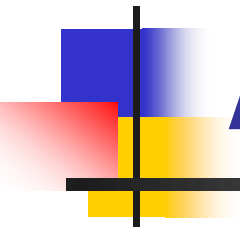
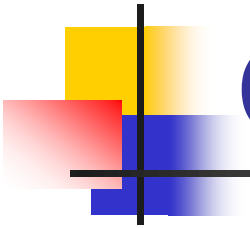


SCC5895 – Análise de Agrupamento de Dados

Validação de Agrupamento: Parte II

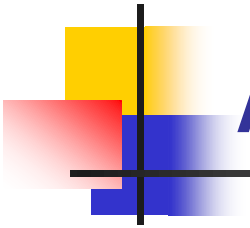
Prof. Eduardo R. Hruschka

PPG-CCMC / ICMC / USP



Créditos

- Este material consiste de adaptações e extensões dos originais elaborados por Eduardo R. Hruschka e Ricardo J. G. B. Campello



Aula de Hoje

- Introdução à Validação Estatística
- Revisão de Testes de Significância
- Análise de Monte Carlo
- Hipóteses Nulas e Alternativas Comuns em Agrupamento de Dados
- Estatísticas Úteis em Análise de Agrupamento
- Validação de Rotulações, Partições e Hierarquias
- Tendência de Agrupamento

Validação Estatística

"This chapter (c. 4) is based on the premise that the problems of cluster validity are inherently statistical"

"A clustering structure is 'valid' if it is 'unusual' in some sense"

Jain & Dubes, Algorithms for Clustering Data, 1988

Testando Hipóteses

- Não é muito difícil propor índices de validade
- Porém, é difícil estabelecer limiares que definam quando um valor para um determinado índice é *alto* ou *baixo*
- Métodos estatísticos nos fornecem uma estrutura de trabalho para abordar esse assunto
- Considere uma determinada *Estatística T*
  - no presente contexto, trata-se de algum critério de avaliação de hierarquias, partições ou grupos individuais
- T pode ser vista como uma variável aleatória
  - distribuição reflete a freqüência relativa com que seus valores ocorrem sob determinada hipótese

Testes de Significância (breve revisão)

Inferência Estatística: oferece métodos para tirarmos conclusões a partir de dados

- Probabilidades expressam a força das nossas conclusões
  - Indicam o que aconteceria se utilizássemos o método de inferência muitas vezes
- Testes de significância (TS) se baseiam em distribuições amostrais de **estatísticas**
  - No presente contexto, estatística é a medida / critério / índice de interesse, vista(o) como uma variável aleatória

Testes de Significância (breve revisão)

- Um TS testa uma hipótese específica, usando dados amostrais para decidir sobre a validade da hipótese
- Além da estatística (T), precisamos:
  - uma hipótese nula H_0 a ser testada
  - uma hipótese alternativa H_1 para a qual procuramos evidência
- Um TS avaliará a força da evidência contra H_0
- Para avaliar H_0, precisamos da distribuição de probabilidades de T sob esta hipótese

Testes de Significância (breve revisão)

- Lógica de um TS
  - Assumir H_0 verdadeira (embora possa ser falsa)
  - Determinar quão provável seria obter dados *tão extremos* quanto os que dispomos, se H_0 é verdadeira
  - Improvável?
    - Tendemos a duvidar de H_0
  - Provável?
    - Tendemos a acreditar em H_0
- O que significa obter dados *tão extremos* quanto aqueles de que dispomos?

Testes de Significância (breve revisão)

- Necessitamos de um espaço amostral, ou de uma distribuição de referência (*baseline distribution*)
- Premissas sobre a população incorporam nossas expectativas sobre os dados – *e.g.*:
  - Dados estão dispostos de maneira aleatória, ou
  - Dados exibem estruturas de grupos
- Observar valor de T e decidir se a observação é não usual segundo a distribuição de referência para T
  - Por exemplo, distribuição sob a hipótese nula de que os dados estão dispostos de maneira aleatória

Testes de Significância (breve revisão)

- Suponhamos uma estatística de teste T e uma hipótese nula H_0. Assumamos por ora que a distribuição de T sob H_0 é conhecida
- Denotemos por $P(T \geq t \mid H_0)$ a probabilidade de que a estatística T é maior do que um determinado limiar t
- Consideremos agora um número pequeno α (e.g. 0,05 ou 0,01). Podemos estabelecer um limiar t_α resolvendo:

$$P(T \geq t_\alpha \mid H_0) = \alpha$$

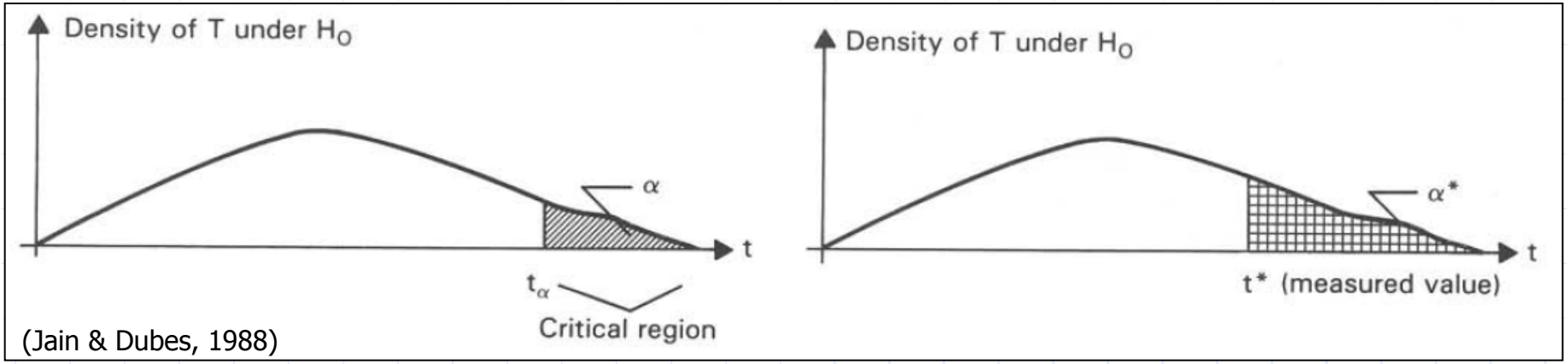
- Assumindo que o valor de T medido em um experimento é t^*, rejeitamos H_0 ao nível de significância α se $t^* \geq t_\alpha$
- Raciocínio análogo para *valores pequenos* de T

Testes de Significância (breve revisão)

- Outra forma de avaliar a significância é resolver a seguinte equação para α^*, lembrando que t^* é o valor medido:

$$P(T \geq t^* \mid H_0) = \alpha^*$$

- O valor α^* é o valor crítico que levaria à rejeição de H_0
  - usualmente denominado *p-value*
- Quanto menor esse valor, maior é a evidência contra H_0



(Jain & Dubes, 1988)

Validação Estatística de Agrupamento

- H_0 mais comum em análise de agrupamento é:
  - **Posições aleatórias equiprováveis**:
    - A distribuição de *T* sob H_0 é gerada considerando que todos os conjuntos de *N* posições dos *N* objetos são igualmente prováveis
      - posições dentro do hiper-cubo *n*-dimensional delimitado pelos possíveis valores dos *n* atributos que descrevem os dados
    - Espera-se que muitos conjuntos de posições (dados) estejam associados a valores usuais de *T* e poucos associados a valores não usuais de *T*
    - Testar a hipótese de posições aleatórias significa avaliar a probabilidade de *T* ser produzida a partir de dados puramente aleatórios, regidos por uma população que não exibe estrutura de grupos

Validação Estatística de Agrupamento

- Outras hipóteses podem ser mais apropriadas em certos contextos:
  - **Rótulos aleatórios equiprováveis**:
    - A distribuição de T sob H_0 é gerada considerando que todas as permutações dos rótulos dos N objetos são igualmente prováveis
    - Espera-se que muitas rotulações devam estar associadas a valores usuais de T e poucas associadas a valores não usuais de T
    - Nesse caso, se as rotulações são equiprováveis, a distribuição de T sob H_0 deve exibir prob. maiores para faixas de valores usuais e prob. menores para faixas de valores não usuais
    - Testar a hipótese de rótulos aleatórios significa avaliar a probabilidade de T ser produzida ao acaso por uma mera rotulação aleatória dos objetos

Validação Estatística de Agrupamento

- Outras hipóteses podem ser mais apropriadas em certos contextos:
  - **Grafos de proximidade aleatórios equiprováveis**:
    - A distribuição de T sob H_0 é gerada considerando que todas as matrizes de proximidade $N \times N$ ordinais (baseadas em *ranks*) são equiprováveis
    - Veremos posteriormente um exemplo
    - Veremos também que a hipótese nula mais apropriada depende do contexto e que um erro na escolha pode levar a conclusões totalmente equivocadas
    - Antes, porém, consideremos por um momento a hipótese alternativa...

Validação Estatística de Agrupamento

- Em muitas aplicações de TS, H_1 é necessariamente verdadeira quando H_0 é falsa
  - Ex.: se H_0 é "temp. média do corpo humano = 37ºC", então H_1 dada por "temp. média do corpo humano ≠ 37ºC" deve ser verdadeira se H_0 é falsa
- Em análise de agrupamento, no entanto, isso não é trivial pois, geralmente, H_0 se refere simplesmente à **ausência de estrutura**:
  - nos dados (problema de **tendência de agrupamento**), ou
  - nos "grupos" encontrados (problema de **validação de agrupamento**)
- Logo, H_0 "conta apenas metade da estória"...

Validação Estatística de Agrupamento

- É preciso H_1 de presença de estrutura
  - Mas podemos ter diferentes hipóteses para estruturas específicas...
- **Força de um TS**:
  - Dada H_0, T e H_1, define-se a força do TS associado a H_0, T e H_1 como:

  $P(T \geq t_\alpha \mid H_1)$

  ou seja, como a probabilidade de se chegar à conclusão correta quando a hipótese alternativa H_1 é de fato verdadeira
- Buscamos por estatísticas T que se caracterizam por levar a testes fortes em diferentes contextos de análise de agrupamento
- Algumas estatísticas são reconhecidas na literatura como tal

Estatística Γ (Estatística de Hubert)

- Sejam X e Y duas matrizes de proximidade $N \times N$ dos mesmos N objetos. A estatística Γ é definida como:

$$\Gamma = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j) \cdot Y(i,j)$$

não normalizada

$$\Gamma = \frac{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} [X(i,j) - \mu_X][Y(i,j) - \mu_Y]}{\sigma_X \sigma_Y}$$

normalizada

- Por exemplo, X pode ser a matriz de dissimilaridades dos objetos e Y o complemento da matriz de conectividade binária obtida a partir de uma rotulação desses objetos:
  - $Y(i,j) = 0$ se objetos i e j possuem o mesmo rótulo ou 1 caso contrário

Estatística Γ (Estatística de Hubert)

- **Questão de um TS**:
  - valores em uma das matrizes (X ou Y) foram inseridos ao acaso?
- Vamos assumir que Y é uma rotulação arbitrária (externa)
- H_0: rótulos aleatórios equiprováveis
  - todas as permutações dos rótulos dos N objetos são equiprováveis
    - uma permutação pode ser vista como uma reordenação dos rótulos
    - se aplicada sobre Y, mantém as categorias, inter-cambiando os objetos
- Para valores pequenos de N, pode-se obter a distribuição de Γ avaliando todas as $N!$ permutações de linhas/colunas de Y
- Vejamos um exemplo...

Exemplo 1 (Jain & Dubes, 1988)

$$X = \begin{bmatrix} 0 & 1.2 & 0.6 & 0.2 \\ & 0 & 0.3 & 0.4 \\ & & 0 & 0.1 \\ & & & 0 \end{bmatrix} \qquad Y = \begin{bmatrix} 0 & 1 & 0 & 1 \\ & 0 & 1 & 0 \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}$$

- X = dissimilaridades entre 4 objetos
- Y = rotulação em duas categorias: {1,3} e {2,4} $\rightarrow$ Γ=1.8
- Exemplo de Permutação: {1,2,3,4} $\rightarrow$ {2,4,1,3} $\rightarrow$ Γ=1.5
  - Aplicando 1 $\rightarrow$ 2, 2 $\rightarrow$ 4, 3 $\rightarrow$ 1 e 4 $\rightarrow$ 3 tem-se as categorias {2,1} e {4,3}

$$X = \begin{bmatrix} 0 & 0.6 & 0.1 & 0.3 \\ & 0 & 0.2 & 1.2 \\ & & 0 & 0.4 \\ & & & 0 \end{bmatrix} \quad \underline{\textbf{OU}} \quad Y = \begin{bmatrix} 0 & 0 & 1 & 1 \\ & 0 & 1 & 1 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}$$

Exemplo 1 (Jain & Dubes, 1988)

$$X = \begin{bmatrix} 0 & 1.2 & 0.6 & 0.2 \\ & 0 & 0.3 & 0.4 \\ & & 0 & 0.1 \\ & & & 0 \end{bmatrix} \qquad Y = \begin{bmatrix} 0 & 1 & 0 & 1 \\ & 0 & 1 & 0 \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}$$

- Como N é pequeno, podemos calcular a distribuição de Γ avaliando as 4! = 24 possíveis permutações dos rótulos

- Resultado:

Γ	1.5	1.8	2.3
No. de Ocorrências	8	8	8

- O que se conclui sobre a rotulação sob avaliação (Γ=1.8) ... ?

Análise de Monte Carlo

- No exemplo anterior, avaliamos 24 possíveis permutações de rótulos para calcular de forma exata a distribuição de Γ sob H_0

- Porém... note que se N aumenta de 4 para apenas 12 objetos, N! aumenta de 24 para mais de 470 milhões !

  - Imagine para quantidades realistas de N ...

- Nesse cenário, o melhor que podemos fazer é estimar a distribuição de forma aproximada

- Para isso, podemos utilizar **Análise de Monte Carlo**

Análise de Monte Carlo

"É uma classe de métodos para estimar parâmetros e probabilidades através de amostragem e simulações computacionais quando as grandezas são difíceis ou impossíveis de calcular diretamente"

"Pode aproximar uma distribuição desconhecida se um procedimento computacional de amostragem puder ser programado em um computador tal que simule o processo de interesse"

tradução livre de (Jain & Dubes, 1988)

Exemplo (Revisão)

- Estimativas de distribuições de referência de índices externos (ARI e Jaccard) para avaliação de partições
- Experimento (Jain & Dubes, 1988):
  - Quatro conjuntos de 100 pontos em 5 dimensões
    - dados estruturados (misturas de Gaussianas) e aleatórios (distrib. uniforme)
    - dados com 2 e 8 grupos (puramente arbitrários no caso de distrib. uniforme)
  - Dados agrupados com single- e complete-linkage (SL e CL)
  - Cortes realizados no número correto de grupos (2 ou 8)
  - Partições obtidas comparadas com os rótulos
  - Experimento repetido 100 vezes (simulação de Monte Carlo)

Exemplo (Revisão)

TABLE 4.8 Comparison of External Indices for Partitions

	Jaccard		Corrected Rand	
k^a	Mean	Std. Dev.	Mean	Std. Dev.
Clustered data				
SL				
2	0.934	0.169	0.870	0.336
8	0.597	0.172	0.683	0.169
CL				
2	0.989	0.043	0.988	0.050
8	0.859	0.116	0.908	0.082
Random data				
SL				
2	0.496	0.007	0.00027	0.005
8	0.118	0.004	0.00053	0.004
CL				
2	0.354	0.029	−0.00069	0.017
8	0.068	0.008	−0.00102	0.015

Note: [a] k is the true number of clusters for clustered data and the number of pseudo clusters for random data.

sumário das distribuições estimadas

Teste de Monte Carlo

- Considere:
  - S_1 : valor de um índice a ser validado (e.g. Γ)
  - S_2, ..., S_m : $m - 1$ valores obtidos via Monte Carlo sob alguma hipótese nula
- Assumindo que valores elevados de S são desejados, como determinar se o valor observado é significativamente alto ou não ?

- Teste de Monte Carlo (MC):
  - Estabeleça um nível de significância α (e.g. 0,01 ou 0,05) para rejeição de H_0
  - Selecione um inteiro r tal que $r / m = \alpha$
  - Se S_1 estiver entre os r maiores dos m valores (S_1, ..., S_m), rejeite H_0 ao nível α

Teste de Monte Carlo

- A estimativa de Monte Carlo aproxima a região crítica do teste, quando se compara com o caso em que a distribuição exata é usada
  - torna incerto o limiar t_α verdadeiro associado a um dado nível α
  - para reduzir esta incerteza e tornar mais precisa a estimativa da distribuição e deste limiar, é preciso aumentar o tamanho m da amostra
    - mas isso vem acompanhado de um elevado custo computacional...
- A probabilidade de que o teste irá rejeitar H_0 é a probabilidade que não mais que $r - 1$ valores amostrais excedam S_1
  - ou seja, que pelo menos $m - r$ valores (já excluído S1) não excedam S_1

Teste de Monte Carlo

- Seja *p* a prob. real de que um dado valor amostral, gerado pela distribuição verdadeira em questão, não exceda S_1
- Dada esta prob., tem-se que, independentemente da distribuição em questão, a prob. que o teste de Monte Carlo rejeite H_0 sendo esta verdadeira é (Jain & Dubes, 1988):

$$P(r, p) = \sum_{i=0}^{r-1} \binom{m-1}{i} p^{m-i-1} (1-p)^i$$

- $P(r,p)$ faz-se função de r e p pois, para α cte., tem-se $m = r / \alpha$
- Valores menores desta probabilidade são desejáveis

Teste de Monte Carlo

- Características do teste MC:
  - Se *r* for mínimo (e.g. 1 ou 2), minimiza-se *m* e o esforço computacional, mas a probabilidade de rejeição de H_0 sendo esta verdadeira é maximizada
  - Aumentar *r* implica aumentar *m* e diminuir a probabilidade de rejeição de H_0 verdadeira, mas a taxas cada vez menores para um crescente custo computacional
- Estudos experimentais são reportados na literatura para valores típicos de α, a saber, $\alpha = 0{,}01$ e $\alpha = 0{,}05$ (vide Jain & Dubes, 1988)
  - Evidência empírica sugere que *r* em torno de 5 provê um bom compromisso
    - Para $\alpha = 0{,}05 \rightarrow m = 100$ valores amostrais
    - Para $\alpha = 0{,}01 \rightarrow m = 500$ valores amostrais

Exemplo 2 (Jain & Dubes, 1988):

- Base de dados 8OX:
  - Caracteres manuscritos digitalizados
  - 8 atributos
  - 45 objetos: 15 de cada caractere (8, O, X)
    - Primeiros 15 objetos: "8"
    - 15 seguintes: "O"
    - 15 últimos: "X"
  - Rotulação é externa: tomemos novamente a hipótese de *rótulos aleatórios*
    - rótulos se ajustam *não usualmente bem* aos dados?

TABLE 2.1 Pattern Matrix for 8OX Data

7	13	5	5	6	13	2	3
5	13	6	4	6	13	3	13
9	10	6	6	8	10	2	3
7	7	6	6	8	7	2	3
8	7	6	6	8	7	2	0
7	7	6	7	7	7	1	1
6	10	7	8	8	9	4	4
6	7	7	7	9	8	4	5
5	5	5	12	10	7	2	3
7	8	4	4	7	6	2	3
8	7	5	4	6	10	1	0
6	10	5	2	6	8	1	2
7	10	5	5	8	7	1	20
7	10	6	6	6	8	3	3
6	6	7	7	8	8	3	2
7	7	5	6	3	3	4	6
7	6	7	6	3	4	6	5
6	6	5	5	4	3	4	5
8	8	7	6	5	7	5	5
6	8	7	5	5	6	2	2
7	7	7	7	6	6	3	2
7	7	7	7	5	6	5	5
7	7	8	7	5	6	4	5
8	7	7	6	6	5	4	5
7	8	6	5	4	6	2	4
9	7	7	6	8	6	4	3
8	8	7	6	7	6	4	4
7	6	6	5	3	2	5	4
7	6	7	5	5	5	4	3
0	8	7	6	6	7	2	4
10	7	6	6	9	9	7	10
10	4	4	4	8	8	3	10
10	7	4	4	9	9	3	9
7	7	6	5	10	10	10	8
6	10	6	10	8	8	13	4
8	10	7	10	9	8	11	4
5	7	8	7	8	9	9	8
5	6	7	7	9	9	9	7
6	7	11	6	8	11	7	10
6	12	10	4	8	11	8	3
7	12	8	6	9	11	9	1
11	8	7	10	11	10	6	9
9	5	6	7	10	9	7	5
10	5	6	4	9	9	6	11
0	5	11	6	9	11	5	9

- Novamente, consideremos uma matriz de proximidades entre objetos, $X_{N \times N}$, e uma matriz $Y_{N \times N}$ definida como:
  - *Y(i,j)* = 0 se objetos *i* e *j* possuem o mesmo rótulo de categoria
  - *Y(i,j)* = 1 caso contrário
- Histograma a partir de 100 permutações de rótulos:

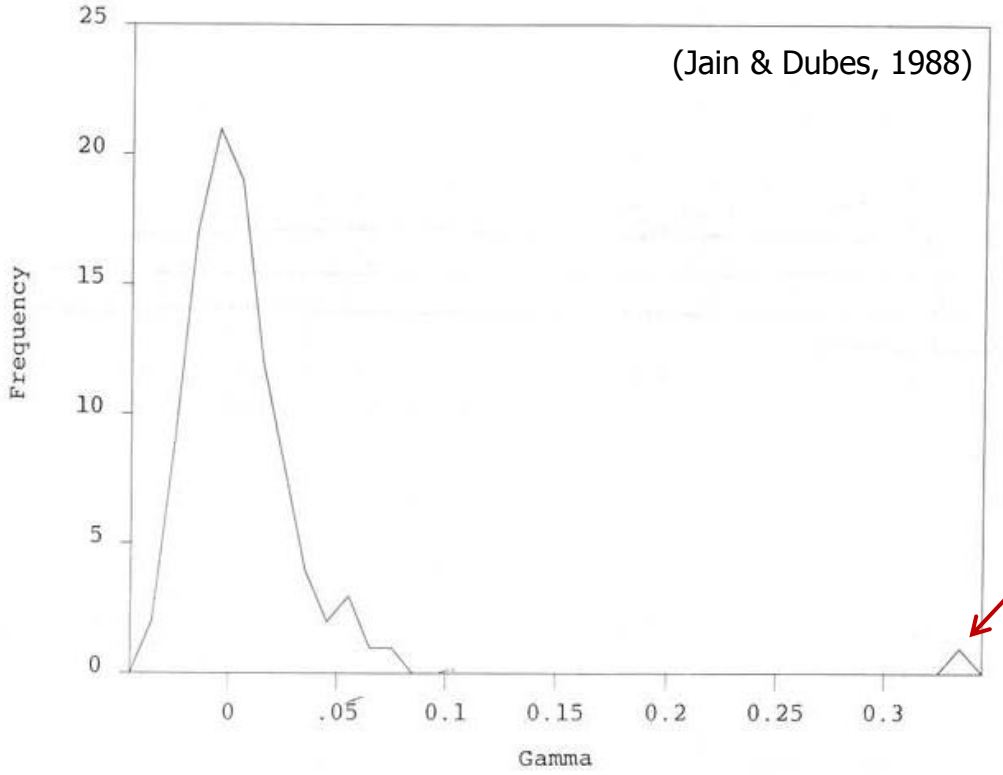


Figure 4.2 Histogram of Γ under the random label hypothesis for the 80X data with category labels.

Γ (normalizado) = 0,33

- Mesmo procedimento repetido para um conjunto de 45 objetos escolhidos aleatoriamente dentro de um hiper-cubo de 8 dimensões análogo àquele que continha os dados originais
  - rótulos mantidos: 1 a 15 = “8”, 16 a 30 = “0”, 31 a 45 = “X”
- Histograma a partir de 100 permutações de rótulos:

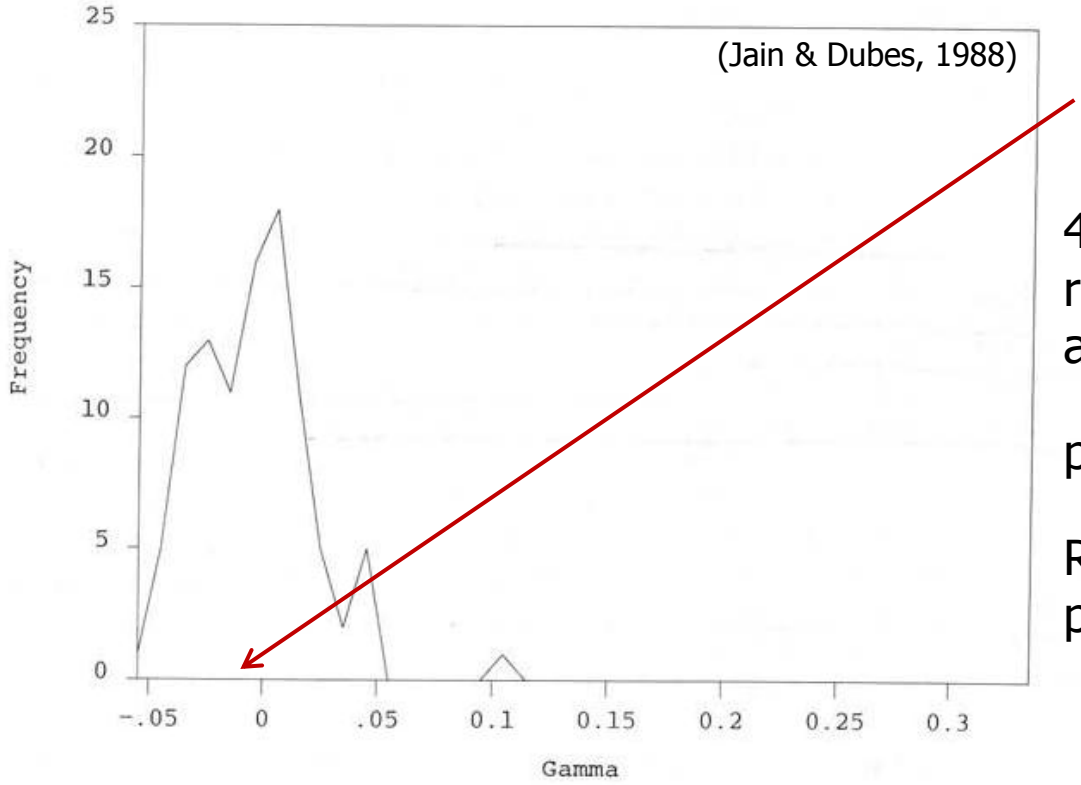


Figure 4.3 Histogram of Γ under the random label hypothesis for 45 random patterns in 8 dimensions with random category labels.

Γ (norm.) = –0,0014

42 permutações produziram resultados maiores ou iguais a Γ = –0,0014

p-value = 0,42

Rótulos das categorias não possuem *significado*

- **Estatística Γ pode ser usada como Critério Interno / Relativo**
  - Avaliação da classificação não supervisionada dos objetos produzida por meio de um algoritmo de agrupamento
- **Exemplo 3 (Jain & Dubes, 1988):**
  - Base 80X → Corte em $k = 3$ do dendrograma por vinculação completa

TABLE 4.2 Cluster by Category Table for 80X Data from Three-Cluster Complete-Link Classification

		Category 8	O	X	
Cluster	1	13	15	4	32
	2	2	0	0	2
	3	0	0	11	11
		15	15	15	

- Hipótese de "rótulos aleatórios" é apropriada para esse cenário?

- Γ (normalizado) = 0,567
- Histograma:

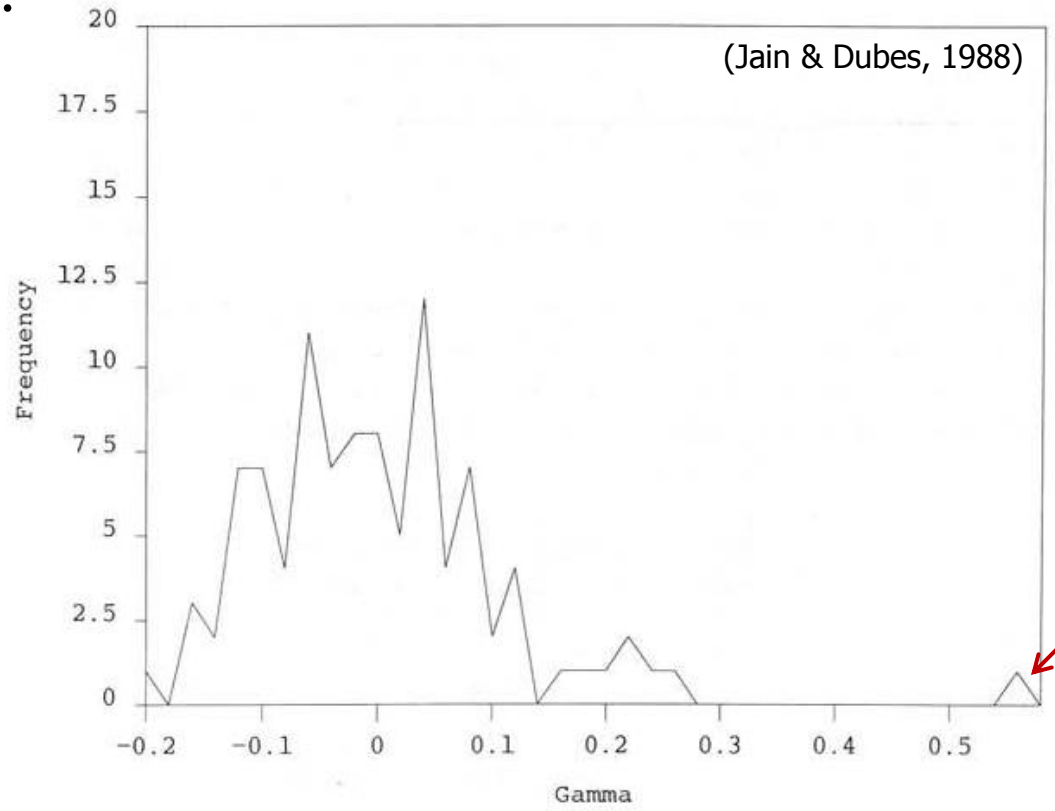


Figure 4.5 Histogram of Γ for 80X data under random label hypothesis from three-cluster complete-link classification.

- Podemos concluir que a partição se ajusta de maneira não usualmente bem com uma estrutura de grupos presente nos dados?

- Antes de responder esta questão, apliquemos o mesmo procedimento (k = 3) aos 45 objetos aleatórios dentro do hiper-cubo de 8 dimensões já discutido antes
- Γ (normalizado) = 0,291
- Tabela de confusão e histograma (rótulos aleatórios):

TABLE 4.3 Cluster by Category Table for random Data from Three-Cluster Complete-Link Classification

		Category			
		1	12	13	
Cluster	1	10	6	6	22
	2	1	3	5	9
	3	4	6	4	14
		15	15	15	

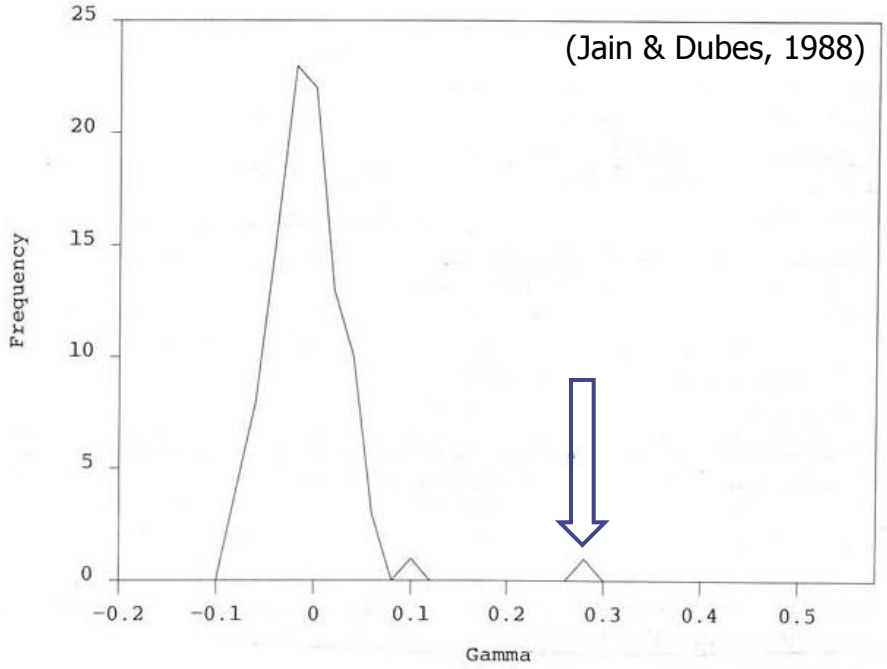


Figure 4.6 Histogram of Γ for random data under random label hypothesis when classification is obtained from cluster analysis.

- Isto significa um agrupamento estatisticamente válido ?
  - Certamente não, afinal os objetos são puramente aleatórios !
- Mas, então, por que a hipótese de rótulos aleatórios não é apropriada aqui !?

➢ Procedimento apropriado para gerar uma distribuição de referência para um critério interno:

- Gerar um grande número de conjuntos de dados (de mesma natureza) sob a hipótese de **posições aleatórias**, obter partições para cada conjunto e calcular os respectivos valores de Γ

➢ Para o exemplo anterior:

1) Gerar um conjunto de 45 objetos no hiper-cubo de 8 dimensões já discutido e obter as respectivas matrizes de dissimilaridade X
2) Aplicar vinculação completa e cortar o dendrograma em $k = 3$
3) Formar as matrizes binárias de conectividade Y
4) Repetir $m = 100$ vezes os passos 1 – 3 e obter um histograma

Analisemos os resultados obtidos...

- Γ (normalizado) = 0,567 para base de dados 8OX
- O que se pode concluir?

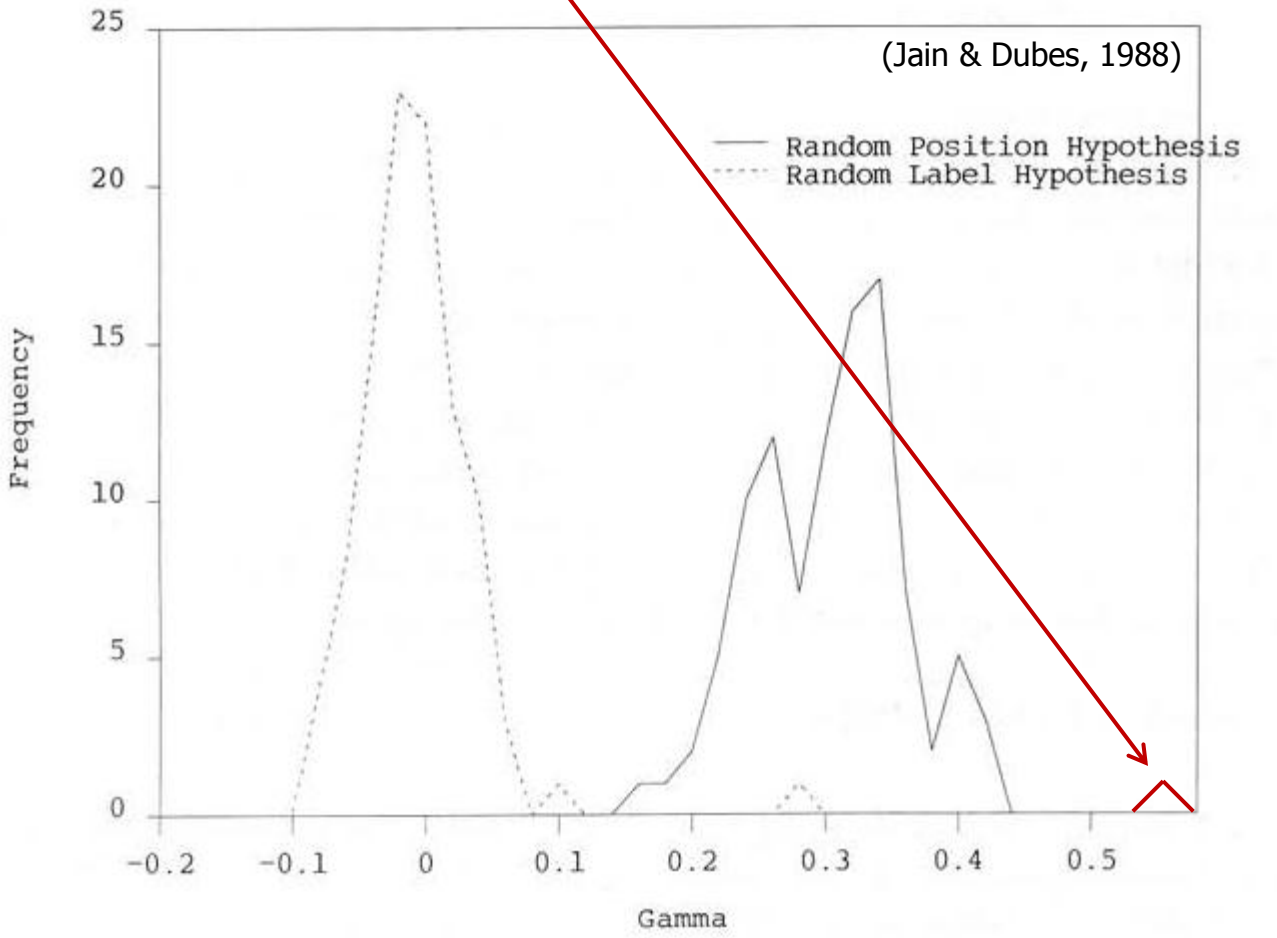


Figure 4.7 Histogram of Γ for 45 random patterns under random position and random label hypotheses from three-cluster complete-link classification.

Estatística γ (Estatística de Goodman-Kruskal)

- Correlação entre *ranks* de duas seqüências
  - $A=\{a_1, a_2, \ldots, a_M\}$ e $B=\{b_1, b_2, \ldots, b_M\}$
- Número de pares concordantes e discordantes
  - Par (a_i, a_j) e (b_i, b_j) é **concordante** se:
    - $(a_i < a_j \ \& \ b_i < b_j)$ ou $(a_i > a_j \ \& \ b_i > b_j)$
  - Par (a_i, a_j) e (b_i, b_j) é **discordante** se:
    - $(a_i < a_j \ \& \ b_i > b_j)$ ou $(a_i > a_j \ \& \ b_i < b_j)$
  - Demais casos são considerados **neutros**

$$\gamma = \frac{S_+ - S_-}{S_+ + S_-}$$

- S_+ e S_- se referem respectivamente às contagens (quantidades) de pares concordantes e discordantes

Exemplo (Goodman-Kruskal)

$A = \{ 0,\ 8, 10\}$

$B = \{ 3, -2,\ 5\}$

(0,8) e (3,-2) são pares discordantes

(0,10) e (3,5) são concordantes

(8,10) e (-2,5) são concordantes

$S_+ = 2$ e $S_- = 1$

$\gamma = 1/3$

Estatística γ de Goodman-Kruskal

- O índice de Goodman-Kruskal é uma correlação
  - $\gamma \in [-1,+1]$
- Uma diferença essencial para a estatística Γ é que γ só leva em consideração os *ranks*, desprezando as magnitudes
- As implicações dessa característica, particularmente no contexto de análise de agrupamento, são discutidas em:

Campello, R. J. G. B. & Hruschka, E. R. "On Comparing Two Sequences of Numbers and Its Applications to Clustering Analysis", **Information Sciences**, Vol. 179, 1025-1039, 2009

- Veremos um exemplo a seguir onde a aplicação de γ pode ser interessante

Exemplo 4 (Jain & Dubes, 1988)

- Considere a seguinte matriz de dissimilaridades **ordinais**:

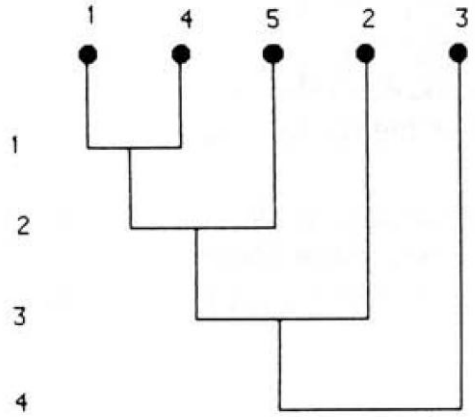
$$\boldsymbol{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 4 & 6 & 1 & 5 \\ — & 8 & 9 & 3 \\ — & — & 7 & 10 \\ — & — & — & 2 \end{bmatrix} \end{array}$$

- Podemos usar a estatística γ como um critério interno / relativo para avaliar agrupamentos hierárquicos desses dados

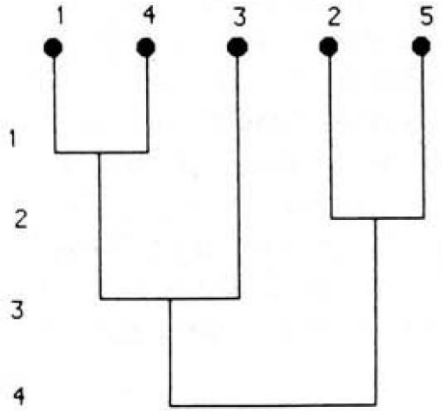
- Espera-se uma correlação "elevada" entre a matriz **D** e a *cophenetic matrix* de um agrupamento adequado dos dados

Exemplo 4 (cont.)

- Os agrupamentos de vinculação simples e completa são:



Single Link Dendrogram



Complete Link Dendrogram

- Os valores de correlação entre **D** e as *cophenetic matrices* correspondentes são $\gamma_{\text{single-link}} = 0{,}714$ e $\gamma_{\text{complete-link}} = 0{,}517$

  - Esse resultado sugere que, em termos relativos, o agrupamento de vinculação simples é mais compatível com os dados. Mas...

  - ... podemos afirmar que o agrupamento é "não usual" (válido) ?

Validação Estatística de Hierarquias

- Para responder esta questão, precisamos de uma distribuição de referência para a estatística γ

- Note que estamos avaliando γ como um critério interno, logo a hipótese nula de rótulos aleatórios não é apropriada

- Note ainda que **D** é uma matriz ordinal que não necessariamente se refere a objetos numéricos (pontos)

  - Logo, a hipótese nula de grafos (de proximidades ordinais) aleatórios pode ser mais apropriada que a hipótese de posições aleatórias

- Um algoritmo para gerar a distribuição referência de γ sob a hipótese nula de grafos aleatórios é apresentado a seguir

Validação Estatística de Hierarquias

ALGORITHM FOR BASELINE DISTRIBUTION OF γ UNDER RANDOM GRAPH HYPOTHESIS

Step 1. For fixed N (number of objects) form a dissimilarity matrix under the random graph hypothesis; that is, fill in the $N(N-1)/2$ entries with a randomly chosen permutation of the integers from 1 to $N(N-1)/2$.

Step 2. Cluster the N objects by a clustering method, such as the single-link or complete-link method.

Step 3. Form the cophenetic matrices for the dendrogram resulting from the clustering method.

Step 4. Compute γ between the dissimilarity and cophenetic matrices.

Repeat steps 1 to 4 on a Monte Carlo basis to create a baseline distribution for γ specific to the clustering method and value of N. (Jain & Dubes, 1988)

- No exemplo anterior (Exemplo 4), Monte Carlo com $m = 1000$ valores amostrais produziu distribuições de referência para single- e complete link tais que o 70º percentil de SL é 0,76 e o 50º percentil de CL é 0,72 p/ $N = 5$
  - Logo, nenhum dos agrupamentos (γ_{SL} = 0,714 e γ_{CL} = 0,517) pode ser considerado se casar com os dados de forma não usual !

Discussão

- O método anterior não poderia usar Γ ao invés de γ ?
  - Sim, mas dado que a natureza do problema é ordinal, pode ser mais apropriado utilizar uma estatística de mesma natureza
- E se a matriz **D** não for naturalmente ordinal ?
  - Lembre que os algoritmos SL e CL são invariantes a alterações na matriz **D** que não alteram as ordens relativas entre seus elementos
    - Produzem precisamente os mesmos dendrogramas
    - Logo, podemos transformar **D** em ordinal sem modificar o agrupamento
    - Infinitas matrizes com mesma ordem relativa mapeadas em uma única
    - Estimativas de Monte Carlo (m finito) podem ser mais precisas

Tendência de Agrupamento

- Refere-se a análise de **tendência de agrupamento** ao problema de avaliar se os dados exibem uma pré-disposição ao agrupamento em grupos naturais **sem identificar os grupos propriamente ditos** (i.e., a priori)

- Trata-se de uma área essencialmente estatística, pouco trivial por diferentes razões. Por exemplo:

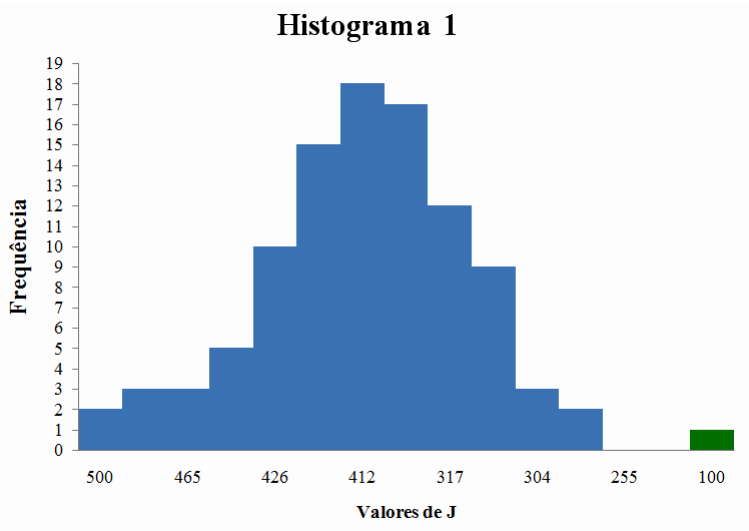
  - Existem variados métodos na literatura
  - Métodos podem depender fortemente da natureza dos dados
  - ...

Tendência de Agrupamento

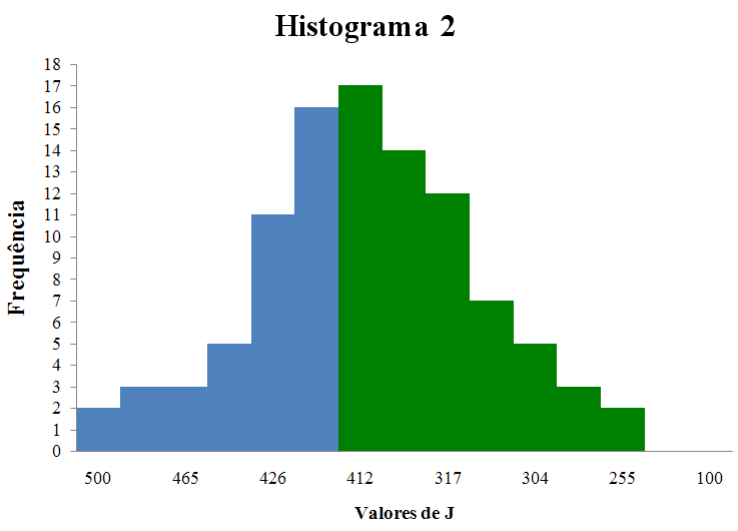
- Não discutiremos métodos de análise (a priori) de tendência de agrupamento
- No entanto, os métodos de validação que estudamos podem nos ajudar a investigar **a posteriori** se existe uma pré-disposição ao agrupamento em grupos naturais
- De fato, podemos tomar um critério interno / relativo como estatística de validação e comparar o valor deste critério para o agrupamento obtido frente a uma distribuição de referência sob hipótese de dados aleatórios
- Hipótese alternativa: um agrupamento válido sugere que existem grupos naturais nos dados (que foram encontrados)

Exemplo

- **Estatística *J* (função objetivo do FCM)**
  - H_0: posições aleatórias (dados da mesma natureza daqueles sob análise)
  - 2 simulações de Monte Carlo com $m = 100$
  - 1 cenário sugere dados com estrutura (Histograma 1) e o outro não



p-value = r/m = 1/100 = 0,01

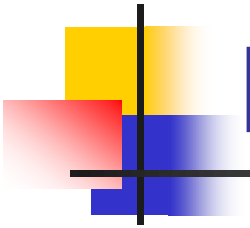


p-value = r/m = 60/100 = 0,6

Baseado no original de Luiz F. S. Coletta

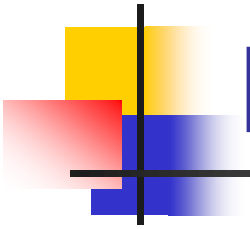
Exercícios

- Tome os resultados dos exercícios de execução de algoritmos hierárquicos propostos nas aulas correspondentes e calcule as correlações de Hubert normalizada e Goodman-Kruskal entre as matrizes de dissimilaridade e as *cophenetic matrices* produzidas pelos algoritmos
- Estime computacionalmente distribuições de referência apropriadas para alguns cenários através de simulações de Monte Carlo e valide os valores obtidos no item anterior contra essas distribuições



Leitura Recomendada

- Fortemente Recomendada...
  - Seções 4.1 a 4.4 de (Jain & Dubes, 1988)
- Sugerida (opcional)
  - Campello, R. J. G. B. & Hruschka, E. R. "On Comparing Two Sequences of Numbers and Its Applications to Clustering Analysis", Information Sciences, Vol. 179, 1025-1039, 2009



Referências

- Jain, A. K. & Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988