

DISTÂNCIA DE MATUSITA E A COMPARAÇÃO
DE DUAS AMOSTRAS

OSCAR VILLANUEVA ROJAS



UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA DA COMPUTAÇÃO

CAMPINAS - SÃO PAULO
BRASIL

UNICAMP
BIBLIOTECA CENTRAL

UMA CONTRIBUIÇÃO AO PROBLEMA DE DUAS AMOSTRAS

Este exemplar corresponde a redação da tese defendida pelo Sr. OSCAR VILLANUEVA ROJAS e aprovada pela comissão julgadora.

Campinas, 12 de Setembro 1984.



PROF. DR. JOSÉ ANTONIO CORDEIRO

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Estatística.

SETEMBRO/1984.

À meus pais
e à minha filha
Paola Andrea

AGRADECIMENTOS

- Ao Professor Doutor José Antonio Cordeiro, orientador e amigo.
- Ao Professor Doutor José Norberto Dachs pelo importantíssimo apoio com que me brindou.
- Aos meus colegas e amigos em geral que colaboraram, especialmente na correção do português e dos programas de computador.
- À Universidade de la Amazonia de Florencia (Caquetá), Colombia, que me deu fundamental apoio e facilitou meu estudo.
- Aos Diretores de Instituto, chefes de Departamento e companheiros, professores da Universidad de la Amazonia (Colombia) que souberam suportar as consequências de minha ausência.
- A minha família, especialmente minha filha Paola Andrea pela compreensão.
- A Lourdes, pelo eficiente trabalho datilográfico

A todos, meus agradecimentos

OSCAR VILLANUEVA ROJAS

INDICE

INTRODUÇÃO	i
CAPÍTULO I - ALGUNS TESTES PARA AJUSTAMENTO E DUAS AMOS- TRAS	1
Introdução	1
Teste de Wald-Wolfowitz	2
Teste da Mediana	3
Teste de Wilcoxon	5
Teste de Mann-Whitney	6
Teste de Cramer, Vom-Mises e Watson	7
Teste de Smirnov e suas Generalizações	9
Teste de Kolmogorov	10
Teste Qui-Quadrado	11
Teste Razão de Verossimilhança	18
Teste de Espaços de Valores Amostrais	19
CAPÍTULO II - OS RESULTADOS DE MATUSITA E ALGUMAS APLICA- ÇÕES	22
Introdução	22
Definições e Propriedades	22
Teoremas Fundamentais	27
Problema de Duas Amostras	33
Outras Pesquisas	38
1 - Coeficiente de Associação	38

2 - Teste de Ajustamento	40
3 - Teste de Independência em Tabelas de Contingência. .	42
4 - Análise Discriminante no Caso Multinomial	45
CAPÍTULO III - CRÍTICA AOS RESULTADOS DE MATUSITA E SIMU -	
LAÇÕES	49
Introdução	49
Exemplos e Críticas	49
Simulações	60
Tabelas Simuladas	68
BIBLIOGRAFIA	77

INTRODUÇÃO

A parte de Inferência Estatística não-paramétrica, por sua importância, tem sido muito discutida; embora os problemas envolvidos ao fazer o teste de bom ajuste para duas amostras tenham sido considerados repetidamente, as dificuldades fundamentais continuam sem solução.

Ao nos referirmos a testes de ajustamento para duas amostras temos necessariamente enunciar os testes de Student, Kolmogorov-Smirnov, Qui-quadrado, etc. como os mais importantes. Eles fazem parte dos problemas essenciais tratados em testes não-paramétricos com seus supostos e características particulares: Qui-quadrado, mesmo sendo um dos mais usados, é sensível a modificações no tamanho amostral, assimétrico e dando maior peso a observações de menor probabilidade, dificuldades na determinação do número de classes e de frequências por classes; Kolmogorov-Smirnov bom para determinar diferenças em funções de distribuição acumulada, supõe distribuições contínuas e o teste t supõe normalidade das populações (este teste é pouco sensível a não-normalidade).

Tentando resolver alguns destes inconvenientes, Matusita 1955 propõe uma distância, de excelentes características, e baseado nela faz uma proposta para o problema de duas amostras. Seguindo nesta direção procuramos fazer uma contribuição na

solução de tal assunto.

No Capítulo I fazemos um levantamento bibliográfico dos testes propostos para o bom ajuste de duas amostras, destacando particularmente, alguns problemas que cada um deles apresenta.

No Capítulo II tratamos em detalhe o artigo de Kameo Matusita, publicado no Annals Mathematical Statistics em 1955 onde aparecem propostas alternativas para o problema de duas amostras.

No Capítulo III aparecerão as críticas ao citado artigo, ao mesmo tempo em que fazemos nossa proposta de teste baseado na distância de Matusita e apresentamos conclusões junto aos nossos exemplos numéricos.

CAPÍTULO I

ALGUNS TESTES PARA AJUSTAMENTO E DUAS AMOSTRAS

INTRODUÇÃO

A comparação de duas amostras aleatórias é uma questão muito discutida na área de estatística não-paramétrica. A hipótese nula, geralmente de igualdade de populações com distribuição comum, completamente não especificada, exceto na suposição de que ela é uma função de distribuição contínua, faz parte do problema geral de teste de hipótese. Solucionar o teste $H_0: F_Y(x) = F_X(x)$ onde $F(x)$, $F(y)$ são funções de distribuição de duas populações independentes, motivou estudos para desenvolver processos estatísticos que dependem do menor número possível de suposições. Assim, foram evoluindo métodos hoje conhecidos como não paramétricos ou equivalentemente chamados métodos livres de distribuição, pois sua aplicação independe da distribuição da variável aleatória, da qual as observações são obtidas. Isto devido ao grande número de suposições feitos na literatura de testes clássicos. Por exemplo, o teste t de Student e o teste F supõem normalidade, os quais nem sempre são sensíveis a violações de tais suposições.

Consideraremos, rapidamente, alguns testes não-paramétricos

ênfatizando suas características e propriedades, tentando apresentar um número razoável deles, sem chegarmos a considerar, obviamente, todos os existentes.

No final desta seção serão incluídos dois testes de ajustamento para uma amostra, que nos pareceram relevante considerar, dado a sua importância.

1. TESTE DE WALD-WOLFOWITZ.

Sejam dois conjuntos de variáveis aleatórias independentes $X_1, \dots, X_m$ e $Y_1, Y_2, \dots, Y_n$ com funções de distribuição contínuas F_1 e F_2 . As $m+n = N$ observações das duas amostras são colocadas numa amostra ordenada em ordem crescente. Cada observação da primeira amostra é substituída por 0 e da segunda por 1, obtendo assim uma sequência de 0's e 1's. Supondo que desejamos testar $H_0 : F_Y(x) = F_X(x)$ para todo x esperamos que as variáveis X, Y se misturam completamente na amostra ordenada. Como todos os $\frac{(n+m)!}{n!m!}$ arranjos possíveis de 0's e 1's têm igual probabilidade de acontecer, então, a distribuição baseada em tais sequências pode ser determinada. Definindo um "RUN" como cada uma das sequências completas de 0's ou 1's, o número de "RUNS" na amostra ordenada é indicativo do grau de mistura.

Quando temos que os X 's são estocasticamente menores ou

maiores que Y 's, temos somente dois "RUNS", sendo que, um critério do teste baseado unicamente no número total de RUNS, pode não ser distinguível nesses casos. O teste de Wald-Wolfowitz é apropriado quando as alternativas são totalmente gerais e bicaudais como

$$H_1 : F_Y(x) \neq F_X(x) \text{ para algum } x,$$

sendo consistente contra qualquer diferença em populações [vide Wald-Wolfowitz, 1940]. Esta generalidade faz com que o teste seja bom contra alternativas especificadas. O poder exato ou assintótico pode ser calculado somente com base de alternativas completamente especificadas.

2. TESTE DA MEDIANA.

Duas amostras aleatórias independentes $X_1, \dots, X_m$ e $Y_1, \dots, Y_n$ serão classificadas segundo sejam menores ou maiores que um número π arbitrariamente escolhido. Aqui comparamos somente as proporções de cada amostra que são estritamente menores que π .

Sendo T, S números respectivos de observações X e Y menores que π , então, T, S tem distribuição binomial com parâmetros:

$$P_X = P_X(X < \pi) \quad e \quad P_Y = P_Y(Y < \pi) \quad (1.1)$$

logo sua distribuição conjunta é

$$f_{T,S}(t,s) = \binom{m}{t} \binom{n}{s} P_X^t P_Y^s (1 - P_X)^{m-t} (1 - P_Y)^{n-s} \quad (1.2)$$

com $t = 0, 1, \dots, m$, $S = 0, 1, 2, \dots, n$, [vide Mood, 1950] .

Como $\frac{T}{m}$, $\frac{S}{n}$ são estimadores não-viciados de P_X e P_Y a diferença $\frac{T}{m} - \frac{S}{n}$ é apropriado para testar a hipótese nula $H_0 : P_X - P_Y = 0$. Podemos determinar a distribuição exata de $\frac{T}{m} - \frac{S}{n}$. Quando m, n crescem a distribuição desses estimadores pode ser feita através da aproximação pela normal.

O teste estatístico depende, neste caso, do valor geral $P = P_X = P_Y$, mas o teste pode ser feito substituindo P por seu estimador não viciado $(T+S)/(m+n)$. Uma outra maneira, seria construir um teste, também aproximado, baseado no critério de diferenças de proporções de observações menores que π . Este seria essencialmente um teste de sinal modificado, bastando para tanto supor que μ_0 é mediana das $m+n=N$ observações com r das $(x_1 - \mu_0), \dots, (x_m - \mu_0)$, $(y_1 - \mu_0), \dots, (y_n - \mu_0)$ negativas e S positivas com $r+S \leq N$.

Assim a distribuição de r dado $r+S$ é binomial com probabilidade de sucesso $\pi = \frac{1}{2}$, logo nosso problema agora é

testar a hipótese $\pi = \frac{1}{2}$ dado que r sucessos tem ocorrido entre $r + S$ ensaios de Bernoulli.

No teste da mediana a hipótese real é que π é o P -ésimo quantil nas duas populações, onde P não é especificado, mas estimado dos dados. O teste é aproximado e frequentemente não é apropriado para o problema geral de duas amostras, onde estamos principalmente interessados na hipótese de igualdade de populações. Se as duas populações são as mesmas, os pontos P -ésimos quantis são iguais para todo valor de P . Contudo, duas populações podem ser inteiramente diferentes mesmo tendo alguns quantis iguais. O valor π que é supostamente escolhido sem conhecimento das observações, altera a sensibilidade do critério do teste. Se o π escolhido é menor ou maior, T, S terão também, menor campo de variação para serem confiáveis. Concluindo, o teste é sensível a diferenças de locações mas é de pouco interesse.

3. A) TESTE DE WILCOXON.

As duas amostras são colocadas como no teste de Wald-Wolfowitz.

Determina-se os postos da segunda amostra e considera-se como critério do teste a soma dos mesmos (vide Wilcoxon, 1945). Se a hipótese nula é certa, qualquer seleção de n postos de

$1, 2, \dots, m+n$ terá a mesma probabilidade e então a distribuição da soma de postos pode ser determinada. Considerando-se alternativas unilaterais, grandes valores da estatística indicará desvio da hipótese nula. Se forem bilaterais serão considerados valores grandes e pequenos na região crítica.

B) TESTE DE MANN-WHITNEY

As duas amostras são colocadas como no teste de Wold-Wolfovitz. Proposto por Mann-Whitney 1947, o critério do teste são as posições dos Y 's na sequência misturada e ordenada das $m+n = N$ observações crescentes em magnitude, tomadas de distribuições contínuas.

A estatística de Mann-Whitney é definida como o número de vezes que Y precede X no arranjo misturado e ordenado das duas amostras independentes.

Os dois testes considerados aqui, são equivalentes. Para sua aplicação, somente necessitamos de independência e distribuições contínuas para testar a hipótese nula de igualdade de populações contra qualquer tipo de alternativa.

O teste de Mann-Whitney é especialmente sensível para diferenças de locação e é particularmente bom como teste de médias ou medianas iguais, por isso é comparável com o teste t de Student para média.

4. TESTES DE CRAMER, VON-MISES E WATSON

A hipótese de que as duas amostras independentes $X_1, \dots, X_m$, e $Y_1, \dots, Y_n$ vêm da mesma população, pode ser testada usando os critérios W^2 de Von-Mises e U^2 de Watson 1961. As estatísticas W^2 e U^2 foram consideradas por E.J. Burr 1963 e G.S. Watson 1961, respectivamente, com suas formas originais seguintes:

$$W^2 = \frac{mn}{(m+n)^2} \sum d_i^2 \quad \text{e} \quad U^2 = \frac{mn}{(m+n)^2} \sum (d_i - \bar{d})^2 \quad (1.3)$$

onde d_i é a diferença entre as funções de distribuição amostral no i -ésimo ponto da amostra tomada, isto é, têm-se m_i elementos da primeira amostra e n_i da segunda contidos no conjunto dos primeiros i -elementos tomados de todos os $m+n$ valores arranjados em ordens de magnitude, então

$$d_i = \frac{n_i}{n} - \frac{m_i}{m}, \quad i=1, 2, \dots, m+n$$

e

$$\bar{d} = \sum \frac{d_i}{m+n} \quad (1.4)$$

5. Neste item serão considerados os testes de Kolmogorov, de Smirnov e Smirnov modificado. Para efeitos de notação, em geral,

concordamos com as definições a seguir. Para duas amostras aleatórias de tamanhos m e n de populações contínuas F_X , F_Y , suas estatísticas de ordem são

$$X_{(1)}, X_{(2)}, \dots, X_{(m)} \quad \text{e} \quad Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$$

com suas correspondentes funções de distribuição empíricas denotadas por $S_m(x)$, $T_n(y)$ definidas como se segue:

$$S_m(x) = \begin{cases} 0 & x < X_{(1)} \\ K/m & X_{(K)} \leq x < X_{(K+1)} \\ 1 & x \geq X_{(m)} \end{cases} \quad K = 1, \dots, m-1$$

$$T_n(y) = \begin{cases} 0 & x < Y_{(1)} \\ K/n & Y_{(K)} \leq x < Y_{(K+1)} \\ 1 & x > Y_{(n)} \end{cases} \quad K = 1, \dots, n-1$$

se a hipótese nula $H_0 : F_X(x) = F_Y(x)$ para todo x é verdadeira, as distribuições populacionais são idênticas e teríamos duas amostras da mesma população. Por conseguinte admitindo-se variações amostrais, sob H_0 será razoável um ajuste entre as duas distribuições empíricas.

a.- TESTE DE SMIRNOV ω^2 :

Modificando a estatística proposta por Cramer e Von Mises, Smirnov 1936, compara $F_Y(x)$ com $F_X(x)$ através de

$$\omega^2 = n \int_{-\infty}^{\infty} [S_n(x) - T_n(x)]^2 dT_n(x) \quad (1.5)$$

onde supomos tamanhos amostrais iguais. Depois de alguma álgebra, 1.5 pode ser escrita como:

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^n \left[T_n(x_i) - \frac{2i-1}{n} \right]^2 \quad (1.6)$$

b.- TESTE W_n^2 :

T.W. Anderson e D.A. Darling 1952 consideram uma generalização de 1.5,

$$W_n^2 = n \int_{-\infty}^{\infty} [S_m(x) - T_n(x)]^2 \Psi[T_n(x)] dT_n(x) \quad (1.7)$$

onde $\Psi(t) \geq 0$, para $0 \leq t \leq 1$, é uma função de pesos. Esta estatística pode se escrever como:

$$W_n^2 = 2 \sum_{J=1}^n \left\{ \Phi_2[T_n(x_J)] - \frac{2J-1}{2n} \Phi_1[T_n(x_J)] \right\} + K_{n,\Psi} \quad (1.8)$$

onde, $\Phi_1(t)$, $\Phi_2(t)$ são funções determinadas de maneira única por $\Psi(t)$. Com $\Psi(t) = 1$ obtemos 1.5

c.- TESTE W_n^2 COM $\Psi(t) = \frac{1}{t(1-t)}$:

Como para x pequeno, as funções de distribuição $F_Y(x)$, $F_X(x)$ são ambas aproximadas a 0 e agora para x grande ambas aproximadas a 1, 1.5 não determina as discrepâncias nas caudas da distribuição.

Para casos, onde tais diferenças são importantes, Anderson e Darling 1952 consideram a função peso $\Psi(t) = \{t(1-t)\}^{-1}$. Uma tabulação de tal distribuição é difícil.

d.- TESTE DE KOLMOGOROV :

Kolmogorov 1953, considera a estatística

$$D_{m,n} = \max_{-\infty < x < \infty} |S_m(x) - T_n(x)| \quad (1.9)$$

onde, a magnitude e não os sentidos das diferenças são consideradas. $D_{m,n}$ é apropriado para uma alternativa geral $H_1: F_Y(x) \neq F_X(x)$ para algum x . Wold-Wolfowitz 1939, consideram 1.9 a classe mais geral de estatísticas não paramétricas e demonstram como calcular sua distribuição de probabilidade para pequenas amostras.

Os testes de Kolmogorov-Smirnov são de fácil aplicação sendo consistentes contra alternativas gerais mais também sensível a qualquer diferença entre as funções de distribuição acumuladas. Para amostras pequenas podemos usar as tabelas conhecidas e para amostras grandes usamos sua distribuição assintótica.

A importância deste teste em aplicações com grandes amostras varia consideravelmente, dependendo da população amostrada quando se testam locação ou parâmetros de escala [vide Kolmogorov-Smirnov, Psychological Bull. 1954, 51-160] O teste de Kolmogorov é adequado para determinar diferenças verticais entre as funções acumuladas de probabilidade [vide Birbaum 1952].

6. TESTE χ^2 :

Este critério é um dos mais usados em aplicações. Proposto por Karl Pearson em 1900, este teste para duas amostras compara

as distribuições empíricas determinadas pelas amostras na base da estatística χ^2 .

A distribuição limite χ^2 é conhecida e amplamente tabulada.

Muitas pesquisas têm sido feitas procurando condições ótimas para um ótimo comportamento do teste χ^2 , isto por causa da variedade de problemas implícitos na estrutura do mesmo. Alguns desses problemas enunciaremos a seguir:

- i) É um teste que depende diretamente do tamanho amostral, isto é, fazendo um aumento no tamanho amostral sempre poderemos rejeitar a hipótese H_0 de igualdade de distribuições.

EXEMPLO: Na tabela a seguir, agrupamos em classes os dados obtidos numa pesquisa sobre esquistossomose efetuada no município Pedro de Toledo (S.P.).

Para classificar os elementos da amostra em positivos ou negativos, considera-se como critério o exame de fezes (Kato).

Para nosso objetivo comparamos a prevalência por faixa etária das populações feminina e masculina.

TABELA 1.1.: PREVALÊNCIA POR FAIXA ETÁRIA SEGUNDO
SEXO DO EXAME DE FEZES (KATO).
PEDRO DE TOLEDO (S.P.)

CLASSES	IDADE (ANOS)	Freq _M	Freq _F	P _{1M}	q _{1F}
01	0 - 4	12	6	0.022	0.022
02	5 - 9	72	63	0.133	0.235
03	10 - 14	145	63	0.268	0.235
04	15 - 19	116	43	0.214	0.160
05	20 - 24	52	25	0.096	0.093
06	25 - 29	27	16	0.050	0.060
07	30 - 34	28	10	0.052	0.037
08	35 - 39	21	7	0.039	0.026
09	40 - 44	23	13	0.043	0.049
10	45 - 49	22	5	0.041	0.019
11	50 - 54	7	8	0.013	0.030
12	55 - 59	9	7	0.017	0.026
13	60 -	7	2	0.013	0.007
TOTAL		541	268	1.00	1.00

FONTE: Pesquisa Departamento de Parasitologia
IB/UNICAMP e SUCEN - SP.

$$\chi^2_2 = 62.326716$$

TABELA 1.2

Tomamos a tabela 1.1 e usando os seus dados originais vamos dividir o seu tamanho amostral por 10. Em seguida, multiplicamos esses mesmos dados originais por 10, calculando, para cada caso, a estatística X^2 .

CLASSES	IDADE	$\text{freq}_M^{1/10}$	$\text{freq}_F^{1/10}$	$P_{iM}^{1/10}$	$q_{iF}^{1/10}$	freq_M^{10}	freq_F^{10}	P_{iM}^{10}	q_{iF}^{10}
01	0-4	1	1	0.018	0.034	120	60	0.022	0.022
02	5-9	7	6	0.130	0.210	720	630	0.133	0.235
03	10-14	15	6	0.240	0.210	1450	630	0.268	0.235
04	15-19	12	4	0.220	0.140	1160	430	0.214	0.160
05	20-24	5	3	0.091	0.100	520	250	0.096	0.093
06	25-29	3	2	0.054	0.069	270	160	0.050	0.060
07	30-34	3	1	0.054	0.034	280	100	0.052	0.037
08	35-39	2	1	0.036	0.034	210	70	0.039	0.026
09	40-44	2	1	0.036	0.034	230	130	0.043	0.049
10	45-49	2	1	0.036	0.034	220	50	0.041	0.019
11	50-54	1	1	0.018	0.034	70	80	0.013	0.030
12	55-59	1	1	0.018	0.034	90	70	0.017	0.026
13	60-	1	1	0.018	0.034	70	20	0.013	0.007
Totais		55	29	1.00	1.00	5410	2680	1.00	1.00

$$1/10 X^2 = 7.764891$$

$$10 X^2 = 623.2671$$

Em geral, se $n^* = Kn$ temos:

$$\begin{aligned} \chi^2 &= \sum_i \frac{(Kn_i - Kn p_i)^2}{Kn p_i} = \sum_i \frac{K^2 (n_i - n p_i)^2}{Kn p_i} = K \sum_i \frac{(n_i - n p_i)^2}{n p_i} \\ &= K \chi^2 \end{aligned}$$

ii) Por causa de sua assimetria, sempre que são trocados valor amostrado e valor esperado, teremos diferentes valores para a estatística.

EXEMPLO: Sobre os dados da tabela 1.1 foi calculado $\chi^2 = 62.326$ usando sua expressão

$$\chi^2 = \sum_i \frac{(n_i - 541 q_i)^2}{541 q_i} .$$

Agora, se fazemos aqueles cálculos usando:

$$\chi^2 = \sum_i \frac{(m_i - 268 p_i)^2}{268 p_i} \text{ obteremos } \chi^2 = 38.722$$

o que de fato é uma diferença considerável.

iii) Dá maior peso às observações de probabilidade pequena, isto é, observações com baixa probabilidade dão maior contribuição ao valor calculado forçando a rejeição da hipótese H_0 .

iv) O fato dos dados amostrais terem de ser agrupados em categorias, faz com que percamos informação, suspeitando-se que se ja maior quando se trata de distribuições contínuas. Além disso, o teste χ^2 não é único por causa da liberdade que existe na escolha do número de intervalos.

v) A estatística χ^2 está baseada no quadrado de desvios de frequências observadas, sendo portanto insensível ao modelo de sinais.

Em vista dos problemas anteriores, muitas sugestões e pesquisas têm sido feitas para eliminar tais dificuldades, principalmente quando fazemos aplicações em tabelas de contingência, onde foram desenvolvidos importantes trabalhos práticos.

Como anota Dadakis 1977 em relação à aproximação de χ^2 por χ^2 , esta aproximação não é boa se as frequências esperadas nas celas, determinadas pela distribuição hipotética, são pequenas. Neste caso, a estatística χ^2 tem vício positivo, isto é, tende a ser maior que o valor teórico χ^2 . Assim, tendemos a rejeitar a distribuição hipotética com probabilidade maior do que o nível de significância selecionado.

Naturalmente, desejamos saber quando as frequências são suficientemente grandes para evitar tal situação no teste qui-quadrado.

Muitos autores têm apresentado trabalhos dando sugestões

para escolher tamanhos amostrais e determinar números de categorias para agrupar os dados. Entre os mais citados estão: Cochran 1952, 1954, Manns e Wold 1942 e Crámer 1946.

Acontece que o cumprimento das regras de Cochran 1952 com um tamanho amostral suficientemente grande é antieconômico e pouco prático, daí a manipulação dos dados no sentido de se obter o menor número de categorias.

Seja como for, o problema ainda continua sem solução. Nas tentativas de solucionar cada problema foram aparecendo diferentes testes, cada um deles padecendo de algum defeito. Alguns deles bons para testar diferença em funções de distribuição acumulada, outros bons para testar parâmetros de escala ou locação, outros bons para testar hipóteses gerais, etc, mas com eles vêm problemas adicionais, tais como difícil computação, pouco econômicos, muitas suposições, aplicações extremamente particulares entre outros. Tudo isso leva a uma complicada análise sobre qual teste usar, que seja mais adequado para a hipótese a testar, em termos de ser de maior poder e de menor complexidade, pois é claro que se nosso desejo é determinar, por exemplo, diferenças em nossas distribuições empíricas, escolheríamos a χ^2 e não o teste de Kolmogorov-Smirnov. Isto inclusive como resultado de uma análise mais profunda do que nós queremos na base daquela distância que fundamenta o teste a usar.

Concluindo, teríamos mais algumas coisas a dizer:

a.- Comparando o teste de Kolmogorov-Smirnov $D_{m,n}$ com χ^2 , como mais conhecidos e usados, vemos que para o ajuste de uma distribuição contínua, $D_{m,n}$ é mais sensível precisando assintoticamente de tamanhos amostrais de ordem $n^{4/5}$ comparado com n para o teste χ^2 sendo então mais eficiente. Além disso, $D_{m,n}$ é mais poderoso particularmente quando a amostra e as frequências esperadas nas celas são pequenas. Por outro lado, $D_{m,n}$ é um teste exato e para pequenas amostras será preferencialmente usado (vide Massey 1951).

b.- Alguns testes de ajustamento para uma amostra são de importância, como:

1.) RAZÃO DE VEROSSIMILHANÇA

Suponha que temos duas distribuições de probabilidade multinomial com parâmetros $(p_1, \dots, p_k)$ e $(q_1, \dots, q_k)$ com uma observação $(n_1, \dots, n_k)$. A estatística razão de verossimilhança

$$R = \left(\frac{p_1}{q_1}\right)^{n_1} \dots \left(\frac{p_k}{q_k}\right)^{n_k} \quad (1.10)$$

é exatamente a razão das probabilidades do evento observado sob as suas distribuições.

Em termos práticos, geralmente se conhece a distribuição de $(p_1, \dots, p_k)$ e os parâmetros da distribuição alternativa $(q_1, \dots, q_k)$, desconhecidos, são substituídos por seus estimadores M.L.

$$\hat{q}_i = \frac{n_i}{N}, \quad i = 1, \dots, k \quad (1.11)$$

A estatística de ajuste da razão de verossimilhança L é o logaritmo de $1/R^2$

$$L = -2 \ln R = -2 \sum n_i (\ln N P_i - \ln n_i) \quad (1.12)$$

L é assintoticamente distribuído qui-quadrado com $k-1$ graus de liberdade, mas muito pouco é conhecido do comportamento de L para n pequeno.

ii.) ESPAÇOS DE VALORES AMOSTRAIS

Sejam $X_1 \leq X_2 \leq \dots \leq X_n$ elementos de uma amostra ordenada, de uma variável aleatória com função distribuição $H(x)$ conhecida. A esperança de $H(x_{i+1}) - H(x_i)$ é $1/(n+1)$, $i=0,1,\dots,n$ com as notações $H(X_0) = 0$, $H(X_{n+1}) = 1$.

Para testar:

$H_0 : x_1, \dots, x_n$ pertencem a $H(x)$

$H_1 : x_1, \dots, x_n$ pertencem a $\mathcal{A}$, onde $\mathcal{A}$ é uma classe de distribuições diferentes de $H(x)$, Kimball 1947 considera a estatística:

$$\sum_{i=1}^{n+1} \left[H(X_i) - H(X_{i-1}) - \frac{1}{n+1} \right]^2 \quad (1.13)$$

Moran 1947 usa a estatística:

$$\sum_{i=1}^{n+1} [H(X_i) - H(X_{i-1})]^2 \quad (1.14)$$

e demonstra que sua distribuição é assintoticamente normal. Sherman 1950 considera

$$\frac{1}{2} \sum_{i=1}^{n+1} |H(X_i) - H(X_{i-1})| \quad (1.15)$$

e demonstra que é assintoticamente normal.

c.- Muitos testes ficaram fora de comentário. É bom dizer que existem outros de menor interesse que não foram discutidos aqui: Fisher-Yate consideram como teste a soma de postos da segunda amostra. Van der Warden usa como critério $\sum \phi^{-1}(s_i/(m+n+1))$ onde ϕ é distribuição acumulada da variável normal padrão. Mood 1954

propôs $W = \sum_{i=1}^{n_1} \{X_i - \frac{1}{2}(n+1)\}^2$. Capón 1961 dá um equivalente àquele de Van der Warden. Além destes, há testes propostos por Lehmann 1951, Klotz 1962, Siegel e Tukey 1960, Kruskall e Wallis 1952, Puri 1964, etc.

No próximo Capítulo discutiremos a distância de Matusita, bem como o teste de bom ajuste proposto por Kirmain em 1971 além de outros tópicos, como medida de associação e teste de In dependência em tabelas de Contingência fundamentados na referida distância.

CAPÍTULO II

OS RESULTADOS DE MATUSITA E ALGUMAS APLICAÇÕES

INTRODUÇÃO

Kameo Matusita, em 1955, publicou um artigo, no qual ele apresentou uma solução alternativa para o problema do bom ajuste. Consideraremos aqui, partes dessa publicação, ampliando suas demonstrações e dando maior ênfase às partes de nosso particular interesse e nas quais baseia-se a melhor estrutura dos testes derivados desta distância.

No final do Capítulo faremos um resumo destacando os principais fatos de algumas pesquisas desenvolvidas baseadas na distância de Matusita. Vale ressaltar que tais citações são importantes por serem aplicações em análise discriminante, Testes de Independência, Testes de Ajuste e Medida de Associação.

DEFINIÇÕES E PROPRIEDADES

Sejam F_1 e F_2 funções de distribuição simultaneamente discretas ou contínuas. A distância d entre F_1 e F_2 é dada por

$$d^2(F_1, F_2) = \|F_1 - F_2\|^2 = \int_R (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 d_m \quad (2.1)$$

onde m é uma medida adequada, $p_1(x)$, $p_2(x)$ suas densidades e R é espaço total.

d é distância pois satisfaz as seguintes propriedades:

$$\begin{aligned}
 1 - \quad & a) \quad d(F,G) \geq 0 \quad \quad \quad d(F,F) = 0 \\
 & b) \quad d(F,G) = d(G,F) \\
 & c) \quad d(F,G) + d(G,H) \geq d(F,H)
 \end{aligned}
 \tag{2.2}$$

para toda distribuição F, G, H onde $d(F,G)$ denota a distância entre F e G .

2 - Dado um número positivo η pode-se achar uma sequência de números $\{B_n\}$, talque,

$$P_r \{d(F, S_n) > \eta\} \leq B_n \tag{2.3}$$

para algum F na classe de funções de distribuição em consideração, onde S_n denota a função de distribuição empírica, baseada em n observações de uma variável aleatória com distribuição F e tal que $B_n \rightarrow 0$, se $n \rightarrow \infty$.

3 - A distância representa bem a diferença entre distribuições em todo ponto ou, para o problema de ajuste, ela representa a diferença nos pontos com probabilidade maior (densidade) melhor que naqueles pontos com probabilidade menor.

DEMONSTRAÇÕES:

$$2.2.b - d^2(F,G) = \int_R (\sqrt{P_1(x)} - \sqrt{P_2(x)})^2 dm =$$

$$= \int_{\mathbb{R}} (\sqrt{P_2(x)} - \sqrt{P_1(x)})^2 d_m = d^2(G, F)$$

$$2.2.c - d(F, G) = \left(\int_{\mathbb{R}} (\sqrt{P_1(x)} - \sqrt{P_2(x)})^2 d_m \right)^{1/2}$$

$$d(G, H) = \left(\int_{\mathbb{R}} (\sqrt{P_2(x)} - \sqrt{P_3(x)})^2 d_m \right)^{1/2}$$

$$d(F, G) + d(G, H) = \left(\int_{\mathbb{R}} (\sqrt{P_1(x)} - \sqrt{P_2(x)})^2 d_m \right)^{1/2} + \left(\int_{\mathbb{R}} (\sqrt{P_2(x)} - \sqrt{P_3(x)})^2 d_m \right)^{1/2}$$

pela desigualdade de Minkowsky:

$$\begin{aligned} d(F, G) + d(G, H) &\geq \left(\int_{\mathbb{R}} ((\sqrt{P_1(x)} - \sqrt{P_2(x)}) + (\sqrt{P_2(x)} - \sqrt{P_3(x)}))^2 d_m \right)^{1/2} = \\ &= \left(\int_{\mathbb{R}} (\sqrt{P_1(x)} - \sqrt{P_3(x)})^2 d_m \right)^{1/2} = d(F, H). \end{aligned}$$

2 - $P_r\{\|F - S_n\|^2 \geq \eta\}, \eta > 0$ basta tomar $\eta = \frac{k-1}{n} t_n$ e fazendo uso da desigualdade de Markov obtemos o seguinte:

$$P_{r_F}(\|F - S_n\|^2 \geq \frac{k-1}{n} t) \leq 1/t \quad (\text{Markov})$$

$$P_r(\|F - S_n\|^2 \geq \eta) \leq \frac{1}{t_n} = B_n \quad \text{isto é,}$$

$$P_r\{\|F - S_n\| \geq \sqrt{\eta}\} \leq B_n$$

Da definição de T_n , vê-se que $\lim_{n \rightarrow \infty} B_n = 0$

A afinidade entre F_1 e F_2 é

$$\rho = \int_R \sqrt{P_1(x)} \cdot \sqrt{P_2(x)} d_m \quad (2.4)$$

Algumas das propriedades da distância $\| \cdot \|$ e afinidade ρ são:

$$a) \quad 0 \leq \rho \leq 1$$

$$b) \quad \|F_1 - F_2\|^2 = 2(1 - \rho(F_1, F_2)) \leq 2 \quad (2.5)$$

$$c) \quad \|F_1 - F_2\|^2 \leq \int_R |P_1(x) - P_2(x)| d_m \leq 2 \|F_1 - F_2\|$$

DEMONSTRAÇÕES:

$$2.5.a - \rho = \int_R \sqrt{P_1(x)} \sqrt{P_2(x)} d_m \quad \text{como } P_1(x) \geq 0, P_2(x) \geq 0$$

então $\rho \geq 0$.

Agora:

$$\int_R \sqrt{P_1(x)} \sqrt{P_2(x)} d_m \leq \left(\int_R (\sqrt{P_1(x)} \sqrt{P_2(x)})^2 d_m \right)^{1/2} =$$

$$\left(\int_R P_1(x) P_2(x) d_m \right)^{1/2} \leq \left(\int_R P_1(x) d_m \right)^{1/2} \left(\int_R P_2(x) d_m \right)^{1/2} = 1$$

obtido a partir da desigualdade de Cauchy-Schwarz.

$$\begin{aligned}
 2.5.b \quad \|F_1 - F_2\|^2 &= \int_{\mathbb{R}} (\sqrt{P_1(x)} - \sqrt{P_2(x)})^2 d_m = \int_{\mathbb{R}} P_1(x) d_m + \int_{\mathbb{R}} P_2(x) d_m - \\
 &- 2 \int_{\mathbb{R}} \sqrt{P_1(x)} \sqrt{P_2(x)} d_m = 2(1 - \rho(F_1, F_2)), \text{ como } 0 \leq \rho \leq 1,
 \end{aligned}$$

então $2(1 - \rho(F_1, F_2)) \leq 2$. Destas propriedades segue-se que:

$$|F_1(E) - F_2(E)| \leq 2\|F_1 - F_2\| \text{ pois para qualquer conjunto}$$

$$\text{mensurável } E, |F_1(E) - F_2(E)| = \left| \int_E P_1(x) d_m - \int_E P_2(x) d_m \right| \leq$$

$$\leq \int_E |P_1(x) - P_2(x)| d_m \leq \int_{\mathbb{R}} |P_1(x) - P_2(x)| d_m \leq 2\|F_1 - F_2\|.$$

Considerando funções de distribuição discreta, enunciaremos:

TEOREMAS FUNDAMENTAIS:

Seja F função de distribuição discreta com probabilidades $P_1, P_2, \dots, P_k$ para os $A_1, A_2, \dots, A_k$ eventos respectivamente. Seja n_i o número de ocorrências do evento i em n observações, denotamos

$$S_n = \left(\frac{n_1}{n}, \dots, \frac{n_k}{n} \right)$$

a distribuição empírica, logo por definição,

$$\|F - S_n\|^2 = \sum_{i=1}^k \left(\sqrt{\frac{n_i}{n}} - \sqrt{p_i} \right)^2 = 2 \left(1 - \sum_{i=1}^k \sqrt{\frac{n_i}{n}} \sqrt{p_i} \right) \quad (2.6)$$

$$\text{Assim} \quad \rho = \sum_{i=1}^k \sqrt{\frac{n_i}{n}} \sqrt{p_i} \quad (2.7)$$

TEOREMA 1: Quando a variável aleatória tem distribuição F então:

$$P_r \{ \|F - S_n\|^2 < \frac{k-1}{n} t \} \geq 1 - \frac{1}{t} \text{ para todo } t > 0 \quad (2.8)$$

PROVA: $p_1, \dots, p_{k'}, > 0$ $p_{k'+1} = \dots = p_k = 0$ $k' \leq k$ logo de 2.6

$$\Delta^2 = \|F - S_n\|^2 = \sum_{i=1}^k \left(\sqrt{\frac{n_i}{n}} - \sqrt{p_i} \right)^2 \text{ mas}$$

$$\frac{n_i}{n} - p_i = \left(\sqrt{\frac{n_i}{n}} + \sqrt{p_i} \right) \left(\sqrt{\frac{n_i}{n}} - \sqrt{p_i} \right) \quad (2.9)$$

$$\text{donde } \sqrt{\frac{n_i}{n}} - \sqrt{p_i} = \frac{\frac{n_i}{n} - p_i}{\sqrt{\frac{n_i}{n}} + \sqrt{p_i}} \quad (2.10)$$

$$\text{assim } \Delta^2 = \sum_{i=1}^{k'} \left(\frac{\frac{n_i}{n} - p_i}{\sqrt{\frac{n_i}{n}} + \sqrt{p_i}} \right)^2 + \sum_{i=k'+1}^k \frac{n_i}{n} \quad (2.11)$$

$$= \sum_{i=1}^{k'} \frac{\left(\frac{n_i}{n} - p_i\right)^2}{\frac{n_i}{n} + p_i + 2\sqrt{\frac{n_i p_i}{n}}} + \sum_{i=k'+1}^k \frac{n_i}{n} \quad (2.12)$$

$$\leq \sum_{i=1}^{k'} \frac{1}{p_i} \left(\frac{n_i}{n} - p_i\right)^2 + \sum_{i=k'+1}^k \frac{n_i}{n} \quad (2.13)$$

$$E(\Delta^2) \leq \sum_{i=1}^{k'} \frac{1}{p_i} E\left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{k'} \frac{1}{n^2 p_i} E(n_i - n p_i)^2 \quad (2.14)$$

$$= \sum_{i=1}^{k'} \frac{1 - p_i}{n} \quad (2.15)$$

$$= \frac{k'-1}{n} \leq \frac{k-1}{n} \quad (2.16)$$

onde E representa a esperança em relação a F .

Usando a desigualdade de Markov

$$P_r\{\Delta^2 < E(\Delta^2)t\} \geq 1 - \frac{1}{t} \quad (2.17)$$

Para todo $t > 0$, portanto:

$$P_r\left\{\Delta^2 < \frac{k-1}{n} t\right\} \geq 1 - \frac{1}{t}, \quad t > 0 \quad (2.18)$$

Sabemos $\chi^2 = \sum_{i=1}^k \frac{(n_i - n_{p_i})^2}{n_{p_i}} \approx \chi_{k-1}^2$ e além disso

de 2.12:

$$\begin{aligned} \Delta^2 &= \sum_{i=1}^k \frac{\left(\frac{n_i - n_{p_i}}{n} \right)^2}{\left(\frac{\sqrt{n_i} + \sqrt{n_{p_i}}}{\sqrt{n}} \right)^2} \\ &= \sum_{i=1}^k \frac{n(n_i - n_{p_i})^2}{n^2 (\sqrt{n_i} + \sqrt{n_{p_i}})^2} \end{aligned} \quad (2.19)$$

$$= \frac{1}{n} \sum_{i=1}^k \frac{(n_i - n_{p_i})^2 / n_{p_i}}{\left(1 + \sqrt{\frac{n_i}{n_{p_i}}} \right)^2} \quad (2.20)$$

então temos:

TEOREMA 2: Quando a variável aleatória tem distribuição discreta F , com $P_i > 0$ $i = 1, \dots, k$ e n é grande obtemos:

$$P_r \{ \|F - S_n\|^2 < \delta^2 \} \approx P_r \{ \chi_{k-1}^2 < 4n \delta^2 \} \quad (2.21)$$

para todo $\delta \geq 0$, onde χ_{k-1}^2 é a distribuição qui-quadrado com

$k-1$ graus de liberdade.

PROVA:

$$P_r \{ \|F - S_n\|^2 < \delta^2 \} = P_r \left\{ \frac{1}{n} \sum_{i=1}^k \left[(n_i - n_{P_i})^2 / n_{P_i} \right] \left[1 + \sqrt{\frac{n_i}{n_{P_i}}} \right]^{-2} < \delta^2 \right\} \quad (2.22)$$

$$= P_r \{ \chi_{k-1}^2 < 4n\delta^2 \} \quad (2.23)$$

Depois da aplicação do teorema de Slutsky e do fato que

$$\sum_{i=1}^k \frac{(k_i - n_{P_i})^2}{n_{P_i}}$$

se distribue como uma χ_{k-1}^2 .

Nosso objetivo é determinar uma regra tal que, para toda distribuição discreta F_0 dado $\delta_0 > 0$, possamos decidir se uma variável aleatória tem distribuição F_0 ou uma distribuição F com $\|F - F_0\| > \delta_0$.

Antes de enunciar o teorema seguinte faremos as seguintes considerações: Quando X tem distribuição F_0 , temos por 2.8

$$P_r \{ \|F_0 - S_n\|^2 < \frac{k-1}{n} t \} \geq 1 - \frac{1}{t} \quad (2.24)$$

e quando tem distribuição F com $\|F - F_0\| \geq \delta_0$,

$$\|F_0 - S_n\| \geq \|F_0 - F\| - \|F - S_n\|$$

$$\geq \delta_0 - \|F - S_n\| \quad (2.25)$$

então,

$$P_r \{ \|F_0 - S_n\| > \delta_0 - \sqrt{\frac{k-1}{n}} t \} \geq 1 - \frac{1}{t} \quad (2.26)$$

$$\text{e além disso quando } n \geq 4(k-1)\delta_0^{-2} t \quad (2.27)$$

$$\text{temos } \delta_0 - \sqrt{\frac{k-1}{n}} t \geq \sqrt{\frac{k-1}{n}} t \quad (2.28)$$

TEOREMA 3: Para todo $t > 0$ e $n \geq 4(k-1)\delta_0^{-2} t$ temos:

a) Quando χ tem distribuição F_0 :

$$P_r \{ \|F_0 - S_n\|^2 < \frac{k-1}{n} t \} \geq 1 - \frac{1}{t} \quad (2.29)$$

b) Quando χ tem distribuição F com $\|F - F_0\| > \delta_0$,

$$P_r \{ \|F_0 - S_n\|^2 \geq \frac{k-1}{n} t \} \geq 1 - \frac{1}{t} \quad (2.30)$$

PROVA:

a) Segue-se do teorema 1.

b) $\|F - S_n\| > \delta_0$

$$\|F_0 - S_n\| \geq \|F_0 - F\| - \|F - S_n\| \geq \delta_0 - \|F - S_n\| \quad (2.31)$$

de 2.26 e 2.28 então concluímos.

$$\begin{aligned}
 P_r \{ \|F_0 - S_n\| \geq \sqrt{\frac{k-1}{n}} t \} &\geq P_r \{ \|F - S_n\| \geq \\
 &\geq \delta_0 - \sqrt{\frac{k-1}{n}} t \} \geq 1 - \frac{1}{t}
 \end{aligned}
 \quad (2.32)$$

Portanto:

$$P_r \{ \|F_0 - S_n\|^2 \geq \frac{k-1}{n} t \} \geq 1 - \frac{1}{t} \quad (2.33)$$

TEOREMA 4: As distribuições F e F_0 com $\|F - F_0\| \geq \delta_0$ estão numa classe (C_n) de distribuições tal que χ^2 baseada em n observações da variável aleatória tem assintoticamente uma χ^2_{k-1} . Seja η todo número positivo menor que δ_0 .

a) Quando X tem distribuição F_0 temos:

$$P_r \{ \|F_0 - S_n\|^2 < \eta^2 \} \doteq P_r \{ \chi^2_{k-1} < 4n\eta^2 \} \quad (2.34)$$

b) Quando X tem distribuição F temos

$$P_r \{ \|F_0 - S_n\|^2 \geq \eta^2 \} \leq P_r \{ \chi^2_{k-1} < 4n(\delta_0 - \eta)^2 \} \quad (2.35)$$

PROVA:

a) Segue-se do teorema 2.

b) $\|F - F_0\| > \delta_0$, de 2.25 temos que

$$P_r \{ \|F_0 - S_n\| > \eta \} \leq P_r \{ \delta_0 - \|F - S_n\| > \eta \} \quad (2.36)$$

$$= P_r \{ \|F - S_n\| < \delta_0 - \eta \} \quad (2.37)$$

$$= P_r \{ \|F - S_n\|^2 < (\delta_0 - \eta)^2 \} \quad (2.38)$$

Segue-se o resultado do teorema 2.

PROBLEMA DE DUAS AMOSTRAS

Sejam X, y , não necessariamente independentes, variáveis aleatórias as quais têm distribuições (marginais) discretas:

$$F = (p_1^0, p_2^0, \dots, p_k^0) \quad \text{e} \quad G = (q_1^1, q_2^1, \dots, q_k^1)$$

sobre os mesmos eventos $1, 2, \dots, k$. Sejam $n_1, \dots, n_k$ com $\sum n_i = n$ e $m_1, \dots, m_k$ com $\sum m_i = m$ observações em X e y .

Sejam também $S_n = (\frac{n_1}{n}, \dots, \frac{n_k}{n})$ e $S'_m = (\frac{m_1}{m}, \dots, \frac{m_k}{m})$ as

distribuições empíricas. Então temos:

TEOREMA 5: Seja η todo número positivo menor que δ_0 .

a) Quando $\|F - G\| = 0$ temos

$$P_r \{ \|S_n - S'_m\| < \eta \} \geq 1 - \frac{k-1}{n^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)^2 \quad (2.39)$$

b) Quando $\|F - G\| > \delta_0$ temos

$$P_r \{ \|S_n - S'_m\| \geq \eta \} \geq 1 - \frac{k-1}{(\delta_0 - \eta)^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)^2 \quad (2.40)$$

Ademais, se X, Y são independentes com $\|F - G\| = 0$ temos

$$c) \quad P_r \{ \|S_n - S'_m\| < \eta \} \geq 1 - \frac{4(k-1)}{\eta^2} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{16(k-1)^2}{\eta^4 nm} \quad (2.41)$$

Quando X, Y são independentes com $\|F - G\| \geq \delta_0$ temos

$$d) \quad P_r \{ \|S_n - S'_m\| \geq \eta \} \geq 1 - \frac{4(k-1)}{(\delta_0 - \eta)^2} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{16(k-1)^2}{(\delta_0 - \eta)^4 nm} \quad (2.42)$$

PROVA:

$$a) \quad \|S_n - S'_m\| \leq \|S_n - F\| + \|G - S'_m\| \quad \text{Se } \|F - G\| = 0 \quad \text{então}$$

$$P_r \{ \|S_n - S'_m\| < \eta \} \geq P_r \{ \|S_n - F\| + \|G - S'_m\| < \eta \}$$

Por 2.17:

$$\geq 1 - \frac{1}{\eta^2} E (\|S_n - F\| + \|G - S'_m\|)^2 \quad (2.43)$$

$$\geq 1 - \frac{1}{\eta^2} (\sqrt{E(\|F - S_n\|^2)} + \sqrt{E(\|G - S'_m\|^2)})^2 \quad (2.44)$$

$$\geq 1 - \frac{1}{\eta^2} \left(\sqrt{\frac{k-1}{n}} + \sqrt{\frac{k-1}{m}} \right)^2 \quad (2.45)$$

$$> 1 - \frac{k-1}{\eta^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)^2 \quad (2.46)$$

$$b) \|S_n - S'_m\| \geq \|F - G\| - \|F - S_n\| - \|G - S'_m\| \geq \delta_0 - \|F - S_n\| - \|G - S'_m\|$$

logo:

$$P_r\{\|S_n - S'_m\| > \eta\} \geq P_r\{\delta_0 - \|F - S_n\| - \|G - S'_m\| > \eta\} \quad (2.47)$$

$$= P_r\{\|F - S_n\| + \|G - S'_m\| < \delta_0 - \eta\} \quad (2.48)$$

$$\leq 1 - \frac{1}{(\delta_0 - \eta)^2} E(\|F - S_n\| + \|G - S'_m\|)^2 \quad (2.49)$$

$$\leq 1 - \frac{1}{(\delta_0 - \eta)^2} (\sqrt{E(\|F - S_n\|^2)} + \sqrt{E(\|G - S'_m\|^2)})^2 \quad (2.50)$$

$$\leq 1 - \frac{1}{(\delta_0 - \eta)^2} \left(\sqrt{\frac{k-1}{n}} + \sqrt{\frac{k-1}{m}} \right)^2 \quad (2.51)$$

$$\leq 1 - \frac{k-1}{(\delta_0 - \eta)^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)^2 \quad (2.52)$$

$$c) P_r\{\|S_n - S'_m\| < \eta\} \geq P_r\{\|F - S_n\| < \frac{1}{2}\eta, \|G - S'_m\| < \frac{1}{2}\eta\} \quad (2.53)$$

$$\geq P_r\{\|F - S_n\| < \frac{1}{2}\eta\} P_r\{\|G - S'_m\| < \frac{1}{2}\eta\} \quad (2.54)$$

$$\geq \left(1 - \frac{4(k-1)}{n\eta^2}\right) \left(1 - \frac{4(k-1)}{m\eta^2}\right)$$

$$= 1 - \frac{4(k-1)}{\eta^2} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{16(k-1)^2}{\eta^4 nm} \quad (2.56)$$

d) $\|S_n - S'_m\| \geq \|F-G\| - \|G-S'_m\| - \|F-S_n\| \geq \delta_0 - \|G-S'_m\| - \|F-S_n\|$ Logo:

$$P_r \{ \|S_n - S'_m\| \geq \eta \} \geq P_r \{ \delta_0 - \|G - S'_m\| - \|F - S_n\| \geq \eta \} \quad (2.57)$$

$$= P_r \{ \|F-S_n\| + \|G-S'_m\| < \delta_0 - \eta \} \quad (2.58)$$

$$= P_r \left\{ \|F-S_n\| < \frac{\delta_0 - \eta}{2}, \|G-S'_m\| < \frac{\delta_0 - \eta}{2} \right\} \quad (2.59)$$

$$= P_r \left\{ \|F-S_n\| < \frac{\delta_0 - \eta}{2} \right\} P_r \left\{ \|G-S'_m\| < \frac{\delta_0 - \eta}{2} \right\} \quad (2.60)$$

$$\geq \left[1 - \frac{1}{\left(\frac{\delta_0 - \eta}{2} \right)^2} E(\|F-S_n\|)^2 \right].$$

$$\left[1 - \frac{1}{\left(\frac{\delta_0 - \eta}{2} \right)^2} E(\|G-S'_m\|)^2 \right] \quad (2.61)$$

$$\geq \left[1 - \frac{4}{(\delta_0 - \eta)^2} \left(\sqrt{\frac{k-1}{n}} \right)^2 \right].$$

$$\left[1 - \frac{4}{(\delta_0 - \eta)^2} \left(\sqrt{\frac{k-1}{m}} \right)^2 \right] \quad (2.62)$$

$$\doteq 1 - \frac{4(k-1)}{(\delta_0 - \eta)^2} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{16(k-1)^2}{(\delta_0 - \eta)^2 nm} \quad (2.63)$$

TEOREMA 6: Sejam X, y independentes e F, G pertencentes a classe (C_n) . Seja η todo número positivo menor que δ_0 . Então quando $\|F - G\| = 0$ temos:

$$a) P_r\{\|S_n - S'_m\| < \eta\} \leq P_r\left\{\frac{1}{2n} \chi_{k-1}^2 + \frac{1}{2m} \chi_{k-1}'^2 < \eta^2\right\} \quad (2.64)$$

$$\leq 1 - \frac{k-1}{2\eta^2} \left(\frac{1}{n} + \frac{1}{m} \right) \quad (2.65)$$

$$b) P_r\{\|S_n - S'_m\| < \eta\} \leq P_r\{\chi_{k-1}^2 < n\eta^2\} P_r\{\chi_{k-1}'^2 < m\eta^2\} \quad (2.66)$$

Quando $\|F - G\| \geq \delta_0$ temos:

$$c) P_r\{\|S_n - S'_m\| \geq \eta\} \leq P_r\left\{\frac{1}{2n} \chi_{k-1}^2 + \frac{1}{2m} \chi_{k-1}'^2 \leq (\delta_0 - \eta)^2\right\} \quad (2.67)$$

$$\leq 1 - \frac{k-1}{2(\delta_0 - \eta)^2} \left(\frac{1}{n} + \frac{1}{m} \right) \quad (2.68)$$

$$d) P_r\{\|S_n - S'_m\| \geq \eta\} \leq P_r\{\chi_{k-1}^2 \leq n(\delta_0 - \eta)^2\} P_r\{\chi_{k-1}'^2 \leq m(\delta_0 - \eta)^2\} \quad (2.69)$$

onde χ_{k-1}^2 , $\chi_{k-1}'^2$ são variáveis aleatórias independentes, cada uma delas com distribuição qui-quadrado com $k-1$ graus de

liberdade.

OUTRAS PESQUISAS

A distância e coeficiente de afinidade de Matusita têm sido muito usados em problemas de aplicação. Entre outras, foram usados por W. Hoeffding 1965, Allan W. Stoner 1979, G.J. Kobyluiski 1979, Ibrahim A. Ahmad 1980, Ned Glick 1974, além de ter sido usado por A.A. Khan e S.M. Ali 1973 para sugerir um novo coeficiente de associação, por S.N.U.A. Kirmani 1974 para propor um teste de bom ajuste, por Mathew Goldstein e outros 1976 para desenvolver um teste de independência e por Dillon e Goldstein 1978 para fazer análise discriminante no caso multinomial.

A seguir faremos um resumo das quatro últimas referências citadas.

1 - COEFICIENTE DE ASSOCIAÇÃO:

Sejam (x, A) , (y, B) espaços mensuráveis tal que $(x \times y, A \times B)$ é seu espaço produto. Seja (X, Y) variável aleatória com distribuição de probabilidade λ naquele espaço produto onde os espaços de variação de X, Y estão em x, y respectivamente. Definindo as distribuições marginais de X, Y por μ, ν respectivamente, definimos sua medida produto como $\lambda^* = \mu \times \nu$ em $(x \times y, A \times B)$. Agora, se m é uma medida em $(x \times y, A \times B)$ que faz λ e λ^* absolutamente contínuas, então, λ e λ^* têm suas respectivas fun-

ções de densidade $f(x,y), f^*(x,y)$ com relação a m .

Sobre estas definições A.H.Khan e S.M.Ali em 1973 propõem a seguinte medida de associação:

$$M^2(x,y) = \frac{1}{2} \int_{x \times y} (\sqrt{f} - \sqrt{f^*})^2 d_m = \frac{1}{2} \|\lambda - \lambda^*\|^2 \quad (2.70)$$

$$= 1 - \int_{x \times y} \sqrt{ff^*} dm = 1 - \rho(\lambda, \lambda^*) \quad (2.71)$$

onde $M(X,y)$ expressa uma medida de associação entre as variáveis X,y .

Se X,y são variáveis aleatórias discretas, tal que sua função de probabilidade conjunta possa ser arranjada numa tabela $r \times S$ com:

$$\pi_{ij} = P(X = x_i, Y = y_j) \quad i = 1, 2, \dots, r \quad j = 1, 2, \dots, S \quad (2.72)$$

então, um coeficiente de associação é da forma:

$$\left\{ \frac{1}{2} \sum_{i=1}^r \sum_{j=1}^S (\sqrt{\pi_{ij}} - \sqrt{\pi_{i.} \pi_{.j}})^2 \right\}^{1/2} \quad (2.73)$$

onde $\sum_{j=1}^S \pi_{ij} = \pi_{i.}$ e $\sum_{i=1}^r \pi_{ij} = \pi_{.j}$

Se X,y são variáveis aleatórias contínuas, admitindo densidades com relação a medida de Lebesgue, então o coeficiente é:

$$\frac{1}{2} \int_y \int_x (\sqrt{f(x,y)} - \sqrt{f_1(x)} \sqrt{f_2(y)})^2 dx dy = \frac{1}{2} \|F_1 - F_2\|^2 \quad (2,74)$$

onde $f_1(x)$, $f_2(y)$ são densidades marginais de x, y e F_1, F_2 suas funções de distribuição.

Em vista das exigências de Renyi 1952, o coeficiente de associação $M(X, y)$ satisfaz algumas propriedades importantes:

- a) $0 \leq M(X, y) \leq 1$
- b) $M(X, y) = 0$ se e somente se X, y são estocasticamente independentes.
- c) $M(X, y) = 1$ se e somente se λ, λ^* são mutuamente singulares
- d) É simétrico
- e) É invariante sob transformações um a um.
- f) Se X, y têm distribuição Normal Bivariada com coeficiente de correlação r , então, $M(X, y)$ é função crescente de $|r|$.
- g) Se $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ é uma partição de $X \sim N_n(\alpha, \Sigma^{-1})$ onde X_1 é p -vetor, X_2 é q -vetor, $p + q = n$, $q \geq p$ então $M(X_1, X_2)$ é uma função monotônica-crescente de cada correlação canônica.

Estas propriedades fazem o coeficiente $M(X, y)$ uma razoável alternativa para o teste qui-quadrado.

2 - TESTE DE AJUSTAMENTO:

S.N.U.A.Kirmani 1974 sugere o seguinte critério para testar $H_0 : F = F_0$ V.S. $H_1 : F \neq F_0$ com F_0 função distribui

ção contínua em $(-\infty, \infty)$ completamente especificada e $F(x)$ contínua em $(-\infty, \infty)$ e absolutamente contínua com relação a F_0 .

Tomando uma amostra aleatória $X_1, \dots, X_n$ de $F(X)$ determinam-se suas estatísticas de ordem, assumindo além disso que $X_{(0)} = -\infty$ e que $X_{(n+1)} = \infty$. Seja $V_i = F_0(X_{(i)}) - F_0(X_{(i-1)})$ para $i=1, 2, \dots, n+1$, logo $\Pi_1 = (V_1, V_2, \dots, V_{n+1})$ é uma distribuição multinomial.

Também sabemos que para todo i :

$$E_{H_0}(V_i) = \frac{1}{n+1} \quad (2.75)$$

tal que $\Pi_2 = (E_{H_0}(V_1), E_{H_0}(V_2), \dots, E_{H_0}(V_{n+1}))$

$$= \left(\frac{1}{n+1}, \frac{1}{n+1}, \dots, \frac{1}{n+1} \right)$$

é outra distribuição multinomial tendo $(n+1)$ categorias. Logo a medida de distância de Matusita entre Π_1 e Π_2 é:

$$M^2(X_1, \dots, X_n) = \sum_{i=1}^{n+1} (\sqrt{V_i} - \sqrt{1/(n+1)})^2 = \|\Pi_1 - \Pi_2\|^2 \quad (2.76)$$

$$= 2 \left(1 - \frac{1}{\sqrt{n+1}} \sum_{i=1}^{n+1} \sqrt{V_i} \right) = 2(1 - \rho(\Pi_1, \Pi_2)) \quad (2.77)$$

rejeitando H_0 se $M(X_1, X_2, \dots, X_n)$ é grande.

O teste proposto baseado em espaços amostrais V_i , pode ser usado para detectar diferenças entre funções de densidade. Considera, ademais a significância de toda observação nas caudas, não requer agrupamento de dados e é significativo sem referência a qualquer conjunto específico de alternativas.

A hipótese H_0 poderia ser rejeitada no nível de significância α se $\gamma \leq K(\alpha)$ onde $\gamma = \sum_{i=1}^{n+1} \sqrt{V_i}$ e $K(\alpha)$ é escolhido tal que $P(\gamma \leq K(\alpha)/H_0) = \alpha$. Kirmani usando os resultados de Darling 1953 e LeCam 1958 prova a normalidade assintótica de γ . Comparando o poder do teste com outros baseados em espaços amostrais como $\sum V_i^2$ de Greenwood, K^* de Yale, K e J onde

$$K = \frac{1}{n+1} \sum \log((n+1)V_i) \quad (2.78)$$

$$K^* = \sum V_i \log\left(\frac{1}{(n+1)V_i}\right) \quad (2.79)$$

$$J = K^* + K \quad (2.80)$$

o teste baseado em γ é superior aos anteriores. Mesmo assim os testes de Greenwood e K^* são assintoticamente mais poderosos.

3 - TESTE DE INDEPENDÊNCIA EM TABELAS DE CONTIGÊNCIA

Mathew Goldstein e outros 1976 baseados na distância de Matusita propõem um teste de independência para tabelas de con-

gência multinomias.

Supondo X, Y, Z variáveis aleatórias com distribuição conjunta dada por:

$$P(X = x_i, Y = y_J, Z = z_k) = \theta_{iJk} \quad (2.81)$$

$$i=1,2,\dots, r_1 \quad J = 1,2,\dots, r_2 \quad k = 1,2,\dots, r_3 \quad e$$

$$\theta_{i..} = \sum_J \sum_K \theta_{iJk} \quad \theta_{.J.} = \sum_i \sum_k \theta_{iJk} \quad \theta_{..K} = \sum_i \sum_J \theta_{iJk}$$

A afinidade entre θ_{iJk} , e, $\theta_{i..}$, $\theta_{.J.}$, $\theta_{..K}$ será dada por:

$$\sum_i \sum_J \sum_K \sqrt{\theta_{iJk}} \sqrt{\theta_{i..} \theta_{.J.} \theta_{..K}} =$$

$$\rho(\theta_{iJk}, \theta_{i..}, \theta_{.J.}, \theta_{..K}) = \rho \quad (2.82)$$

e a distância entre θ_{iJk} , e, $\theta_{i..}$, $\theta_{.J.}$, $\theta_{..K}$ por

$$\begin{aligned} \|\theta_{iJk} - \theta_{i..} \theta_{.J.} \theta_{..K}\|^2 &= \sum_i \sum_J \sum_K \{ \sqrt{\theta_{iJk}} - \sqrt{\theta_{i..} \theta_{.J.} \theta_{..K}} \}^2 = \\ &= 2(1 - \rho) \end{aligned} \quad (2.83)$$

Geralmente ρ é desconhecido devendo portanto ser esti-

mado. Grandes valores de $\hat{\rho}$, isto é, próximo de 1, nos leva a concluir que a hipótese de independência é consistente.

Usando estimadores M.L. de $\theta_{ijk}, \theta_{i..}, \theta_{.j.}, \theta_{...k}$ digamos $P_{ijk}, P_{i..}, P_{.j.}, P_{...k}$, obtemos:

$$\hat{\rho} = \frac{\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} \sqrt{P_{ijk} P_{i..} P_{.j.} P_{...k}}}{\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} P_{ijk}} \quad (2.84)$$

para o caso de uma tabela de contingência $r_1 \times r_2 \times r_3$. Usando o argumento de Khan e Ali 1973 prova-se também que assintoticamente.

$$\hat{\rho} = 1 - \frac{1}{8n} \chi^2_{(r_1 r_2 r_3 - r_1 - r_2 - r_3 + 2)} \quad (2.85)$$

que também pode ser expressado como:

$$\hat{\rho} = \frac{1}{n} \sum_i \sum_j \sum_k \sqrt{\frac{O_{ijk}}{E_{ijk}}} \quad (2.86)$$

onde O_{ijk} é frequência observada na cela ijk contanto que E_{ijk} é a frequência esperada sob independência.

Então a região crítica para rejeitar a hipótese de independência no nível α está definida por:

$$M = 8n \left\{ 1 - \frac{1}{n} \sum_i \sum_j \sum_k \sqrt{\frac{O_{ijk}}{E_{ijk}}} \right\} > \chi^2_{r_1 r_2 r_3 - r_1 - r_2 - r_3 + 2} \quad (2.87)$$

onde pode-se ver que $\hat{p}$ e χ^2 são assintoticamente equivalentes com exceção de constantes.

A estatística M pode facilmente generalizar-se para ordens superiores. Em um estudo de Monte Carlo feito para tamanhos amostrais 100, 150, 200 em tabelas 2×3 e $2 \times 2 \times 2$ se nota pouca diferença entre os valores calculados para LR (razão de verossimilhança), χ^2 e M . [Vide Goldstein e Outros, 1976].

Concluindo: A estatística χ^2 (Pearson) difere de LR por termos da ordem $n^{-1/2}$ enquanto que M difere de χ^2 por termos dependendo da ordem de $(P_{iJK} - P_{i..}P_{.J.}P_{..K}) / (P_{i..}P_{.J.}P_{..K})$.

4 - ANÁLISE DISCRIMINANTE NO CASO MULTINOMIAL

William Dillon e Mathew Goldstein 1978 comparam diferentes métodos de Análise Discriminante para classificação Multinomial, discutem o modelo de Martin-Bradley, a representação de Bahadur a função discriminante linear, além do método proposto, baseado na distância de Matusita. Mesmo seu desempenho sendo bom, em estudos de Monte Carlo, nos limitaremos apenas a apresentar sua forma.

$$\|F - G\|^2 = \sum_{i=1}^S (\sqrt{p_i} - \sqrt{q_i})^2$$

é a distância de Matusita definida para duas distribuições amostrais F, G com probabilidades p_i, q_i para $i = 1, 2, \dots, S$ respectivamente.

Suponhamos que as amostras aleatórias independentes de tamanho n e m são obtidas de duas populações Π_1 e Π_2 respectivamente e que além disso S_n e S_m representam suas distribuições empíricas.

Seja $X = x$ um vetor observação gerado de Π_1 ou Π_2 . Consideramos a regra:

$$\text{Classifica } X \text{ em } \Pi_1 \text{ se } \|S_{n+1} - S_m\| > \|S_n - S_{m+1}\| \quad (2.88)$$

$$\text{Classifica } X \text{ em } \Pi_2 \text{ se } \|S_{n+1} - S_m\| < \|S_n - S_{m+1}\|$$

Se as duas distâncias são iguais a observação, classifica-se aleatoriamente. Por outro lado a regra classifica X naquela população com maior distância distribucional, baseada na amostra.

Escrevendo de novo a regra obtemos:

Classifica X em Π_1 se:

$$\sum_{i=1}^S \{ (n_i^*/(n+1))^{1/2} - (m_i^*/m)^{1/2} \}^2 > \sum_{i=1}^S \{ (n_i/n)^{1/2} - (m_i^*/(m+1))^{1/2} \}^2 \quad (2.89)$$

onde $n_i^* = n_i$ e $m_i^* = m_i$ para $i \neq j$ e $n_i^* = n_i + 1$, $m_i^* = m_i + 1$ para $i = j$ se $X = x$ é uma observação pertencente ao estado J .

Suponhamos que X pertença realmente ao estado J , logo depois de alguma álgebra, nossa inequação fica:

$$\begin{aligned}
& [(n_i+1)/(n+1)]^{1/2} (m_J/m)^{1/2} + \sum_{i \neq J} [n_i/(n+1)]^{1/2} (m_i/m)^{1/2} < \\
& < (n_J/n)^{1/2} [(m_J+1)/(m+1)]^{1/2} + \sum_{i \neq J} (n_i/n)^{1/2} [m_i/(m+1)]^{1/2} \quad (2.90)
\end{aligned}$$

Agora, se $m = n$ obtemos:

$$\text{Classifica } X \text{ em } \Pi_1 \text{ se } n_J > m_J. \quad (2.91)$$

Note-se que a última desigualdade é exatamente aquela produzida pela regra Bayes quando é substituída por sua análoga baseada na amostra usando estimadores não-paramétricos das densidades f_1, f_2 ou estimadores M.L. dos parâmetros multinomiais desconhecidos.

A equivalência é baseada na equiprobabilidade a priori. Se os tamanhos amostrais não são iguais, então a regra não é equivalente ao método da razão de verossimilhança baseada na amostra. Em geral, pode-se escrever como

$$\frac{[m_J(n_J+1)]^{1/2} + \sum_{i \neq J} (n_i m_i)^{1/2}}{[n_J(m_J+1)]^{1/2} + \sum_{i \neq J} (n_i m_i)^{1/2}} < \left[\frac{m(n+1)}{n(m+1)} \right]^{1/2} \quad (2.92)$$

Um problema que aparece quando fazemos classificação multinomial, é que, se temos 0 no estado J de Π_1 , isto pode representar uma maior frequência teórica inteiramente diferente de 0 no estado J de Π_2 , especialmente se o tamanho amostral de Π_1 é muito menor comparado com o tamanho de Π_2 . Sob o suposto de igualdade de probabilidades a priori a regra usual multinomial, X pertence ao estado J classificando X em Π_1 se $\frac{n_J}{n} > \frac{m_J}{m}$, baseada na amostra, é insensível neste problema. Assim, se J é zero em Π_1 e não é zero em Π_2 , X será classificado em Π_2 .

A regra 2.92 é sensível neste caso. Suponhamos $n < m$, tal que $m(n+1)/n(m+1) > 1$ e $n_J=0$ mas $m_J > 0$, então a regra classifica X em Π_1 somente se:

$$\sqrt{m_J} < \sum_{i \neq J} (n_i m_i)^{1/2} (\{m(n+1)/n(m+1)\}^{1/2} - 1) \quad (2.93)$$

CAPÍTULO III

CRITICA AOS RESULTADOS DE MATUSITA E SIMULAÇÕES

INTRODUÇÃO

Apresentamos nos capítulos anteriores, as características gerais das estatísticas propostas para o ajuste de duas amostras, particularmente consideramos em detalhe a publicação de Matusita feita no Annals Mathematical Statistics no ano de 1955. A seguir faremos algumas críticas ao citado artigo mostrando o mau comportamento das desigualdades propostas ao mesmo tempo que exploramos as boas características da distância de Kameo Matusita para formular a nossa proposta de teste.

EXEMPLOS E CRÍTICAS

A partir da tabela 1.1 obtemos:

TABELA 3.1

CLASSES	IDADE	freq _M	freq _F	P _{i_n}	q _{i_F}	(P _{i_n}) ^{1/2}	(q _{i_F}) ^{1/2}
01	0-4	12	6	0.022	0.022	0.148	0.148
02	5-9	72	63	0.133	0.235	0.365	0.485
03	10-14	145	63	0.268	0.235	0.518	0.485
04	15-19	116	116	0.214	0.160	0.462	0.400
05	20-24	52	25	0.096	0.093	0.310	0.305
06	25-29	27	16	0.050	0.060	0.223	0.245
07	30-34	28	10	0.052	0.037	0.228	0.192
08	35-39	21	7	0.039	0.026	0.197	0.161
09	40-44	23	13	0.043	0.049	0.207	0.221
10	45-49	22	5	0.041	0.019	0.202	0.139
11	50-54	7	8	0.013	0.030	0.114	0.173
12	55-59	9	7	0.017	0.026	0.130	0.161
13	60-	7	2	0.013	0.007	0.114	0.083
TOTAL		541	268	~ 1.00	~ 1.00		

Seja

$$\chi^2 = \|S_n - S'_m\|^2 = \sum (\sqrt{P_i} - \sqrt{q_i})^2 \quad (3.1)$$

Dos resultados obtidos na tabela 3.1, temos que

$$\nabla = 0.178889$$

Consideramos agora, o teorema 5 do capítulo II com os dados apresentados na tabela 3.1, isto é, $n = 541$, $m = 268$, $k = 13$, $\eta = 0.179$ e tomemos $\delta_0 = 0.199$.

$$a) \quad \Pr\{\|S_n - S'_m\| \leq \eta\} \geq 1 - \frac{k-1}{\eta^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)^2 \quad (3.2)$$

$$\Leftrightarrow P_r\{\|S_{541} - S'_{268}\| < 0.179\} \geq -3.06209 \quad (3.3)$$

$$b) \quad P_r\{\|S_n - S'_m\| \geq \eta\} \geq 1 - \frac{k-1}{(\delta_0 - \eta)^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)^2 \quad (3.4)$$

$$\Leftrightarrow P_r\{\|S_{541} - S'_{268}\| \geq 0.179\} \geq -323.94643 \quad (3.5)$$

Supondo independência entre as duas amostras, teríamos:

$$c) \quad P_r\{\|S_n - S'_m\| < \eta\} \geq 1 - \frac{4(k-1)}{\eta^2} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{16(k-1)^2}{\eta^4 nm} \quad (3.6)$$

$$\Leftrightarrow P_r\{\|S_{541} - S'_{268}\| < 0.179\} \geq 8.123 \text{ absurdo!!?} \quad (3.7)$$

$$d) \quad P_r \{ \|S_n - S'_m\| \geq \eta \geq 1 - \frac{4(k-1)}{(\delta_0 - \eta)^2} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{16(k-1)^2}{(\delta_0 - \eta)^4 nm} \} \quad (3.8)$$

$$\Leftrightarrow P_r \{ \|S_{541} - S_{268}\| \geq 0.179 \} \geq 98649.906 \quad (3.9)$$

absurdo!!?

Em caso nenhum o teorema produz boa informação. Nos dois casos a) e b) mesmo sendo satisfeita a desigualdade não diz quanto se afasta da verdadeira probabilidade, e nos casos c) e d) supondo independência nos leva a um resultado absurdo.

Também em 3.2 e 3.6 vamos supor $n = m = 1$ daí obtemos:

$$P_r \{ \|S_n - S'_m\| < \eta \} = 1 \quad (3.10)$$

portanto, se testamos $H_0 : F = G$ contra $H_1 : F \neq G$ onde F e G são duas funções qualquer, certamente sempre aceitaríamos H_0 , independente do valor de η .

Ao analisar o teorema 6, em 2.66 supondo $m = n$ obtemos:

$$P_r \{ \|S_n - S'_m\| < \eta \} \geq P_r^2 (X_{k-1}^2 < n\eta^2) \quad (3.11)$$

usando 3.11 calculamos as probabilidades acumuladas da χ_{k-1}^2 e da distância de Matusita obtidas por simulação* para $k=10$,

* Os detalhes da simulação serão apresentados mais adiante.

tamanhos amostrais $NP = 25, 100$ e 400 e diferentes valores η^{**} [vide tabela 3.2]. As figuras 3.1 e 3.2 comparam os resultados obtidos.

Da tabela 3.2 e das figuras 3.1 e 3.2 concluimos que aproximação proposta por 3.11 não é boa, isto é, a probabilidade acumulada da distância de Matusita mesmo satisfazendo a desigualdade do teorema, fica muito longe da probabilidade acumulada da χ^2_{k-1} , sendo portanto em termos de uso prático, de escassa validade.

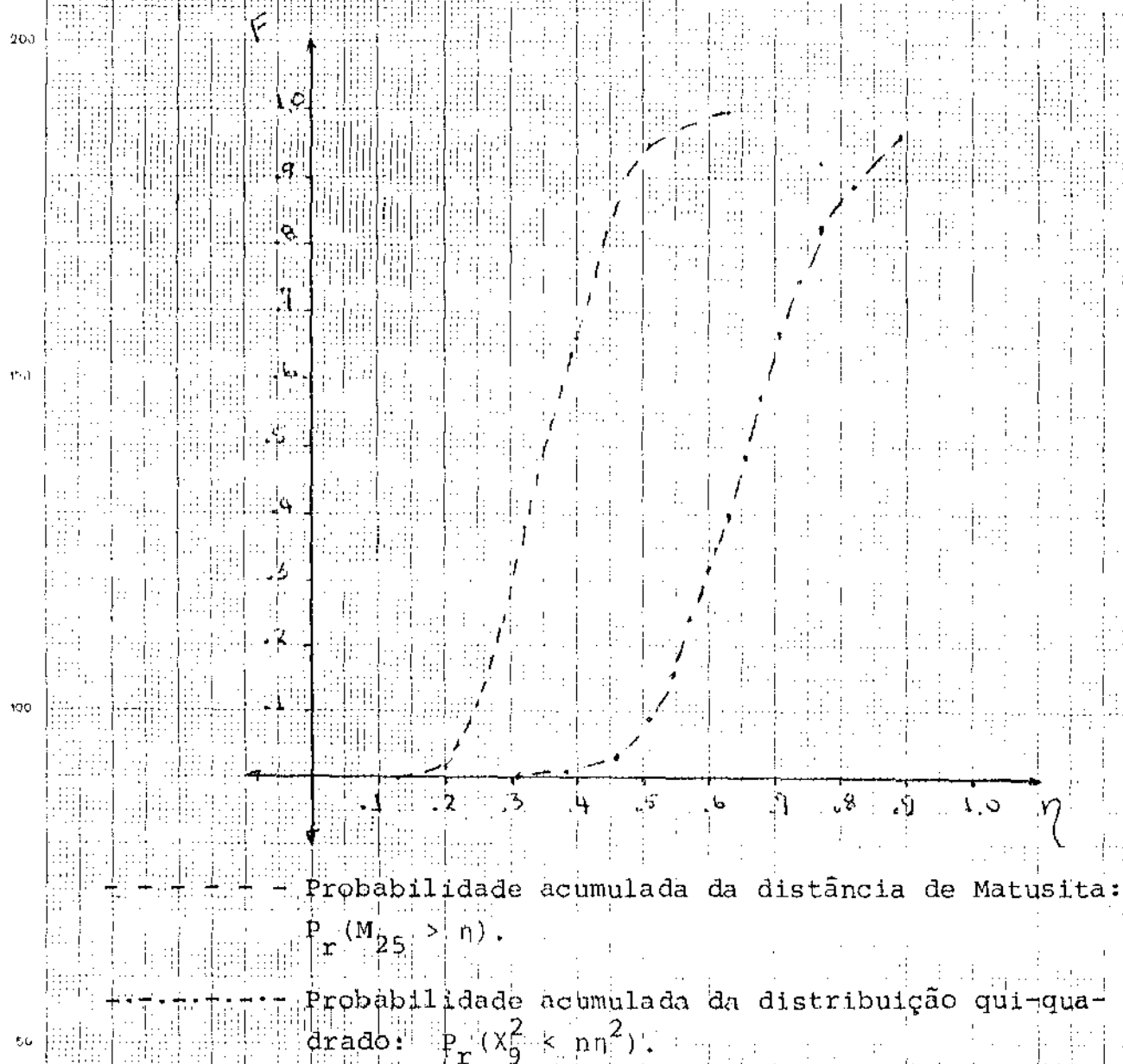
^{**} η é definido por $\eta = \sqrt{\frac{x^2}{N_p}}$.

TABELA 3.2

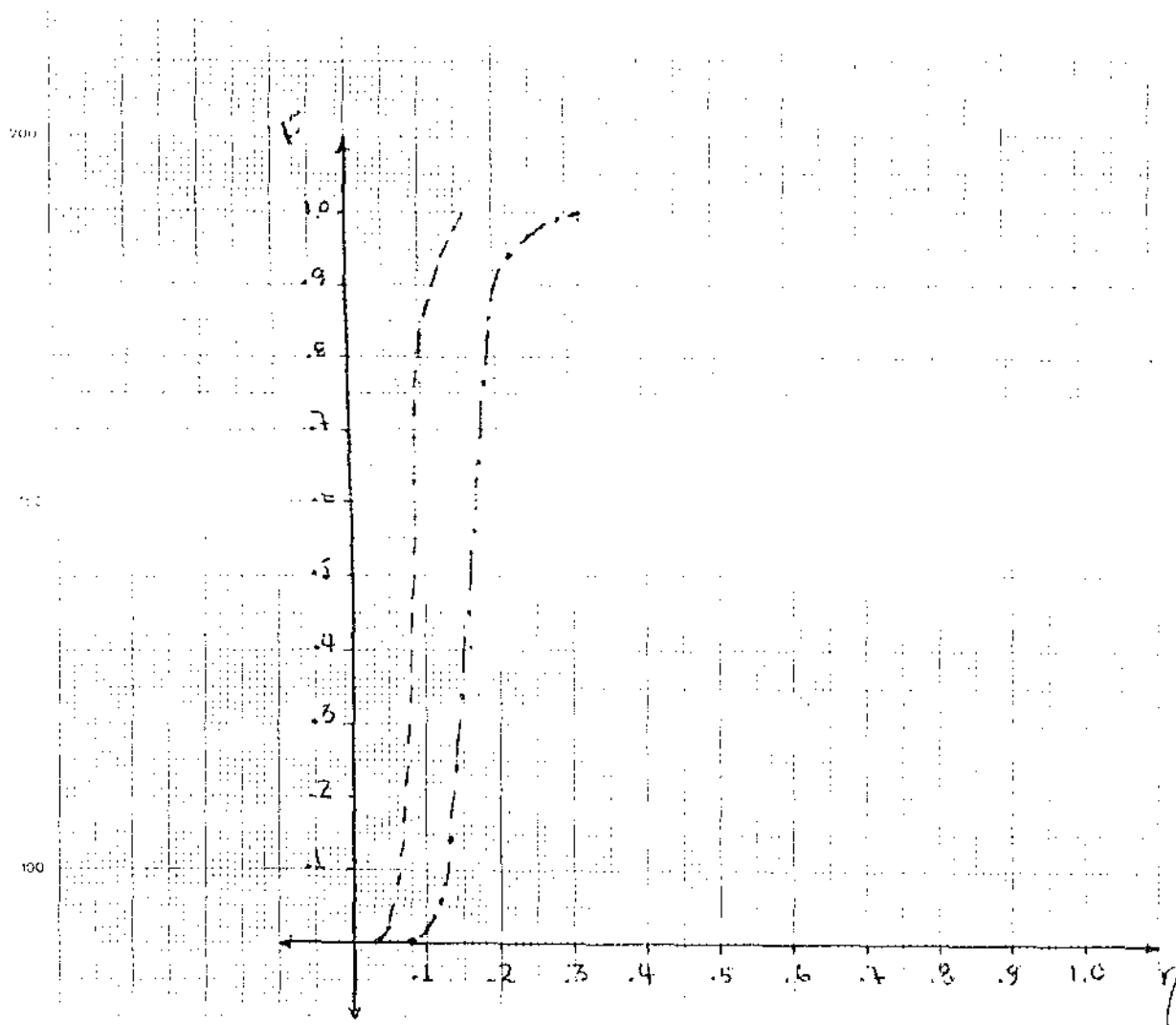
Tabela comparativa das probabilidades acumuladas da distribuição χ^2 e da distância de Matusita segundo simulação feita para tamanhos amostrais 25, 100 e 400, e $k = 10$.

χ^2	η_{25}	η_{100}	η_{400}	$P^2(\chi^2_q < 2n\eta^2)$	$P(M_{25} < \eta)$	$P(M_{100} < \eta)$	$P(M_{400} < \eta)$
0.4	0.13	0.06	0.03	0	0	0	0
0.6	0.15	0.08	0.04	0	0.003	0.007	0.002
0.8	0.18	0.09	0.04	0	0.007	0.017	0.002
1.0	0.20	0.10	0.05	0	0.015	0.041	0.022
1.2	0.22	0.11	0.05	0	0.045	0.081	0.022
1.4	0.24	0.12	0.06	0	0.088	0.141	0.094
1.6	0.25	0.13	0.06	0	0.113	0.225	0.094
1.8	0.27	0.13	0.07	0	0.161	0.225	0.229
2.0	0.28	0.14	0.07	0	0.199	0.328	0.229
2.4	0.31	0.15	0.08	0	0.327	0.434	0.447
2.8	0.33	0.17	0.08	0.001	0.396	0.652	0.447
3.2	0.36	0.18	0.09	0.002	0.526	0.738	0.671
3.6	0.38	0.19	0.09	0.004	0.589	0.810	0.671
4.0	0.40	0.20	0.10	0.008	0.661	0.874	0.835
4.4	0.42	0.21	0.10	0.014	0.728	0.926	0.835
4.8	0.44	0.22	0.11	0.022	0.805	0.953	0.937
5.2	0.46	0.23	0.11	0.034	0.858	0.974	0.937
5.6	0.47	0.24	0.12	0.049	0.889	0.981	0.979
6.0	0.49	0.24	0.12	0.068	0.922	0.981	0.979
6.4	0.51	0.25	0.13	0.090	0.947	0.986	0.995
6.8	0.52	0.26	0.13	0.117	0.961	0.993	0.995
7.2	0.54	0.27	0.13	0.147	0.967	0.998	0.995
7.6	0.55	0.28	0.14	0.181	0.974	0.998	0.999
8.0	0.57	0.28	0.14	0.217	0.987	0.998	0.999
8.4	0.58	0.29	0.14	0.256	0.993	0.999	0.999
8.8	0.59	0.30	0.15	0.296	0.996	0.999	0.999
9.2	0.61	0.30	0.15	0.338	0.998	0.999	0.999
9.6	0.62	0.31	0.15	0.380	0.998	1.000	0.999
9.8	0.63		0.16	0.401	1.000		1.000

FIGURA 3.1 - CURVAS COMPARATIVAS DAS PROBABILIDADES ACUMULADAS DA DISTRIBUIÇÃO QUI-QUADRA-DO E DA DISTÂNCIA DE MATUSITA PARA $K = 10$ e $NP = 25$.



ACUMULADAS DA DISTRIBUIÇÃO QUI-QUADRA-
DO E DA DISTÂNCIA DE MATUSITA PARA
 $K = 10$ e $NP = 400$.



----- Probabilidade acumulada da distância de Matusita:

$$F_r(M_{400} < \eta).$$

..... Probabilidade acumulada da distribuição Qui-qua-
drado: $F_r(\chi^2_9 < n\eta^2).$

As causas de tais resultados é a mesma forma dos teoremas, isto é, cada proposição em cada teorema é a forma final de uma desigualdade tomada sobre uma cadeia de desigualdades o que faz em cada passo irem-se afastando do valor exato.

Porém, os teoremas podem-se melhorar fazendo algumas suposições. Consideremos 3.2, assim:

$$P_r\{\|S_n - S'_m\| < \eta\} \geq 1 - \frac{k-1}{\eta^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)^2$$

ou

$$P_r\{\|S_n - S'_m\| \geq \eta\} \leq \frac{k-1}{\eta^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)^2 \quad (3.12)$$

Assim em 3.12 desejamos que

$$0 \leq \frac{k-1}{\eta^2} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)^2 \leq 1 \quad (3.13)$$

daí: se $n=m$ obtemos, usando a desigualdade do lado direito:

$$\frac{4(k-1)}{n\eta^2} \leq 1 \quad (3.14)$$

ou

$$n \geq \frac{4(k-1)}{\eta^2} \quad (3.15)$$

Para nosso caso com $\hat{\eta} = 0.179$ e $k = 13$ precisaríamos um $n \geq 1499$ o que seria uma amostra bem grande. De maneira geral cada variação do η na ordem de 10^{-n} , $n \geq 1$ faz variar n na ordem de 10^{2n} , $n \geq 1$. Se $\hat{\eta}_0 = 0.0179$ e $k = 13$, então $n \geq 149809$.

Considerando 3.6 chegamos nas mesmas conclusões:

$$\begin{aligned} \Pr\{\|S_n - S'_m\| < \eta\} &\geq 1 - \frac{4(k-1)}{\eta^2} \left(\frac{1}{n} + \frac{1}{m}\right) + \frac{16(k-1)^2}{\eta^4 nm} \\ \Leftrightarrow \Pr\{\|S_n - S'_m\| > \eta\} &< \frac{4(k-1)}{\eta^2} \left(\frac{1}{n} + \frac{1}{m}\right) - \frac{16(k-1)^2}{\eta^4 nm} \end{aligned} \quad (3.16)$$

supondo $n = m$, desejamos ter:

$$0 \leq \frac{4(k-1)}{\eta^2} \left(\frac{2}{n}\right) - \frac{16(k-1)^2}{\eta^4 n^2} \leq 1 \quad (3.17)$$

$$\Leftrightarrow 0 \leq 8(k-1)\eta^2 n - 16(k-1)^2 \leq \eta^4 n \quad (3.18)$$

$$\Leftrightarrow 16(k-1)^2 \leq [8(k-1)\eta^2 - \eta^4]n \leq 16(k-1)^2 \quad (3.19)$$

portanto

$$n = \frac{16(k-1)^2}{8(k-1)\eta^2 - \eta^4} \quad (3.20)$$

Nestas condições temos que concluir dizendo que mesmo com estas suposições, claramente o problema continua, pois trabalhamos com desigualdades que exigem para sua validade amostras muito grandes que são antieconômicas e pouco práticas.

O único caminho que fica, é explorar as boas características da distância de Matusita, a saber:

i) A distância ∇^2 em 3.1 independe do tamanho amostral isto é, se $n^* = k_1 n$ e $m^* = k_2 m$, $k_1, k_2 \geq 1$ então

$$\nabla_{*}^2 = \|S_{n^*} - S_{m^*}\|^2 = \sum (\sqrt{p_i^*} - \sqrt{q_i^*})^2 \quad (3.21)$$

$$= \sum \left(\sqrt{\frac{k_1 n_i}{k_1 n}} - \sqrt{\frac{k_2 m_i}{k_2 m}} \right)^2 \quad (3.22)$$

$$= \sum (\sqrt{p_i} - \sqrt{q_i})^2 \quad (3.23)$$

$$= \nabla^2 .$$

ii) A distância V^2 é simétrica, ou seja, conserva seu valor quando fazemos a mudança entre valor esperado e valor observado.

iii) Não precisamos agrupar os dados, nem determinar limites no número de frequências por classe; somente precisamos ter o mesmo número de classes para cada amostra tomada.

iv) A facilidade do cálculo de 3.1 a faz preferível para efeitos computacionais.

Por esta razão, nossa proposta é usar

$$V^2/2 = \frac{1}{2} \sum (\sqrt{P_i} - \sqrt{q_i})^2 \quad (3.24)$$

isto é:

$$V^1 = (\frac{1}{2} \sum (\sqrt{P_i} - \sqrt{q_i})^2)^{1/2} \quad (3.25)$$

para gerar por simulação tabelas que permitam efetuar o teste de ajustamento $H_0 : F = G$ V.S. $H_1 : F \neq G$. Isto devido ao fato de não conhecer-se a distribuição exata de 3.1.

SIMULAÇÕES

Para gerar as tabelas definitivas, foi primeiramente mostrado por simulação, que o efeito de assimetria, mudava em muito pouco o comportamento de 3.1 como aparece nas tabelas 3.2, 3.3 e 3.4 geradas para $K = 10, 15, 20$ e tamanhos amostrais $n_p = 25, 50, 100, 200, 400$ segundo o programa N° 1, modificado.

```

C MATOSI:SIMULACAO DA DISTANCIA DE MATUSITA
C VF( ) : VETOR COM LOS LIMITES SUPERIORES DOS INTERVALOS
C LF( ) : MATRIZ COM AS FREQUENCIAS DOS INTERVALOS
C DM( ) : VETOR COM AS DISTANCIAS
C IDM( ) : VETOR COM AS FREQUENCIAS DAS DISTANCIAS
C NP : TAMANHO DA AMOSTRA
C LJ : NUMERO DE AMOSTRAS
C LI : NUMERO DE GRUPOS DE INTERVALOS
      DIMENSION XLF(24,2),VF(24),LF(24,2),DM(1500),IDM(100)
      Y=0.77
      LJ=16
      CALL SETRAN
C LEITURA DO NUMERO DE GRUPOS DE INTERVALOS
      READ(22,16)LM
      16 FORMAT(12)
      DO 17 I=1,LM
C LEITURA DE LOS LIMITES SUPERIORES DOS INTERVALOS
      DO 1 K=10,24,2
      READ(22,2)(VF(I),I=1,K)
      2 FORMAT(24F5.3)
C IMPRESSAO DOS INTERVALOS
      WRITE(23,18)(VF(I),I=1,K)
      18 FORMAT(3(/),20X,'INTERVALOS UTILIZADOS:',4(/,43X,6F7.3),3(/))
C GERACAO DAS AMOSTRAS
      JL=1
      NP=30
      12 DO 3 I=1,1500
          DO 20 L=1,2
          DO 20 J=1,K
      20 LF(J,L)=0
          DO 4 L=1,2
          DO 4 J=1,NP
              Y=RAN(Y)
              IJ=1
      6 IF(Y.GT.VF(IJ))GO TO 5
          LF(IJ,L)=LF(IJ,L)+1
          GO TO 4
      5 IJ=IJ+1
          GO TO 6
      4 CONTINUE
          IF(I.LE.1498)GO TO 50
          WRITE(23,30)I
      30 FORMAT(20X,'MULTINOMIAL GERADA',3X,I4,/)
          DO 32 L=1,2
          WRITE(23,31)(LF(J,L),J=1,K)
      31 FORMAT(24I4)
      32 CONTINUE
C CALCULO DA DISTANCIA DE MATUSITA
      50 S=0.0
          DO 8 L=1,K
              XLF(L,2)=FLOAT(LF(L,2))
              XLF(L,1)=FLOAT(LF(L,1))
              S=S+(SQRT(XLF(L,2))-SQRT(XLF(L,1)))**2
      8 CONTINUE
          DM(I)=SQRT(S/(2.0*NP))
C VERIFICACAO DO INTERVALO NO QUAL ESTA A DISTANCIA
      Z=0.01
      IL=1
      10 IF(DM(I).GT.Z)GO TO 9

```

```

ID*(11)=ID*(11)+1
GO TO 3
9 Z=Z+0.01
IL=IL+1
GO TO 10
3 CONTINUE
C IMPRESSAO DAS FREQUENCIAS DAS DISTANCIAS
WRITE(23,19)NP,(IDM(J),J=1,100)
19 FORMAT(20X,'TAMANHO ANOSTRAL:',1X,I3,/,20X,'FREQUENCIAS:',
*5(/,13X,20I5),3(/))
DO 14 L=1,100
    IDM(L)=0
14 CONTINUE
IF(JL.GE.LJ)GO TO 1
IF(NP.LT.50)GO TO 33
IF(NP.LT.100)GO TO 34
IF(NP.LT.250)GO TO 35
NP=NP+100
GO TO 41
33 NP=NP+5
GO TO 41
34 NP=NP+10
GO TO 41
35 NP=NP+50
41 JL=JL+1
GO TO 12
1 CONTINUE
17 CONTINUE
STOP
END

```

TABELA 3.3: Comparação do efeito de assimetria para $K = 1.0$ e diferentes tamanhos amostrais, nas frequências da distância de Matusita obtida por simulação, em seus correspondentes pontos percentuais.

	SIMETRIA	ASSIMETRIA À DIREITA	ASSIMETRIA À ESQUERDA
	0.100 0.200 0.300 0.400 0.500	0.100 0.250 0.450 0.600	0.020 0.060 0.130 0.210
	0.600 0.700 0.800 0.900 1.000	0.700 0.780 0.850 0.910	0.300 0.400 0.550 0.750
		0.960 1.000	0.900 1.000
	K = 10	K = 10	K = 10
	25 50 100 200 400	25 50 100 200 400	25 50 100 200 400
.1	5 81 466 937	7 78 439 940	2 5 75 438 918
.2	27 381 845 534 63	24 336 823 560 60	37 332 809 562 82
.3	300 513 74	230 536 99 1	260 551 115
.4	365 91	462 116	441 105 1
.5	255 5	237 5	214 7
.6	51	46	43
.7	2	1	3
.8			
.9			
.0			

	SIMETRIA					ASSIMETRIA À DIREITA					ASSIMETRIA À ESQUERDA				
	0.050	0.100	0.150	0.200	0.250	0.006	0.023	0.075	0.178	0.366	0.001	0.003	0.005	0.010	0.018
	0.300	0.350	0.400	0.450	0.500	0.521	0.650	0.761	0.834	0.879	0.028	0.042	0.061	0.087	0.121
	0.550	0.600	0.650	0.700	0.750	0.913	0.939	0.957	0.971	0.982	0.166	0.239	0.350	0.479	0.633
	0.800	0.850	0.900	0.950	1.000	0.990	0.995	0.997	0.999	1.000	0.822	0.925	0.977	0.994	1.000
	K = 20					K = 20					K = 20				
	25	50	100	200	400	25	50	100	200	400	25	50	100	200	400
.1	7 335					6 229					9 211				
.2	5 239 936 665					3 11 186 885 771					3 19 201 882 789				
.3	163 705 57					38 320 764 109					51 316 751 109				
.4	23 526 56					287 599 50					302 571 48				
.5	195 284					485 69					454 94				
.6	476 22					171 1					167				
.7	274					16					23				
.8	32														
.9															
.0															

Por simulação em computador, conforme Programa Nº 1, foram geradas as tabelas 3.6 a 3.14 para $K = 8, 10, 12, 14, 16, 18, 20, 22, 24$ e tamanhos amostrais $NP = 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 150, 200, 250, 350, 450$ e 550 para serem usadas ao se efetuar o teste de ajustamento para duas amostras $H_0 : F = G$, $H_1 : F \neq G$.

As tabelas 3.6 a 3.14 são o valor crítico da distância η entre duas amostras, de forma geral:

$$\Pr(\|S_n - S'_m\| \leq \eta) = \alpha \quad (3.26)$$

Por exemplo, na tabela 3.6

$$\Pr(\|S_{45} - S'_{45}\| \leq 0.259) = 0.85 \quad (3.27)$$

Assim, com tamanhos amostrais iguais rejeitaremos H_0 sempre que o valor calculado da distância de Matusita ∇^1 em 3.25 fosse maior do que o valor tabelado. Para tamanhos amostrais diferentes decidiremos pelo menor tamanho amostral.

Para gerar as tabelas foi levado em conta o procedimento a seguir: o intervalo $(0,1)$ foi subdividido segundo K e para cada tamanho amostral NP foram geradas aleatoriamente 1500 pares de amostras multinomiais, sendo que para cada par delas foi calculada a distância $\sqrt{1}$ segundo 3.25. Logo depois, o intervalo $(0,1)$ foi subdividido em caselas com comprimento 0.01 cada uma, nelas foram classificadas segundo o valor obtido as 1500 distâncias $\sqrt{1}$. O procedimento o executamos para os diferentes valores de K e NP .

As tabelas são apresentadas a seguir:

TABELA 3.6

K = 8								
n \ α	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.106	0.141	0.160	0.173	0.335	0.361	0.394	0.477
35	0.094	0.130	0.147	0.163	0.309	0.336	0.375	0.434
40	0.086	0.119	0.139	0.152	0.277	0.292	0.318	0.369
45	0.086	0.109	0.126	0.139	0.259	0.276	0.304	0.357
50	0.078	0.105	0.122	0.132	0.249	0.264	0.289	0.340
60	0.073	0.096	0.110	0.121	0.217	0.226	0.253	0.301
70	0.071	0.092	0.105	0.115	0.200	0.214	0.232	0.267
80	0.061	0.081	0.094	0.104	0.187	0.199	0.217	0.244
90	0.059	0.078	0.091	0.100	0.180	0.189	0.205	0.236
100	0.058	0.075	0.080	0.093	0.168	0.178	0.191	0.220
150	0.046	0.061	0.069	0.075	0.135	0.142	0.154	0.179
200	0.040	0.053	0.060	0.065	0.120	0.127	0.137	0.153
250	0.035	0.046	0.052	0.056	0.105	0.112	0.120	0.137
350	0.031	0.040	0.045	0.049	0.088	0.094	0.102	0.117
450	0.026	0.034	0.040	0.043	0.078	0.083	0.091	0.105
550	0.021	0.031	0.035	0.039	0.066	0.074	0.079	0.088

TABELA 3.7

K = 10								
n \ α	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.134	0.181	0.203	0.221	0.394	0.420	0.446	0.508
35	0.131	0.166	0.187	0.203	0.360	0.381	0.414	0.482
40	0.120	0.151	0.167	0.181	0.326	0.346	0.374	0.440
45	0.111	0.142	0.160	0.172	0.301	0.321	0.348	0.404
50	0.102	0.131	0.150	0.161	0.282	0.298	0.320	0.373
60	0.095	0.119	0.135	0.146	0.247	0.263	0.280	0.336
70	0.088	0.120	0.125	0.136	0.227	0.239	0.257	0.302
80	0.085	0.104	0.113	0.131	0.212	0.224	0.238	0.275
90	0.076	0.097	0.110	0.119	0.198	0.207	0.223	0.253
100	0.074	0.092	0.102	0.111	0.186	0.195	0.211	0.239
150	0.059	0.075	0.084	0.090	0.151	0.160	0.173	0.198
200	0.053	0.066	0.073	0.078	0.133	0.140	0.149	0.169
250	0.044	0.056	0.063	0.067	0.117	0.124	0.132	0.148
350	0.036	0.047	0.054	0.060	0.100	0.106	0.112	0.128
450	0.031	0.042	0.048	0.052	0.078	0.093	0.099	0.112
550	0.031	0.040	0.043	0.047	0.077	0.080	0.087	0.098

TABELA 3.8

K = 12								
$n \backslash \alpha$	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.179	0.221	0.245	0.266	0.444	0.470	0.590	0.557
35	0.163	0.199	0.221	0.234	0.400	0.420	0.448	0.501
40	0.143	0.179	0.207	0.216	0.367	0.386	0.412	0.468
45	0.140	0.161	0.184	0.198	0.337	0.359	0.389	0.444
50	0.124	0.154	0.174	0.186	0.311	0.327	0.358	0.420
60	0.111	0.141	0.155	0.168	0.278	0.293	0.316	0.371
70	0.110	0.133	0.146	0.156	0.250	0.265	0.284	0.325
80	0.101	0.126	0.138	0.147	0.234	0.245	0.264	0.299
90	0.094	0.118	0.131	0.139	0.216	0.228	0.244	0.277
100	0.091	0.108	0.120	0.128	0.206	0.215	0.229	0.264
150	0.071	0.085	0.095	0.103	0.165	0.173	0.184	0.205
200	0.064	0.077	0.084	0.089	0.141	0.148	0.159	0.178
250	0.054	0.068	0.075	0.081	0.127	0.133	0.141	0.158
350	0.047	0.056	0.062	0.066	0.106	0.110	0.119	0.133
450	0.041	0.050	0.054	0.059	0.094	0.098	0.105	0.119
550	0.038	0.045	0.051	0.054	0.087	0.089	0.097	0.108

TABELA 3.9

K = 14								
$n \backslash \alpha$	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.222	0.270	0.298	0.315	0.486	0.509	0.542	0.593
35	0.191	0.229	0.260	0.277	0.446	0.464	0.496	0.563
40	0.173	0.212	0.239	0.254	0.403	0.425	0.451	0.503
45	0.164	0.198	0.218	0.231	0.370	0.388	0.419	0.460
50	0.150	0.183	0.199	0.214	0.352	0.367	0.389	0.439
60	0.138	0.164	0.181	0.192	0.310	0.325	0.345	0.402
70	0.125	0.151	0.164	0.174	0.276	0.292	0.312	0.353
80	0.114	0.139	0.153	0.162	0.253	0.266	0.284	0.324
90	0.109	0.132	0.144	0.151	0.239	0.252	0.269	0.298
100	0.105	0.124	0.137	0.146	0.221	0.230	0.246	0.276
150	0.081	0.098	0.109	0.116	0.178	0.186	0.198	0.221
200	0.071	0.086	0.094	0.100	0.156	0.163	0.174	0.192
250	0.061	0.076	0.084	0.089	0.137	0.142	0.149	0.167
350	0.051	0.063	0.071	0.074	0.116	0.119	0.128	0.145
450	0.047	0.057	0.063	0.067	0.102	0.100	0.113	0.127
550	0.042	0.051	0.057	0.060	0.090	0.095	0.099	0.112

TABELA 3.10

K = 16								
$n \backslash \alpha$	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.259	0.306	0.335	0.353	0.523	0.545	0.568	0.629
35	0.231	0.277	0.302	0.318	0.484	0.504	0.527	0.579
40	0.199	0.247	0.273	0.287	0.436	0.455	0.483	0.530
45	0.191	0.231	0.252	0.266	0.409	0.425	0.449	0.493
50	0.170	0.213	0.230	0.244	0.379	0.395	0.419	0.470
60	0.157	0.185	0.201	0.213	0.335	0.348	0.371	0.415
70	0.142	0.167	0.182	0.193	0.299	0.314	0.333	0.383
80	0.134	0.158	0.171	0.179	0.274	0.288	0.307	0.345
90	0.130	0.147	0.161	0.168	0.254	0.266	0.283	0.319
100	0.120	0.141	0.153	0.161	0.239	0.248	0.265	0.303
150	0.096	0.112	0.122	0.128	0.189	0.197	0.209	0.233
200	0.085	0.098	0.105	0.112	0.163	0.172	0.184	0.206
250	0.074	0.085	0.092	0.097	0.145	0.150	0.159	0.178
350	0.060	0.071	0.078	0.082	0.121	0.126	0.134	0.149
450	0.053	0.063	0.070	0.073	0.108	0.112	0.118	0.129
550	0.049	0.057	0.062	0.066	0.098	0.102	0.108	0.119

TABELA 3.11

K = 18								
$\alpha \backslash n$	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.295	0.346	0.373	0.391	0.558	0.577	0.606	0.656
35	0.254	0.308	0.334	0.353	0.508	0.526	0.552	0.598
40	0.244	0.285	0.307	0.323	0.477	0.496	0.522	0.563
45	0.213	0.258	0.282	0.298	0.441	0.463	0.481	0.524
50	0.197	0.233	0.256	0.271	0.405	0.423	0.445	0.494
60	0.181	0.207	0.224	0.235	0.352	0.383	0.402	0.445
70	0.160	0.189	0.204	0.214	0.320	0.334	0.357	0.400
80	0.147	0.172	0.186	0.196	0.297	0.308	0.327	0.366
90	0.134	0.161	0.173	0.182	0.270	0.282	0.303	0.328
100	0.127	0.152	0.165	0.174	0.256	0.268	0.288	0.315
150	0.107	0.123	0.132	0.138	0.202	0.211	0.224	0.242
200	0.090	0.104	0.112	0.117	0.171	0.177	0.187	0.208
250	0.080	0.092	0.100	0.105	0.155	0.160	0.169	0.188
350	0.065	0.077	0.084	0.089	0.130	0.136	0.143	0.159
450	0.059	0.070	0.075	0.079	0.115	0.118	0.126	0.138
550	0.053	0.063	0.069	0.072	0.104	0.107	0.113	0.124

TABELA 3.12

K = 20								
$n \backslash \alpha$	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.330	0.381	0.409	0.428	0.591	0.610	0.635	0.683
35	0.296	0.345	0.371	0.389	0.546	0.563	0.585	0.645
40	0.277	0.314	0.338	0.353	0.496	0.516	0.545	0.591
45	0.238	0.288	0.308	0.322	0.467	0.485	0.512	0.550
50	0.218	0.256	0.278	0.296	0.435	0.444	0.475	0.518
60	0.200	0.224	0.246	0.259	0.383	0.400	0.422	0.468
70	0.180	0.206	0.225	0.237	0.344	0.358	0.379	0.413
80	0.172	0.194	0.207	0.216	0.313	0.327	0.348	0.394
90	0.154	0.177	0.191	0.200	0.291	0.305	0.328	0.366
100	0.141	0.165	0.181	0.189	0.273	0.286	0.303	0.336
150	0.113	0.137	0.142	0.149	0.216	0.228	0.235	0.269
200	0.100	0.114	0.123	0.128	0.185	0.197	0.203	0.224
250	0.086	0.100	0.108	0.113	0.163	0.169	0.177	0.197
350	0.072	0.084	0.091	0.095	0.138	0.147	0.149	0.169
450	0.064	0.074	0.081	0.084	0.110	0.126	0.133	0.145
550	0.060	0.068	0.073	0.076	0.108	0.111	0.117	0.128

TABELA 3.13

K = 22								
$n \backslash \alpha$	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.366	0.408	0.438	0.458	0.615	0.631	0.659	0.702
35	0.324	0.378	0.406	0.428	0.564	0.582	0.602	0.658
40	0.301	0.339	0.365	0.382	0.529	0.547	0.569	0.614
45	0.267	0.311	0.333	0.351	0.490	0.506	0.531	0.571
50	0.243	0.287	0.312	0.327	0.460	0.476	0.501	0.550
60	0.217	0.251	0.269	0.281	0.409	0.424	0.446	0.485
70	0.192	0.218	0.237	0.251	0.367	0.383	0.404	0.444
80	0.176	0.206	0.219	0.230	0.335	0.348	0.368	0.399
90	0.166	0.192	0.205	0.215	0.309	0.321	0.341	0.376
100	0.153	0.177	0.189	0.198	0.287	0.301	0.319	0.348
150	0.122	0.422	0.152	0.161	0.225	0.234	0.245	0.268
200	0.107	0.122	0.131	0.136	0.191	0.197	0.207	0.226
250	0.096	0.111	0.118	0.123	0.170	0.177	0.188	0.205
350	0.081	0.092	0.097	0.101	0.143	0.147	0.155	0.167
450	0.071	0.081	0.085	0.090	0.127	0.131	0.137	0.149
550	0.052	0.062	0.068	0.072	0.105	0.108	0.114	0.126

TABELA 3.14

K = 24								
n \ α	0.01	0.05	0.10	0.15	0.85	0.90	0.95	0.99
30	0.382	0.441	0.468	0.485	0.634	0.652	0.676	0.724
35	0.364	0.406	0.422	0.448	0.595	0.613	0.639	0.689
40	0.330	0.375	0.397	0.414	0.553	0.566	0.592	0.639
45	0.284	0.336	0.360	0.381	0.512	0.523	0.551	0.591
50	0.269	0.311	0.332	0.350	0.484	0.501	0.524	0.575
60	0.234	0.270	0.292	0.308	0.435	0.449	0.468	0.507
70	0.211	0.238	0.256	0.272	0.389	0.405	0.429	0.483
80	0.191	0.221	0.238	0.249	0.355	0.368	0.387	0.428
90	0.183	0.203	0.218	0.229	0.330	0.342	0.362	0.396
100	0.167	0.191	0.203	0.213	0.305	0.316	0.334	0.370
150	0.132	0.152	0.162	0.169	0.233	0.241	0.254	0.280
200	0.114	0.131	0.140	0.145	0.201	0.208	0.219	0.239
250	0.101	0.115	0.123	0.129	0.177	0.184	0.194	0.214
350	0.086	0.097	0.103	0.108	0.149	0.154	0.150	0.175
450	0.075	0.085	0.091	0.095	0.130	0.135	0.142	0.152
550	0.067	0.076	0.082	0.085	0.118	0.123	0.128	0.138

BIBLIOGRAFIA

1. AHMAD, I.A. (1980) - Nonparametric Stimation of Matusi-
ta's Measure of affinity between absolutely continuous dis-
tributions, Annals Institute Statistical Mathematics 32-
241, Part A.
2. AHMAD, I.A. (1980) - Nonparametric Stimation of an af-
finity measure between two absolutely continuous distri-
butions with hypothesis testing applications. Ann. Inst.
Stat. Math. 32-223, Part A.
3. BIRBAUM, Z.W. (1952) - Distribution-free tests of fit for
continuous distributions functions. Ann. Mat. Stat. 24-1.
4. BURR, E.J. (1963) - Small Sample Distributions of the two
Sample Cramér-Von Mises w^2 , and Watson U^2 . Ann. Mat.
Stat. 34-95.
5. CRAMER, H. (1946) - Mathematical Methods of Statistics ,
Princeton, N.Y., Princeton University Press.
6. COCHRAN, W.G. (1952) - The χ^2 test of goodness of fit,
Ann. Mat. Stat., 23-315.

7. DADAKIS, H.S. (1977) - Goodness of fit test for discrete data: A review and an application to a Heath Impairment Scale, Biometrics, 33-237.
 8. DICKINSON, G.J. (1971) - Nonparametric Statistical Inference, Mc. Graw-Hill.
 9. DILLON, R.W. and GOLSTEIN, M. (1978) - On the performance of some Multinomial Classification rules, JASA, 73-305.
 10. GOLSTEIN, M. and others (1976) - On a test of Independence for contingency tables. Commun. Statistics-Theory Math. , A5 (2), 159-169.
 11. HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions, Ann. Math. Stat., 36-369.
 12. KENDALL, M.G. and A. STUART - The advanced theory of Statistics, Vol. 2, Mc Millan Editors 2^a Edition, 1979.
 13. KHAN, A.H. and ALI, S.M. (1973) - A new coefficient association, Ann. Inst. Stat. Math., 25-41.
 14. KIMBALL (1947) - Some basic theorems for developing tests of fit for the case of the nonparametric probability distribution function, Ann. Math. Stat., 18-540.
-

15. KIRMANI, S.N.U.A. (1974) - On a goodness of fit test based on Matusita's Measure of distance. Ann. Inst.Stat.Math. 25-493.
16. KOBYLINSKI, G.I. and SHERIDAN, P.F. (1979) - Distribution, abundance feeding and longterm fluctuation of spot, eciostonus Xanthurus and, croaker, micropogonias undulatus, in apalachiola bay, Florida 1972-1977, Contributions in Marine Science, volume 22.
17. KOLMOGOROV (1953) - Sulla determinazione empirica di una legge di distribuzione. Giorn. Ist. Ital. Attuari , 4-83.
18. KOLMOGOROV-SMIRNOV (1954)-Tests for Psychological Research Psychological Bull. 51-60.
19. LEHMANN, E.L. and D'ABRERA, H.J.M. (1971) - Nonparametrics: Statistical Methods Based on Ranks: Mc Graw-Hill.
20. LEVINE, B. (1979) - Fondements théoriques de la radio-technique Statistique. Tomo III. Mir Editions Moscov.
21. MANN-WHITNEY (1947) - On a test whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18-50.

22. MASSEY (1975) - The Kolmogorov-Smirnov teste of goodness of fir. JASA, 46-68.
23. MATUSITA (1955) - Decision rules, based on the distance for problema of fit, two samples and stimation. Ann.Math. Stat. 26-631.
24. MOOD (1950) - Introduction to the theory of statistics , McGraw-Hill Book Company, N.Y., 394
25. MORAN (1947) - The random division of an interval. Journal Royal Stat. Soc. Suppl., 9-12.
26. NED, GLICK (1974) - Separation and probability of correct classification among two or more distributions. Ann. Inst. Stat. Math. 25-373.
27. PEARSON, K. (1900) - On the criterion that a given system of deviations from the probable in the case of a correlated system variables in such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. Series 5, 50-157.
28. SHERMAN (1950) - A random variable related to the spacing of sample values. Ann. Math. Stat. 21-339.

29. STONER A.W. (1979) - Feeding Ecology of lagodon Rhomboides (Pisces - Sparidae) - Variation and Functional responses. Fishery Bull. Vol. 78 № 2.
30. WALD WOLFOWITZ (1939) - Confidence limits for continuous distribution functions. Ann. Math. Stat. 10-105.
31. WALD WOLFOWITZ (1940) - On a test whether two samples are from the same populations. Ann. Math. Stat. 11-147.
32. WATSON, G.S. (1961) - Goodness of fit on a circle, Biometrika 48, 1 and 2, 109.
33. WILCOXON (1945) - Individual comparison by ranking methods. Biometrics Bull. 1-80.