

TABELA 7

TIPO DE CRUZAMENTO	TIPO DE DESCENDENTE			PROPORÇÃO DOS CRUZAMENTOS
	G_1G_1	G_1G_2	G_2G_2	
$G_1G_1 \times G_1G_1$	p_1^2	0	0	p_1^2
$G_1G_1 \times G_1G_2$	p_1p_2	p_1p_2	0	$2p_1p_2$
$G_1G_1 \times G_2G_2$	0	$2p_1p_3$	0	$2p_1p_3$
$G_1G_2 \times G_1G_2$	$\frac{1}{4}p_2^2$	$\frac{1}{2}p_2^2$	$\frac{1}{4}p_2^2$	p_2^2
$G_1G_2 \times G_2G_2$	0	p_2p_3	p_2p_3	$2p_2p_3$
$G_2G_2 \times G_2G_2$	0	0	p_3^2	p_3^2
PROPORÇÕES NA OUTRA GERAÇÃO	q_1	q_2	q_3	1

Se $p = p_1 + \frac{1}{2}p_2$, podemos escrever

$$q_1 = p^2, \quad q_2 = 2p(1-p) \text{ e } q_3 = (1-p)^2.$$

Com estes valores de q_i substituindo os p_i na

tabela, encontraremos as proporções da geração seguinte que serão:

G_1G_1 : $p^4 + 2p^3(1-p) + p^2(1-p)^2 = p^2 = q_1$

G_1G_2 : $2p^3(1-p)+2p^2(1-p)^2 + 2p^2(1-p)^2 + 2p(1-p)^3=2p(1-p)=q_2$

G_2G_2 : $p^2(1-p)^2 + 2p(1-p)^3 + (1-p)^4 = (1-p)^2 = q_3$.

O teorema então fica demonstrado.

Não é difícil entendermos que o resultado acima pode ser estendido para o caso de polialelismo. Note que no caso de equilíbrio, a proporção do gene do tipo G_1 será $p = \sqrt{p_1}$ e o do tipo G_2 será $1-p=1-\sqrt{p_1} = \sqrt{p_3}$. No caso de três aletos, G_1, G_2 e G_3, teríamos: para G_1, $p=\sqrt{p_1}$; para G_2, $q=\sqrt{p_3}$ e para G_3, $1-p-q=1-\sqrt{p_1}-\sqrt{p_3}$ = $= \sqrt{p_6}$. Aqui, p_1, p_2, p_3, p_4, p_5 e p_6 são, respectivamente, as proporções das classes G_1G_1, G_1G_2, G_2G_2, G_1G_3, G_2G_3 e G_3G_3.

A propriedade de equilíbrio define então uma relação especial entre as proporções p_1,p_2 e p_3. Assim, para o geneticista decidir-se, baseado na amostra descrita anteriormente, sobre a existência ou não do equilíbrio, ele deve testar a hipótese nula, H_0, que (p_1,p_2,p_3) pertença à corda $\Omega_0=\{(x,y,z);\ x\in(0,1), y=2\sqrt{x}(1-\sqrt{x}),\ z=(1-\sqrt{x})^2\}$ contra a alternativa, H_1, de que (p_1,p_2,p_3) pertença ao complemento de Ω_0 em relação ao simplex $\Omega=\{(x,y,z);\ x\geq 0,\ y\geq 0,\ z\geq 0$ e $x+y+z=1\}$. Esses dois conjuntos do R^3 podem ser

representados no R^2 por

$$\Omega_0' = \{(x,z);\ x\in(0,1) \text{ e } z=(1-\sqrt{x})^2\}$$

e

$$\Omega' = \{(x,z);\ x\geq 0,\ z\geq 0 \text{ e } x+z \leq 1\}\ .$$

A figura 8 apresenta os conjuntos Ω_0' e Ω' e a figura 9 ilustra os conjuntos Ω_0 e Ω.

O problema estatístico que o geneticista deve resolver é a construção de um teste para confrontar

$$H_0:\ (p_1,p_2,\ p_3)\ \in\ \Omega_0$$

e

$$H_1:\ (p_1,\ p_2,\ p_3)\ \in\ \Omega-\Omega_0.$$

Definido o processo de seleção amostral, a verossimilhança é dada por

$$v(\omega|d) = h(d)\ p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

onde $d=(n_1,n_2,n_3)$ e $\omega = (p_1,\ p_2,\ p_3)$. A densidade a priori escolhida na classe conjugada é uma $D_3(a_1,a_2,a_3)$ isto é

$$g(\omega) = \frac{\Gamma(a_1+a_2+a_3)}{\Gamma(a_1)\Gamma(a_2)\,\Gamma(a_3)}\ p_1^{a_1-1}\ p_2^{a_2-1}\ p_3^{a_3-1}$$

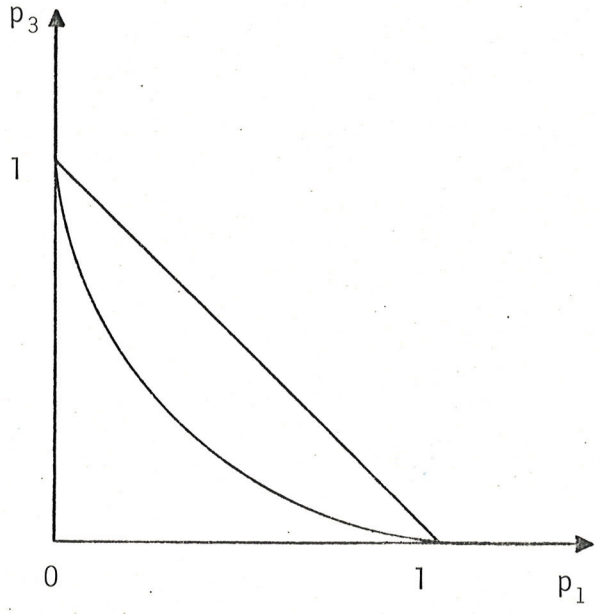


Figura 8 - Representação no R^2. O espaço paramétrico Ω' é a região interna do triângulo. A corda interna que liga os vértices é o espaço Ω'_0.

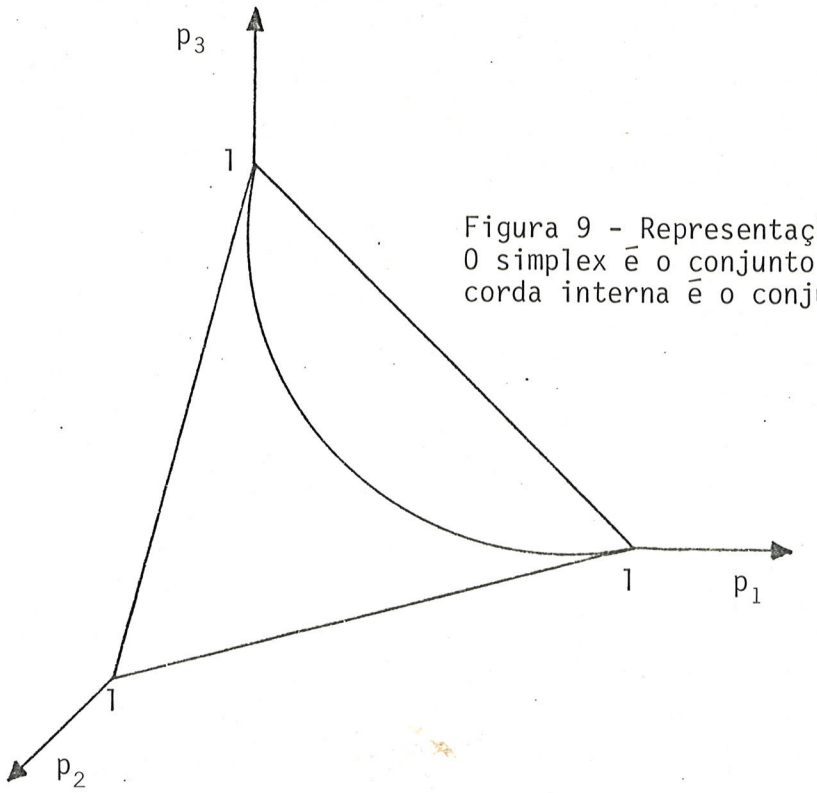


Figura 9 - Representação no R^3. O simplex é o conjunto Ω e a corda interna é o conjunto Ω_0.

onde $(p_1, p_2, p_3) \in \Omega$. Lembremos que a posteriori teríamos $\omega|d \sim D_3(a_1+n_1, a_2+n_2, a_3+n_3)$.

A notação abaixo simplificará nossas fórmulas:

$$B(a_1,a_2,a_3) = \frac{\Gamma(a_1)\Gamma(a_2)\ (a_3)}{\Gamma(a_1+a_2+a_3)}$$

$$E(a_1,a_2,a_3) = \int_0^1 \sqrt{1-3x(1-x)}\ \ x^{\alpha_1-3}(1-x)^{\alpha_2-3}\ dx$$

onde $\alpha_1 = 2a_1+a_2$ e $\alpha_2 = 2a_3+a_2$.

Note que a corda Ω_0 define o sistema de equações $p_1=p^2$, $p_2 = 2p(1-p)$ e $p_3 = (1-p)^2$ cujo vetor de derivadas é

$$(D_1,D_2,D_3) = (2p,\ 2-4p,\ 2p-2).$$

O comprimento de Ω_0 será, então,

$$\int \sqrt{D_1^2+D_2^2+D_3^2}\ dp = 2\sqrt{2} \int_0^1 \sqrt{1-3p(1-p)}\ dp = 2\sqrt{2}\,E(1,1,1) =$$

$$= 1,95186.$$

As funções de verossimilhança modificadas (veja def. 3 do capítulo I) são

$$f_0(n_1,n_2,n_3) = \frac{\int g(\omega)v(\omega|d)d\Omega_0}{\int g(\omega)\ d\Omega_0}$$

e

$$f_1(n_1,n_2,n_3) = \frac{\int g(\omega)v(\omega|d)\ d\Omega}{\int g(\omega)\ d\Omega} .$$

Por outro lado, se $A_1 = a_1+n_1$, $A_2=a_2+n_2$ e $A_3 = a_3+n_3$ temos:

$$\int g(\omega)d\Omega_0 = 2^{a_2}\sqrt{2}\ \frac{E(a_1,a_2,a_3)}{B(a_1,a_2,a_3)} ,$$

$$\int g(\omega)v(\omega|d)d\Omega_0 = 2^{A_2}\sqrt{2}\ \frac{E(A_1,A_2,A_3)}{B(a_1,a_2,a_3)}\ h(n_1,n_2,n_3),$$

$$\int g(\omega)\ d\Omega = 1 \quad e$$

$$f_1 = \int g(\omega)v(\omega|d)\ d\Omega = \frac{B(A_1,A_2,A_3)}{B(a_1,a_2,a_3)}\ h(n_1,n_2,\ n_3)$$

Finalmente temos:

$$f_0(n_1,n_2,n_3) = 2^{n_2}h(n_1,n_2,n_3)\ \frac{E(A_1,A_2,A_3)}{E(a_1,a_2,a_3)}$$

A razão das verossimilhanças modificadas pode então ser escrita como:

$$\frac{f_0}{f_1} = 2^{n_2}\ \frac{E(A_1,A_2,A_3)}{B(A_1,A_2,A_3)}\ \frac{B(a_1,a_2,a_3)}{E(a_1,a_2,a_3)}$$

que multiplicada por $\frac{\xi}{1-\xi}$ nos dá a razão de Bayes para o teste de H_0 contra H_1. O resultado abaixo resume estes aspectos.

Resultado 4

Com os entes considerados, se $\frac{\xi}{1-\xi}$ é a razão de probabilidades a favor de H_0, o teste de Bayes para equilíbrio consiste em comparar

$$R_{01} = \frac{\xi}{1-\xi} \quad \frac{E(A_1,A_2,A_3)}{B(A_1,A_2,A_3)} \quad \frac{B(a_1,a_2,a_3)\ 2^{n_2}}{E(a_1,a_2,a_3)}$$

com uma constante k pré-determinada.

No caso de $\xi = 1-\xi = \frac{1}{2}$ e $a_1 = a_2 = a_3 = 1$, teríamos:

$$R_{01} = \frac{E(n_1+1,n_2+1,n_3+1)}{n_1!\ n_2!\ n_3!} \quad \frac{(n+2)!\ 2^{n_2}}{0,69}\ .$$

Se o teste consiste em comparar R_{01} com a unidade, estaremos com um teste que minimiza $\alpha + \beta$, a soma das médias dos erros de 1a. e 2a. espécies.

Os resultados que apresentaremos se referem ao caso uniforme descrito acima.

Vamos supor que o tamanho da amostra, n, é pré

fixado. Isto é,

$$h(n_1,n_2,n_3) = \frac{n!}{n_1!n_2!n_3!}$$

Com tal processo, as preditivas serão

$$f_0(n_1,n_2,n_3) = 2^{n_2} \frac{n!}{n_1!n_2!n_3!} \frac{E(n_1+1,n_2+1,n_3+1)}{0,69} \text{ e}$$

$$f_1(n_1,n_2,n_3) = \frac{(n+1)(n+2)}{2}$$

Vamos definir a região crítica do teste como

$$C = \{(n_1,n_2,n_3); \frac{f_0}{f_1} < 1 \}$$

e os erros serão

$$\alpha = \sum_{C} f_0(n_1,n_2,n_3)$$

e

$$\beta = 1 - \sum_{C} f_1(n_1,n_2,n_3).$$

A tabela 8 apresenta os valores desses erros para alguns tamanhos de amostra.

TABELA 8

Erros \ n	10	20	30	40	50	60	70	80	90	100
α	0,207	0,131	0,110	0,094	0,085	0,077	0,073	0,065	0,063	0,058
β	0,379	0,355	0,321	0,298	0,275	0,265	0,250	0,243	0,233	0,227

Para ilustrar o tipo de região crítica obtido pelo teste, a tabela 9 descreve o espaço amostral onde os valores de n_1 e n_3 são as entradas das colunas e linhas, respectivamente. O corpo da tabela é preenchido pela parte inteira de R_{01}. Assim os pontos com zeros formam a região crítica e os com inteiros positivos formam a região de aceitação. Se a figura 5, com dimensões adaptadas, sobrepor esta tabela visualizaremos o ótimo ajuste da corda de equilíbrio com a região de aceitação.

Neste ponto devemos nos perguntar se seria possível testarmos equilíbrio no caso de dominância do gene A_1. A amostra observada neste caso será (n_1+n_2, n_3) e a verossimilhança seria proporcional a

$$p_3^{n_3}(p_1+p_2)^{n_1+n_2}.$$

Para evitar mais um desenvolvimento algébrico, vamos analisar este problema apenas no seu lado intuitivo.

Tabela 9 - Espaço amostral particionado
pelos valores inteiros da razão de Bayes.
n = 50

Evidentemente, para todo valor de p_3, vai existir um de $p\in(0,1)$ tal que $p_3=(1-p)^2$ e logo $p_1+p_2=1-(1-p)^2$ o que não implica que $p_1=p^2$. Assim, a amostra pode trazer nenhuma informação sobre o equilíbrio. Todas as informações, possíveis de serem analizadas, estão contidas na priori. Assim, um estudo sobre as relações entre a_1, a_2 e a_3 é que poderá auxiliar nas conclusões sobre equilíbrio.

Existem ainda exemplos onde apenas parte dos n_1 elementos e parte dos n_2 elementos, são classificados em suas respectivas classes, visto que o processo de identificação é muito caro. A verossimilhança nesses casos será proporcional a

$$p_1^{N_1}\ p_2^{N_2}\ p_3^{n_3}\ (p_1+p_2)^{N_3}$$

onde $N_1+N_2+N_3=n_1+n_2$. Problemas como este são tratados em Basu & Pereira (1982). Contudo, não temos conhecimento de trabalhos que desenvolvem testes para tais casos de dados incompletos. Com a metodologia aqui desenvolvida, os testes são construídos seguindo as mesmas linhas daqueles desenvolvidos até aqui. Evidentemente, as integrais envolvidas deverão ser calculadas numericamente.

Finalizamos esta seção com um cálculo aproximado da razão de Bayes, R_{01}. Utilizando o desenvolvimento em série de Taylor da função $[1-3x(1-x)]^{1/2}$ no ponto $x=\frac{1}{2}$

e utilizando apenas os três primeiros termos (até a segunda derivada), vamos considerar a aproximação

$$[1-3x(1-x)]^{1/2} \sim \frac{1}{2} + 3(x-\frac{1}{2})^2 = \frac{5}{4} - 3x(1-x).$$

O cálculo de $E(a_1,a_2,a_3)$, usando esta expressão, será reduzido à seguinte forma:

$$E(a_1,a_2,a_3) \sim \frac{5}{4} B_2(2a_1+a_2-2,2a_3+a_2-2) -$$

$$-3B_2(2a_1+a_2-1,2a_3+a_2-1),$$

onde $B_2(x,y)$ é a função beta no ponto (x,y).

Como normalmente a_1,a_2 e a_3 são inteiros, podemos escrever

$$E(a_1,a_2,a_3) \sim B_2(2a_1+a_2-2,2a_3+a_2-2)[\frac{5}{4} - 3AR(1-R)]$$

onde $A = \dfrac{2(a_1+a_2+a_3)-4}{2(a_1+a_2+a_3)-3}$ e $R = \dfrac{2a_1+a_2-2}{2(a_1+a_2+a_3)-4}$.

No caso de uma densidade uniforme a priori teríamos a seguinte aproximação para R_{01}:

$$R_{01}(\frac{1-\xi}{\xi}) \cong \frac{(n+2)!}{n_1!n_2!n_3!} \frac{(2n_1+n_2)!(2n_3+n_2)!\ 2^{n_2-1}}{(2n+1)!\ 3} \times$$

$$\left(5 - 12 \frac{2n+2}{2n+3} \left[\frac{2n_1+n_2+1}{2n+2} \frac{2n_3+n_2+1}{2n+2} \right] \right)$$

Note que $2n$ é o número total de genes na amostra, $2n_1+n_2$ e $2n_3+n_2$ são as freqüências (gênicas) amostrais dos genes G_1 e G_2, respectivamente, e finalmente n_1, n_2 e n_3 são as freqüências genotípicas da amostra. O nosso teste, então, relaciona freqüências gênicas com freqüências genotípicas e isto é uma propriedade esperada em qualquer bom teste.

3 - SISTEMA DE GENES LIGADOS AO SEXO, MONOGÊNICO E DIALÉLICO.

Nesta seção vamos supor que os genes G_1 e G_2 são ligados ao sexo. Isto equivale a dizer que a população masculina é particionada em apenas duas classes genotípicas, representadas por G_1 e G_2, e a população feminina, como antes, é particionada nas três classes G_1G_1, G_1G_2 e G_2G_2. Esta estrutura define duas subpopulações distintas (masculina e feminina) que através de seu cruzamento surgirão duas novas subpopulações. Na população feminina, como antes, p_1, p_2 e p_3 vão representar as proporções fenotípicas. Na população masculina q_1 e q_2, $q_1+q_2=1$, são as proporções das classes genotípicas G_1 e G_2.

Vamos agora supor que uma amostra casual de tamanho m é retirada da população feminina e uma amostra de tamanho n é retirada da população masculina. Repre-

sentando as freqüências das classes genotípicas por m_1, m_2 e m_3 na amostra de fêmeas (mulheres) e por n_1 e n_2 na amostra de machos (homens), a verossimilhança será

$$v(\omega|d) \alpha p_1^{m_1} p_2^{m_2} p_3^{m_3} q_1^{n_1} q_2^{n_2}$$

onde $=(p_1,p_2,p_3,q_1,q_2)$, $d=(m_1,m_2,m_3,n_1,n_2)$ e α substitui "proporcional à".

A distribuição a priori conjugada é descrita como

$$(p_1,p_2,p_3) \perp\!\!\!\perp (q_1,q_2) ,$$

$$(p_1,p_2,p_3) \sim D_3(a_1,a_2,a_3) \quad e$$

$$(q_1q_2) \sim D_2(b_1,b_2).$$

Lembremos que $D_2(b_1,b_2)$ é a representação singular da Beta com parâmetro (b_1,b_2). A posteriori teremos

$$(p_1,p_2,p_3) \perp\!\!\!\perp (q_1,q_2) \,|d ,$$

$$(p_1,p_2,p_3)|d \sim D_3(A_1,A_2,A_3) \quad e$$

$$(q_1,q_2)|d \sim D_2(B_1,B_2)$$

onde $A_1 = a_1+m_1$, $A_2= a_2+m_2$, $A_3= a_3+m_3$, $B_1=b_1+n_1$ e $B_2=b_2+n_2$.

Como já vimos, o interesse do geneticista é verificar se a população em estudo é pan-mítica e não se afasta das leis mendelianas.

O resultado abaixo mostra que no presente caso esta hipótese não é equivalente à hipótese de equilíbrio.

Teorema 5

Em uma população pan-mítica que não se afasta das leis mendelianas as proporções genotípicas satisfazem à seguinte propriedade:

$\exists s \in (0,1)$ e $t \in (0,1)$ tais que

$p_1 = st$, $p_2 = (1-s)t + s(1-t)$, $p_3 = (1-s)(1-t)$

$q_1 = t$ e $q_2 = 1-t$.

O equilíbrio ocorre se e somente se $s=t$.

Demonstração

Representando por $(p_1, p_2, p_3, q_1, q_2)$ o vetor de proporções genotípicas de uma determinada geração, seja $(P_1, P_2, P_3, Q_1, Q_2)$ o vetor correspondente na geração subsequente.

O processo de geração de descendentes é descrito pelo seguinte quadro:

TABELA 10

TIPO DE CRUZAMENTO	TIPO DE DESCENDENTE					PROPORÇÃO DOS CRUZAMENTOS
	G_1G_1	G_1G_2	G_2G_2	G_1	G_2	
$G_1G_1 \times G_1$	$\frac{1}{2} p_1q_1$	0	0	$\frac{1}{2} p_1q_1$	0	p_1q_1
$G_1G_1 \times G_2$	0	$\frac{1}{2} p_1q_2$	0	$\frac{1}{2} p_1q_2$	0	p_1q_2
$G_1G_2 \times G_1$	$\frac{1}{4} p_2q_1$	$\frac{1}{4} p_2q_1$	0	$\frac{1}{4} p_2q_1$	$\frac{1}{4} p_2q_1$	p_2q_1
$G_1G_2 \times G_2$	0	$\frac{1}{4} p_2q_2$	$\frac{1}{4} p_2q_2$	$\frac{1}{4} p_2q_2$	$\frac{1}{4} p_2q_2$	p_2q_2
$G_2G_2 \times G_1$	0	$\frac{1}{2} p_3q_1$	0	0	$\frac{1}{2} p_3q_1$	p_3q_1
$G_2G_2 \times G_2$	0	0	$\frac{1}{2} p_3q_2$	0	$\frac{1}{2} p_3q_2$	p_3q_2
PROPORÇÃO NA OUTRA GERAÇÃO	$\frac{1}{2} P_1$	$\frac{1}{2} P_2$	$\frac{1}{2} P_3$	$\frac{1}{2} Q_1$	$\frac{1}{2} Q_2$	1

Se $t = p_1 + \frac{1}{2} p_2$ e $s = q_1$, temos

$P_1 = st$, $P_2 = (1-s)\,t + s(1-t)$, $P_3 = (1-s)(1-t)$

$Q_1 = t$ e $Q_2 = (1-t)$, o que prova a primeira parte do teorema. Note também que se $s=t$, então, $p_1 = s^2$, $p_2 = 2s(1-s)$, $p_3 = (1-s)^2$, $q_1 = s$ e $q_2 = 1-s$, assim, $P_1 = p_1$, $P_2 = p_2$, $P_3 = p_3$, $Q_1 = s^2 + s(1-s) = q_1$ e $Q_2 = q_2$. Isto prova que o equilíbrio ocorre se e só se esse sistema é satisfeito.

Este resultado introduz dois tipos de hipótese: Pan-mixia e Equilíbrio. Os conjuntos envolvidos são apresentados a seguir.

Para possibilitar uma representação gráfica, os conjuntos envolvidos serão definidos no R^3, visto que $p_2=1-p_1-p_3$ e $q_2=1-q_1$. O espaço paramétrico para (p_1, p_3, q_1) é o conjunto

$$\Omega = \{(x,y,z);\ x\geq 0,\ y\geq 0,\ x+y\leq 1 \text{ e } 0\leq z\leq 1\} .$$

A hipótese de pan-mixia é representada pela superfície

$$\Omega_1 = \{(x,y,z);\ 0\leq x,y,z\leq 1,\ \ y=(1-z)(1-\frac{x}{z}),\ x+y\leq 1\}.$$

Finalmente a hipótese de equilíbrio é representada pela corda

$$\Omega_0 = \{(x,y,z);\ 0\leq z\leq 1,\ \ x=z^2 \text{ e } y=(1-z)^2\} .$$

As figuras 10 e 11 descrevem os conjuntos Ω_0, Ω_1 e Ω.

Aqui, o problema estatístico do geneticista é a construção de um teste para confrontar as hipóteses

$H_0:(p_1,p_2,q_1) \in \Omega_0$ - Equilíbrio

$H_1:(p_1,p_3,q_1) \in \Omega_1-\Omega_0$ - Pan-mixia

$H_2:(p_1,p_3,q_1) \in \Omega-\Omega_1$ - Alternativa

Voltando ao espaço original, R^5, a corda Ω_0 é caracterizada pelas equações $p_1=s^2$, $p_2=2s(1-s)$, $p_3=(1-s)^2$,

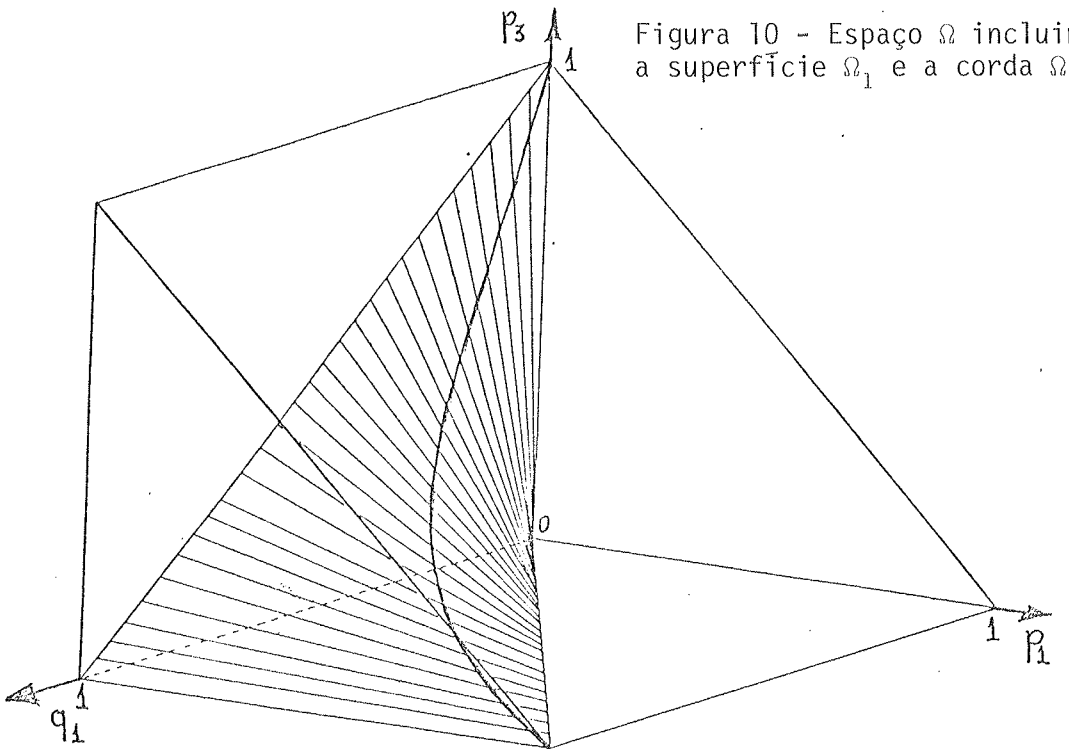


Figura 10 - Espaço Ω incluindo a superfície Ω_1 e a corda Ω_0.

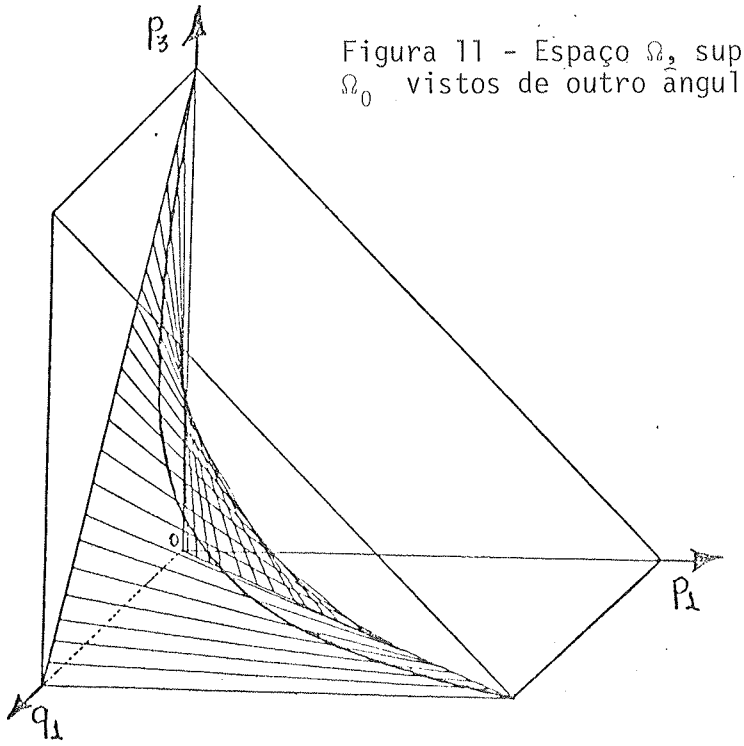


Figura 11 - Espaço Ω, superfície Ω_1 e a corda Ω_0 vistos de outro ângulo.

$q_1 = s$ e $q_2 = 1-s$ que produzem o seguinte vetor de derivadas: $(2s, 2-4s, -2(1-s), 1, -1)$, cuja soma de quadrados pode ser escrita como $2\{1+4[1-3s(1-s)]\}$. O comprimento da corda, no R^5, será então

$$\sqrt{2}\int_0^1 \sqrt{1+4[1-3s(1-s)]}\,ds = \sqrt{2}\,E(1,1,1,1,1)=2,424$$

onde $E(a_1,a_2,a_3,b_1,b_2) = \int_0^1 \sqrt{1+4[1-3s(1-s)]}\; s^{\alpha-4}(1-s)^{\beta-4}ds,$

$$\alpha=2a_1+a_2+b_1 \quad e \quad \beta=2a_3+a_2+b_2.$$

Ainda no R^5, a superfície Ω_1 é caracterizada pelo sistema

$$p_1=st,\ p_2=s(1-t)+(1-s)t,\ p_3=(1-s)(1-t)\ ,$$

$$q_1 = t \quad e \quad q_2 = 1-t\ ,$$

cuja matriz de derivadas é dada por:

$$\begin{bmatrix} t & 1-2t & t-1 & 0 & 0 \\ s & 1-2s & s-1 & 1 & -1 \end{bmatrix}$$

Os determinantes de ordem 2 serão

$$D_1= t-s\ ,\ D_2= s-t,\ D_3=t\ ,\ D_4= -t\ ,$$

$$D_5= t-s \quad D_6=1-2t \quad D_7= 2t-1$$

$$D_8= t-1 \quad D_9=1-t \quad D_{10}=0$$

cuja soma de quadrados será

$$4\left[1-3t(1-t)+3\left(\frac{t-s}{2}\right)^2\right].$$

A área da superfície Ω_1, no R^5, será o valor da integral

$$2\int_0^1\int_0^1\sqrt{1-3t(1-t)+3\left(\frac{t-s}{2}\right)^2}\,ds\,dt = 2F(1,1,1,1,1) = 3,76$$

onde

$$F(a_1,a_2,a_3,b_1,b_2) = \int_0^1\int_0^1\sqrt{1-3t(1-t)+3\left(\frac{t-s}{2}\right)^2}\left[(1-t)s+t(1-s)\right]^{a_2-1} x$$

$$x\ t^{a_1+b_1-2}(1-t)^{a_3+b_2-2}s^{a_1-1}(1-s)^{a_3-1}\,ds\,dt.$$

Definido o processo de seleção da amostra, a verossimilhança é dada por

$$v(\omega|d) = h(d)\ p_1^{m_1}p_2^{m_2}p_3^{m_3}q_1^{n_1}q_2^{n_2}$$

com $(p_1,p_3,q_1) \in \Omega$ e $p_1+p_2+p_3 = q_1+q_2 = 1$.

Usando a distribuição conjugada a priori, as verossimilhanças modificadas, correspondentes a H_0, H_1 e H_2 serão:

$$f_0 = 2^{m_2}h(d)\ \frac{E(A_1,A_2,A_3,B_1,B_2)}{E(a_1,a_2,a_3,b_1,b_2)}$$

$$f_1 = h(d)\ \frac{F(A_1,A_2,A_3,B_1,B_2)}{F(a_1,a_2,a_3,b_1,b_2)} \quad e$$

$$f_2 = h(d)\ \frac{B(A_1,A_2,A_3)B_2(B_1,B_2)}{B(a_1,a_2,a_3)B_2(b_1,b_2)}$$

Com estas funções, o teste é obtido usando-se o item b do Lema 1. Este teste está incluído no resultado abaixo.

Resultado 5

Sendo ξ_0, ξ_1 e ξ_2 as probabilidades a priori das hipóteses H_0, H_1 e H_2, respectivamente, o teste de Bayes para equilíbrio e pan-mixia consiste em comparar, respectivamente,

$$R_{01} = \frac{\xi_0}{\xi_1} \frac{E(A_1,A_2,A_3,B_1,B_2)}{F(A_1,A_2,A_3,B_1,B_2)} \frac{F(a_1,a_2,a_3,b_1,b_2)}{E(a_1,a_2,a_3,b_1,b_2)} 2^{m_2},$$

$$R_{02} = \frac{\xi_0}{\xi_2} \frac{E(A_1,A_2,A_3,B_1,B_2)}{B(A_1,A_2,A_3)B_2(B_1,B_2)} \frac{B(a_1,a_2,a_3)B_2(b_1,b_2)}{B(a_1,a_2,a_3,b_1,b_2)} 2^{m_2} \text{ e}$$

$$R_{12} = \frac{\xi_1}{\xi_2} \frac{F(A_1,A_2,A_3,B_1,B_2)}{B(A_1,A_2,A_3)B_2(B_1,B_2)} \frac{B(a_1,a_2,a_3)B_2(b_1,b_2)}{E(a_1,a_2,a_3,b_1,b_2)}$$

com as constantes C_{01}, C_{02} e C_{12}, pré-fixadas de acordo com o teorema 3a. do capítulo I.

No caso onde m e n são previamente fixados, a verossimilhança é o produto de uma trinomial por uma binomial. A tabela 11 apresenta os erros para alguns casos onde m=n, $\xi_0=\xi_1=\xi_2=\frac{1}{3}$ e $a_1=a_2=a_3=b_1=b_2=1$. As razões neste caso são reduzidas às seguintes expressões:

$$R_{01} = \frac{E(m_1+1, m_2+1, m_3+1, n_1+1, n_2+1)}{F(m_1+1, m_2+1, m_3+1, n_1+1, n_2+1)} \quad \frac{1,881}{1,714}^{m_2}$$

$$R_{02} = \frac{E(m_1+1, m_2+1, m_3+1, n_1+1, n_2+1)}{m_1!\ m_2!\ m_3!\ n_1!\ n_2!} \quad \frac{[(m+2)!]^2 2^{m_2}}{3,427(m+2)}$$

A decisão será favorável a H_0 se $R_{01} \geq 1$ e $R_{02} \geq 1$; será favorável a H_1 se $R_{01} < 1$ e $R_{02} \geq R_{01}$ e; será favorável a H_2 se $R_{02} < 1$ e $R_{02} < R_{01}$. As tabelas de 12 a 22 descrevem as decisões de cada ponto amostral, para n = m = = 10, ao utilizarmos esta regra. Evidentemente, as fórmulas acima podem ser reduzidas caso se utilizem métodos padrões de aproximação.

TABELA 11

ERRO	TAMANHO DA AMOSTRA n=m					
	5	10	15	20	25	30
α_1	0,2073	0,2116	0,1977	0,1867	0,1822	0,1699
α_2	0,2857	0,2534	0,2151	0,1960	0,1782	0,1556
β_0	0,0301	0,0136	0,0131	0,0133	0,0150	0,0180
β_2	0,0476	0,0441	0,0469	0,0511	0,0584	0,0676
γ_0	0,2147	0,1313	0,1015	0,0761	0,0592	0,0498
γ_1	0,1371	0,1089	0,1043	0,0931	0,0838	0,0781
TOTAL	0,9225	0,7629	0,6786	0,6163	0,5768	0,5390

TABELA 12: $n_1 = 0$

m_3 ↑	0	1	2	3	4	5	6	7	8	9	10
10	E										
9	E	A									
8	E	E	A								
7	E	E	A	A							
6	E	E	A	A	A						
5	P	A	A	A	A	A					
4	P	A	A	A	A	A	A				
3	P	A	A	A	A	A	A	A			
2	P	A	A	A	A	A	A	A	A		
1	P	A	A	A	A	A	A	A	A	A	
0	P	P	A	A	A	A	A	A	A	A	A

m_1 →

E = Equilíbrio, P=Pan-mixia e A=Alternativa.

Tabela 13: $n_1 = 1$

	0	1	2	3	4	5	6	7	8	9	10
10	E										
9	E	A									
8	E	E	A								
7	E	E	A	A							
6	E	E	A	A	A						
5	E	E	A	A	A	A					
4	E	E	A	A	A	A	A				
3	P	P	A	A	A	A	A	A			
2	P	P	A	A	A	A	A	A	A		
1	P	P	A	A	A	A	A	A	A	A	
0	P	P	A	A	A	A	A	A	A	A	A

TABELA 14 : $n_1 = 0$

	0	1	2	3	4	5	6	7	8	9	10
10	E										
9	E	A									
8	E	E	A								
7	E	E	A	A							
6	E	E	A	A	A						
5	E	E	E	A	A	A					
4	E	E	E	A	A	A	A				
3	E	E	E	A	A	A	A	A			
2	A	P	A	A	P	A	A	A	A		
1	A	A	A	A	A	A	A	A	A	A	
0	A	A	A	A	A	A	A	A	A	A	A

Tabela 15: $n_1 = 3$

	0	1	2	3	4	5	6	7	8	9	10
10	A										
9	E	A									
8	E	A	A								
7	E	E	A	A							
6	E	E	A	A	A						
5	E	E	E	A	A	A					
4	E	E	E	A	A	A	A				
3	E	E	E	E	A	A	A	A			
2	A	E	E	E	A	A	A	A	A		
1	A	A	A	A	A	A	A	A	A	A	
0	A	A	A	A	A	A	A	A	A	A	A

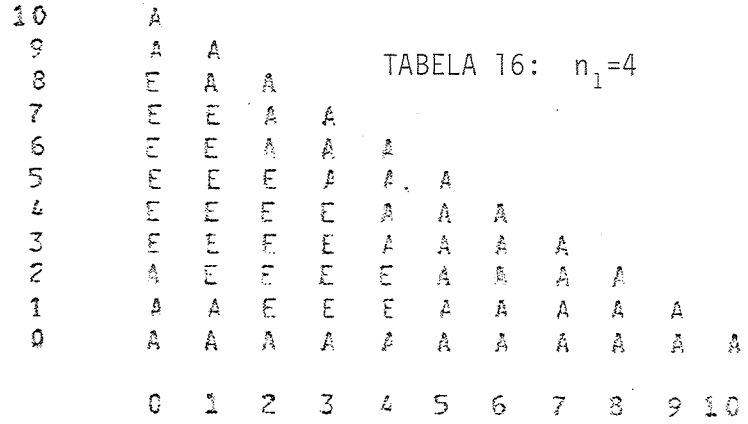


TABELA 16: n_1=4

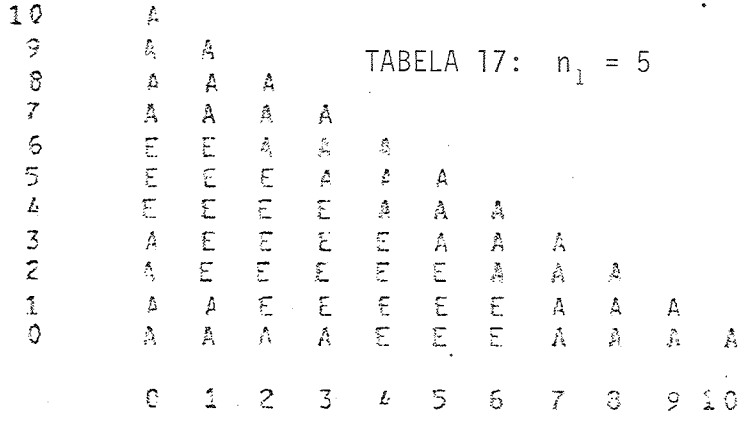


TABELA 17: n_1 = 5

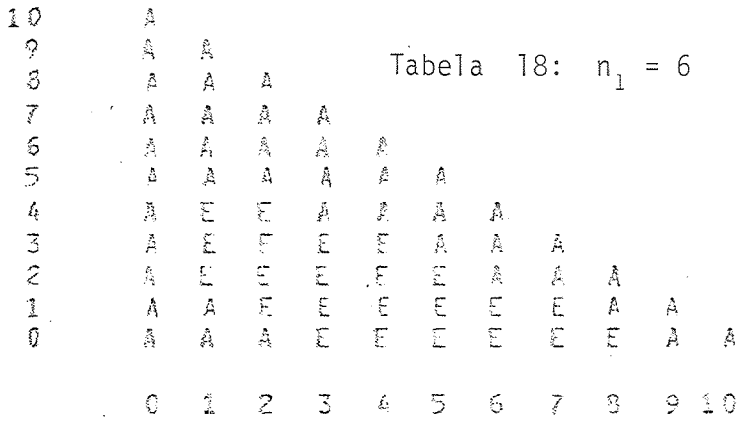


Tabela 18: n_1 = 6

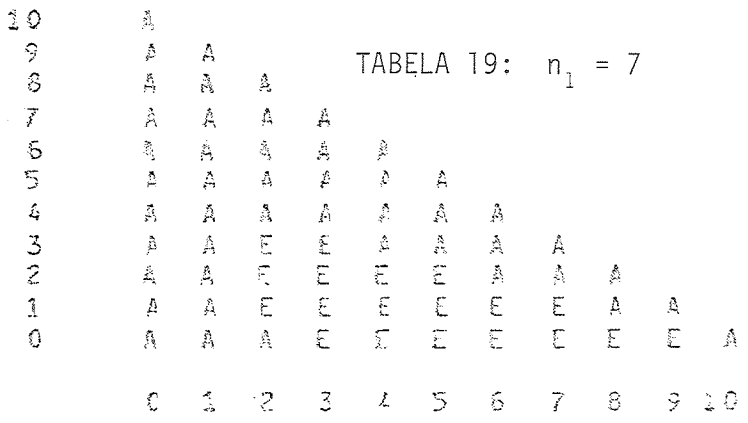


TABELA 19: n_1 = 7

TABELA 20: $n_1 = 8$

10	A										
9	A	A									
8	A	A	A								
7	A	A	A	A							
6	A	A	A	A	A						
5	A	A	A	A	A	A					
4	A	A	A	A	A	A	A				
3	A	A	A	A	A	A	A	A			
2	A	A	A	E	E	E	A	A	A		
1	A	A	P	E	E	E	E	E	E	A	
0	A	A	A	E	E	E	E	E	E	E	E
	0	1	2	3	4	5	6	7	8	9	10

TABELA 21: $n_1 = 9$

10	A										
9	A	A									
8	A	A	A								
7	A	A	A	A							
6	A	A	A	A	A						
5	A	A	A	A	A	A					
4	A	A	A	A	A	A	A				
3	A	A	A	A	A	A	A	A			
2	A	A	A	A	A	A	A	A	A		
1	P	P	P	P	E	E	E	E	E	A	
0	P	P	P	P	E	E	E	E	E	E	E
	0	1	2	3	4	5	6	7	8	9	10

TABELA 22: $n_1 = 10$

10	A										
9	A	A									
8	A	A	A								
7	A	A	A	A							
6	A	A	A	A	A						
5	A	A	A	A	A	A					
4	A	A	A	A	A	A	A				
3	A	A	A	A	A	A	A	A			
2	A	A	A	A	A	A	A	A	A		
1	P	A	A	A	A	A	E	E	E	A	
0	P	P	P	P	P	P	E	E	E	E	E
	0	1	2	3	4	5	6	7	8	9	10

OBSERVAÇÃO: Nestas tabelas, as entradas das colunas são os valores de m_1 e das linhas os de m_3. Se para uma trinca (n_1, m_1, m_3) corresponder a letra E, aceita-se o equilíbrio; a letra P, aceita-se pan-mixia e, a letra A, rejeita-se pan-mixia e equilíbrio.

Note que todos os testes aqui descritos são testes exatos. Para que esses testes possam ser largamente utilizados e difundidos necessário se torna o desenvolvimento de uma forte infra-estrutura computacional. É importante aqui lembrarmos das dificuldades que encontraríamos para a utilização de um método clássico, que para este problema específico é construído para grandes amostras, não sendo portanto exato. Veja Elandt - Jonson (1971) para uma descrição desses métodos.

4 - GRUPO SANGÜINEO ABO

Como foi comentado na seção 2, não seria difícil provar que, em um sistema autossômico, monogênico e polialélico, equilíbrio e pan-mixia são conceitos equivalentes. Isto é, a lei de Hardy-Weinberg é satisfeita, mesmo em sistema polialélicos. Assim, a construção do teste de Bayes em tal sistema, no caso de codominância total, seria um simples exercício algébrico, pois, as etapas a seguir são aquelas descritas na seção 2. Assim, decidimos descrever aqui apenas o caminho que deve ser seguido no caso onde exista mistura de relações de dominância e codominância. Mais especificamente, apresentaremos o problema do teste de equilíbrio do grupo sangüíneo conhecido como ABO.

No grupo sangüíneo ABO, os três alelos, não relacionados a sexo, A,B e O produzem os 6 genótipos AA, AO, BB, BO, AB e OO. Contudo, apenas 4 fenótipos são obtidos da seguinte forma:

TABELA 23

FENÓTIPO	GENÓTIPO	PROPORÇÃO FENOTÍPICA
A	AA ou AO	p_1
B	BB ou BO	p_2
AB	AB	p_3
O	OO	p_4

Com referência a este grupo sangüíneo, uma população estará em equilíbrio H-W se, considerando s,t e (1-s-t) como as freqüências gênicas, as freqüências genotípicas são dadas por

Genótipo	AA	AO	BB	BO	AB	OO
Proporção	s^2	$2s(1-t-s)$	t^2	$2t(1-t-s)$	$2st$	$(1-s-t)^2$

Como apenas as classes fenotípicas são observáveis, considere uma amostra de tamanho n onde as freqüências referentes a estas classes são n_1, n_2, n_3 e n_4, $n_1+n_2+n_3+n_4=n$. A verossimilhança será então proporcional a $\prod_{i=1}^{4} p_i^{n_i}$ mostrando que o parâmetro identificável é

$$\omega = (p_1, p_2, p_3, p_4) \in \Omega$$

onde Ω é o simplex

$$\{(p_1, p_2, p_3, p_4) \; ; \quad p_i \geq 0 \, , \, p_1 + p_2 + p_3 + p_4 = 1\} \; .$$

Testar o equilíbrio, corresponde a testar a hipótese H_0: $\omega \in \Omega_0$, onde $\Omega_0 (\subset \Omega)$ é um conjunto definido pelo sistema

$$p_1 = 2s - s^2 - 2st$$

$$p_2 = 2t - t^2 - 2st$$

$$p_3 = 2st$$

$$p_4 = (1-s-t)^2 \; ,$$

contra a alternativa $H_1 : \omega \in \Omega - \Omega_0$. A matriz das derivadas do sistema será

$$\begin{bmatrix} 2(1-s-t) & -2t & 2t & -2(1-s-t) \\ -2s & 2(1-s-t) & 2s & -2(1-s-t) \end{bmatrix}$$

e para a soma dos quadrados dos determinantes de ordem 2 teremos

$$16\{3(1-s)^2 (1-t)^2 - 4(1-s-t)[s(1-s)^2 + t(1-t)^2]\} = 16\Delta^2$$

O volume do simplex Ω é dado por

$$\int d\Omega = \frac{\sqrt{4}}{3!} = \frac{1}{3} \; .$$

A área da superfície Ω_0 será então

$$\int d\Omega_0 = 4 \int_0^1 \int_0^{1-s} \Delta ds\, dt = 1{,}06564$$

Ao considerarmos como priori para ω uma distribuição de Dirichlet, $D_4(a_1,a_2,a_3,a_4)$, podemos calcular as verossimilhanças modificadas. Seja

$$E(a_1,a_2,a_3,a_4) = \int_0^1 \int_0^{1-s} \Delta g(s,t)\, ds\, dt \quad \text{onde}$$

$$g(s,t) = (2s-s^2-2st)^{a_1-1}(2t-t^2-2st)^{a_2-1}(2st)^{a_3-1}(1-s-t)^{2a_4-2}.$$

Assim, teremos

$$f_0(d) = \frac{h(d)E(A_1,A_2,A_3,A_4)}{E(a_1,a_2,a_3,a_4)}$$

e

$$f_1(d) = \frac{h(d)B(A_1,A_2,A_3,A_4)}{B(a_1,a_2,a_3,a_4)}$$

onde $A_1 = a_1+n_1$, $A_2 = a_2+n_2$, $A_3 = a_3+n_3$ e $A_4 = a_4+n_4$. Caso a distribuição uniforme seja, como antes, a distribuição escolhida como priori das freqüências genotípicas, teríamos $a_1=a_2=2$ e $a_3=a_4=1$. Se além disto, n é pré-fixado, isto é, $h(d) = \dfrac{n!}{n_1!n_2!n_3!n_4!}$, obteríamos

$$f_1(d) = (n_1+1)(n_2+1)\binom{n+5}{5}^{-1}$$

Resultado 6

No caso de $\xi_0=\xi_1 = \frac{1}{2}$ e de considerar-se uma uniforme para as proporções genotípicas, teríamos como razão de Bayes a expressão

$$R_{01} = \frac{E(n_1+2,n_2+2,n_3+1,n_4+1)\ (n+5)\ !}{E(2,2,1,1)(n_1+1)!(n_2+1)!n_3!n_4!5!} .$$

Para ilustrar o uso de R_{01}, considere o exemplo abaixo:

Exemplo

Em uma amostra da população nordestina brasileira encontrou-se as seguintes freqüências genotípicas: $n_1= 72$, $n_2=26$, $n_3= 7$ e $n_4= 107$. Com esta amostra encontramos

$$R_{01} = 2,3029$$

o que favorece a hipótese de equilíbrio. Assim, caso, decidamos que o teste deve minimizar $\alpha+\beta$, o equilíbrio não deve ser rejeitado.

Com este exemplo fica transparente as dificuldades de utilização de nossa metodologia, caso não se tenha à disposição um computador de grande porte. Contudo, métodos de aproximações podem ser desenvolvidos com o objetivo de viabilizar o uso de pequenas calculadoras nos

cálculos descritos.

Na próxima seção o teste da seção 2 é comparado com o teste padrão do qui-quadrado.

5 - COMPARAÇÃO DOS TESTES DE BAYES E QUI-QUADRADO: SISTEMA AUTOSSÔMICO, MONOGÊNICO E DIALÉLICO.

O objetivo desta seção é avaliar a performance do teste de Bayes em relação ao teste do qui - quadrado. Evidentemente, devemos caracterizar as condições de comparação entre os dois testes. Consideramos aqui apenas o sistema autossômico, monogênico e dialélico.

Ao efetuar um teste de hipótese, o estatístico indica a decisão a ser tomada e apresenta os valores dos erros de 1a. e 2a. espécies. Evidentemente esses valores são calculados com base em suposições e afirmações teóricas usadas na construção do teste. Contudo, os valores dos erros que efetivamente são cometidos, podem ser diferentes daqueles que são informados pelos estatísticos. Os valores dos erros efetivamente cometidos são definidos como:

Definição 6

Suponha que um teste δ fosse aplicado um gran-

de número de vezes.

a) $\alpha^*(\delta)$ é a proporção esperada de vezes que rejeitaríamos H_0 quando H_0 é verdadeira.

b) $\beta^*(\delta)$ é a proporção esperada de vezes que aceitaríamos H_0 quando H_0 é falsa.

Note que $\alpha(\delta)$ e $\beta(\delta)$ são as probabilidades dos erros, informado pelo estatístico com base em seus cálculos. A comparação que será feita aqui, é baseada na seguinte definição:

Definição 7

Um teste δ_1 é considerado superior ao teste δ_2 se as duas condições abaixo são satisfeitas.

i) $\alpha^*(\delta_1)+k\beta^*(\delta_1) \leq \alpha^*(\delta_2)+k\beta^*(\delta_2)$

ii) $|\alpha(\delta_1)-\alpha^*(\delta_1)| \leq |\alpha(\delta_2)-\alpha^*(\delta_2)|$ e

$|\beta(\delta_1)-\beta^*(\delta_1)| \leq |\beta(\delta_2)-\beta^*(\delta_2)|$

O valor de k depende da relação de importância entre os dois tipos de erro e assim depende do problema de aplicação que se está resolvendo. O critério (i) esta na mesma direção do teorema 1a do capítulo I. O critério (ii) indica o melhor teste como aquele cujos valores dos erros efetivos estão mais próximos dos valores calculados pelo estatístico. Isto é, o melhor teste é

aquele que além de errar menos é o que menos "mente".

Para confrontar o teste da seção 2 com o teste qui-quadrado, considerou-se k=1 (erros igualmente importantes) e o tamanho de amostra razoavelmente grande, n = = 50. Com estes valores, o teste de Bayes produz a tabela 9 que informa para cada amostra possível a decisão a ser tomada. Os valores calculados para os erros do teste de Bayes são α= 0,085 e β = 0,275. Com este valor de α, construímos a tabela 24 que informa, para cada ponto amostral, se a hipótese de equilíbrio, H_0, é rejeitada (F) ou aceita (T) pelo teste do qui-quadrado com nível de significância α = 0,085.

Acreditamos assim, que os testes são construídos sob as mesmas condições e assim a comparação é plausível.

No espaço paramétrico Ω' (veja figura 8), considerou-se o conjunto dos pontos, $\overline{E}$, formados pelos encontros das linhas $p_1 = i$ e $p_3 = j$

onde i,j = 0;0,1;0,2;...; 0,9;1, excluindo-se aqueles que estariam na corda de equilíbrio. Para cada um dos pontos de $\overline{E}$, foram geradas 200 amostras de tamanho 50 e registrou-se o número $\overline{e}_2$ de amostras que aceitariam equilíbrio com o teste qui-quadrado e o número e_2 de amostras que aceitariam o equilíbrio com o teste de Bayes. Com a soma dos $\overline{e}_2$, calcula-se a proporção das amostras

TABELA 24: Decisões indicadas pelo teste qui-quadrado com $\alpha = 0{,}085$ e $n = 50$. T= aceita equilíbrio
F= rejeita equilíbrio

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
50	F																																																	
49	T	F																																																	
48	T	F	F																																															
47	T	F	F	F																																														
46	T	F	F	F	F																																												
45	T	F	F	F	F	F																																											
44	T	T	F	F	F	F	F																																										
43	T	T	F	F	F	F	F	F																																									
42	T	T	F	F	F	F	F	F	F																																								
41	T	T	F	F	F	F	F	F	F	F																																							
40	T	T	F	F	F	F	F	F	F	F	F																																						
39	T	T	T	F	F	F	F	F	F	F	F	F																																					
38	T	T	T	F	F	F	F	F	F	F	F	F	F																																				
37	T	T	T	F	F	F	F	F	F	F	F	F	F	F																																			
36	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F																																		
35	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F																																	
34	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F																															
33	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F																														
32	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F																													
31	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F																											
30	F	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																										
29	F	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																									
28	F	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																								
27	F	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																							
26	F	T	T	T	T	T	T	F	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																						
25	F	F	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																					
24	F	F	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																				
23	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																			
22	F	F	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																		
21	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																	
20	F	F	F	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F																
19	F	F	F	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F															
18	F	F	F	F	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F														
17	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F													
16	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F												
15	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F											
14	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F										
13	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F									
12	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F								
11	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F							
10	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F						
9	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F					
8	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F				
7	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F			
6	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F		
5	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
4	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
3	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
2	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
1	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
0	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	F

que levam a aceitação de H_0. Esta proporção $\bar{e}$ a estimativa, $\hat{\beta}^*$, do erro efetivo β^* do teste qui-quadrado. Com a soma dos e_2 calcula-se a proporção das amostras que levam a aceitação de H_0. Este $\bar{e}$ a estimativa do erro efetivo β^* do teste de Bayes. Estas estimativas estão na tabela 25.

Na corda Ω_0' (figura 8) considerou-se os 11 pontos obtidos com p_1=0; 0,1;0,2; ...; 0,9;1. Da mesma forma, para cada um desses pontos foram geradas 200 amostras de tamanho 50 e registrou-se o número $\bar{e}_1$ de amostras que rejeitariam equilíbrio pelo teste qui-quadrado e o número e_1 pelo teste de Bayes. Com a soma dos $\bar{e}_1$ e com a soma dos e_1 obtêm-se respectivamente as estimativas do erro efetivo α^* do teste qui-quadrado e do teste de Bayes. Os resultados estão na tabela 25.

TABELA 25

	EFETIVO(SIMULADO)			TEÓRICO(CALCULADO)		
	α^*	β^*	TOTAL	α	β	TOTAL
BAYES	0,1005	0,3175	0,4180	0,085	0,275	0,360
χ^2	0,2191	0,3292	0,5483	0,085	*	—

* Valor desconhecido

Os valores apresentados na tabela 25 sugerem uma superioridade do teste de Bayes em relaçao ao teste

qui-quadrado pois o critério (i) da definição 7 é satisfeito em favor do teste de Bayes. A falta do valor de β para o teste de qui-quadrado prejudica um pouco a análise do critério (ii). Contudo, com apenas o valor de α, concluímos a superioridade do teste de Bayes visto que a distância entre α e α^* é muito menor para este teste. Há que se ressaltar, no entanto, o fato de o teste qui-quadrado estar sendo aplicado mesmo em amostra com pequenas freqüências esperadas. Para que um julgamento justo possa ser feito, necessitamos incluir em nossas simulações um dos testes exatos descritos na literatura para pequenas amostras. A nossa decisão de considerar apenas o teste qui-quadrado foi devido ao uso indiscriminado deste teste por uma parcela relevante da nossa comunidade de usuários.

B I B L I O G R A F I A

BASU, D. 1975. Statistical information and likelihood. *Sankhyã. Series* A, *37*(1):1-71.

BASU, D. 1977. On the elimination of nuisance parameters. *Journal of the American Statistical Association*, *72*(358): 355-366.

BEIGUELMAN, B. 1977. *Genética médica*. São Paulo, EDART - Editora da Universidade de São Paulo. v.2. 390p.

BERGER, J.O. c1980. *Statistical decision theory: foundations, concepts and methods*, New York, Springer. 425p.

BOX, G.E.P. & TIAO, G.C. c1973. *Bayesian inference in statistical analysis*. Reading, Addison-Wesley. 588p.

CANNINGS, C. & EDWARDS, A.W.F. 1969. Expected genotypic frequencies in a small sample: deviation from Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, *21*:245-247.

CHAPCO, W. 1976. An exact test of the Hardy-Weinberg law. *Biometrics*, *32*(1):183-189.

COURANT, R. & JOHN, F. c1974. *Introduction to calculus and analysis*. New York, John Wiley. v.2. 954p.

COX, D.R. 1958. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, *29*(1):357-372.

DEGROOT, M.H. c1975. *Probability and statistics*. London, Addison-Wesley. 607p.

ELANDT-JOHNSON, R.C. c1971. *Probability models and statistical methods in genetics*. New York, John Wiley. 592p.

ELSTON, R.C. & FORTHOFER, R. 1977. Testing for Hardy-Weinberg equilibrium in small samples. *Biometrics*, *33*(3):536-542.

EMIGH, T.H. & KEMPTHORNE, O. 1975. A note on goodness of fit of a population to Hardy-Weinberg structure. *American*

Journal of Human Genetics, 27:778-783.

EVERITT, B.S. c1977. *The analysis of contingency tables*. London, Chapman and Hall. 128p.

FIENBERG, S.E. c1980. *The analysis of cross-classified categorical data*. 2.ed. Cambridge, MIT Press. 198p.

HALDANE, J.B.S. 1954. An exact test for randommess of mating. *Journal of Genetics*, 52:631-635.

HARDY, G.H. 1908. Mendelian propositions in a mixed population. *Sciences*, 28:49-50.

IRONY, T.Z. 1984. *Testes exatos para tabelas 2×2: Bayes X Fisher*. São Paulo. 133p. Dissertação (Mestrado) - IME USP.

JEFFREYS, H. 1961. *Theory of probability*. 3.ed. Oxford, Clarendon. 447p.

KREWSKI, D.; BRENNAN, J.; BICKIS, M. 1984. The power of the Fisher permutation test in 2Xk tables. *Communications in Statistics: Simulation and Computation*, 13(4):433-448.

LECOUTRE, B. *Reconsideration of the F-test of the analysis of variance: the semi-Bayesian significance test*. 9p. (to appear).

LEONARD, T. *The Bayesian analysis of categorical data*. 30p. (to appear).

LEVENE, H. 1949. On a matching problem arising in genetics. *Annals of Mathematical Statistics*, 20:91-94.

LINDLEY, D.V. c1965. Inference. Part II. In: *Introduction to probability and statistics: from a Bayesian viewpoint*. Cambridge, University Press. v.2. 292p.

LINDLEY, D.V. 1982. The Bayesian approach to statistics. In: OLIVEIRA, T. & EPSTEIN, B., eds. *Some recent advances in statistics*. New York, Academic Press. p.65-87.

LINDLEY, D.V. 1983. *Lectures on Bayesian statistics*. São Paulo, IME-USP. 48p. (Publicações do Instituto de Matemática e Estatística da Universidade de São Paulo).

LINDLEY, D.V. & PHILLIPS, L.D. 1976. Inference for a Bernoulli process: a Bayesian view. *The American Statisti-*

cian, *30*(3):112-119.

MARIOTTO, A.B. 1983. Inferência parcial. São Paulo. 107p. Dissertação (Mestrado) - IME-USP.

PEREIRA, C.A. de B. 1971. Estimativa da probabilidade a priori em um problema de classificação. *Ciência e Cultura*, *23*(6):773-786.

PEREIRA, C.A. de B. 1983. *Stopping rules and conditional inference in 2×2 contingence tables*. São Paulo, IME-USP. 7p. (RT-MAE-8301).

PEREIRA, C.A. de B. & BASU, D. 1982. On the Bayesian analysis of categorical data: the problem of nonresponse. *Journal of Statistical Planning and Inference*, *6*(4):345-362.

PEREIRA, C.A. de B. & LINDLEY, D.V. 1983. *Examples questioning the use of partial likelihood*. São Paulo, IME-USP. 9p. (RT-MAE-8306).

PEREIRA, C.A. de B. & ROGATKO, A. 1984. The Hardy-Weinberg equilibrium under a Bayesian perspective. *Revista Brasileira de Genética*, *7*(4):689-707.

PEREIRA, C.A. de B. & VIANA, M.A.G. 1982. *Elementos de inferência Bayesiana*. São Paulo, 98p. Trab. apres. ao 5º Simpósio Nacional de Probabilidade e Estatística, São Paulo, 1982.

VITHAYASAI, C. 1975. Exact critical values of the Hardy-Weinberg test statistics for two alleles. *Communications in Statistics*, *1*:229-242.

WEINBERG, W. 1908. Uber den Nachweis der vererbung bein menschen. *Jahresh. Verein Vaterl. Naturk. in Wurttemberg*, *64*:368-382.

WILKS, S.S. c1962. *Mathematical statistics*. New York, John Wiley. 644p.

ÍNDICE

pág.

pág.

pág.