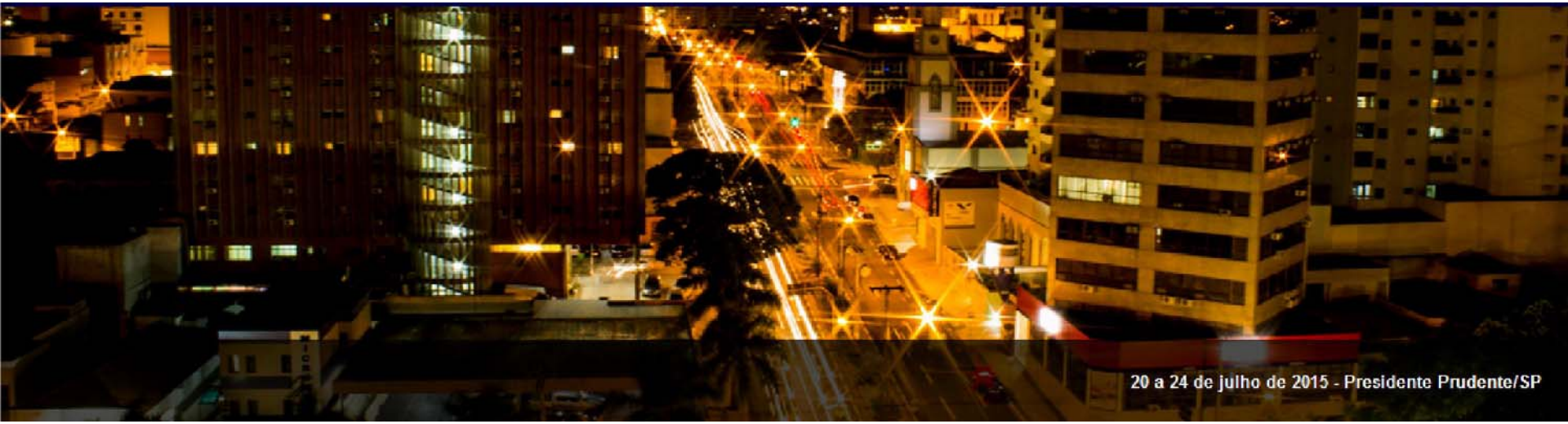
60ª RBras
Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria
SEAGRO
16º Simpósio de Estatística Aplicada à Experimentação Agronômica
A Estatística e os Novos Desafios: Tratamento e Modelagem da Informação



Aprendizado Estatístico em Medicina

Basilio de Bragança Pereira – FM e COPPE/UFRJ
Emilia Matos do Nascimento - UEZO e ICES/UFRJ

Objetivos

• Apresentar a metodologia estatística para estudantes de medicina e médicos, de forma conceitual sem ênfase em matemática (segundo preconizado por um dos criadores da Medicina Baseada em Evidência , Dr David Sackett).

• Desmistificar o uso inadequado, automatizado e indiscriminado de métodos estatísticos na pesquisa médica.

• Dar uma visão geral introdutória de métodos estatísticos modernos que contribuem para o conhecimento médico.

Ementa

• Papel da estatística na medicina. Tipos de medidas. Estatística fisiológica. Probabilidade, distribuições e propriedades. Medidas descritivas: posição, dispersão e forma. Representação gráfica.

• Inferência frequentista e Bayesiana: estimação, testes e intervalos. Análise de sobrevivência Modelos de regressão: linear logística , Poisson e Cox.

• Aplicações Médicas de métodos de aprendizado estatístico (redes neurais, árvores de classificação, regressão e de sobrevivência. Máquinas de vetores suporte) e de estatística multivariada (modelos log lineares, e grafos não orientados, regressão LASSO para escolha de variáveis)

• Apresentação do R-project (sistema computacional)

Índice

CAPÍTULO 1

ESTATÍSTICA EM MEDICINA

Estatística

Um assunto que a maioria dos estatísticos acha difícil porém que quase todos os médicos são especialistas.

Estatísticos são vistos como:

- Desnecessários: alguém que sabe usar pacotes estatísticos.
- Técnicos necessários: digitadores de números.
- Demônios necessários: a benção do estatístico é necessária para publicação.
- Mágicos necessários : pode obter significância manipulando os dados (Lies, Damn lies, and Statistics – Disraeli).

Ou como:

- Deus: Salvador, responde as rezas.
- Bispo: Abençoa, ouve aos pecados.
- Padre : Companheiro.
- Sacristão: Servo, faz o que é mandado.

Estatístico como colega:

- Coletando informação adequadamente.
- Interpretando informação adequadamente.
- Analisando informação adequadamente.
- Podendo pescar ou ensinar como pescar.
- Estar disposto a ensinar os **conceitos** atrás da metodologia.
- Precisa ter um conhecimento da área de aplicação para ser um consultor eficiente.

Efeito da revolução do computador:

- Liberou os cálculos cansativos
- Facilita a análise exploratória de dados
- Permite trabalhar com grande massa de dados
- Permitiu trabalhar com métodos multivariados complexos
- Permite o uso de métodos computacionalmente intensivos
- Permite a possibilidade de estudar convergência assintótica e revolucionou o ensino

- Sistemas computacionais comerciais: caros , tornando-se inviável

Sistemas gratuitos:

- Sistema R
- WinBugs
- Etc.

Desenvolvimento histórico:

Começo do século 20 (antes de 1950):
- Aplicações a agricultura
- Modelos paramétricos (Gaussianos)
- Univariados

Anos 1960-1980:
- Aplicações biomédicas
- Modelos lineares
- Multivariado

Anos após 1990 e século 21:
- Genética
- Métodos computacionais intensivos
- Modelos longitudinais, multidimensionais , complexos , não lineares, métodos robustos etc
- Aprendizado Estatístico

Exemplos: tese de doutorado da Clinica Medica (CART) – Dra Fernanda Mello e da COPPE (redes neurais) – Dra Alcione Miranda dos Santos, ambas desenvolvidas na UPT – Unidade de Pesquisa de Tuberculose do HUCFF - Prêmios 2002 e 2004 de Ciência e Tecnologia do SUS.

Outras: Amália Reis (Doutorado Medicina, Redes Neurais), Emília Nascimento (Mestrado Produçao, Redes Neurais), Rodrigo Collazo (Mestrado Produçao – Support Vector Machine), Alfredo Passos (Mestrado Produçao – Redes Neurais Probabilísticas) – todas na área de Medicina.

Interação

Medicina x Estatística

Como promovê-la? Por que?

Por que?

Dificuldades com as diversas fórmulas estatísticas para o clínico-futuro-realizador de um ensaio clínico:

Causas:

- Elas assustam e dão medo de usar
- Elas são difíceis de lembrar
- Elas requerem um conhecimento de matemática e estatística muito longe do conhecimento e experiência do clínico (would-be-trialist)
- O tempo necessário para entender suas nuances será feito às expensas de manter competência clínica, vida social, uma auto-imagem positiva e um senso de humor
- Elas existem isoladas e sem relação com cada uma das outras

(Tenha cuidado com o homem que trabalha duro para aprender algo , aprende , e no final não está mais competente do que antes. Ele está cheio de re-sentimento criminoso com as pessoas que não são competentes, mas que não chegaram à sua situação da maneira difícil).

Como?

David Sackett (2001):

Solução é uma introdução a **Estatística Fisiológica**

Esqueça as fórmulas (eu sei menos fórmulas hoje do que quando planejei meu primeiro RCT em 1963)

Nunca trabalhe sozinho, porém sempre com um estatístico (a grande maioria de clínicos que eu encontrei sabem suficiente estatística para arranjar problemas, porém não o suficiente para sair deles)

Empregue “estatística fisiológica”:

A importância das formulas estatísticas não está na sua individualidade mas sim na sua combinação criteriosa. Clínicos as entenderão bem melhor se pensarem nelas em termos fisiológicos, análogos a combinar os determinantes do sistema sanguíneo de pressão arterial.

A única formula da estatística fisiológica é ridiculamente simples:

Quão curto é o intervalo de confiança

Diferença entre os efeitos do tratamento experimental e do controle

$$confiança = \frac{\text{sinal}}{ruído}\sqrt{n}$$

Soma de todos os fatores que podem afetar o sinal (Incerteza)

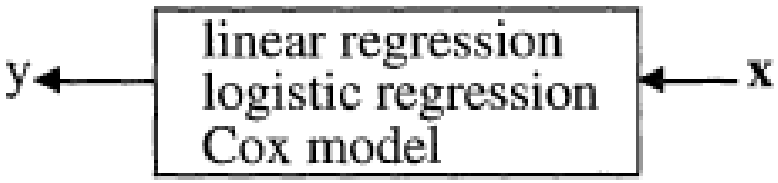
Nº de pacientes na amostra

e se for Bayesiano

$$P(\theta / X) \ \alpha \ P(\theta)P(X / \theta)$$

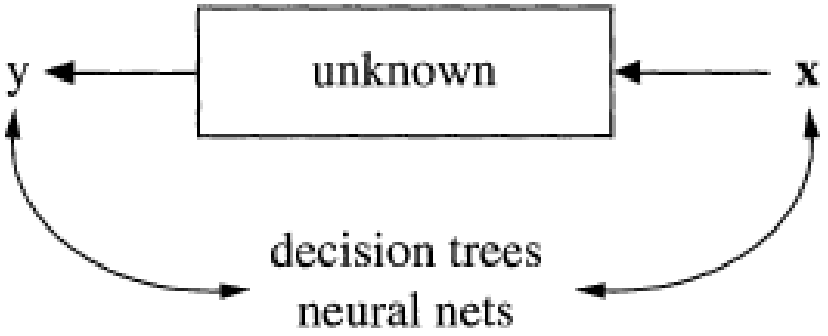
Duas Culturas

Cultura de Modelagem dos Dados



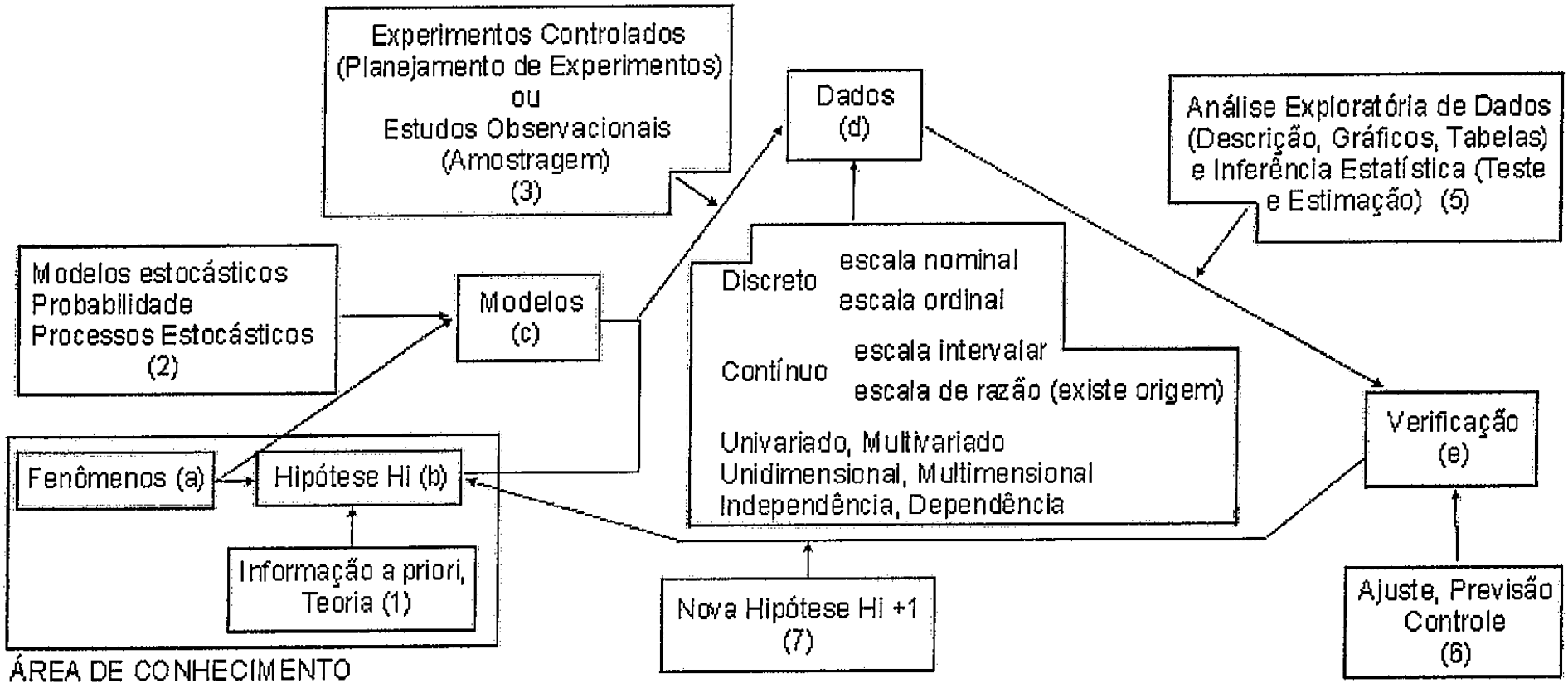
Estatística:Teoria em busca de dados

Cultura Algorítmica



Data Mining:Dados em busca de teoria

Paradigma I - Estatística



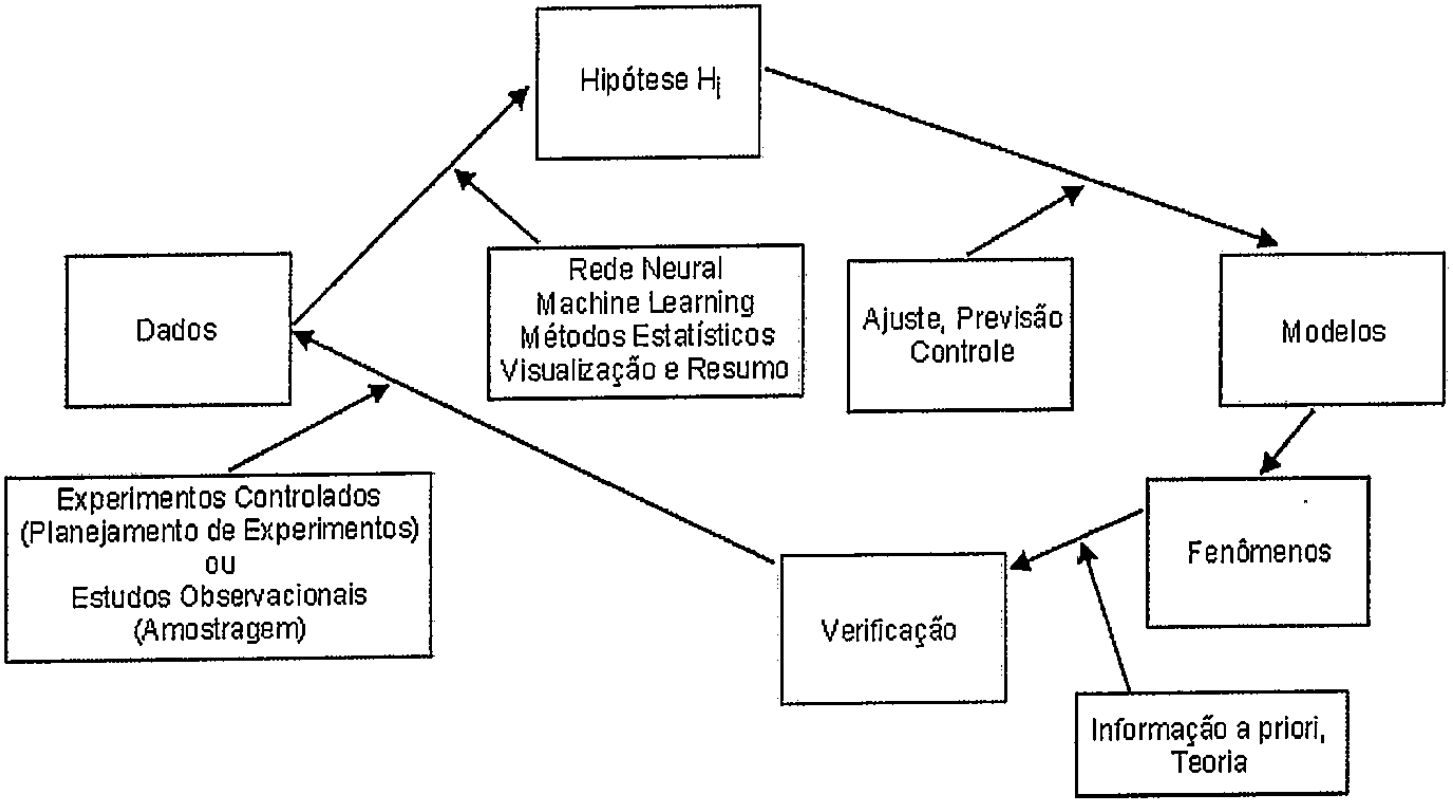
A estatística dispõe de diversos métodos para auxiliar um cientista:

- Modelos
  - Estocásticos (fenômenos dinâmicos)
  - Probabilísticos (fenômenos estáticos)
- Coleta de dados
  - Planejamento de Experimentos (experimentos controlados)
  - Amostragem (estudos não controlados)
- Verificação das Hipóteses
  - Análise Exploratória de Dados
    - Listagens e Relatórios
    - Tabelas
    - Visualização Gráfica

- Inferência Estatística
  - Testes e Estimação
    - Séries Temporais
    - Estatística Multivariada (Cluster, Classificação, Componentes Principais, Análise de Correspondência, Análise Fatorial, Correlação Canônica, Escalonagem etc.)
    - Estatística não Paramétrica (Robustez, Regressão em Árvore, Projection Pursuit, GAM, etc.)
    - Regressão (Linear, Logística, Log-linear, GLIM etc.)
    - Etc.

- Verificação do Modelo
  - Análise de Resíduos
  - Teste de Adeqüabilidade
  - Previsão e Controle

Paradigma II - Data Mining



- Técnicas Estatísticas
  - Componentes Principais
  - Análise Fatorial
  - Equações Estruturais Lineares
  - Mínimos Quadrados Parciais
  - Correlação Canônica
  - Análise de Correspondência
  - Escalonagem Multidimensional
  - Análise de Conglomerados
  - Classificação e Discriminação
  - Modelo Linear Generalizado
  - Modelo Log-linear
  - Regressão Não Paramétrica
  - Modelo Aditivo Generalizado
  - Árvores de Regressão
  - Séries Temporais
- *Machine Learning*
  - *K-nearest neighbour*
  - Árvores de Classificação
  - Redes Neurais Artificiais
  - Algoritmos Genéticos
  - Aquecimento Simulado
  - Indução de Regras
- Visualização e Resumo

Estas técnicas são, em geral, utilizadas com o objetivo de :

- Classificação
- Estimação
- Previsão
- Análise de Associação
- Analise de Agrupamentos

“A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original”

Albert Einstein.

O Problema do Epidemiologista

Um epidemiologista foi enviado a uma região para conferir a prevalência de uma doença. Ele foi informado que os casos foram numerados seqüencialmente, e durante um período ele observou uma amostra aleatória de 5 doentes. Não querendo consultar os prontuários, de difícil acesso, será que ele pode fazer algumas afirmações sobre o número de casos baseado nos números de registros dos pacientes vistos no período (amostra) : 405, 280, 73, 440, 179 ?

i)Inicialmente consideremos o problema de estimação

Ordenemos os pontos em uma linha reta

73 179 280 405 440

_____ N = ?

O início da linha é o número 1, qual será o ponto final N à direita que corresponderá ao numero de casos prevalentes ? Sabemos que o ponto deve ser maior ou igual a 440.

Podemos argumentar que, se temos 72 números menores que o menor valor observado (73), é razoável supor que podemos ter também 72 números acima de 440. Em linguagem estatística, uma estimativa razoável para a prevalência seria 440 + 72 = 512.

Um outro argumento seria considerar que se temos 279 números menores que a mediana 280 seria razoável supor que também teríamos 279 acima da mediana. Uma outra estimativa seria então 280 + 279 = 559.

Temos duas estimativas, a primeira 512,denominada **estimativa pelo extremo-**(**ee**) e a segunda 559,denominada **estimativa pela mediana-**(**em**). Qual delas escolher ? Bioestatísticos tem métodos para responder essas questões, que ilustraremos a seguir.

Suponha que o verdadeiro numero dos casos prevalentes seja 550. Neste caso, os erros são

erro (**ee**) = |550 - 512| = 38

erro (**em**) = |550 - 559| = 9

Para conferir se esta diferença entre os erros tem algum padrão, observamos mais três amostras com os resultados:

Amostra	ee (erro)	em (erro)
1 – (405, 280, 73, 440, 179)	512 (38)	559 (9)
2 – (72, 132, 189, 314, 290)	385 (165)	377 (173)
3 – (191, 124, 460, 256, 401)	583 (33)	511 (39)
4 – (450, 485, 56, 383, 399)	540 (10)	797 (247)

Verificamos que a média dos erros são:

ee: (38+165+33+10)/4=61,5
em: (9+173+39+247)/4=117

Pode-se mostrar que, se continuássemos a tirar amostras a média dos erros de **ee** seriam menores.

Uma outra razão para escolher **ee** é que em alguns casos **em** produz resultados inconsistentes. Por exemplo, se na nossa amostra inicial o maior número fosse 650 em vez de 440, **em** continuaria a ser 559, o que é uma estimativa ruim já que observamos 650.

Bioestatísticos, através da teoria das probabilidades desenvolveram métodos e critérios para escolher entre estimativas, a serem apresentados na Seção 3.

É interessante mencionar que estimativas estatísticas semelhantes as anteriores, sobre o número de tanques produzidos pelos alemães na Segunda Guerra Mundial, eram muito mais precisas do que as baseadas em fontes de inteligência.

ii) Consideremos agora o problema de testar uma hipótese

Suponha que não sabemos o valor do numero de casos prevalentes e que desejamos testar a hipótese de que o mesmo é 1000, baseado na amostra: 405, 280, 73, 440, 179. Isto é, a amostra obtida permite que duvidemos que N = 1000? Por que?

Para avaliar a evidência experimental (amostra) com a afirmação da hipótese (N = 1000) façamos primeiro uma analogia com o lançamento de uma moeda.

Sob a suposição de que N = 1000, associemos números menores que 500 com C – cara, e maiores que 500 com K – coroa, esquematicamente.

x_____ x _____ x

0 500 1000

Cara – C Coroa = K

p(C) = 1/2 p(K) = 1/2

É fácil verificar que lançando a moeda:

2 vezes, temos os resultados possíveis: CC, CK, KC, KK e logo como são equiprováveis p(CC) = 1/4 = 1/22

3 vezes, temos os resultados: CCC, KKK, CCK, CKC, KCC, CKK, KCK, KKC, e logo p(CCC) = 1/8 = 1/23

… … …

5 vezes, temos p(CCCCCC) = 1/25 = 1/32 = 0,031

Logo se N = 1000 a probabilidade da amostra observada é 1/32, já que os números observados são menores que 500. Portanto temos duas alternativas: a afirmação (N = 1000) é verdadeira e um evento raro ocorreu ou a afirmação não é verdadeira. A segunda afirmativa parece mais razoável.

iii) Finalmente consideremos estimação por intervalos ou intervalos de confiança

Inicialmente observe que na analogia anterior, “CCCCC” e “todos os 5 números são menores que 500” eram equivalentes com probabilidade p(CCCCCC) = 1/32 = 1/25 = 1/2 . 1/2 . 1/2 . 1/2 . 1/2 = 0.031. Na realidade os 5 números são menores ou iguais a 440 e portanto a probabilidade de escolher um número menor que 440 entre os números menores ou igual a 1000 é 440/1000. Logo a probabilidade exata de escolher 5 números desta forma é:

440/1000 . 440/1000 . 440/1000 . 440/1000 . 440/1000 = 0,016

que é bem menor que a probabilidade aproximada 0.031, isto é, este método indica que se N = 1000 a amostra é mais rara ainda.

Vamos agora testar as hipóteses: N = 900, 800, 700 etc. De forma análoga teríamos:

N	**p**
1000	(440/1000)^5 = 0,016
900	(440/900)^5 = 0,028
800	(440/800)^5 = **0,05** = **1/20**
700	(440/700)^5 = 0,098

Alguns bioestatísticos consideram p = 0,05 como ponto divisório entre probabilidades “pequenas” que sugerem rejeição da hipótese e probabilidades “grandes” demais para sugerir rejeição. Neste caso valores maiores que 800, para o numero desconhecido de doentes são rejeitados pois tem probabilidades “pequenas” associadas, e valores menores ou iguais a 800 não são rejeitados pois tem probabilidades “grandes” associadas. Neste caso afirmamos que:

$N \leq 800$ com 95% de confiança

O mesmo tipo de raciocínio pode ser usado para obter um limite inferior. Sabemos que o valor mínimo é 440, que foi observado. Caso o número de doentes seja 440 a probabilidade deste doente não ser observado na amostra é (439/440)^5 e logo a probabilidade dele ser observado é:

0.011 = 1 – (439/440)^5

Como é uma probabilidade “pequena”, N = 440 é rejeitado, ou seja N deve ser maior que 440.

De forma análoga temos:

N	**p**
440	$1 - (439/440)5 = 0.011$
441	$1 - (439/441)5 = 0.022$
444	$1 - (439/444)5 =$ **0.05** = **1/20**

e $N \geq 444$ com 95% de confiança, e combinando os dois resultados:

$444 \leq N \leq 800$ com 90% de confiança

Finalmente, é importante mencionar que a regra de valor p = 0,05 não deve ser considerada estritamente. Em aplicações, outros valores de p (0.10, 0.015, ou 0.01) podem ser usados. É mais conveniente determinar o valor p e decidir em cada problema específico se o evento é raro ou não.

Problema	**Dados, Amostras, Informação**	**Parâmetros Envolvidos**	**Estimadores**	**Hipóteses**	**Estatística de Teste**	**Distribuição Amostral**	**p-Valor**	**Intervalo de Confiança (1-α)%**
Do epidemiologista	(405, 280, 73, 440, 179) tamanho = 5	N – n^0. doentes	**ee**, **em** extremo, mediana	H_0: N = 1000	N^{0}s. menores que 500 supondo N = 1000 ≈ (C e K)	Binomial – lançamentos C e K	P(CCCCC) = 0.031 ou melhor 0,016	444 ≤ N ≤ 800, α = 0.10
Uma proporção p	$(x_1,\ldots,x_n)$ x = 1 ou 0 tamanho = n	p – proporção de 1	$p = \frac{\sum x_i}{n}$	$H_0 : p = \frac{1}{2}$	$Z_{obs} = (\hat{p} - 1/2)/\sqrt{n\frac{\hat{p}(1-\hat{p})}{}}$	Aproximadamente N(0,1), Z normal	Ver tabela da normal para Z_{obs}	$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Uma média μ	$(x_1,\ldots,x_n)$ tamanho = n	μ, σ^2	$\bar{x} = \frac{1}{n}\sum x_i \quad s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$	H_0: μ = 0	$T_{obs} = (\bar{x} - 0)/\frac{s}{\sqrt{n}}$	Distribuição t_{n-1}	Ver tabela de t_{n-1}, para t_{obs}	$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$
Razão de variâncias $\frac{\sigma_x^2}{\sigma_y}$	$(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$ tamanhos n e m	μ_x, μ_y σ_x^2, σ_y^2	$\bar{x}$ e $\bar{y}$ $\quad s_x^2$ e s_y^2	$H_0 : \frac{\sigma_x^2}{\sigma_y} = 1$	$F_{obs} = \frac{s_x^2}{s_y^2} \quad s_x^2 > s_y^2$	Distribuição F(n-1, m-1)	Ver tabela $F_{1-\alpha/2}(n_1 - 1, m-1)$ e $F_{\alpha/2}$(n-1, m-1)	$s_x^2/s_y^2 \quad F_{1-\alpha/2} \le \frac{\sigma_x^2}{\sigma_y^2} \le \frac{s_x^2}{s_y^2} F_{\alpha/2}$
Diferença de proporções $p_x - p_y$	$(x_1, \ldots, x_n)$, $(y_1, \ldots, y_m)$ x_1, y_1 = 1 ou 0 tamanhos n e m	p_x, p_y	$\hat{p}_x = \frac{\sum x}{n} \quad \hat{p}_y = \frac{\sum y}{m}$	H_0: $p_x - p_y = 0$	$Z_{obs} = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n} + \frac{\hat{p}_y(1-\hat{p}_y)}{n}}}$	Aproximadamente N(0,1), Z normal	Ver tabela do normal para Z_{obs}	$\hat{p}_x - \hat{p}_y \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n} + \frac{\hat{p}_y(1-\hat{p}_y)}{m}}$
Diferença de médias: $\mu_x - \mu_y$ Variâncias iguais $\sigma_x^2 = \sigma_y^2$ e desconhecidos	$(x_1, \ldots, x_n)$, $(y_1, \ldots, y_m)$ tamanhos n e m	μ_x, μ_y $\sigma^2 = \sigma_x^2 = \sigma_y^2$	$\bar{x}, \bar{y} \quad s_c = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$	H_0: $\mu_x - \mu_y = 0$	$t_{obs} = \frac{\bar{x} - \bar{y}}{s_c\sqrt{\frac{1}{n} + \frac{1}{m}}}$	Distribuição t_{n+m-2}	Ver tabela de t_{n+m-2} para t_{obs}	$\bar{x} - \bar{y} \pm t_{n+m-2,\alpha/2}\, s_c\sqrt{\frac{1}{n} + \frac{1}{m}}$
Diferença de média: $\mu_x - \mu_y$ Variâncias desiguais e desconhecidas $\hat{\sigma}_x \neq \hat{\sigma}_y$	$(x_1, \ldots, x_n)$, $(y_1, \ldots, y_m)$ tamanhos n e m	μ_x, μ_y σ_x^2, σ_y^2	$\bar{x}$ e $\bar{y} \quad s_x^2$ e s_y^2	H_0: $\mu_x - \mu_y = 0$	$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$	Distribuição t_0 com g.l. $\upsilon = \left\{\frac{\frac{s_x^2}{n} + \frac{s_y^2}{m}}{\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{n_2^2(n_2+1)}}\right\}$	Ver tabela de t_0 para t_{obs} se n e m maior que 15 use normal Z para Z_{obs}	$\bar{x} - \bar{y} \pm t_{0,\alpha/2}\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$
Teste pareado de igualdade de médias	$(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$ $(y_1 - x_1, \ldots, y_n - x_n)$ tamanhos n	μ_x, μ_y $d = \mu_y - \mu_x$	$\bar{d} = \frac{\sum d_i}{n} \quad s_d^2 = \frac{\sum (d_i - \bar{d})}{n-1}$	H_0: $\mu_y - \mu_x = 0$	$t_{obs} = \frac{d}{s_d/\sqrt{n}}$	Distribuição t_{n-1}	Ver tabela de t_{n-1} para t_{obs}	$\bar{d} \pm t_{n-1,\alpha/2}\, s_d/\sqrt{n}$
Teste de ajustamento Frequências relativas iguais a frequência de uma distribuição	Histograma com K classes	Parâmetros da distribuição θ se f(x θ)	Estimadores $\hat{\theta}$ e histograma de f(x, $\hat{\theta}$)	H_0: dados seguem f(x θ)	$\chi^2_{obs} = \sum_{i=1}^{k}\left(\frac{O_i - E_i}{E_i}\right)^2$ O_{oi} – frequência observada E_i – frequência esperada segundo f(xθ)	Distribuição χ^2_{k-1-m} M – número de parâmetros da distribuição teórica	Ver tabela de χ^2_{k-1-m} para χ^2_{obs}	-
Teste de independência e homogeneidade em tabelas de contingência	Tabela de contingência p por q	Frequências supondo independência $\theta_{ij} = \theta_{i.}\,\theta_{.j}$	$\hat{\theta}_{ij} = \hat{\theta}_{i.}\,\hat{\theta}_{.j}$ $\hat{\theta}_{ij} = \frac{n_{ij}}{N}$	H_0: classificações independentes	$\chi^2_{obs} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(\frac{O_i - E_i}{E_i}\right)^2$ O_i – frequência observada Ei – frequência esperada supondo independência	Distribuição $\chi^2_{(p-1)(q-1)}$	Ver tabela de $\chi^2_{(p-1)(q-1)}$ para χ^2_{obs}	-
Wilcoxon-Man-Whitney para comparação de duas médias	$(x_1, \ldots, x_n)$ $(y_1, \ldots, y_m)$ suponha n < m N = n + m	-	postos -	H_0: $\mu_x - \mu_y = 0$	$W_{obs} = n(N+1) - R$ R = soma das ordenações da menor amostra ou R_{obs}	Distribuição de W – M – W e se n e m maior que 15 R tem distribuição aproximadamente normal com $\mu_R = \frac{n(n+m+1)}{2} \quad \sigma_R^2 = \frac{m\,\mu_R}{6}$	Ver tabela de W – M – W para W ou tabela da normal para $\frac{R_{obs} - \mu_R}{\sigma_R}$	-

Test 88 Likelihood ratio test for testing the parameter of the rectangular population

Object

To test for one of the parameters of the rectangular population with probability density function (PDF)

$$f_0(X) = \begin{cases} \dfrac{1}{2\beta}, & \alpha - \beta \leqslant X \leqslant \alpha + \beta \\ 0, & \text{otherwise.} \end{cases}$$

Method

Let $X_1, X_2, \ldots, X_n$ be a random sample from the above rectangular population and we are interested in testing $H_0: \alpha = 0$ against $H_1: \alpha \neq 0$. Then, the likelihood ratio (LR) test criterion for testing H_0 is:

$$\lambda = \left(\frac{X_{(n)} - X_{(1)}}{2Z}\right)^n = \left(\frac{R}{2Z}\right)^n,$$

where R is the sample range and $Z = \max[-X_{(1)}, X_{(n)}]$. Then the asymptotic distribution of $2\log_e\lambda$ is χ^2_2.

Example data

Consider $(-0.2, -0.3, -0.4, 0.4, 0.3, 0.5)$ as a random sample from a uniform distribution. Here,
$n = 6$

$R = 0.5 - (-0.2) = 0.7$, $Z = \max[0.4, 0.5] = 0.5$

$$\lambda = \left(\frac{R}{2Z}\right)^6 = (0.7)^6 = 0.1176, \quad 2\log_e\lambda = 0.18588$$

Critical value $\chi^2_2(0.05) = 5.9$ [Table 5]
Hence we reject the hypothesis that $H_0: \alpha = 0$.

Test 89 UMP test for testing the parameter of an exponential population

Object

A test for the parameter (θ) of the exponential population with PDF

$$f(X, \theta) = \theta e^{-\theta X}, \qquad X > 0.$$

Method

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with PDF

$$f(X, \theta) = \theta e^{-\theta x}, \qquad X > 0.$$

Let our null hypothesis to be tested be $H_0: \theta = \theta_0$ against the alternative $H_1: \theta = \theta_1 (\theta_1 \neq \theta_0)$.

Case (a) $\theta_1 > \theta_0$: The MP (most powerful) critical region is given by:

$$W_0 = \left[X = \sum_i X_i \leqslant \chi^2_{1-\alpha,2n} \,|\, 2\theta_0\right].$$

Since W_0 is independent of θ_1, so W_0 is UMP for testing $H_0: \theta = \theta_0$ against $H_1: \theta > \theta_0$.

Case (b) $\theta_1 < \theta_0$: The MP critical region is given by:

$$W_1 = \left[X = \sum X_i > \chi^2_{\alpha,2n} \,|\, 2\theta_0\right].$$

Again W_1 is independent of θ_1 and so it is also UMP for testing $H_0: \theta = \theta_0$ against $H_1: \theta < \theta_0$.

Example data

Let us consider a sample of size 2 from the population $f(X_1\theta) = \theta e^{-\theta X}$, $X > 0$. Consider testing $H_0: \theta = 1$ against $H_1: \theta = 2$, i.e. $\theta > \theta_0$.

Critical region is $W = \left\{\mathbf{X} \,\middle|\, \sum X_i \leqslant \chi^2_{0.95,4|2}\right\}$

$$= \left\{\mathbf{X} \,\middle|\, \sum X_i \leqslant 0.711 \,\middle|\, 2\right\} \quad \text{[Table 5]}$$

$$= \left\{\mathbf{X} \,\middle|\, \sum X_i \leqslant 0.356\right\}$$

Similarly a critical region for the other part can be constructed.

Exemplos

1) Testes de Significância e Testes Diagnósticos

Tabela 1 – Resultado de teste diagnóstico

Teste	**Doença**		**Total**
	Presente D^+	**Ausente D^-**	
Positivo (T^+)	Correto positivo a	Falso positivo c	T^+ positivo
Negativo (T^-)	Falso negativo b	Correto Negativo d	T^- negativo
Total	D – doentes a + b	A – ausência c + d	N = a + b + c + d

Quantidades Associadas

- p(T+/D+) = a/(a+b) = S – sensibilidade
- p(T-/D-) = d/(c+d) = E – especificidade
- p(D+) = (a+b)/N = prevalência
- p(T+) = (a+c)/N = positividade do teste
- p(T-) = (b+d)/N = negatividade do teste
- p(D+/T+) = a/(a+c) = VPP – valor preditivo positivo
- p(D-/T-) = d/(b+d) = VPN – valor preditivo negativo

Tabela 2 – Decisões e erros de teste de hipótese

Decisão do	**Realidade**	
Teste	**H_0: D^- é verdadeira**	**H_A: D^+ é verdadeira**
Não Rejeita H_0 (T^-) (Rejeita H_A)	Decisão correta Probabilidade: 1 - α Nível de confiança	Erro tipo II Probabilidade: β
Não Rejeita H_A (T^+) (Rejeita H_0)	Erro tipo I Probabilidade: α Nível de significância p-p-Valor	Decisão correta Probabilidade: 1 - β Poder

Tabela 3 – Analogias: teste diagnóstico x teste de hipóteses

Proporção	**Símbolo**	**Teste Diagnóstico**	**Teste de Hipótese**
Correto positivo	$S = P(T^+/D^+)$	Sensibilidade	$1-\beta$: poder
Correto negativo	$E = P(T^-/D^-)$	Especificidade	$1-\alpha$: nível de confiança
Falso positivo	$P(T^+/D-)$	1 – Especificidade	α: erro tipo I, valor-p nível de significância
Falso negativo	$P(T^-/D^+)$	1 – Sensibilidade	β: erro tipo II

2) Verossimilhança

- Um determinado medicamento em teste foi utilizado em 10 pacientes, deste 7 ficaram curados
- Se soubermos a eficácia p.ex (π= 0,7) podemos calcular a probabilidade de obter x curas em 10 pacientes

$$P\left(x \mid \pi = 0,7\right) = C_{10}^{x} 0,7^{x} (1-0,7)^{10-x}$$

- O problema é que a real eficácia (π) deste medicamento é desconhecida e então definimos a função de verossimilhança.
- O valor da verossimilhança de cada valor de π é:

$$L\left(\pi \mid x = 7\right) = \begin{pmatrix} 10 \\ 07 \end{pmatrix} \pi^{7} \left(1-\pi\right)^{3}$$

Verossimilhança de π

Estimador de máxima verossimilhança

y	π								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	.349	.107	.028	.006	.001				
1	.387	.269	.121	.04	.01	.002			
2	.194	.302	.234	.121	.044	.01	.002		
3	.057	.201	.267	.215	.117	.043	.009	.001	
4	.011	.088	.2	.251	.205	.111	.036	.005	
5	.002	.027	.103	.201	.246	.201	.103	.027	.002
6		.005	.036	.111	.205	.251	.2	.088	.011
7		.001	.009	.043	.117	.215	.267	.201	.057
8			.002	.01	.044	.121	.234	.302	.194
9				.002	.01	.04	.121	.269	.387
10					.001	.006	.028	.107	.349
	1	1	1	1	1	1	1	1	1

Razão de Verossimilhança

- Corresponde a quantas vezes um determinado valor é mais plausível que outro.

- Exemplo: π=0,7 ou π=0,5

$$\frac{L\left(\pi = 0,7 \mid x = 7\right)}{L\left(\pi = 0,5 \mid x = 7\right)} = \frac{0,267}{0,117} = 2,28$$

3) Inferência Bayesiana

- O **Teorema de Bayes** transforma a crença prévia (distribuição a priori, prevalência antes do teste, risco inical) através da verossimilhança (dados, resultado do teste) em uma crença posterior (distribuição a posteriori, prevalência após resultado do teste).

- Vamos considerar o mesmo caso do remédio experimental.
- Mas temos 6 médicos com crenças prévias na eficiência do remédio
- Temos uma distribuição a priori a eficiência (π) do remédio

Eficiência (π)	Nº de médicos	P(π)
0,4	1	1/6
0,5	2	2/6
0,6	2	2/6
0,7	1	1/6

- A verossimilhança seria a experiência onde 7 de 10 ficaram curados
- Com isto a distribuição a posteriori da eficiência (π) do remédio é:

π	Priori-p(π)	Verossimilhança	Priori x verossimilhança	Posteriori p(π/y=7)
0,4	1/6 = 0,167	0,043	0,167 X 0,043 = 0,007	0,007/,163 = 0,043
0,5	2/6 = 0,333	0,117	0,333 X 0,117 = 0,039	0,039/,163 = 0,239
0,6	2/6 = 0,333	0,215	0,333 X 0,215 = 0,072	0,072/,163 = 0,442
0,7	1/6 = 0,167	0,267	0,167 X 0,267 = 0,045	0,045/,163 = 0,276
total	1,	N,A	0,163	1,

Estimador de máxima probabilidade posterior

Algumas reflexões

1)Exemplo de interpretação incorreta do valor-p

A verificação da falta de entendimento do significado do Valor-P, tem sido testado em turmas de pós-graduação de Medicina e Engenharia usando os seguintes questionários de Diamond e Forrester e Freeman respectivamente.

Questionário 1 – (Diamond e Forrester)

O que você concluiria se um experimento clínico bem planejado, realizado para verificar o efeito de um certo tratamento, resultou em uma resposta benéfica ($p < 0,05$)?

a. de acordo com este resultado, as chances são menos de 5% de que a terapia não tem efeito;

b. as chances são menos de 5% em obter este resultado se a terapia não tem feito;

c. as chances são menos de 5% de não ter obtido esse resultado se a terapia tem efeito;

d. nenhum acima.

Questionário 2 – (Freeman)

Um experimento controlado, realizado para determinar a eficácia de um novo tratamento que o mesmo é significativamente melhor que placebo ($p < 0,05$). Qual das seguintes afirmações você prefere?

a. foi aprovado que o tratamento foi melhor que placebo;

b. se o tratamento não tem efeito, há menos de 5% de chance de se obter tal resultado;

c. o efeito observado do tratamento é tão grande que há menos de 5% de chance do tratamento não seria melhor que placebo;

d. realmente não sei o que é valor – p e não quero adivinhar.

A conclusão obtida com as aplicações destes questionários a alunos de pós-graduação de engenharia e medicina e com presença de alguns estatísticos coincide com as dos autores.

A resposta correta em ambos é b) mas em geral mais de 50% das pessoas respondem incorretamente e todos tem dificuldades de distinguir a diferença entre as escolhas.

Em um curso de doutorado em medicina apliquei estes questionários em alunos que já haviam feito pelo menos um curso de estatística e um curso de analise crítica de artigos médicos com analises estatísticas. Foi desconcertante verificar que nenhum dos 18 participantes respondeu corretamente.

Uma coisa a ser pensada é :porque ensinar e dar tanta importância a algo que confunde tanto?

2) Alguns mal entendidos

Significância

Eu suponho que e nossa falsa realidade e não devíamos nos apropriar da palavra "significância". Ela parece boa, importante, muito desejável pela fraternidade médica.

Se os pioneiros da estatística tivessem chamado de "improbabilidade" eu duvido que teríamos os problemas de interpretação que temos hoje.

(Dr Fisher, 2004)

Comparação com valores críticos tabelados foi arbitrário, embora razoável nos anos 1930, quando os testes estatísticos tinham que ser trabalhosamente tabelados. Asteriscos também datam de uma época que a mais avançada tecnologia em um escritório era a máquina de escrever.

E o destino dos gurus (no caso Sir Ronald Fisher) que o que ele vê como uma opção conveniente porém arbitrária vire uma lei escrita na pedra. É uma filosofia a ser abandonada.

(Allan Reese, 2004)

Nenhum modelo é melhor que os dados na qual ele se baseia. (Piantadosi,1997)

Quando não rejeitamos uma hipótese, na realidade o que ocorre e que a amostra não e suficientemente grande para rejeitar a mesma. Se aumentarmos o número de observações rejeitamos qualquer hipótese.

Todo modelo é errado, alguns são úteis. (G.E.P. Box, 1979)

Quando realizamos um ensaio clínico e testamos o tratamento A contra o tratamento B, é claro que sempre encontraremos diferença estatísticamente significante (basta ter um número grande de pacientes), já que os agentes em A e B são diferentes. O importante é saber se a diferença observada e Clinicamente Significante e não que é estatisticamente significante (para isto basta aumentar o tamanho da amostra)

Eu não sei de nenhuma disciplina além da Estatística na qual seja uma recomendação positiva para um novo livro (ou mesmo um curso) e a ser mencionado na capa, que o mesmo **não** foi escrito por um especialista. Algum leitor médico, alguma editora médica, algum estudante de medicina assistiria minha nova introdução a cirurgia do cérebro – muito mais simples e muito mais claro do que aquelas escritas por neuro-cirurgiões profissionais, com aquelas quantidades de detalhes confusos?

Eu acredito (e espero) que não.

(M.J.R Healy, 1991)

O pesquisador que buscar aconselhamento já com os dados coletados e o experimento realizado, em geral só obterá um atestado de óbito do ensaio. Nenhuma análise estatística sofisticada vai remediar uma coleta mal planejada. Isto é, o trabalho do estatístico começa bem antes da investigação se iniciar.

Eu acho altamente indesejável enviar estatísticos juniores sozinhos para um departamento cheio de médicos renomados. Eles precisam aprender antes trabalhando com outros estatísticos seniores para ganhar experiência. Só assim eles aprendem que ajuda podem melhor oferecer.

(Dr Fisher, 2004)

Experiência não se aprende, se adquire.

Estatísticos juniores devem ensinar cursos avançados e estatísticos seniores devem ensinar cursos introdutórios, porque se os estudantes começam mal eles não serão capazes de avançar.

(Sir David Cox, 2004)

Outra dificuldade para sedimentar o mercado é o fato de na maioria dos lugares não existir estatísticos seniors. .... Muito dos problemas de convencimento dos pesquisadores (médicos) em aceitar as suas sugestões (do estatístico)não é do conhecimento técnico, mas sim o de autoridade . (Wilton Bussad)

Conclusão:

Existe uma velha piada sobre quatro irmãos, com idades de 4, 5, 6 e 18 anos, que viram da janela um homem e uma mulher nus em uma cama.

O garoto de 4 anos: Vejam aquele homem e aquela mulher! Eles estão lutando.

O garoto de 5 anos: Bobo, eles estão fazendo sexo.

O garoto de 6 anos: Sim, mas muito mal.

O jovem de 18 anos: Concorda, e estava preocupado com seu casamento próximo

O garoto de 4 anos não sabia nada sobre sexo. O de 5 anos tinha chegado a um entendimento conceitual, e o de 6 anos sabia suficientemente bem sobre sexo (provavelmente sem ter experimentado), para ser um observador critico. O objetivo desta interação é tornar alguns (Clínicos) em um Estatístico de 6 anos e outros (Epidemiologistas) em um Estatístico de 18 anos.

Bioestatístico ou Epidemiologista - Alguém que não acredita que Colombo descobriu a América porque ele disse que estava procurando a Índia no ensaio original.

Significância Estatística - O oposto do Iraque : todo mundo quer ir lá, mas ninguém está certo como.

Ensaio Clinico - Um experimento que qualquer tolo pode planejar e freqüentemente planeja.

Bayesiano - aquele que esperava vagamente um cavalo (priori), dando uma rápida olhada em um burro (verossimilhança), conclui fortemente que viu uma mula (posteriori).

CAPÍTULO 2

OBSERVAÇÕES E PROBABILIDADE

Tipos de observações

Dependendo do problema, as observações são feitas em diferentes escalas, número de medições, dimensão e estrutura de dependência (4). Relacionado à escala tem-se: escala nominal (ou classificatória) que é a forma mais fraca de mensuração. Números, símbolos ou nomes são usados para classificar os objetos em celas, contando-se quantos são em cada cela. Por exemplo, descrevendo-se a cor de grãos de areia.

branco	amarelo	cinza	pode ser codificado
1	2	3	ou
45	3	20	ou
-	·	+	etc.

Escala ordinal (ou rank) é usada quando os objetos são identificados como diferentes e também colocados em uma seqüência em relação aos outros. Um exemplo seria a resposta ao efeito de anestésico onde existe uma certa ordenação.

nenhuma dor	pouca dor	muita dor	pode ser codificado
0	1	2	ou
1	50	60	ou
4	3	1	etc.

Escala intervalar é utilizada quando há uma igualdade no comprimento dos intervalos entre as classes. Isto faz com que a razão entre dois intervalos seja independente da unidade de medida utilizada e da origem, pois ambos são arbitrários. Como exemplo considera-se as medidas de temperatura em centígrados e Fahrenheit que tem diferentes origens (arbitrárias), a razão entre as diferenças de duas temperaturas (intervalos) é a mesma em ambas as escalas:

$^\circ C$	0	20	50	100
$^\circ F$	32	68	122	212

em $^\circ C \frac{100-50}{20-0} = 2,5$
em $^\circ F \frac{212-122}{68-32} = 2,5$

Escala de razão onde há uma origem não arbitrária além das propriedades da escala intervalar, neste caso a razão entre duas medidas é independente da unidade de medida. Massa, comprimento, velocidade, profundidade estão sempre nesta escala: a razão entre medidas de comprimento em centímetros e em polegadas é sempre constante e igual a 2,54. A escala de razão é a mais versátil e poderosa escala pois contém o máximo de informação.

Dois tipos diferentes de medição, atributos e variáveis foram considerados. Atributos são obtidos em escala nominal ou ordinal e tem valores discretos tais como: presença ou ausência ou cinco. Variáveis têm escala intervalar ou de razão e tem uma escala contínua de valores.

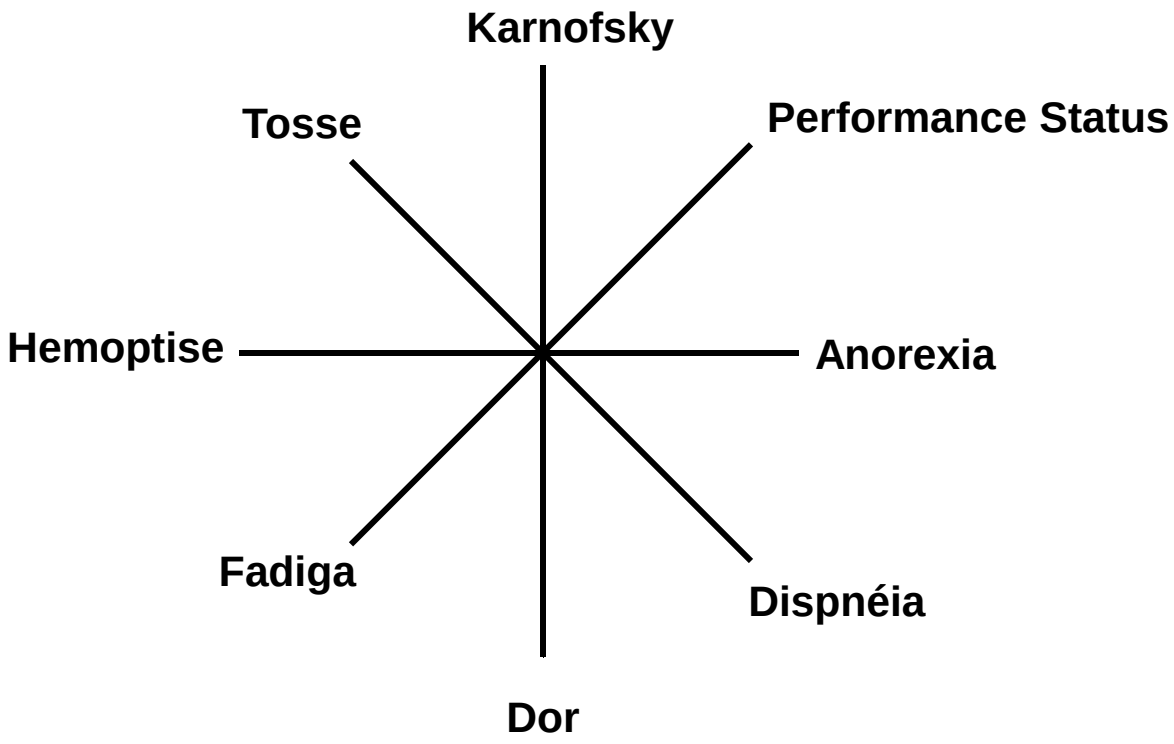
Relacionado com o número de medidas do objeto em estudo podemos ter uma observação univariada, quando tomamos uma só medida, por exemplo o peso do objeto. Alternativamente temos uma observação multivariada quando tomamos várias medidas, por exemplo: peso, altura e volume do objeto.

A dimensão da observação é importante quando tratamos de sistemas dinâmicos que evoluem em relação a outras variáveis (tempo, espaço, profundidade etc.). Teremos uma observação unidimensional quando se tem o sistema variando em relação a uma só variável, por exemplo a evolução do Produto Interno Bruto com o tempo. A observação é multidimensional quando a variação se faz em relação a mais de uma variável, por exemplo a evolução do número de casos de sarampo por semana e por municípios de um estado ou a porosidade do solo por profundidade, latitude ou longitude numa bacia petrolífera.

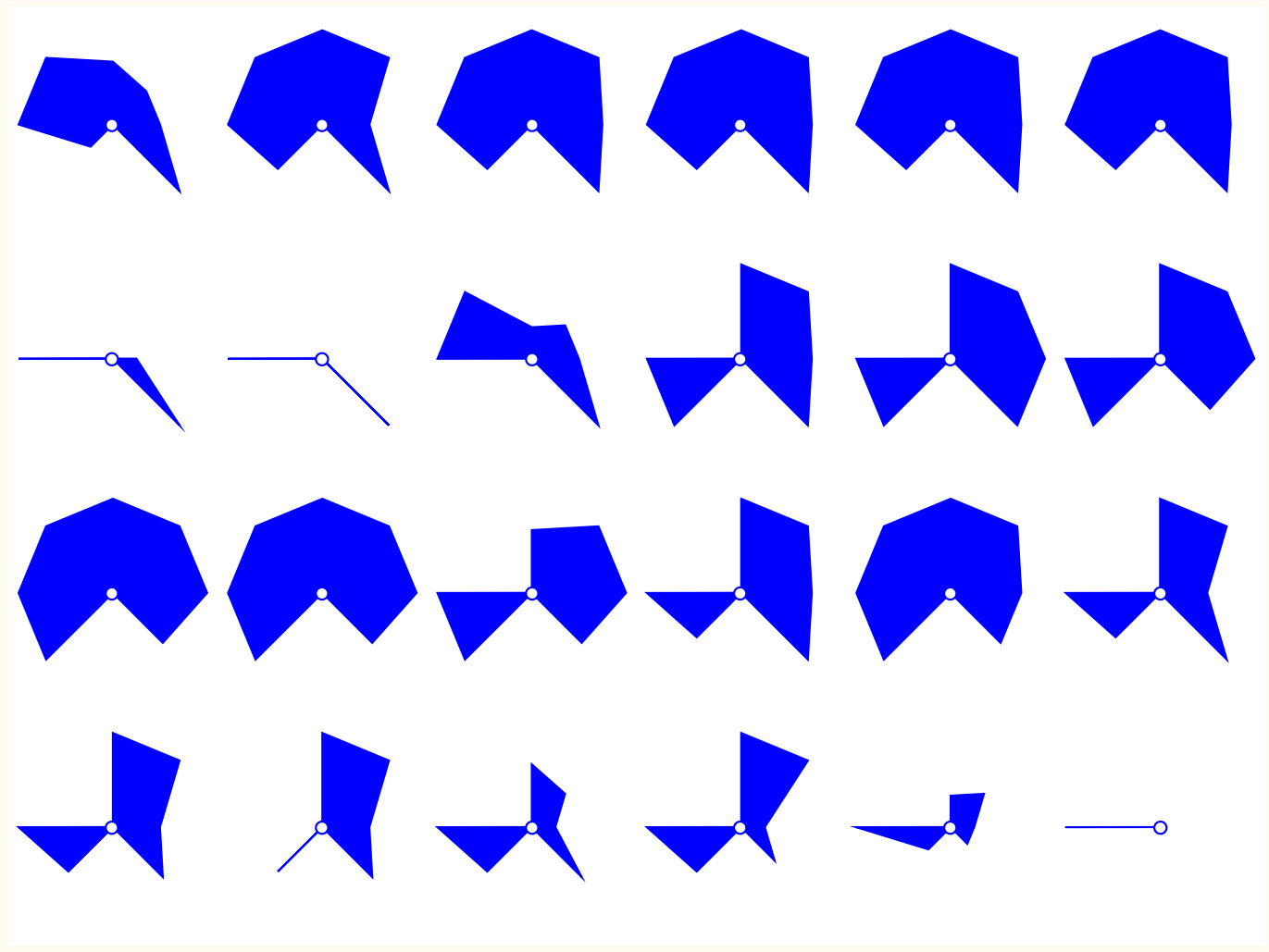
A estrutura de dependência se refere a dependência entre os objetos em estudo. As observações serão independentes se a escolha de um objeto observado não influencia a escolha dos outros objetos observados. Como exemplo temos o caso de pesquisa eleitoral quando a seleção de um eleitor para opinar não é influencida ou influencia a escolha dos outros eleitores a serem ouvidos. As observações serão dependentes quando a ordem ou posição em que foi feita a observação é também importante. Por exemplo no estudo da evolução do Produto Interno Bruto é importante não só ter o valor do mesmo nos diversos anos mas também poder identificar os instantes das observações, no estudo da evolução dos casos de sarampo é importante identificar não só o número de ocorrências como as semanas e os municípios correspondentes.

Visualização

a) Tratamento paliativo



Paciente



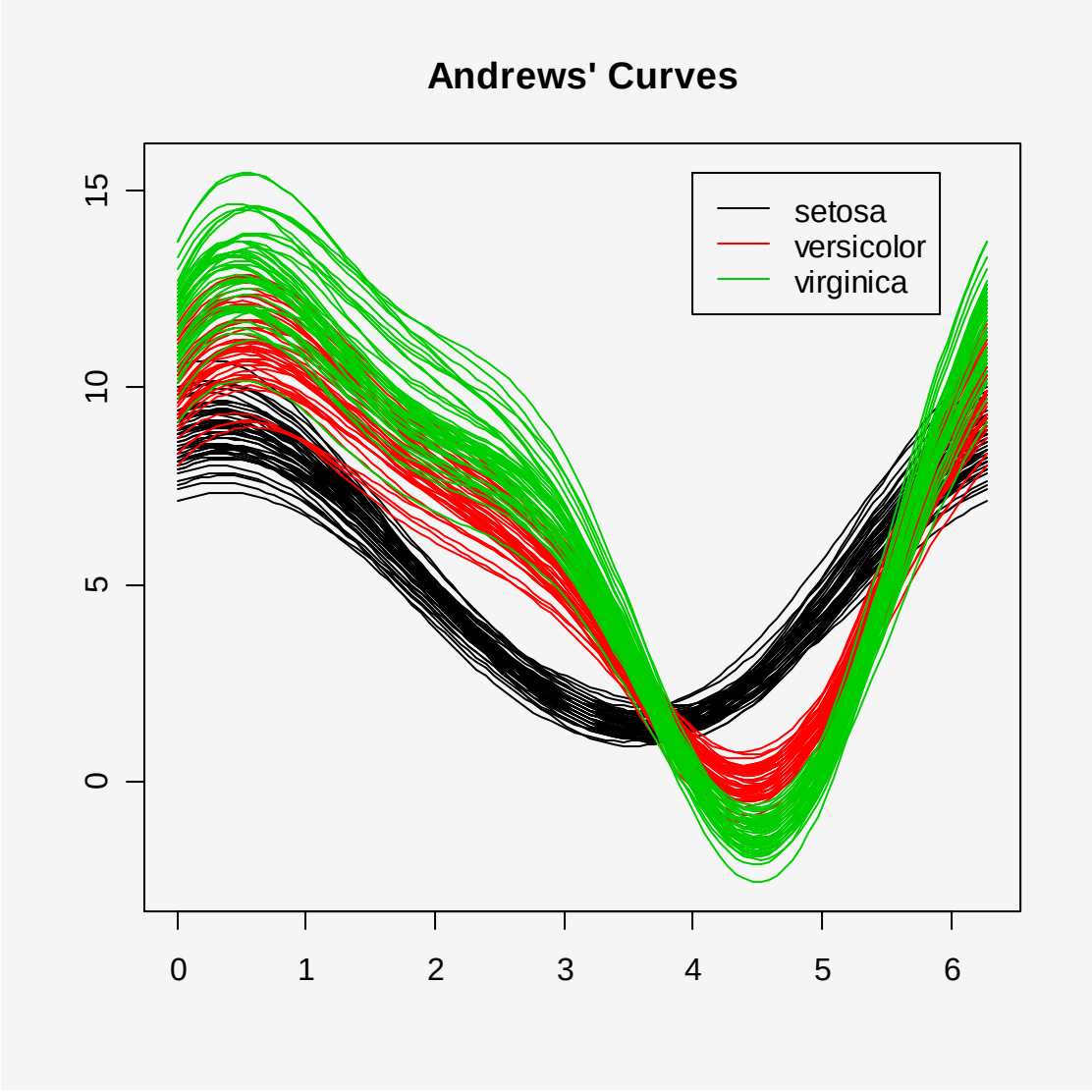
b) Andrew's Plot

Andrews' curves plot each N-dimensional point as a curved line using the function

$$f(t) = x1 / sqrt(2) + x2 \cdot \sin(t) + x3 \cdot \cos(t) + x4 \cdot \sin(2t) + x5 \cdot \cos(2t) + \ldots$$

where the n-dimensional point is X = (x1,x2,….., xn) The function is usually plotted in the interval - Pi < t < Pi. This is similar to a Fourier transform of a data point. One advantage of this visualization is that it can represent many dimensions. A disadvantage is the computational time to display each n-dimensional point for large datasets.

Iris Data


Andrews' Curves
setosa
versicolor
virginica
15
10
5
0
0
1
2
3
4
5
6

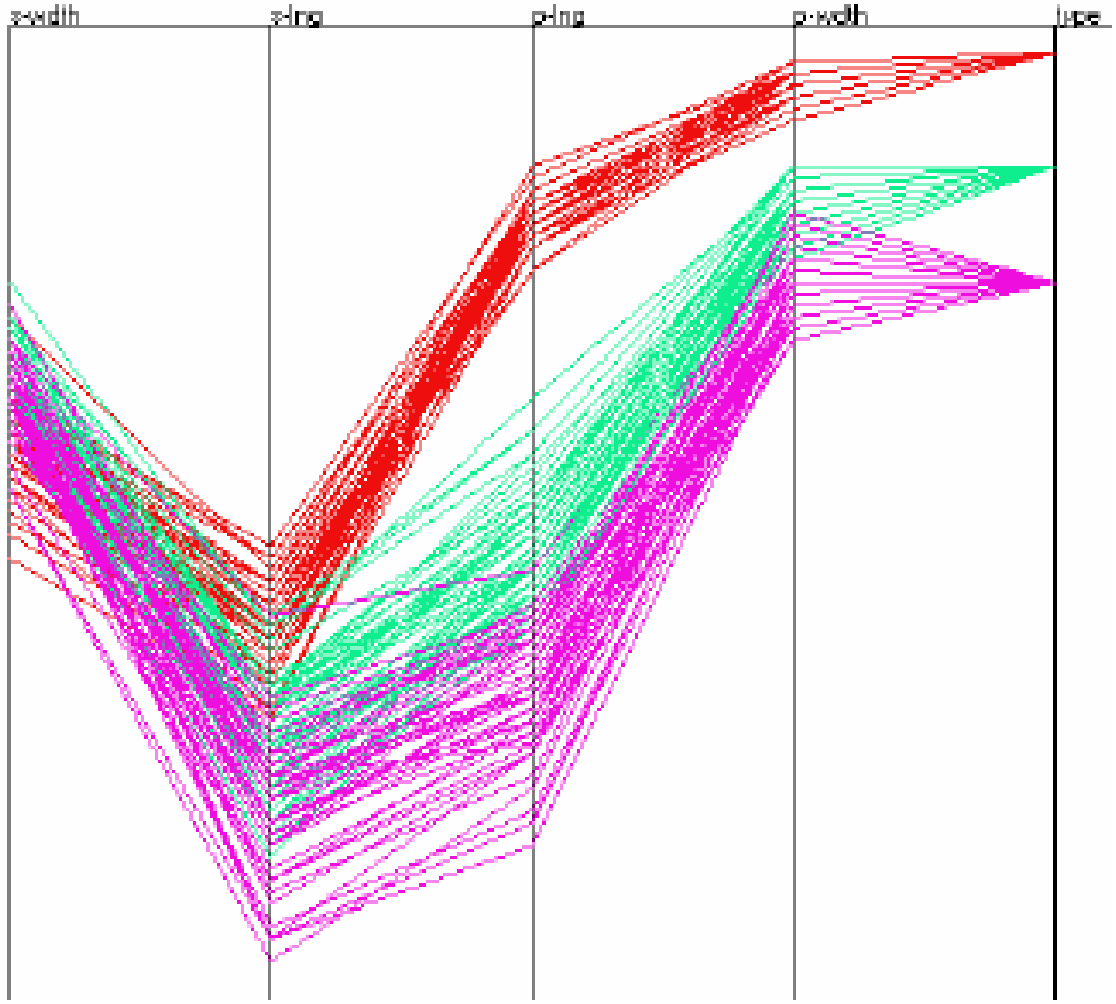
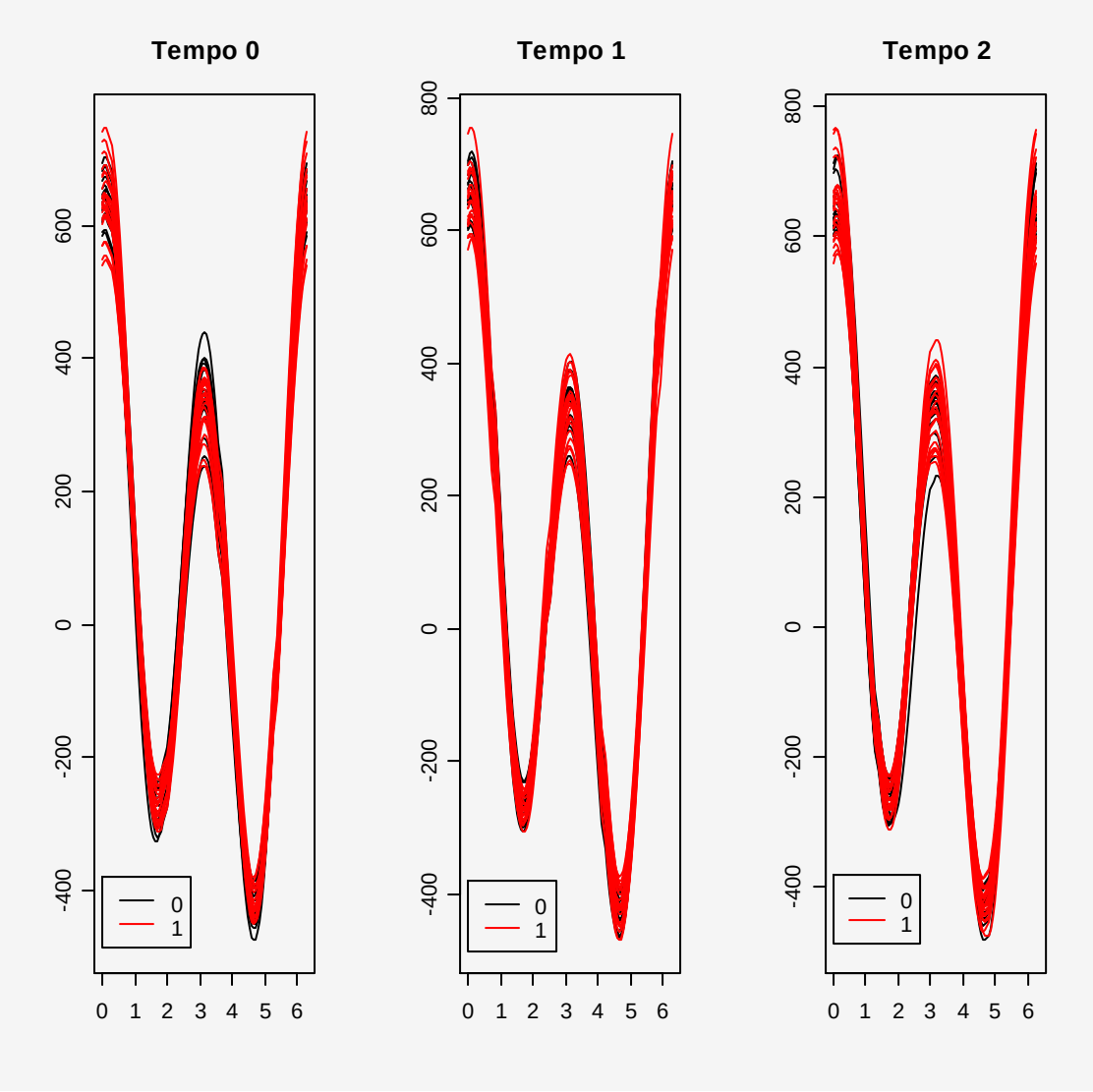


Figure 25 Parallel Coordinates of the Iris dataset (Global Normalization)

Alterações Eletrocardiográficas



0 – Não cardiopata

1 – Cardiopata

Modelos Probabilísticos

1. Introdução

A concepção atual de ciência é de aprendizado sobre um fenômeno em estudo através de:

1. observação
2. construção de um modelo para explicar ou descrever as observações
3. usar o modelo para predizer observações futuras (o comportamento futuro do fenômeno). Se as observações futuras não estão de acordo com as predições, o modelo deve ser modificado.

Os modelos podem ser descritos de forma conceitual (em linguagem corrente), física (protótipo, maquete) ou matemática. A escolha em utilizar um desses tipos de modelos é do cientista. Nada impede que o modelo seja descrito verbalmente ou usando um protótipo. A desvantagem é que a forma verbal não teria a precisão necessária no raciocínio dedutivo e o uso de um protótipo é cara, pesada, perigosa, etc.

Por outro lado a matemática é uma linguagem mais precisa e ao mesmo tempo de grande generalidade. Podemos dizer que a matemática é a linguagem da ciência e a evidência que se tem na ciência, é que de acordo com o desenvolvimento do assunto o mesmo se torna cada vez mais matemático.

Em ciência nem tudo é conhecido com certeza absoluta, a razão é que para muitas coisas da realidade, estamos ainda ignorantes. Isto não significa que não temos informação, porém nos leva a aceitação do acaso ou aleatoriedade e a utilização de modelos probabilísticos. Tais modelos além de representar as regularidades da natureza através de uma componente determinística representa a incerteza através de uma componente probabilística.

O epidemiologista utiliza esse tipo de modelo para verificar a ocorrência ou não de certos eventos na história natural da doença.

A medida de incerteza nos modelos é dada pela noção de probabilidades, cujas definições são:

Definições de Probabilidade

- Clássica

$$p = \frac{número\ de\ casos\ \ favoráveis}{número\ de\ casos\ \ possíveis}$$

Exemplo: P (nascer de sexo masculino) = $\frac{1}{2}$ = 0,5

- Frequentista – p é o valor que parece se estabilizar a freqüência de um evento quando o número de realizações do experimento aumenta.

Exemplo: P (nascer de sexo masculino) = 0,51 ou 0,52 na grande maioria de populações observadas.

- Subjetiva – p é baseado na informação que você tem sobre o evento.

Exemplo: P (habitante ser do sexo feminino) em uma cidade que você atuará como médico e você tem a informação de grande emigração de homens em busca de trabalho.

Exemplo: P (de um novo paciente em seu consultório vir a tratar da hipófise).

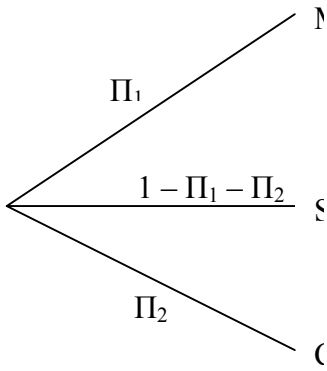
Se você é endocrinologista famoso no tratamento de hipófise, sua avaliação desta probabilidade será diferente da prevalença na população geral.

Cada definição é aplicada em diferentes contextos.

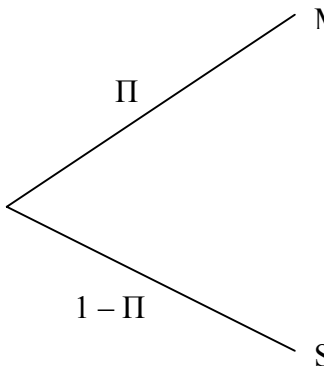
2. Modelo Binário

Suponha um estudo que acompanhou um grupo de pessoas por um período, para estudar a mortalidade por uma causa específica (exemplo: câncer). Temos três possibilidades: M – morte pela causa específica, S – sobrevida e C – censura (morte por outra causa ou abandono do estudo).

Suponha Π_1 = P(M), probabilidade de morte, Π_2 = P(C), probabilidade de censura e $1 - \Pi_1 - \Pi_2$ = P(S), probabilidade de sobrevida e/ou sobrevivência.



Se não houver censura temos o modelo binário



Uma importante alternativa para representar o modelo binário é através dos *odds* ou razão de chances de morte contra sobrevida

$$\Omega = \frac{\Pi}{1-\Pi} \text{ ou invertendo } \Pi = \frac{\Omega}{1+\Omega}$$

Observe que se a doença é rara, isto é, a probabilidade Π é pequena

$$\Omega \approx \Pi$$

Exemplos:

Se P(M) = 0,75 $\Omega = \frac{0,75}{0,25} = 3$

Se P(M) = 0,50 $\Omega = \frac{0,50}{0,50} = 1$

Se P(M) = 0,25 $\Omega = \frac{0,25}{0,75} = 0,333$

Se P(M) = 0,003 $\Omega = \frac{0,003}{0,997} = 0,003009 \approx 0,003$

Se Ω = 0,3 $\Omega = \frac{0,3}{1+0,3} = 0,23$

Se Ω = 3 $\Omega = \frac{3}{4} = 0,75$

Se Ω = 0,003 $\Omega = \frac{0,003}{1+0,003} = 0,00299 \approx 0,003$

As seguintes propriedades são observadas:

I) $0 \leq p \leq 1$
II) P(evento certo) = 1 = P(M ou S ou C)
III) P(M ou C) = P(M) + P(C) = $\Pi_1 + \Pi_2$ se os eventos são mutuamente exclusivos

3. Estimação de Parâmetros

Sem o valor de Π o modelo não serve para predição. Nosso problema mais interessante é usar os dados para estimar Π.

Parece óbvio, que se estudarmos N indivíduos por um período de tempo e M morrem e N – M sobrevivem uma estimativa Π será $\frac{M}{N}$ e a Ω por $\frac{M}{N-M}$. Além disso teremos mais confiança nas estimativas se N = 1000 do que se N = 10.

4. O Modelo é Verdadeiro?

É óbvio que um modelo que supõe que um grupo de pacientes tem a mesma probabilidade de morte, independente, de por exemplo, sexo, idade, condição de vida, etc. não representa verdadeiramente o fenômeno.

Na verdade devemos perguntar se o modelo é útil.

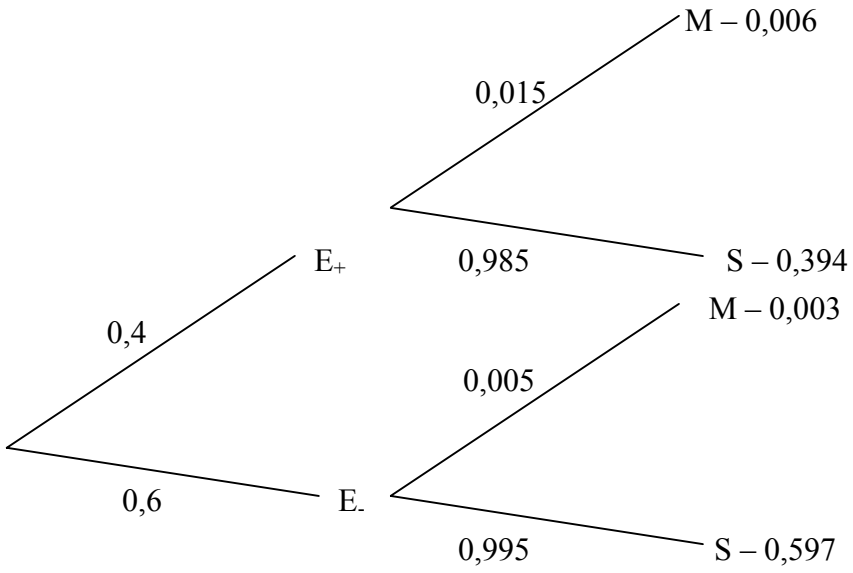
Na realidade: "Todos os modelos são errados porém alguns são úteis".

5. Probabilidade Condicional

As probabilidades de morte por câncer podem ser melhor preditas, se soubermos os hábitos de fumo dos indivíduos. Se as probabilidades são 0,015 para fumantes e 0,005 para não fumantes chamamos estas probabilidades de condicionais.

As causas ou eventos que alteram as probabilidades são chamados exposições.

Podemos representar as probabilidades condicionais às classificações: exposto (E_+) e não exposto (E_-) pela árvore.



Estamos supondo as probabilidades nos ramos conhecidos. As probabilidades das combinações de exposição (E_+, E_-) e resultado (M, S) são obtidos por multiplicação e a notação é

$P(E_+ \text{ e } M) = P(E_+)\, P(M/E_+) = 0,4 \times 0,015 = 0,006$

$P(E_+ \text{ e } S) = P(E_+)\, P(S/E_+) = 0,4 \times 0,985 = 0,394$

$P(E_- \text{ e } M) = P(E_-)\, P(M/E_-) = 0,6 \times 0,005 = 0,003$

$P(E_- \text{ e } S) = P(E_-)\, P(S/E_-) = 0,6 \times 0,995 = 0,597$

A probabilidade de morte é (independente da exposição)

$0,009 = 0,006 + 0,003 = P(E_+ \text{ e } M) + P(E_- \text{ e } M) = P(M)$

A probabilidade de não exposição é $P(E_-) = 0,6$.

A probabilidade de sobreviver é 0,597 + 0,394 = 0,991.

A probabilidade de E_- ou S é

$P(E_- \text{ ou } S) = P(E_-) + P(S) - P(E_- \text{ e } S) = 0,6 + 0,991 - 0,597 = 0,994$

Temos então mais as propriedades:

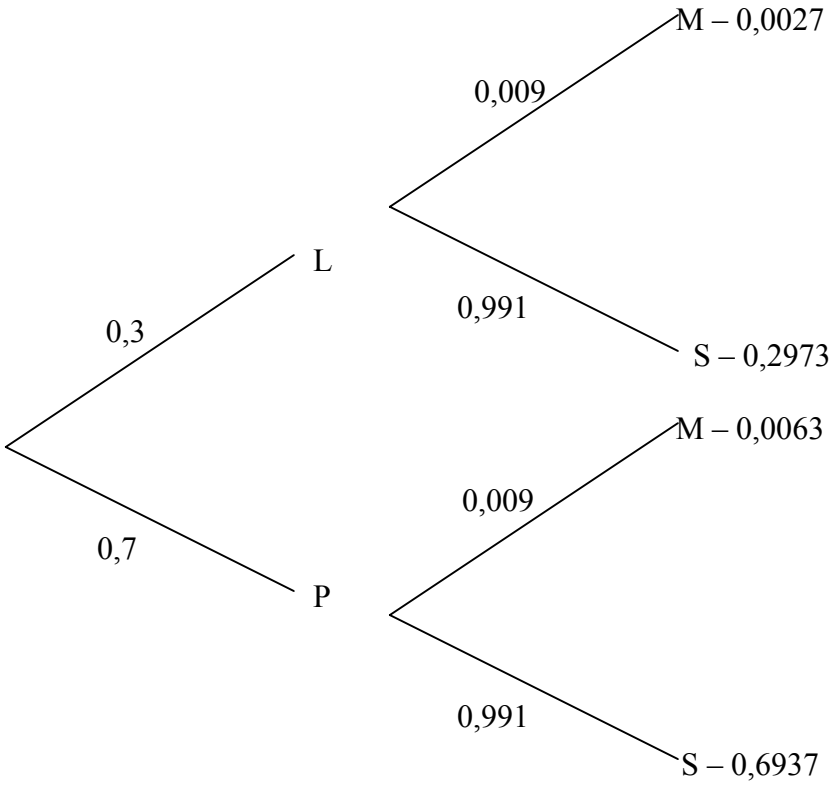
IV) P(E e M) = P(E) P(M/E) (Lei da Multiplicação)
V) P(E ou S) = P(E) + P(S) – P(E e S) se os eventos podem ocorrer juntos, isto é, não são mutuamente exclusivos (Lei de Adição)

6. Independência Estatística

No caso anterior a probabilidade de morte difere de acordo com exposição ou não ao fumo.

Se a probabilidade de morte é a mesma se o indivíduo é exposto ou não aquele fator, dizemos que a morte e o fator de exposição são estatisticamente independentes.

Por exemplo, morte por câncer e cor dos cabelos (louro, preto) teríamos



Observe que:

$0{,}0027 = P(L \text{ e } M) = P(L)\,P(M/L) = P(L)\,P(M/P) = P(L)\,P(M)$

Analogamente

$P(L \text{ e } S) = P(L)\,P(S/L) = P(L)\,P(S)$

$P(P \text{ e } M) = P(P)\,P(M/P) = P(P)\,P(M)$

$P(P \text{ e } S) = P(P)\,P(S/P) = P(P)\,P(S)$

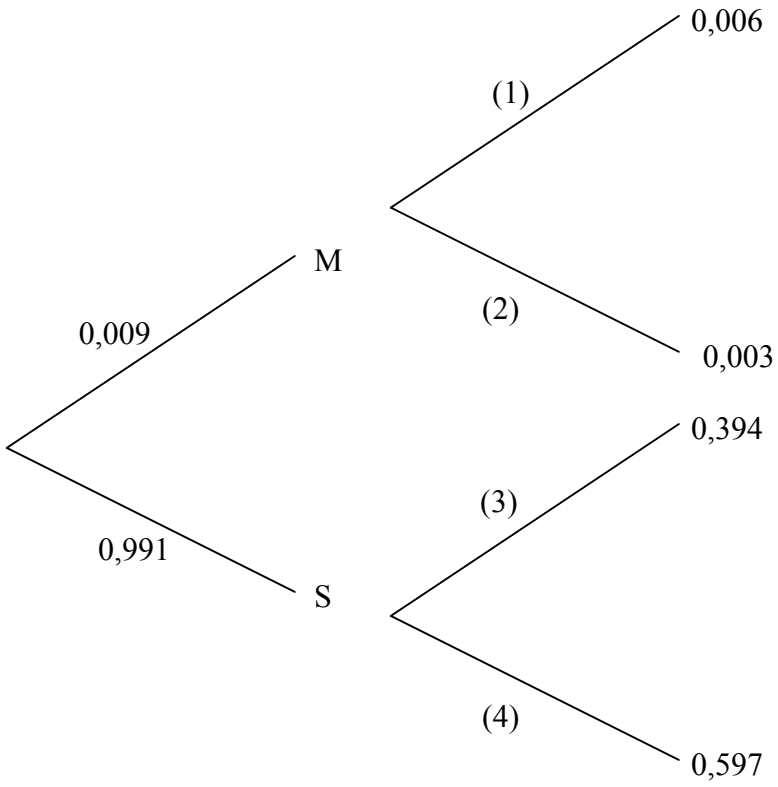
Logo a propriedade

VI) Se L e M são eventos independentes

$$P(L \text{ e } M) = P(L)\,P(M)$$
$$P(M/L) = P(M)$$
$$P(L/M) = P(L)$$

7. Mudando o Condicionamento: Regra de Bayes

A aplicação das propriedades anteriores permite mudar a direção das predições. Por exemplo, um modelo para a probabilidade de morte dado exposição, pode ser transformado em um modelo para a probabilidade de exposição dado morte.



Os ramos são obtidos por:

$$(1) \quad P(E_+ / M) = \frac{P(E_+ \ e \ M)}{P(M)} = \frac{P(E_+ \ e \ M)}{P(E_+ \ e \ M) + P(E_- \ e \ M)} = \frac{0{,}006}{0{,}009} = 0{,}667 = \frac{P(E_+)P(M / E_+)}{P(E_+)P(M / E_+) + P(E_-)P(M / E_-)}$$

Isto é, com os dados da árvore (modelo) condicionada em E_+ e E_- calculamos os dados da árvore (modelo) condicionado em M e S.

Analogamente, para os outros ramos obtemos

(2) $\frac{0,003}{0,009} = 0,33$ (3) $\frac{0,394}{0,991} = 0,3976$ (4) $\frac{0,597}{0,991} = 0,6024$

Regra de Bayes

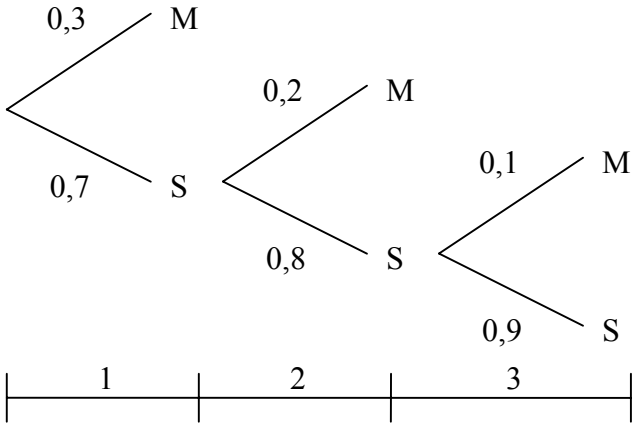
Se E_+ e E_- são eventos mutuamente exclusivos e exaustivos e M pode ocorrer com os E isto é M = (M e E_+) ou (M e E_-)

$$P(E_i / M) = \frac{P(E_i)P(M / E_i)}{P(E_+)P(M / E_+) + P(E_-)P(M / E_-)}, i = +,-$$

8. Construção de Tabelas de Vida ou Curva de Sobrevivência

a) Seqüência de modelos binários

Considere o exemplo de um estudo durante 3 anos e as mortes em cada ano



P(M no 1° ano) = 0,3

P(M no 2° ano) = 0,7 x 0,2 = 0,14

P(M no 3° ano) = 0,7 x 0,8 x 0,1 = 0,056

P(5 aos 3 anos) = 0,7 x 0,8 x 0,9 = 0,504

As probabilidades acumuladas de sobrevida são:

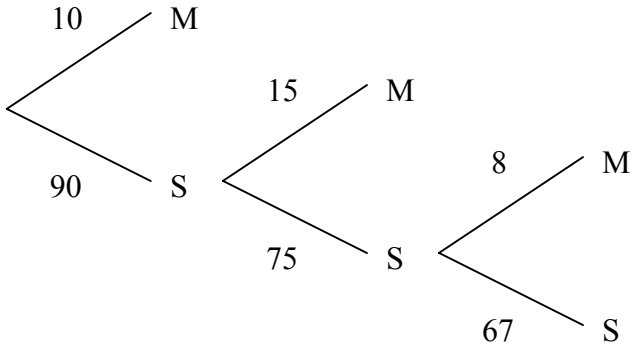
P(S ao 1° ano) = 0,7

P(S ao 2° ano) = 0,7 x 0,8

P(S ao 3° ano) = 0,7 x 0,8 x 0,9

b) Estimando as probabilidades condicionais de morte

Suponha que 100 indivíduos são acompanhados por 3 anos e que ocorreram 10 mortes no primeiro ano, 15 mortes no segundo ano e 8 mortes no terceiro ano, sobrevivendo 67 ao fim de 3 anos.



As probabilidades condicionais de morte são $\frac{10}{100}, \frac{15}{90}, \frac{8}{75}$

A probabilidade acumulada de sobrevida

P(S no 1° ano) = $\frac{90}{100}$

P(S no 2° ano) = $\frac{75}{100} = \frac{90}{100}\frac{75}{90}$

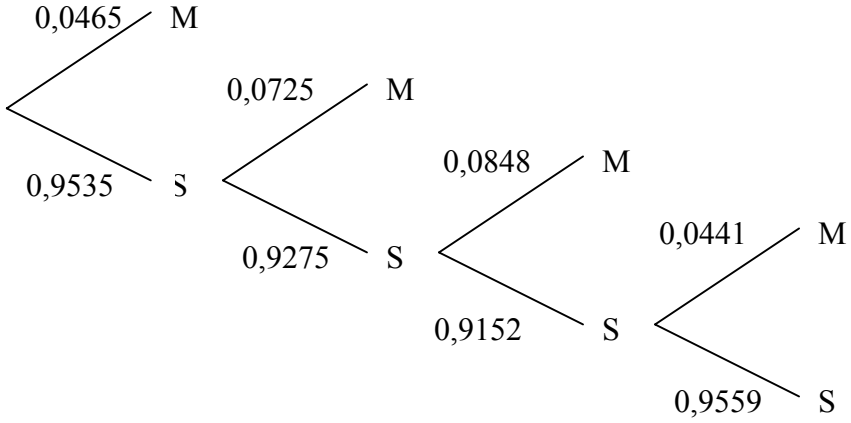
P(S no 3° ano) = $\frac{67}{100} = \frac{90}{100}\frac{75}{90}\frac{67}{75}$

c) Tabela de vida com censura

Neste caso supomos que os indivíduos censurados saíram uniformemente durante o ano ou equivalentemente permaneceram até o meio do ano.

A tabela a seguir considera a sobrevida de um grupo de mulheres com câncer no cérvix diagnosticado como estágio I (M – morte, C – censura, N – população sob risco).

Ano	N	M	C	$\frac{C}{2}$	$N' = N - \frac{C}{2}$	S(t)
1	110	5	5	2,5	107,5	0,9535
2	100	7	7	3,5	96,5	0,9275
3	86	7	7	3,5	82,5	0,9152
4	72	3	8	4,0	68,0	0,9559



d) Tempo de morte e censura exatos. Estimativa de Kaplan – Méier

Se conhecemos os tempos de morte e de censura exatos, não é necessário supor que as censuras ocorrem continuamente no intervalo.

				Probabilidade Condicional		Probabilidade Acumulada de Sobrevivência
Mês	N	M	C	M	S	
0	50	2	-	$\frac{2}{50} = 0,04$	0,96	0,96
1	48	1	-	$\frac{1}{48} = 0,0208$	0,9792	0,96 x 0,9792 = 0,94
2	47	2	-	$\frac{2}{47} = 0,0426$	0,9574	0,94 x 0,9574 = 0,90
3	45	1	1	$\frac{1}{45} = 0,0222$	0,9778	0,90 x 0,9778 = 0,88
8	43	1	-	$\frac{1}{43} = 0,0233$	0,9767	0,80 x 0,9767 = 0,86

Distribuições de Variáveis Aleatórias

1. Variáveis Aleatórias

Em geral os eventos que observamos se apresentam numericamente ou pelo menos podemos representá-los assim. Por exemplo no modelo binário poderíamos associar:

$$M \rightarrow 1$$

$$S \rightarrow 0$$

A esses valores numéricos chamamos variáveis aleatórias que podem ser discretas ou contínuas.

As leis de probabilidade relativas a esses valores numéricos são associadas tendo em vista:

i. o conhecimento da estrutura do fenômeno e portanto o que é razoável supor no caso; e

ii. escolhendo um modelo que melhor se ajusta aos dados observados sobre o fenômeno.

2. Histogramas e Funções de Densidade

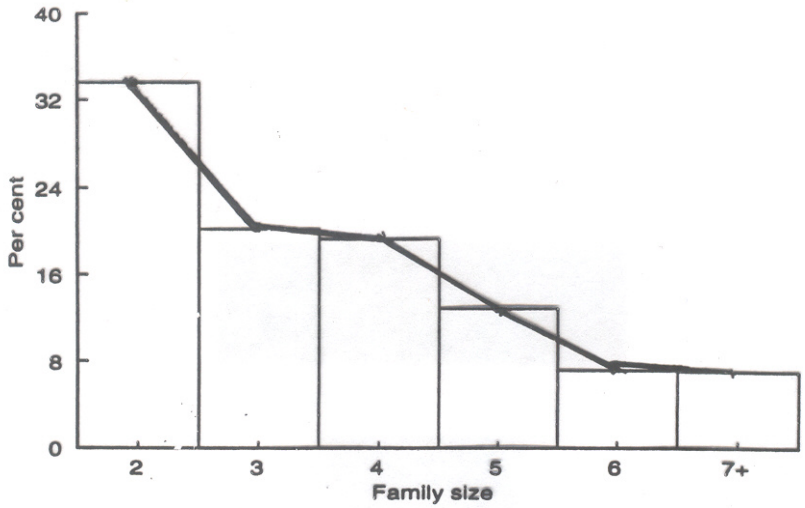
Consideremos a tabela abaixo com a distribuição das famílias dos EUA pelo seu tamanho.

Tamanho (n° de pessoas)	%
2	33,6
3	20,2
4	19,3
5	12,8
6	7,1
7 ou mais	7,0
TOTAL	100,0

Uma tabela de freqüência (relativa) fornece o número (a percentagem) de indivíduos em cada categoria ou intervalo da variável.

Um histograma é um gráfico construído de uma tabela de freqüência (relativa).

Por convenção o histograma é construído de tal forma que a frequência relativa de cada classe corresponda a área do retângulo correspondente aquela classe.



Histogram of the distribution of families by size.

Se $\prod_i$ é a frequência relativa (estimative da probabilidade) e f_i é a altura do retângulo e Δ_i o comprimento da base temos

$$\prod_i = f_i \, \Delta_i \text{ (área)}$$

Como $\prod_i$ e Δ_i são conhecidos, calculamos

$$f_i = \frac{\Pi_i}{\Delta_i}$$

A altura f_i é conhecida como densidade de probabilidade.

Se $\Delta_i = 1$ então $f_i = \prod_i$, que ocorre quando as variáveis são discretas, neste caso chamamos f_i simplesmente distribuição de probabilidades.

Frequentemente, a forma do histograma fica mais fácil de visualizar unindo os pontos médios no alto dos retângulos, criando os polígonos de freqüência ou de probabilidades.

3. Estatísticas de Resumo

É claro que as tabelas de freqüências e histogramas são resumos de protocolos obtidos de levantamentos de todas as unidades, ou de uma amostra.

Podemos ainda querer resumir ainda mais os dados através de algumas estatísticas descritivas ou de resumo. Algumas delas são:

i. M – max, o maior valor observado;

ii. m – min, o menos valor observado;

iii. M_0 – moda, o valor mais observado, isto é, com maior freqüência;

iv. $Q_2 = M_d$ – mediana, o valor que separa as observações em duas partes iguais, metade menor que M_d e metade maior que M_d;

v. Q_1 – primeiro quartil, o valor que separa as observações em duas partes, um quarto menor que Q_1 e três quartos maior que Q_1;

vi. Q_3 – terceiro quartil, o valor que separa as observações em duas partes, três quartos menor que Q_3 e um quarto maior que Q_3;

vii. $\overline{X}$ – média amostral ou empírica:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i ;$$

viii. S^2 – variância amostral ou empírica e S – desvio padrão

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \text{ e } S = DP = \sqrt{S^2} ;$$

ix. R – amplitude total

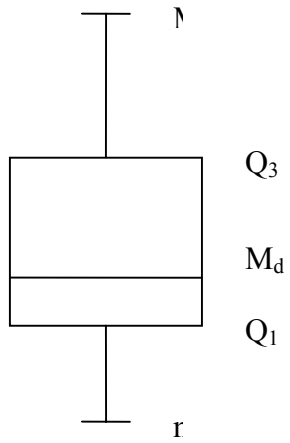
$$R = M - m; \text{ e}$$

x. R_I – amplitude interquartilica

$$R_I = Q_3 - Q_1$$

Note, por exemplo, que M_0, M_d e $\overline{X}$ são medidas de posição ou locação, enquanto S^2 e R são medidas de variação.

Diversas outras medidas de variação e forma podem ser construídas a partir destas medidas básicas, inclusive gráficos com o Box-Plot:



Vamos nos concentrar apenas na média e variância.

4. Funções de Densidades Teóricas

Vimos que o histograma resume as observações sobre um fenômeno (processo gerador dos dados).

Vamos agora descrever algumas funções de densidade com pequeno número de parâmetros que possam representar adequadamente os dados. Isto é, vamos criar um catálogo de funções e suas propriedades e que possam adequadamente representar nossos dados e seja um candidato a representar o processo gerador dos dados. Ou ainda, desejamos uma curva que seja próxima do polígono de freqüência.

Observe que a média e a variância empíricas terão correspondentes teóricos nestas curvas.

5. Distribuições para Contagem

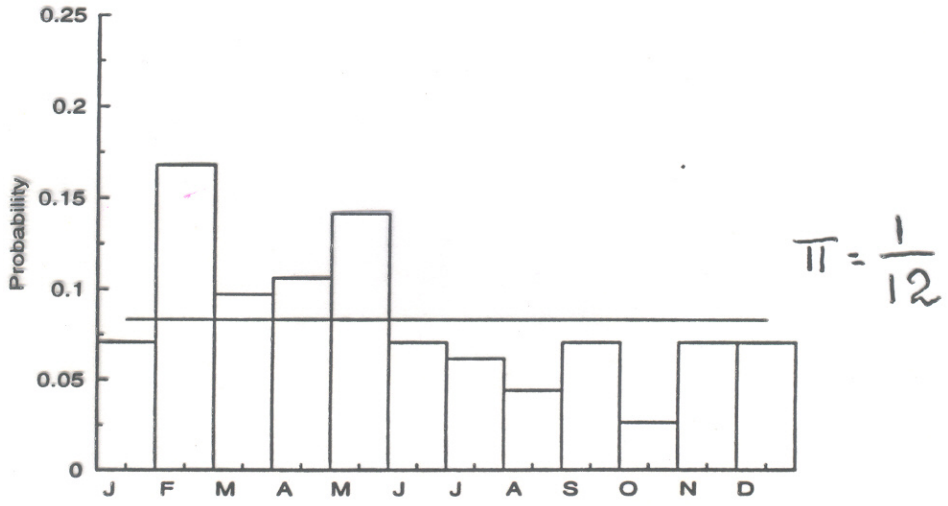
i. Distribuição Uniforme

É uma distribuição em que

$$\Pi_i = \frac{1}{I}, \quad i = 1,\text{K}\ ,I$$

Exemplo: Número de crianças nascidas com certa doença em cada mês.

MÊS	NÚMERO
Janeiro	8
Fevereiro	19
Março	11
Abril	12
Maio	16
Junho	8
Julho	7
Agosto	5
Setembro	8
Outubro	3
Novembro	8
Dezembro	8
TOTAL	113



Histogram and uniform density function for the illness of children over 12 months of the year.

ii. Distribuição Binomial

$$\Pi_i = f_i(y_i, p) = \binom{n}{y_i} p^{y_i}(1-p)^{y_i} \quad 0 \le p \le 1 \quad y_i = 0,1,\text{K}\ ,n$$

a. média = np variância = np (1-p)

b. Estrutura

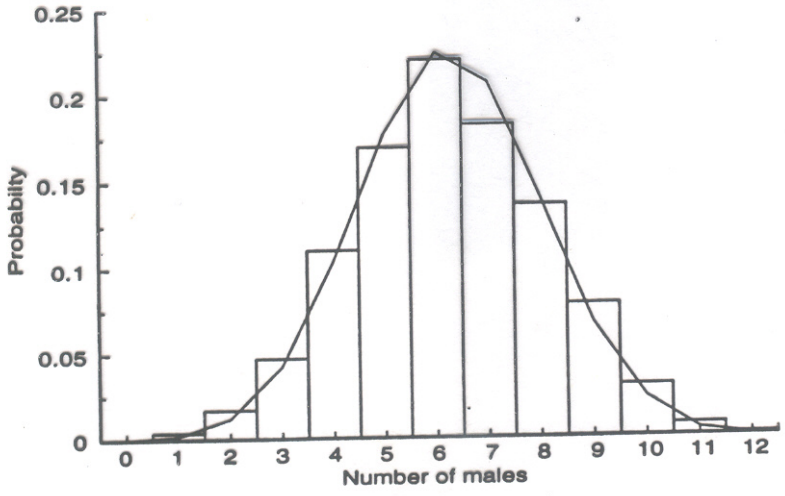
Fixamos n experimentos e verificamos o número y de sucessos onde a probabilidade de sucesso é constante e igual a p em cada realização do experimento e os experimentos são independentes.

Exemplo: Supondo que a probabilidade de nascimento do sexo masculino, e que as probabilidades de nascimento de cada sexo em uma família é constante (dúvida?).

Para famílias de tamanho 12 as probabilidades de nascimento do sexo masculino serão

$$\Pi(y) = \binom{12}{y}\left(\frac{1}{2}\right)^{y}\left(\frac{1}{2}\right)^{12-y} \qquad y = 0,1,\text{K}\ ,12$$

Observou-se 6115 famílias de tamanho 12 na Saxonia obtendo-se $p \approx 0,519$ e as figuras abaixo (a qualidade do ajustamento deve ser julgado!!).



Histogram and binomial density function for the number of male children in families of 12.

iv. Distribuição Geométrica ou de tempo de espera

$$\Pi_i = f(y_i, p) = p(1-p)^{y_i - 1} \quad 0 \le p \le 1 \quad y_i = 0,1,\text{K}$$

a. média = $\frac{1-p}{p}$ variância = $\frac{1-p}{p^2}$

b. Estrutura:

É o número de experimentos até a ocorrência do primeiro sucesso, onde a probabilidade de sucesso é constante em cada realização e as realizações são independentes entre si.

É uma distribuição sem memória, isto é, como as probabilidades são constantes não importa quantos experimentos (tempo) já se realizaram. Como as probabilidades constantes não importa quantos experimentos (tempo) já se realizaram.

Exemplo: Número de dias que um hospital leva até ocupar todos os leitos disponíveis.

Dias de Espera	Frquência	%	Geométrica p = 0,658
0	678	0,671	0,658
1	227	0,225	0,225
2	56	0,055	0,077
3	28	0,028	0,026
4	8	0,008	0,009
5+	14	0,014	0,005

p = 0,658 $\overline{X} = 0,519$

iii. Distribuição de Poisson

$$\Pi_i = f_i(y_i, \lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \quad y_i = 0,1,\text{K}$$

a. média = λ variância = λ

b. Estrutura:

Ocorre nas condições da binomial com n muito grande e p muito pequeno, isto é, a distribuição de probabilidades de eventos raros. Alternativamente é a probabilidade de ocorrência de eventos aleatoriamente distribuídos em intervalos de tempo ou espaço.

Exemplo: Número de pacientes que chegam a dois hospitais I e II para operação de coração.

NÚMERO DE PACIENTES	NÚMERO DE DIAS	
	I	II
0	13	31
1	31	90
2	40	30
3	31	12
4	18	4
5 ou mais	22	4
TOTAL	155	182
Média	2,56	1,23
Variância	2,63	1,25

Como as médias e variâncias empíricas são iguais em cada hospital a distribuição de Poisson deve ser adequada.

6. Distribuições de Variáveis Contínuas

i. Distribuição Normal

$$\Pi_i = f\left(y_i, \mu, \sigma^2\right)\Delta_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)} \Delta_i \quad -\infty < y < \infty$$

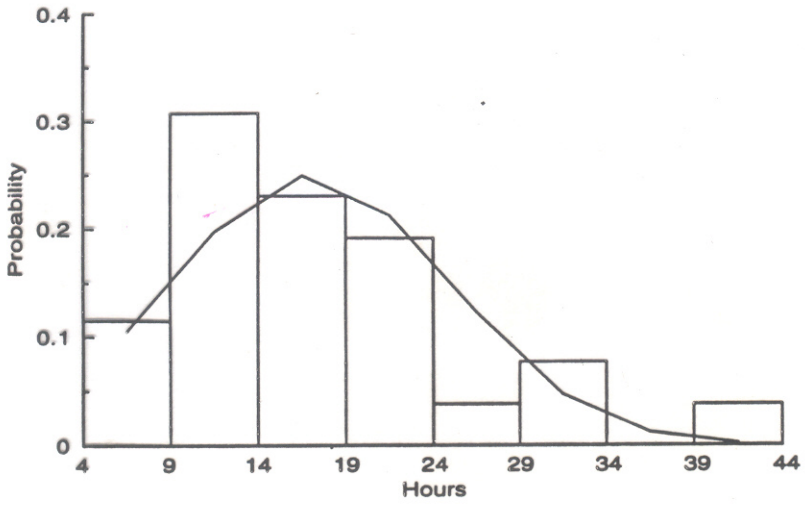
a. média = μ variância = σ^2

b. Estrutura:

Ocorre para fenômenos causados pela soma de um número grande de fatores, cada um com aproximadamente a mesma importância, e independência entre eles (Teorema Central do Limite).

Exemplo: Número de horas para se recuperar de uma operação.

Hours	Students	Multinomial	Normal	Residual
5-9	3	0,115	0,105	0,159
10-14	8	0,308	0,198	1,260
15-19	6	0,231	0,250	-0,199
20-24	5	0,192	0,213	-0,233
25-29	1	0,038	0,123	-1,225
30-34	2	0,077	0,047	0,691
35-39	0	0,000	0,012	-0,567
40-44	1	0,038	0,002	3,974



ii. Distribuição Log Normal

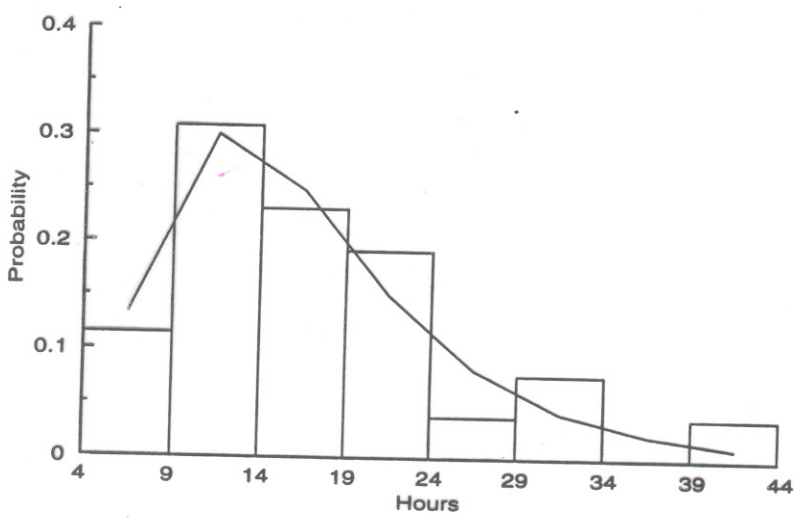
$$\Pi_i = f\left(y_i, \mu, \sigma^2\right)\Delta_i \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log y_i - \mu)^2} \Delta_i \quad y \geq 0$$

a. média = $e^{\mu+\frac{\sigma^2}{2}}$ variância = $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$

b. Estrutura:

A variável log(y) tem distribuição normal. Ocorre para fenômenos causados pelo produto de um número grande de fatores cada um com aproximadamente a mesma importância e independência entre eles.

Hours	Students	Multinomial	Log Normal	Residual
5-9	3	0,115	0,134	0,262
10-14	8	0,308	0,300	0,070
15-19	6	0,231	0,248	-0,179
20-24	5	0,192	0,151	0,547
25-29	1	0,038	0,082	-0,769
30-34	2	0,077	0,042	0,859
35-39	0	0,000	0,022	-0,752
40-44	1	0,038	0,011	1,310



iii. Distribuição Exponencial

$$\Pi_i = f(y_i, \lambda)\Delta_i = \lambda e^{-\lambda y_i}\Delta_i \quad y > 0$$

a. média = $\frac{1}{\lambda}$ variância = $\frac{1}{\lambda^2}$

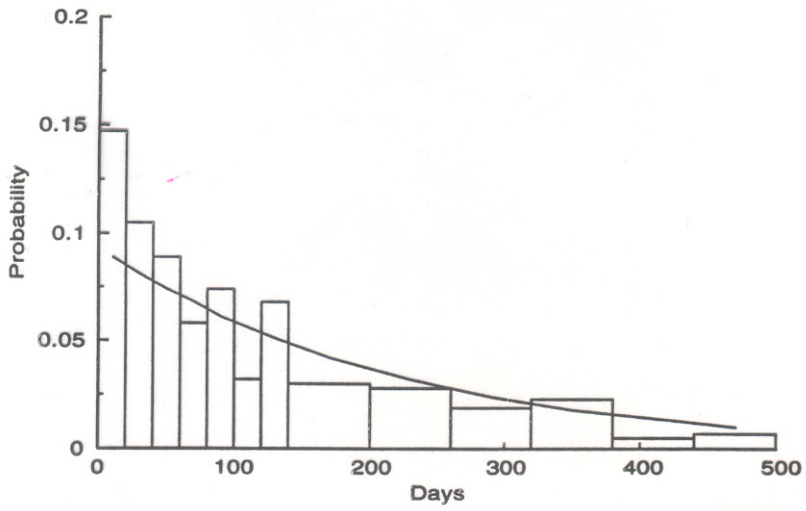
b. Estrutura:

É uma distribuição sem memória. Útil para tempo de vida onde não há envelhecimento.

Exemplo: Duração em dias entre desastres em mina de carvão no Reino Unido.

$\overline{X} = 213{,}4 dias$

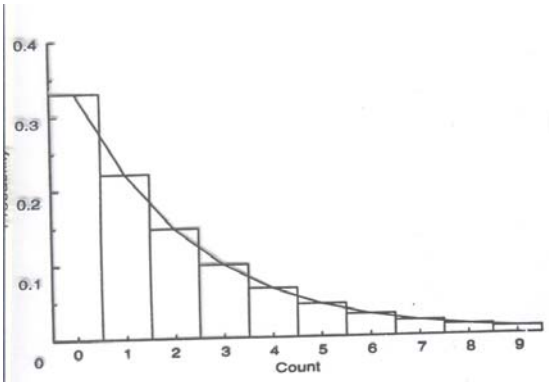
Days	Disasters	Multinominal	Exponential
0-20	28	0,147	0,089
20-40	20	0,105	0,081
40-60	17	0,089	0,074
60-80	11	0,058	0,068
80-100	14	0,074	0,061
100-120	6	0,032	0,056
120-140	13	0,068	0,051
140-200	17	0,089	0,127
200-260	16	0,084	0,096
260-320	11	0,058	0,072
320-380	13	0,068	0,055
380-440	3	0,016	0,041
440-500	4	0,021	0,031
>500	17	0,089	0,098



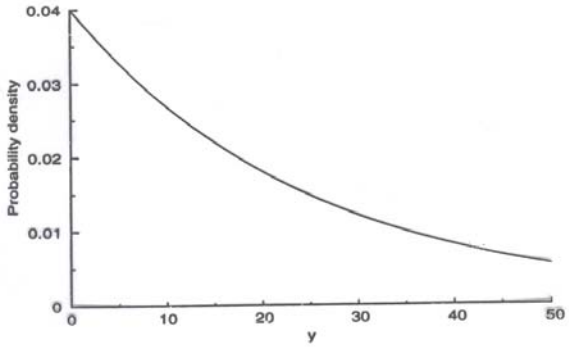
Histogram and exponential density function for the times between mine disasters.

Time	1851-1891			1891-1962		
	Disasters	Multi.	Exp.	Disasters	Multi.	Exp.
0-20	22	0,176	0,157	6	0,092	0,049
20-40	17	0,136	0,132	3	0,046	0,047
40-60	13	0,104	0,111	4	0,062	0,044
60-80	8	0,064	0,094	3	0,046	0,042
80-100	14	0,112	0,079	0	0,000	0,040
100-120	6	0,048	0,067	0	0,000	0,038
120-140	9	0,072	0,056	4	0,062	0,036
140-200	12	0,096	0,120	5	0,077	0,098
200-260	11	0,088	0,072	5	0,077	0,085
260-320	4	0,032	0,043	7	0,108	0,073
320-380	4	0,032	0,026	9	0,138	0,063
380-440	2	0,016	0,015	1	0,015	0,054
440-500	0	0,000	0,009	4	0,062	0,046
>500	2	0,016	0,018	15	0,231	0,286

$$= \frac{\alpha \delta^{\alpha}}{y_i^{\alpha+1}} \Delta_i \quad \alpha > 0 \quad y_i \geq \delta > 0$$



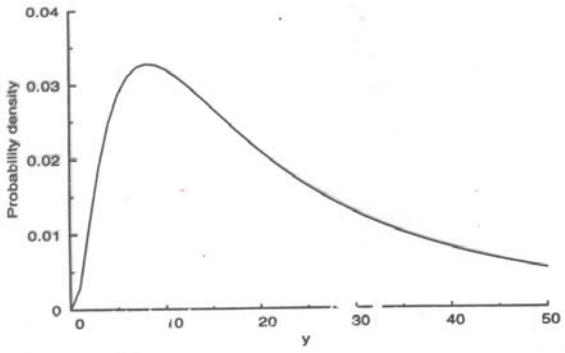
Geometric distribution for $\nu_1 = 0.33$.



Exponential distribution with $\mu = 25$.

$$\Pr(y_i + y_0 | y_0) = \Pr(y_i)$$

This is called the *Markov property.*



Log normal distribution for $\mu = 3$ and $\sigma^2 = 0.9$.

$$VPN = \frac{(1 - prevalência) \times especificidade}{(1 - prevalência) \times especificidade + prevalência \times (1 - sensibilidade)}$$

$$VPN = (1 - prevalência) \times \frac{especificidade}{negatividade\,do\,teste}$$

$$VPN \propto (1 - prevalência) \times especificidade$$

Em termos de inferência as expressões em destaque se traduzem como:

$$\text{Posteriori} \propto \text{priori} \times \text{verossimilhança}$$

Isto é, o teorema de Bayes transforma a crença a priori (distribuição a priori, prevalência antes do teste) através da verossimilhança (dados, resultado do teste) em crença a posteriori (distribuição a posteriori, prevalência após resultado do teste).

Consideramos o problema da seção anterior em que tínhamos 10 pacientes para receber a droga A. Suponha que as crenças a priori (antes de aplicar a droga) sobre a eficácia - π fossem:

Distribuição a priori de π:

π	0,4	0,5	0,6	0,7
P(π)	1/6	2/6	2/6	1/6

Após aplicar a droga verificamos que 7 pacientes ficaram curados. Com os valores da tabela e a fórmula para as posteriores os cálculos das crenças a posteriores pode ser vistas na tabela a seguir.

Cálculos Bayesianos

π	Priori-p(π)	Verossimilhança	Priori x Verossimilhança	Posteriori p(π/y=7)
0,4	1/6 = 0,167	0,043	0,007	0,007/0,163 = 0,043
0,5	2/6 = 0,333	0,117	0,039	0,039/0,163 = 0,239
0,6	2/6 = 0,333	0,215	0,072	0,072/0,163 = 0,442
0,7	1/6 = 0,167	0,267	0,045	0,045/0,163 = 0,276
-	1,000	-	0,163	1,000

Vemos então como a informação a priori é modificada pela informação dos dados, y = 7, tendo havido um deslocamento da distribuição a posteriori, para a direita, em relação a distribuição a priori.

Uma estimativa Bayesiana (de máxima probabilidade a posteriori) seria $\hat{\pi} = 0,6$, que contrasta com o estimador clássico, neste caso $\hat{\pi} = 7/10 = 0,7$.

A hipótese H_0: $\pi = 0,4$ seria rejeitada, pois tem probabilidade a posteriori muito baixa: $p(\pi = 0,4/y = 7) = 0,043$.

Em geral, se $(y_1,...,y_n)$ são observações independentes de uma densidade de probabilidade $f(y/\theta)$, onde θ é um parâmetro com valores contínuos e $g(\theta)$ é a função de densidade a priori de θ, a regra de Bayes fornece

$$f(\theta/y) \propto g(\theta)\, f(y/\theta),$$

que servirá de base para inferências sobre θ. Por exemplo Estimadores Bayesianos, Intervalos Bayesianos e Testes Bayesianos são obtidos a partir da densidade a posteriori que resume a informação dos dados e as informações a priori.

CAPÍTULO 3

INFERÊNCIA ESTATÍSTICA

FREQUENTISTA OU CLÁSSICA

Estimação

Estimação: conjunto de procedimentos que permitem obter dos dados uma aproximação (bem como uma medida da qualidade da aproximação) para uma quantidade de interesse cujo valor é desconhecido, denominado **parâmetro** e denotado genericamente por θ.

Estimação

- **Estatística** é qualquer função dos dados amostrais. Se for utilizada como aproximação de um valor desconhecido é chamada **estimador**.

- O valor numérico do estimador obtido de uma amostra é chamado de **estimativa**.

- No problema do epidemiologista consideramos os estimadores **ee** e **em** e as suas estimativas 512 e 559 .

Distribuição amostral

Se retirarmos diversas amostras de mesmo tamanho de uma população, para cada amostra teremos um valor para o estimador.

Esses valores têm uma média, variância, mediana, etc. e uma distribuição.

O desvio padrão desses valores chama-se **Erro Padrão** (da Estimativa) e a sua distribuição chama-se **Distribuição Amostral** (do Estimador).

No problema do epidemiologista, a distribuição das caras e coroas é uma distribuição amostral .

Para ilustrar, considere-se o caso da média. A Figura 1 apresenta a distribuição amostral da média amostral $\bar{X}=\sum_{i=1}^{n} X_i/n$, para n = 2, n = 5 e n = 10 de diferentes populações. Observe que para n = 10 a distribuição de assemelha-se à distribuição normal, ilustrando um forte efeito do Teorema Central do Limite, que prova que se espera obter uma distribuição normal sempre que a variação nos dados for devida a soma de efeitos independentes e que nenhum deles é predominante.

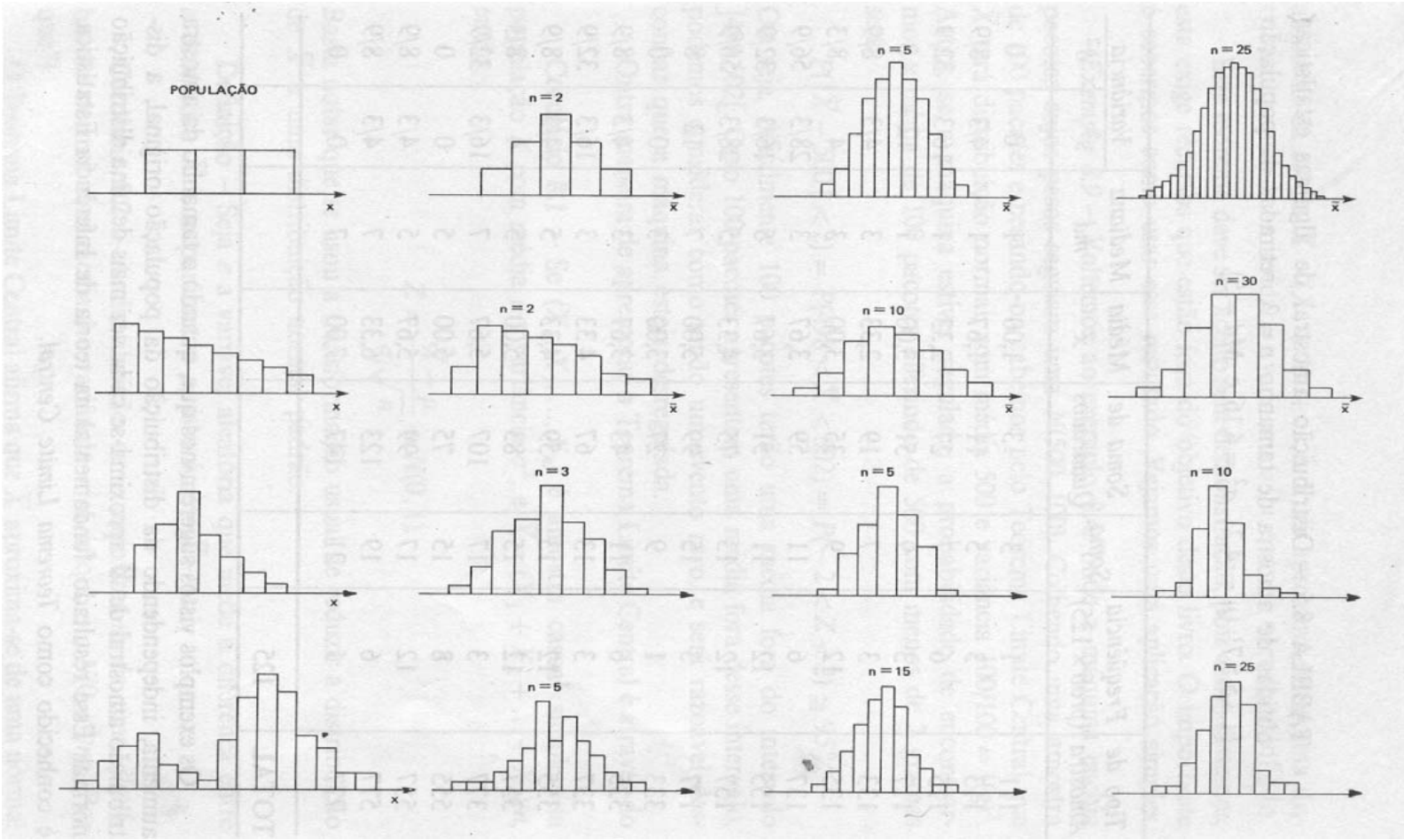
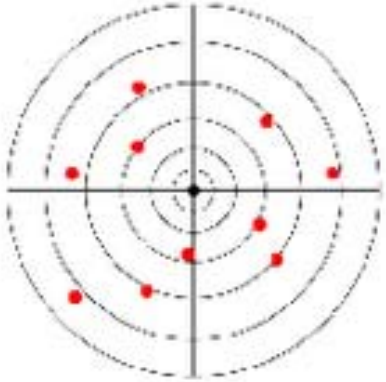


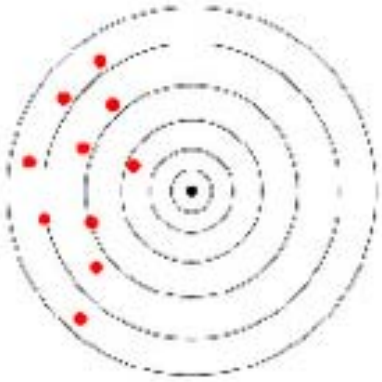
Figura – Histogramas correspondentes à distribuição amostral de para algumas populações

Propriedades dos estimadores

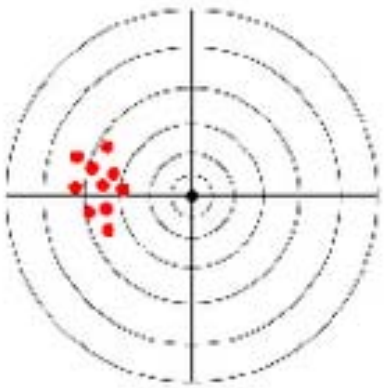
- não viciado ou não tendencioso ;
- preciso ou de pequena variação;
- consistente
- eficiente.



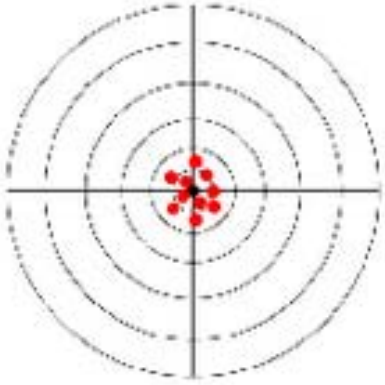
Não tendencioso
e impreciso



Tendencioso
e impreciso



Tendencioso
e preciso



Não tendencioso
e preciso

Estimador Não Tendencioso

- Também chamado de não viciado

- Fornece uma estimativa em torno do valor verdadeiro do parâmetro, sem uma tendência de erro em uma direção especifica.

Estimador Preciso

- Quando a estimativa tem uma pequena variação

- Ou seja tem um pequeno erro padrão

Estimador Consistente

- O estimador é **consistente** quando suas estimativas se aproximam do valor verdadeiro que se quer estimar, à medida que a amostra cresce.

Estimador Eficiente

- Quando comparamos dois estimadores não tendenciosos

- Um é dito mais **eficiente** que outro quando seu erro padrão for menor que o erro padrão do outro

Exemplo

- Considere a amostra da altura de 25 pacientes retirados de uma população com altura média de 1,7 m e 4 cm de variância:

- {1,67; 1,62; 1,74; 1,68; 1,63; 1,70; 1,64; 1,63; 1,65; 1,75; 1,72; 1,64; 1,66; 1,68; 1,71; 1,68; 1,71; 1,64; 1,72; 1,64; 1,74; 1,72; 1,69; 1,69; 1,65}

Estimador

- Neste caso

$$\overline{X} = \frac{1}{25}\sum_{i=1}^{25} x_i$$

•É um estimador preciso pois

$$\overline{X} = \hat{\mu} = 1{,}7$$

com erro padrão de:

$$E\hat{P} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{2}{\sqrt{25}} = \frac{2}{5}$$

•Note que o estimador é consistente pois quanto maior o N menor o EP

Testes Estatísticos

Em testes estatísticos, copiamos a estratégia matemática de provar por contradição. Começando com uma hipótese H_o que se quer rejeitar, supomos que H_o é verdadeiro. Desenvolvendo argumentos de forma correta, se chegamos a uma contradição, então, a hipótese H_o deve ser falsa. Em estatística, copiamos este enfoque, mas em vez de atingir uma contradição, observamos um resultado improvável. Especificamente, começando com uma hipótese nula (por exemplo, que não existe diferença entre dois tratamentos), observamos o resultado de um experimento bem planejado. A seguir, verificamos quão provável é o resultado observado no estudo, supondo não haver diferença entre os tratamentos. Se calculamos através de um procedimento de teste estatístico que o resultado do estudo é improvável, temos então duas alternativas: 1) não há diferença entre os tratamentos, e o que ocorreu foi um resultado muito improvável; 2) há diferença entre os tratamentos (isto é, a premissa inicial era falsa) e o que ocorreu foi um evento muito provável. A decisão mais sensata é considerar a segunda alternativa como a verdadeira.

A distinção entre *teste de hipótese e teste de significância* é que no primeiro especificamos, além de uma hipótese nula, uma hipótese alternativa de interesse específico; no segundo, somente a hipótese nula é de interesse.

Apresentamos a seguir alguns comentários ,que achamos importantes e nem sempre são bem explicitados em livros textos

Testes Estatístico

Basicamente em um teste estatístico supomos que a hipótese (a ser testada) é verdadeira e obtemos a probabilidade (valor-p) da amostra selecionada . Se esta probabilidade é pequena , temos duas possibilidades : a) A hipótese é verdadeira e fomos muito azarados e tiramos uma amostra muito estranha , ou b) A hipótese não é verdadeira . Obviamente a segunda alternativa é mais razoável.

Portanto o valor-p não é a probabilidade da hipótese ser verdadeira P(H verdadeira) mas sim a improbabilidade da mesma isto é P(Amostra observada supondo a hipótese verdadeira).

Logo se os fundadores das estatística frequentista ou clássica (Sir Ronald Fisher,Jersy Neyman e Egon Pearson) tivessem chamado o valor pequeno desta probabilidade de "improbabilidade" em vez de "significante" , ele seria muito menos mal interpretado.

Outro detalhe a esclarecer, é o costume, quase um tabu de se fixar o valo-p em 0.05 (5%) para definir pequeno ou grande , sem levar em conta o problema especifico.

Como exemplo , não se entraria em um avião se a probabilidade de acidente for grande 0.051 (5,1) , mas entraríamos no avião se esta probabilidade for 0.049 (4.9%).

Sir Ronald Fisher sugeriu este valor numa época, anos 1920, em que a ferramenta mais moderna em um laboratório era uma maquina de escrever

Intervalos de confiança

Na estatística clássica para um intervalo de confiança supomos que se tirássemos um numero grande de amostras de mesmo tamanho n e para cada amostra calculássemos um intervalo, por exemplo de 95% de confiança para um parâmetro desconhecido (p. ex. a média), 95% destes intervalos conteriam o o parâmetro desconhecido .

Como observamos apenas uma amostra seriamos muito azarados de ter tirado uma amostra entre as 5% que não contem o parâmetro. Logo é mais razoável supor ou termos confiança que tiramos uma entre as 95% amostras que contém o parâmetro.

Observe que 95% não é a Pr{parâmetro estar no intervalo} mas sim a confiança que temos do intervalo conter o parâmetro.

Relação entre testes e intervalos

Vemos então que testes e intervalos de confiança são relacionados, podemos dizer que um é o dual ao outro. Um intervalo de confiança contém todos os valores do parâmetro que seriam aceitos em um teste.

Finalmente segundo Sr David Cox , devemos distinguir Teste de Significância (Sir Ronald Fisher) que somente considera a hipótese sob teste ao passo que Teste de Hipótese (Neyman-Pearson) considera não só hipótese sob teste mas também uma hipótese alternativa.

- Teste de hipóteses

  - ✓ Hipótese nula

    Exemplo:

    H_0: não há diferença entre tratamentos

  - ✓ Hipótese alternativa

    Exemplo:

    H_1: há diferença entre tratamentos

- Teste de hipóteses

  - ✓ Critério de decisão ⟶ estatística de teste

    - ✓ Estatística de teste ⟶ mede a discrepância entre os valores amostrais observados e os esperados caso H_0 fosse verdadeira.

    - ✓ Uma grande distância medida pela distribuição de probabilidade indica que H_0 não é verdadeira, devendo ser rejeitada.

- Respostas dicotômicas: amostras independentes
  - ✓ Teste Z para comparação de proporções
    - ✓ Estatística de teste:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \quad \text{ou} \quad Z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$\hat{p}$ é a prob. estimada de sucesso nas duas amostras combinadas $\longrightarrow \hat{p} = \frac{m_1 + m_2}{n_1 + n_2}$

n_1 e n_2 é o tamanho de cada amostra

m_1 e m_2 é o nº de sucessos em cada amostra

- Respostas dicotômicas: amostras independentes

  - ✓ Ex: Comparação de drogas contra náusea

    grupo 1 ⟶ n_1 = 200 marinheiros ⟶ pílula A
    grupo 2 ⟶ n_2 = 200 marinheiros ⟶ pílula B

    grupo 1 ⟶ m_1 = 152 marinheiros não enjoaram
    grupo 2 ⟶ m_2 = 132 marinheiros não enjoaram

- A eficácia das pílulas A e B é a mesma?

  - ✓ H_0: Não há diferença entre as pílulas A e B

- Respostas dicotômicas: amostras independentes
  - ✓ Ex: Comparação de drogas contra náusea

Proporções estimadas:

$$\hat{p}_1 = \frac{152}{200} = 0,76 \qquad \hat{p}_2 = \frac{132}{200} = 0,66 \qquad \hat{p} = \frac{152 + 132}{200 + 200} = 0,71$$

1ª Estatística de teste:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0,76 - 0,66}{\sqrt{\frac{0,76(1-0,76)}{200} + \frac{0,66(1-0,66)}{200}}} = 2,22$$

Valor-p: $P(Z > 2,22) = 1 - P(Z \leq 2,22) = 1 - 0,9868 = 0,0132$

- Respostas dicotômicas: amostras independentes

  - ✓ Ex: Comparação de drogas contra náusea

Proporções estimadas:

$$\hat{p}_1 = \frac{152}{200} = 0,76 \qquad \hat{p}_2 = \frac{132}{200} = 0,66 \qquad \hat{p} = \frac{152 + 132}{200 + 200} = 0,71$$

2ª Estatística de teste:

$$Z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,76 - 0,66}{\sqrt{0,71(1-0,71)\left(\frac{1}{200} + \frac{1}{200}\right)}} = 2,20$$

Valor-p: $P(Z > 2,20) = 1 - P(Z \leq 2,20) = 1 - 0,9861 = 0,0139$

✓ Rejeita-se H_0 a um nível de significância de 5% e conclui-se que as pílulas A e B possuem eficácias diferentes

Exemplos para Precauções com o uso de Resultados de Testes de Hipóteses

1) VALOR-P E TAMANHO DA AMOSTRA

Podemos considerar que todo Valor-p, por exemplo 0,041, indica igual evidência contra a hipótese, independentemente da hipótese e do contexto dos dados?

Considere um experimento em que todos os pacientes recebem ambos os tratamentos A e B e são solicitados a explicar suas preferências. Os quatros conjuntos de dados são:

Dados I: 15 pacientes preferem A, 5 preferem B (**r** = 15/20 = 0,75);
Dados II: 114 pacientes preferem A, 86 preferem B (r = 0,57);
Dados III: 1046 pacientes preferem A, 954 preferem B (**r** = 0,523);
Dados IV: 1001445 pacientes preferem A, 998555 preferem B (**r** = 0,5007).

Estes conjuntos de dados produzem um Valor-p de 0,041, obtido de um teste de proporção para a hipótese nula r = 1/2 (que indica igual preferência por A e B) (r é o valor verdadeiro, **r** é o valor calculado dos dados), informando assim que o resultado obtido em cada conjunto (indicando maior preferência por A) é significativamente diferente da hipótese nula (que indica igual preferência por A e B). Porém, estes dados não são igualmente convincentes sobre a maior preferência por A. Os dados I provavelmente serão rejeitados por ser a amostra muito pequena, embora indiquem uma preferência para A de 75%. Os dados IV indicam de forma quase conclusiva que as preferências são iguais, isto é, r = 1/2 (preferência por A de 50,07%). Portanto, o Valor-p de 0,041 não pode ser tomado como evidência independente do contexto e do tamanho da amostra.

2) testes de hipóteses Bayesianos para a hipótese H_o: r = 1/2 (igual preferência) no exemplo visto anteriormente, que concluíram que r = 1/2 com probabilidades:
P(r = 1/2, Dados I: n = 20 e **r** = 0,75) = 0,382

P(r = 1/2, Dados II: n = 200 e **r** = 0,57) = 0,637
P(r = 1/2, Dados III: n = 2000 e **r** = 0,523) = 0,846
P(r = 1/2, Dados IV: n = 2000000 e **r** = 0,5007) = 0,994

Portanto, H_o: r = 1/2 (igual preferência) é bem plausível segundo os dados II, III e IV.

Estas probabilidades são obtidas da expressão

$$P(H_o / \text{Dados}) = \left\{1 + \left[(1+n)^{1/2} \exp\left\{-\frac{n}{n+1} \times \frac{Z_\alpha}{2}\right\}^2\right]^1\right\}^{-1}$$ (onde Z_α é o valor da

tabela da distribuição normal correspondente ao valor α, no caso $\alpha = 0,041$, $Z\alpha = 1,97$) e sua justificativa pode ser vista em Berge e Salke[6]. Observe que aqui a interpretação

2) **Ensinar a pensar nas escolhas de probabilidades de erros Tipo I e Tipo II** (Piantosi, 1997, p 162)

Convencionalmente a maioria dos ensaios clínicos são planejados com nível de significância bilateral $\alpha = 0.05$ e poder $1- \beta = 0.80$ ou 0.90 ($\beta=0.20$ ou 0.10).

Isto é correto se a terapia padrão e efetiva e associada com poucos efeitos colaterais. Quando testamos um tratamento alternativo , associado com sérios efeitos colaterais devem manter a taxa de erro do Tipo I pequena (0.05) para reduzir a chance de falso positivo e podemos permitir a taxa de erro Tipo II ser grande (0.20 ou 0.10).

Em contraste , se estamos estudando a prevenção de alguma doença comum usando um agente seguro como dieta ou suplemento dietético haverá pouco prejuízo na aplicação de tal tratamento , portanto a consequência de erro Tipo I não é séria. De fato alguns benefícios podem ocorrer mesmo se não atue na doença. Por outro lado um erro do Tipo II e mais sério porque um tratamento seguro , barato e possivelmente efetivo seria perdido. Em tal caso devemos usar , por exemplo $\alpha = 0.2$ e $\beta = 0.01$.

3) Suponha que separamos dois grupos de pessoas, um grupo A com 3 homens e o grupo B com 2 homens e 1 mulher. Após dois anos nasce um filho no grupo B (paternidade não definida!!!!).

Testes estatísticos de diferença de proporções, como vistos acima, aplicados a esses dados aceitariam as hipóteses:

i) Não há diferença de proporções quanto ao sexo nos dois grupos A e B

ii) Não há diferença de fertilidade nos dois grupos A e B

Conclui-se, com isso, que mulher não é necessária para procriação?

4) Eu e Michael Jordan lançamos 7 bolas ao cesto. Eu acertei 3 lances, Michael Jordan acertou os 7 lances.

Aplicando um teste estatístico, como visto acima, conclui-se que não há diferença de habilidade em lançar bola ao cesto entre Eu e Michael Jordan.

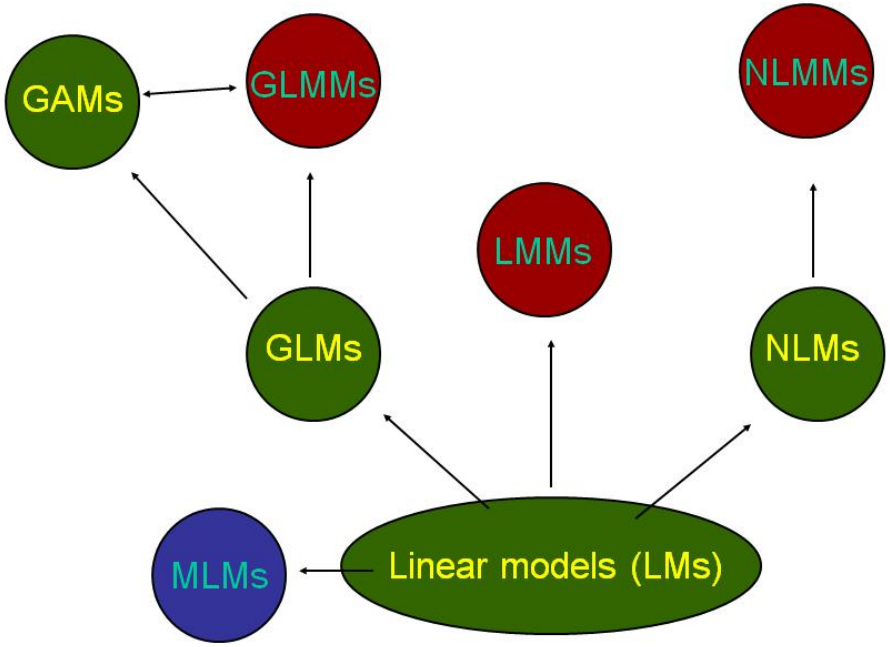
Portanto finalizo aqui minha aula para me dedicar ao basketball e assinar contrato com a NBA nos EUA.

CAPÍTULO 4

MODELOS ESTATÍSTICOS

E APLICAÇÕES

Modelos Lineares Tradicionais



	Acronym
Linear Models	LM
Multivariate LMs	MLM
Generalized LMs	GLM
Linear Mixed Models	LMM
Non-linear Models	NLM
Non-linear Mixed Models	NLMM
Generalized LMMs	GLMM
Generalized Additive Ms	GAM

Modelos de Regressão Usuais em Medicina

Regressão Linear ($-\infty < Y < \infty$) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Regressão Logística (Y=0 ou 1) $\log[P(Y=1)/P(Y=0)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Regressão de Poisson (Taxa de risco: $0 < \lambda < \infty$) $\log[\lambda] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Regressão de Cox (Taxa de risco: $0 < \lambda(t) < \infty$) $\log[\lambda(t)] = \log(\lambda_0) + \beta_1 X_1 + \beta_2 X_2$

Covariável	Estimativa do Coeficiente	Erro Padrão	Estatística de Teste	P-valor
(Intercepto β_0)	3,088	0,132	23,316	<0,0000001
Idade	-0,016	0,003	-5,758	<0,0000001
Sexo	-0,254	0,083	-3,054	0,002
Fumante	-0,206	0,102	-2,012	0,044
PSA	-0,169	0,078	-2,175	0,030

Estatísticas de teste: t –student (regressão linear) e para amostras grandes z-normal (Wald) ou χ^2 (razão de verossimilhança)

Modelos Lineares Iterativos Generalizados - GLIM

Componente Aleatório: distribuição f(y:θ) com média de Y= μ

Componente Sistemático: $\beta_0 + \beta_1 X_1 + \beta_2 X_2$

Função de Ligação: $g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0,1,\ldots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	$0,1,2,\ldots$	μ_i
Gamma	Inverse	$(0,\infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0,\infty)$	$\phi\mu_i^3$

NOTE: ϕ is the dispersion parameter, η_i is the linear predictor, and μ_i is the expectation of Y_i (the response). In the binomial family, n_i is the number of trials.

OBS.: Os resultados são obtidos como na tabela das regressões usuais

Modelos Aditivos Generalizados - GAM

Este modelo generaliza o modelo GLIM modificando a componente sistemática, que passa a ser não paramétrica, com a forma

Função de Ligação: $g(\mu) = \beta_0 + f_1(X_1) + f_2(X_2)$

E agora no lugar de estimativa de coeficientes teremos funções das covariáveis , funções estas estimadas não parametricamente.

Mínimos Quadrados

$$S = \sum_{1=1}^{n} {r_i}^2$$

$$r_i = y_i - f(x_i, \beta)$$

$$f(x, \beta) = \beta_0 + \beta_1 x$$

Ridge Regression e *Elastic Net*

$$\widehat{\beta}(ridge) = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Isto é, minimiza a soma dos erros ao quadrado sujeito a

Lasso

$$\widehat{\beta}(lasso) = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$\widehat{\beta}(enet) = \left(1 + \frac{\lambda_2}{n}\right)\left\{\arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1 \|\beta\|_1\right\}$$

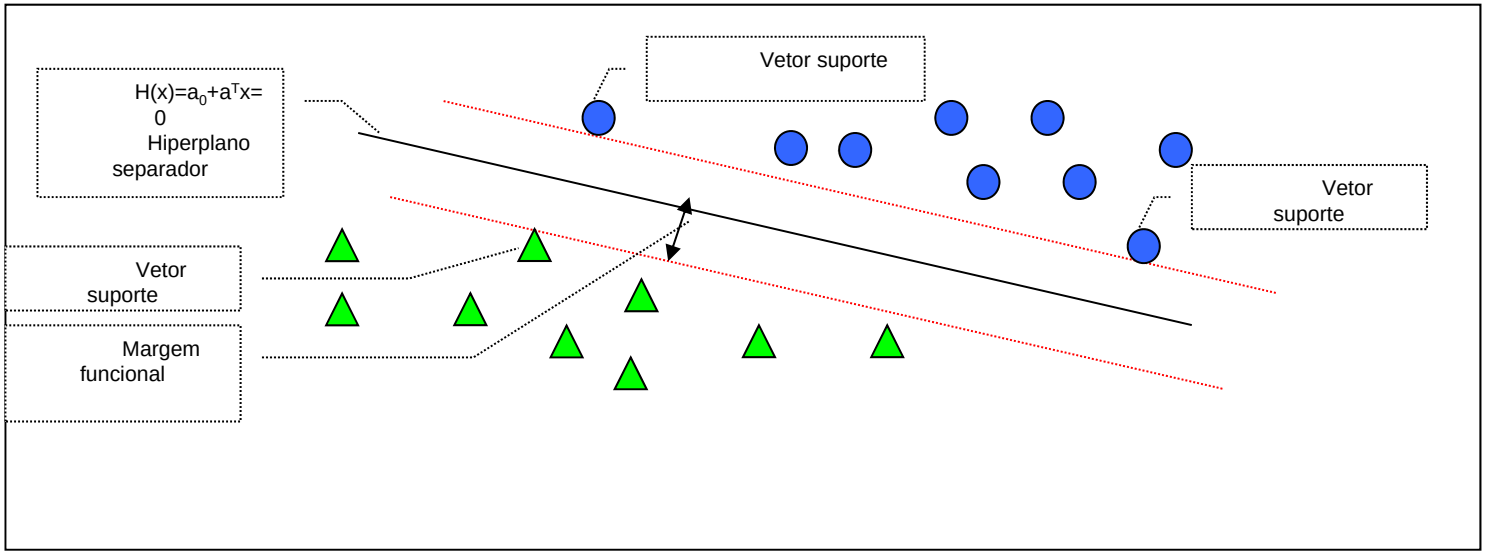
- Mínimos quadrados obtém estimadores de Beta que minimizem

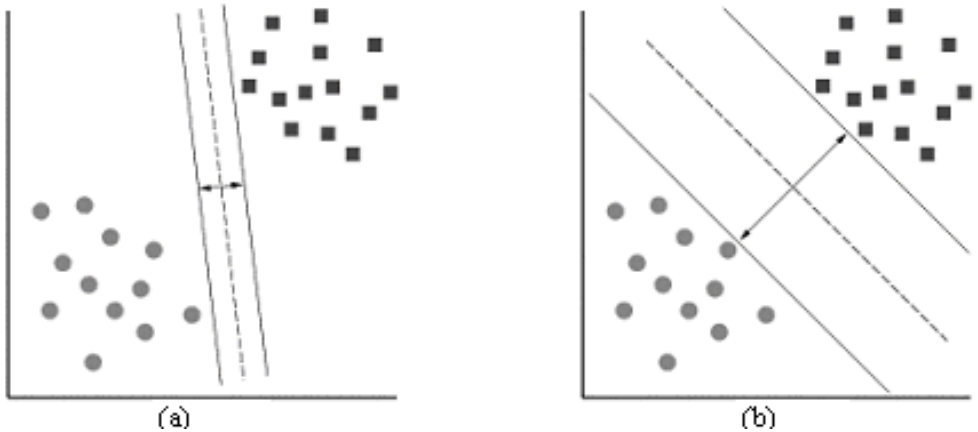
$$S = \sum_{1=1}^{n} r_i^2$$

- Ridge Regression é útil quando há multicolinearidade entre as covariáveis
- Lasso é útil na escolha de variáveis
- Elastic Net é útil para colinearidade, escolha de variáveis e se aplica se p > n

Support Vector Machine (SVM): definição e vantagens

- Treinamento da SVM: consiste na seleção de um hiperplano que minimize o risco estrutural, a partir da resolução de um problema convexo quadrático (PCQ).
- Esta técnica de aprendizado, quando associada à função núcleo, permite a construção de classificadores não-lineares, através do mapeamento dos dados iniciais em um espaço de dimensão superior ao original.
- Boa generalização, ou seja, uma boa capacidade de predizer corretamente o desfecho de indivíduos não utilizados na amostra de treinamento.
- Técnica classificatória de elevado poder distintivo, de custo computacional relativamente baixo e de fácil implementação.

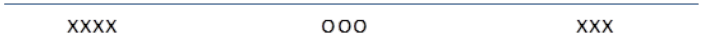


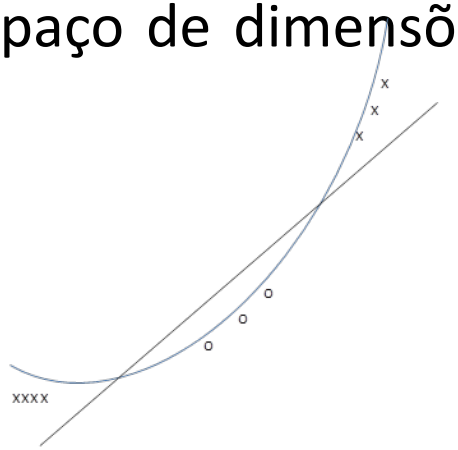


Retas separadoras : a) Pequena margem b) Grande margem.

- O classificador linear em um espaço de dimensão superior corresponderá a um classificador não-linear no espaço original.

- Teorema da Separabilidade de Cover:
  - Um problema de difícil classificação é mais provável de ser linearmente separável em um espaço de dimensões mais elevadas.





- Se o problema for separável os vetores de suporte são obtidos a partir de um problema de otimizaçao do tipo

$$\text{Max} \quad \frac{1}{\| \mathrm{w} \|^2}$$

$$\text{St.} \qquad y_i (x_i w + b) - 1 \geq 0 \,,$$

- Se o problema não for separável os vetores de suporte são obtidos utilizando uma função (núcleo) para trabalharmos em um espaço de maior dimensão e os vetores de suporte são obtidos de um problema de otimização do tipo

$$\text{Max} \quad \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} a_i a_j y_i y_j \varphi(x_i) \cdot \varphi(x_j)$$

$$\text{St.} \quad a_i \geq 0, i = 1, \ldots, l, \text{ and } \sum_{i=1}^{l} a_i y_i = 0 \,,$$

As funções núcleo:

- Vantagens:
  Evita a maldição da dimensionalidade;
  Reduz o custo computacional.

- Consequência:
  Torna os problemas tratáveis, mesmo quando se trabalha em espaço de dimensões elevadas.

Modelos de Classificação e Regressão: CART

CART (*classification and regression trees*)

Particionamento recursivo do espaço de covariáveis

$\rightarrow$ modelo estruturado em árvore

- Árvore de regressão $\rightarrow$ resposta contínua
- Árvore de classificação $\rightarrow$ resposta binária

Árvore de regressão

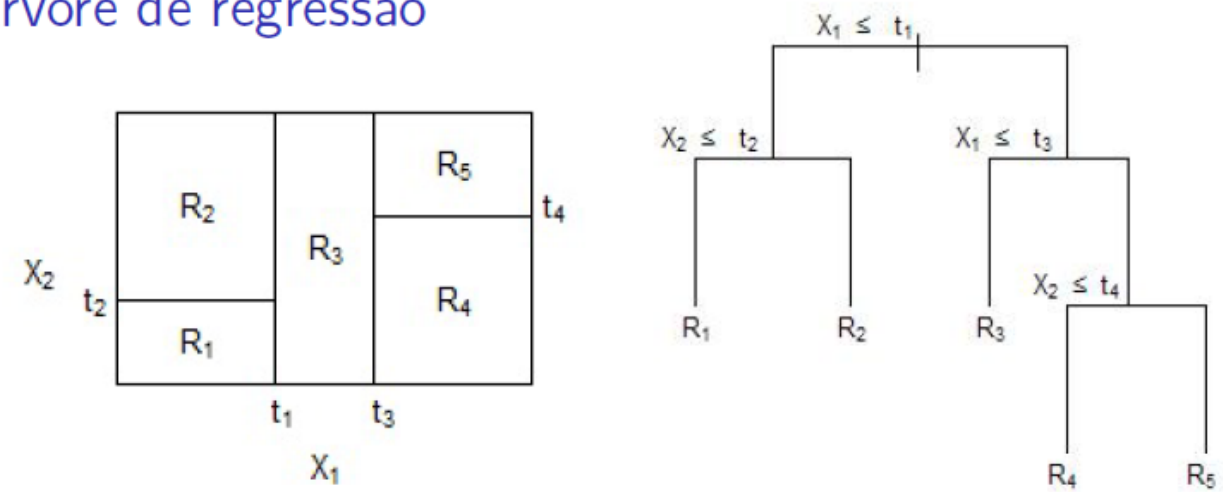


Figura: Partição de χ e árvore binária

- variável X_1 dividida no ponto $X_1 = t_1$;
  subregião $X_1 \leq t_1$ dividida no ponto $X_2 = t_2$;
  subregião $X_1 > t_1$ dividida no ponto $X_1 = t_3$;
  subregião $X_1 > t_3$ no ponto $X_2 = t_4$

$$\hat{y} = \sum_{m=1}^{5} C_m I\{(X_1, X_2) \in R_m\}$$

- Mais que 3 variáveis $\rightarrow$ difícil visualização
  Vantagem da divisão binária $\rightarrow$ fácil interpretação

Árvore de regressão

- Construção da árvore de regressão
- Amostra de aprendizagem $\mathcal{L}$
- Bondade da divisão dos nós $\rightarrow$ menor impureza $\rightarrow$ mse

$$i(t) = \sum_t \frac{(y - \bar{y})^2}{n_t}$$

$$n_t = \text{nº elementos no nó } t$$

- Regra para determinar um nó terminal
- Regra para atribuir o valor $y(t)$ a cada nó terminal t

Árvore de classificação

- $\chi \rightarrow$ espaço de medidas
- $C = \{1, 2, \ldots, J\} \rightarrow$ classes
- Classificador ou regra de classificação $\rightarrow$ função $d(x) = j$, que associa cada vetor x a um ponto no espaço χ

$$A_j = \{x; d(x) = j\}, \qquad A_1, A_2, \ldots, A_j \text{ disjuntos}$$

Redes Neurais Artificiais

- Inspiradas no neurônio biológico
- Assemelham-se ao cérebro humano

  Aquisição de conhecimento $\rightarrow$ aprendizagem

  Pesos sinápticos $\rightarrow$ armazenar conhecimento
- Aplicações
  - diagnóstico médico
  - classificação
  - reconhecimento de padrões
  - diagnóstico de falhas
  - reconhecimento de fala
  - localização de fontes de radar
  - otimização de processos químicos
  - detecção de sinais

Arquitetura

- *Feedforward* ou acíclica

  $\rightarrow$ a saída de um neurônio não é usada como entrada de nodos em camadas anteriores $\rightarrow$ não há realimentação

- *Feedback* ou cíclica

  $\rightarrow$ saída de um neurônio da i-ésima camada usada como entrada de nós em camadas de índice $\leq i$

Tipo *Feedforward*

- *Multilayer perceptron* - MLP
  $\rightarrow$ uma ou mais camadas escondidas

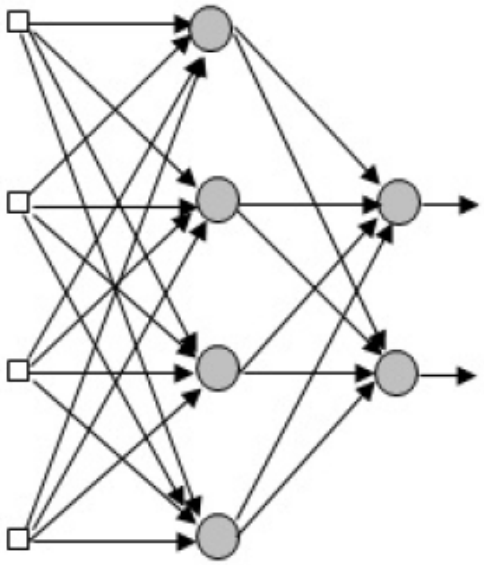


Figura: Rede *feedforward* MLP (topologia 4-4-2)

Mapas Auto-Organizáveis

(*Self-Organizing Map*) - SOM

- Aprendizagem não supervisionada
- Partindo do espaço de pontos original

  $\rightarrow$ comprimem a informação em uma rede

  $\rightarrow$ conjunto de pontos bem menor, representativo, preservando as relações de distância e vizinhança

APLICAÇÕES

1 – Modelo Linear Iterativo Generalizado

Regressão de Poisson

Pereira et al., 2010a

Objetivo

- Identificar alterações gustatórias por comprometimento do nervo corda do tímpano em pacientes com otite média crônica (OMC), ainda não submetidos à cirurgia

Método

- 45 pacientes com colesteatomatosa unilateral ou OMC supurada, um dos ouvidos sem alterações
- Teste gustatório usando tiras gustativas com concentrações de sal, doce, amargo e azedo
- Grupo controle: metade da língua correspondente ao ouvido sadio
- Notas de 0 a 16

Regressão de Poisson

distribuição f(y:θ) com média de Y = λ

link: log[λ] = $\beta_0 + \beta_1 X_1 + \beta_2 X_2$

Tabela: Modelo de Poisson

	Estimate	*Std. Error*	*z value*	$Pr(>\|z\|)$
(*Intercept*)	3.087648	0.132424	23.316	$< 2e-16 ***$
side	−0.396536	0.074669	−5.311	$1.09e-07 ***$
age	−0.015697	0.002726	−5.758	$8.53e-09 ***$
gender	−0.253850	0.083129	−3.054	$0.00226 **$
smoke	−0.205866	0.102337	−2.012	$0.04426*$
cholesteatoma	−0.168641	0.077527	−2.175	$0.02961*$

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2 – Modelo Linear Iterativo Generalizado

Modelo Log-linear

Distribuição multivariada $y=(f_a, f_b \ldots)$,

$f(y:\theta)$ multinomial com média $\mu_i=np_i$ para cada variável f_i

link: $\log(f)=\lambda_a + \lambda_b + \lambda_c + \lambda_{ab} + \lambda_{ac}\ldots..$

Terzi et al., 2010

Objetivo

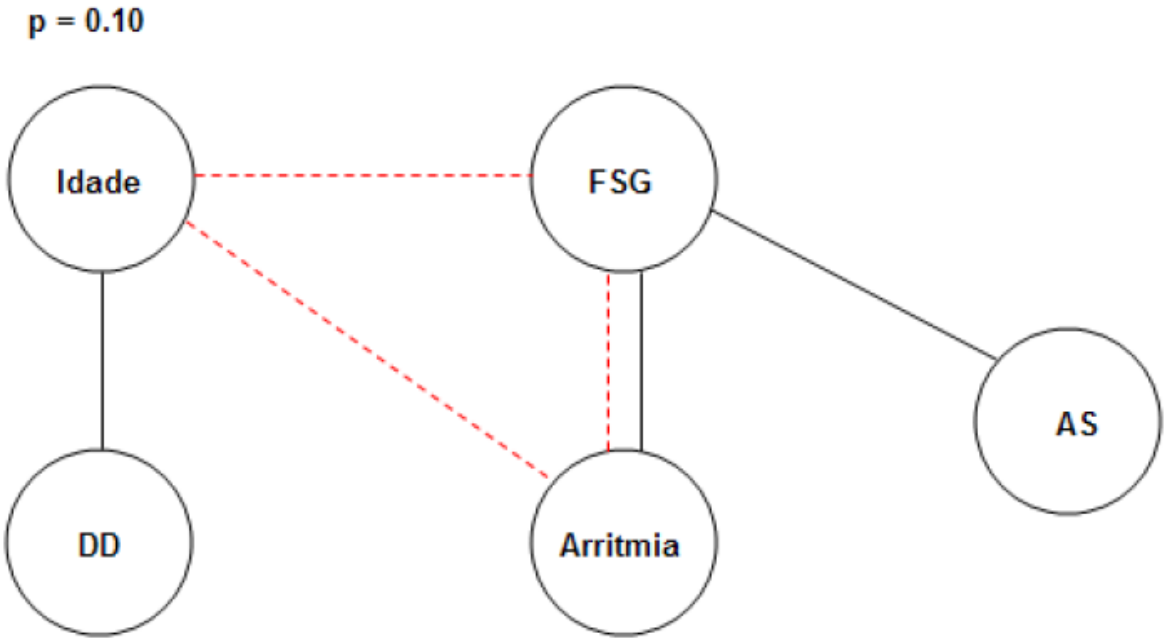
- Determinar a relação entre alteração contrátil e arritmias ventriculares complexas em pacientes chagásicos

Método

- 49 pacientes com Doença de Chagas e eletrocardiograma normal ou borderline
- Exames realizados: ecocardiograma, teste ergométrico e Holter

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			31	134.195	
Idade	1	6.022	30	128.172	0.014
FSG	1	40.220	29	87.952	2.269e-10
AS	1	13.376	28	74.577	2.549e-04
DD	1	0.020	27	74.556	0.886
Arrit	1	21.190	26	53.366	4.158e-06
Idade:FSG	1	0.562	25	52.804	0.453
Idade:AS	1	0.437	24	52.366	0.508
FSG:AS	1	13.793	23	38.574	2.041e-04
Idade:DD	1	21.754	22	16.820	3.100e-06
FSG:DD	1	0.396	21	16.424	0.529
AS:DD	1	0.005	20	16.419	0.941
Idade:Arrit	1	2.702	19	13.717	0.100
FSG:Arrit	1	3.377	18	10.339	0.066
AS:Arrit	1	2.121	17	8.218	0.145
DD:Arrit	1	0.900	16	7.318	0.343
Idade:FSG:Arrit	1	3.220	15	4.098	0.073

- Idade (0 se < 54, 1 c.c)
- FSG - Disfunção sistólica global do VE (0 Normal, 1 leve)
- AS - Alteração segmentar (0 Ausente, 1 Presente)
- DD - Disfunção diastólica (0 Ausente, 1 Presente)
- Arrit - Arritmia ventricular complexa (0 Negativo, 1 Positivo)



- Idade (0 se < 54, 1 c.c)
- FSG - Disfunção sistólica global do VE (0 Normal, 1 leve)
- AS - Alteração segmentar (0 Ausente, 1 Presente)
- DD - Disfunção diastólica (0 Ausente, 1 Presente)
- Arrit - Arritmia ventricular complexa (0 Negativo, 1 Positivo)

3 – Modelo Linear Iterativo Generalizado

Modelo Log-linear

MSc Maria Beatriz Altschüller (Medicina)

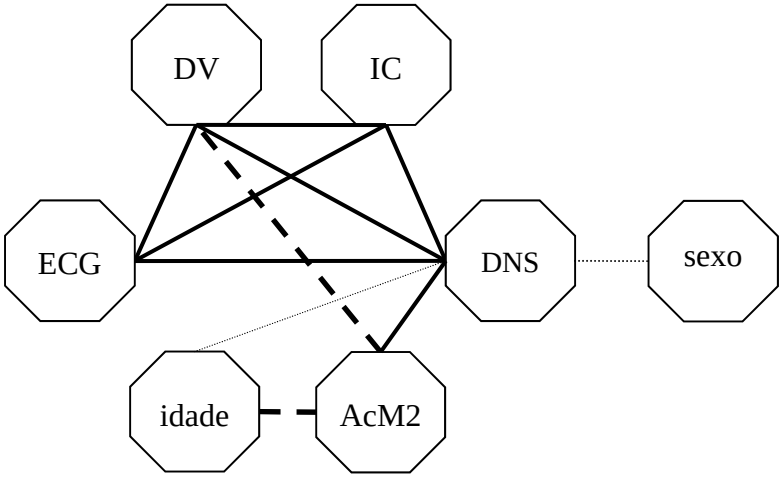
Objetivo

- Avaliar a relação entre as presenças de: disfunção do nódulo sinusal, pelo teste Holter, disfunção ventricular, pelo ecocardiograma, e anticorpos agonistas muscarínicos no soro.

Método

- 69 pacientes chagásicos crônicos, em vários estádios da doença
- Modelo Log-linear

Tabela 2x2



Fonte: SDM/ HUCFF (2003-2005)

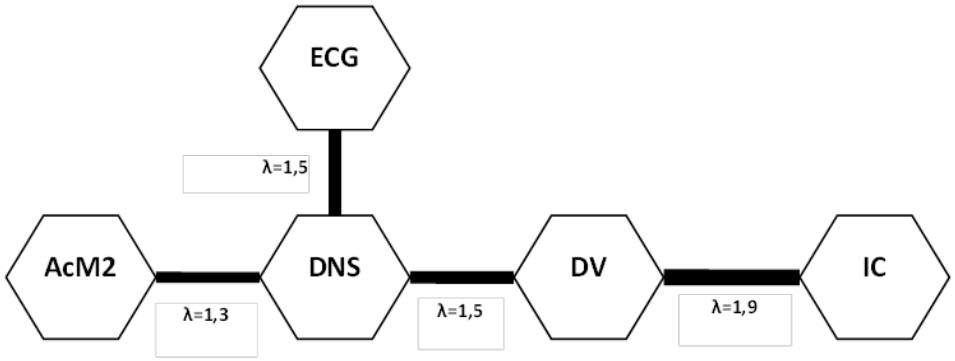
Representação gráfica das relações entre as variáveis pelo teste qui-quadrado. Linhas cheias representam os valores de p <0,05, as tracejadas os valores de p ≥0,05 e ≤0,10 e as pontilhadas os valores de p >0,10 e <0,16.

AcM2: anticorpos com ação agonita de receptores muscarínicos M2; DNS: disfunção do nódulo sinusal; ECG: eletrocardiograma alterado; DV: disfunção ventricular; IC: insuficiência cardíaca.

Modelo log linear Distribuição multivariada $y=(f_a, f_b ...)$,

$f(y:\theta)$ multinomial com média $\mu_i=np_i$ para cada variável f_i

link: $\log(f)=\lambda_a + \lambda_b + \lambda_c + \lambda_{ab} + \lambda_{ac}.....$



Fonte: SDM/ HUCFF (2003-2005)

- Interdependência entre a disfunção nodal e a ventricular (p=0,0005; $\lambda_{DNS:DV}$=1,5) e entre a disfunção nodal e os anticorpos (p=0,002; $\lambda_{AcM2:DNS}$=1,3)
- Não houve relação entre os anticorpos e a disfunção ventricular.
- Sexo e idade não tiveram influência sobre as outras variáveis.

AcM2: anticorpo com ação agonista de receptores muscarínicos M2; DNS: disfunção do nódulo sinusal, DV: disfunção ventricular, ECG: eletrocardiograma alterado, IC: clínica de insuficiência cardíaca.

4 – Regressão de Cox

Nascimento et al., 2012

Objetivo

- Investigar a associação entre o tempo de sobrevida e as variáveis explicativas: sexo, idade, tipo sanguíneo, imc, etiologia da doença, câncer de fígado, MELD

Método

- 529 pacientes acompanhados de nov/1977 a jul/2006
- Regressão de Cox

MELD *Model for End-Stage Liver Disease*

- Sistema de pontuação para avaliar a gravidade da doença hepática crônica
- Ferramenta para estudar a mortalidade na fila de transplante de fígado no curto prazo

$$MELD = 3.8 \log_e(b) + 11.2 \log_e(INR) + 9.6 \log_e(cr) + 6.4 * et$$

- b = bilirrubina sérica [mg/dL]
- INR = Razão Normalizada Internacional para o tempo de protrombina
- cr = creatinina sérica [mg/dL]
- et = etiologia da cirrose (0 colestática ou alcoólica; 1 caso contrário)

Regressão de Cox

Taxa de risco: $0 < \lambda(t) < \infty$

$\log[\lambda(t)] = \log(\lambda_0) + \beta_1 X_1 + \beta_2 X_2$

Tabela: Modelo de Cox para as variáveis MELD, idade e HCC

	coef	*se*(*coef*)	*z*	*p*
age	0.0208	0.00645	3.23	0.0013
meld	0.1097	0.01094	10.03	0.0000
*hcc*1	0.5259	0.21502	2.45	0.0140

5 – LASSO

MSc Raphael Iglesias O. Vidal (Medicina)

Objetivo

- Analisar os fatores prognósticos para sobrevivência e recorrência do CHC em pacientes portadores de infecção crônica pelo VHC submetidos a transplante hepático no Hospital Universitário Clementino Fraga Filho, no período de 1998 a 2008.

Método

- 174 pacientes com cirrose hepática submetidos a transplante
- selecionados 76 casos de pacientes que apresentaram CHC nas análises dos fígados explantados

TODAS AS VARIÁVEIS	VIVOS	ÓBITOS SEM RECORRÊNCIA DE CHC	ÓBITOS COM RECORRÊNCIA DE CHC
Intercepto	*1,5636*	*0,5539*	*-2,1175*
ABO	.	.	.
CHILD	.	.	.
MELD	.	.	.
CHC INCIDENTAL	.	.	.
NÚMERO	.	.	.
TAMANHO	.	.	.
INVASÃO MICROVASCULAR	.	.	1,963.2
α-FP	.	0,0006	.
TRATAMENTO LOCORREGIONAL	.	.	.
CHC COM NECROSE	.	0,1465	.
α-FP EM ASCENSÃO	.	.	.
CHA	-0,0486	0,0655	.

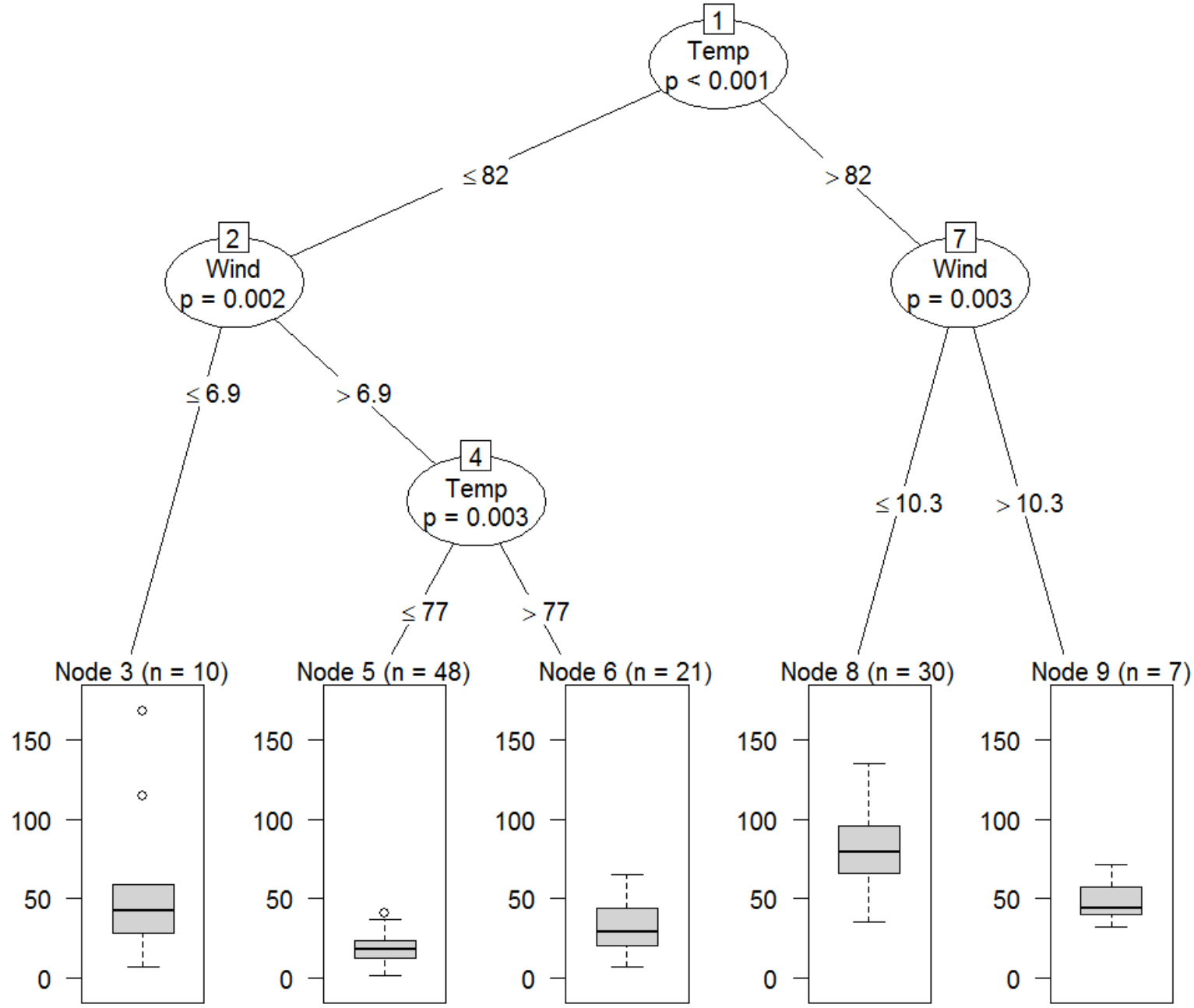
6 – Árvore de Regressão

Dados

- Dados diários de qualidade do ar em NY medidos de maio a setembro/1973 (Chambers et al., 1983).

Variáveis

- *Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island*
- *Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park*
- *Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport*
- *Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.*


1
Temp
p < 0.001
≤ 82
> 82
2
Wind
p = 0.002
7
Wind
p = 0.003
≤ 6.9
> 6.9
4
Temp
p = 0.003
≤ 10.3
> 10.3
≤ 77
> 77
Node 3 (n = 10)
Node 5 (n = 48)
Node 6 (n = 21)
Node 8 (n = 30)
Node 9 (n = 7)
150
100
50
0

7 – Árvore de Classificação

Landesmann et al., 2011

Objetivo

- Detectar precocemente a disfunção adrenérgica cardíaca em pacientes chagásicos assintomáticos, com eletrocardiograma normal ou borderline e função ventricular preservada

Método

- 40 pacientes chagásicos e 19 normais
- Cintilografia com MIBG-Iodo-123, com imagens planares de 20 min e 3 h, e imagens tomográficas de 60 a 90 min após a injeção

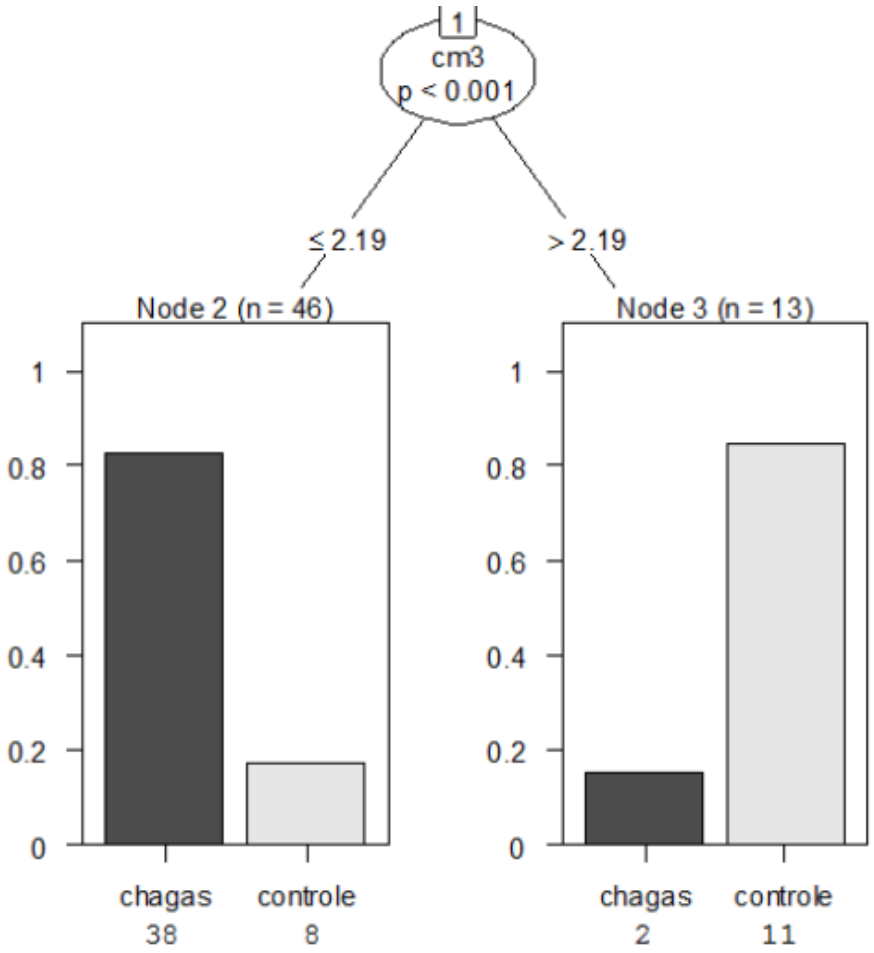


Figura: Idade, CM20, CM3, Washout, SPECT, Gênero

8a – Árvore de Sobrevida

Nascimento et al., 2012

Objetivo

- MELD como preditor da mortalidade no longo prazo e árvores de sobrevida para determinar o ponto de corte

Método

- 529 pacientes acompanhados de nov/1977 a jul/2006
- Variáveis: MELD; sexo; idade; tipo sanguíneo; imc; etiologia da doença; hepatocelular carcinoma; tempo de espera na fila (dias)

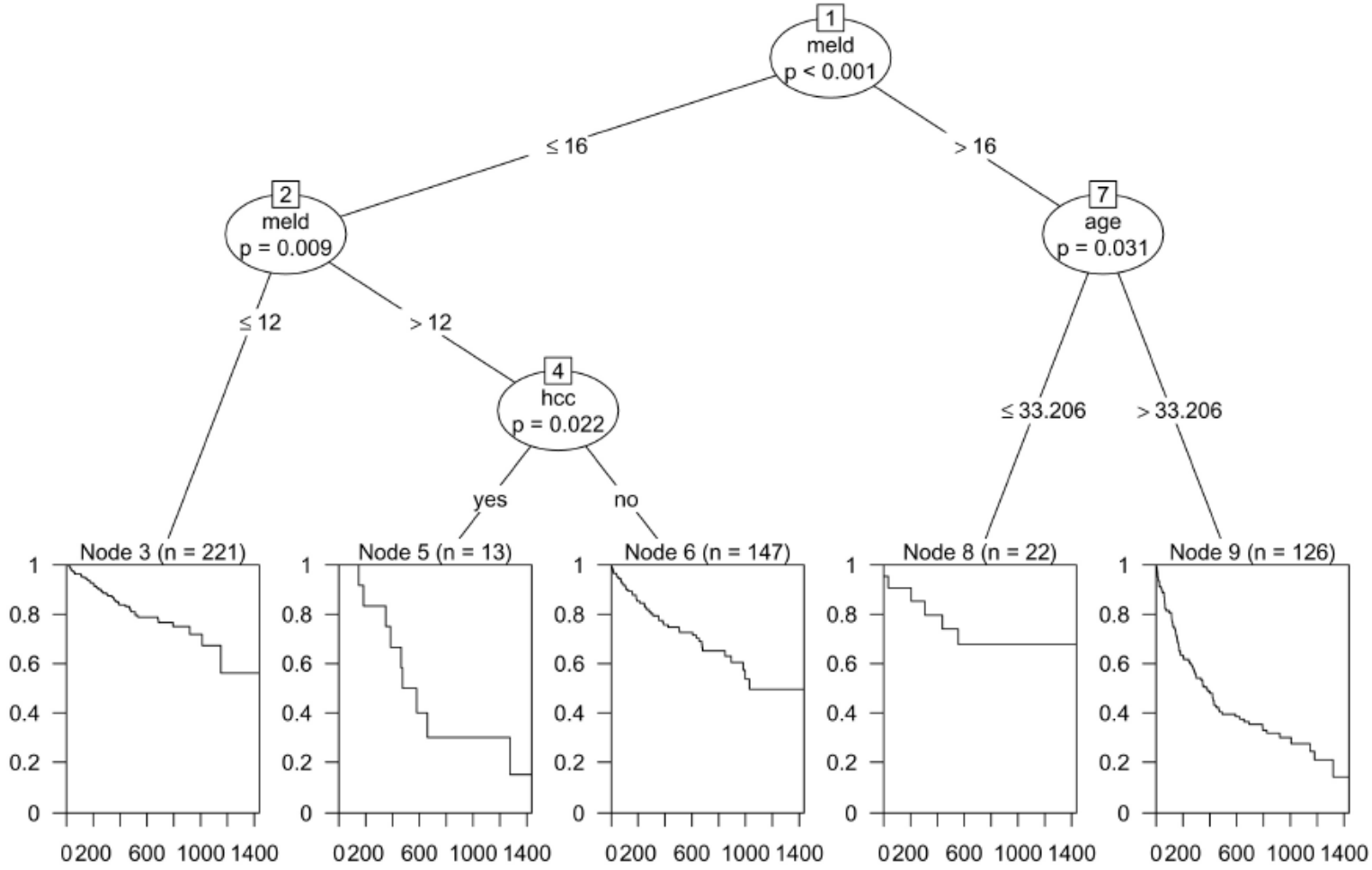


Figura: MELD, idade e HCC

8b – Árvore de Sobrevida

Petriz et al., 2014

Objetivo: Verificar a associação entre os dados de anatomia e magnitude do infarto, obtidos da ressonância magnética cardíaca pós-infarto agudo do miocárdio, e mortalidade em longo prazo.

Métodos: Foram identificados 1.959 laudos com “massa infartada” em 7.119 exames de ressonância magnética cardíaca, dos quais 420 possuíam documentação clínica e laboratorial de infarto agudo do miocárdio prévio.

As variáveis estudadas foram os fatores de risco clássicos, fração de ejeção do ventrículo esquerdo, função ventricular categorizada e localização do infarto agudo do miocárdio.

Massa infartada, extensão e transmuralidade do infarto agudo do miocárdio foram analisadas de maneira isolada e conjuntamente, pela variável denominada “MET-IAM”.

A análise estatística foi feita pelo *elastic net regularization, pelo modelo de Cox e por árvores de sobrevida.*

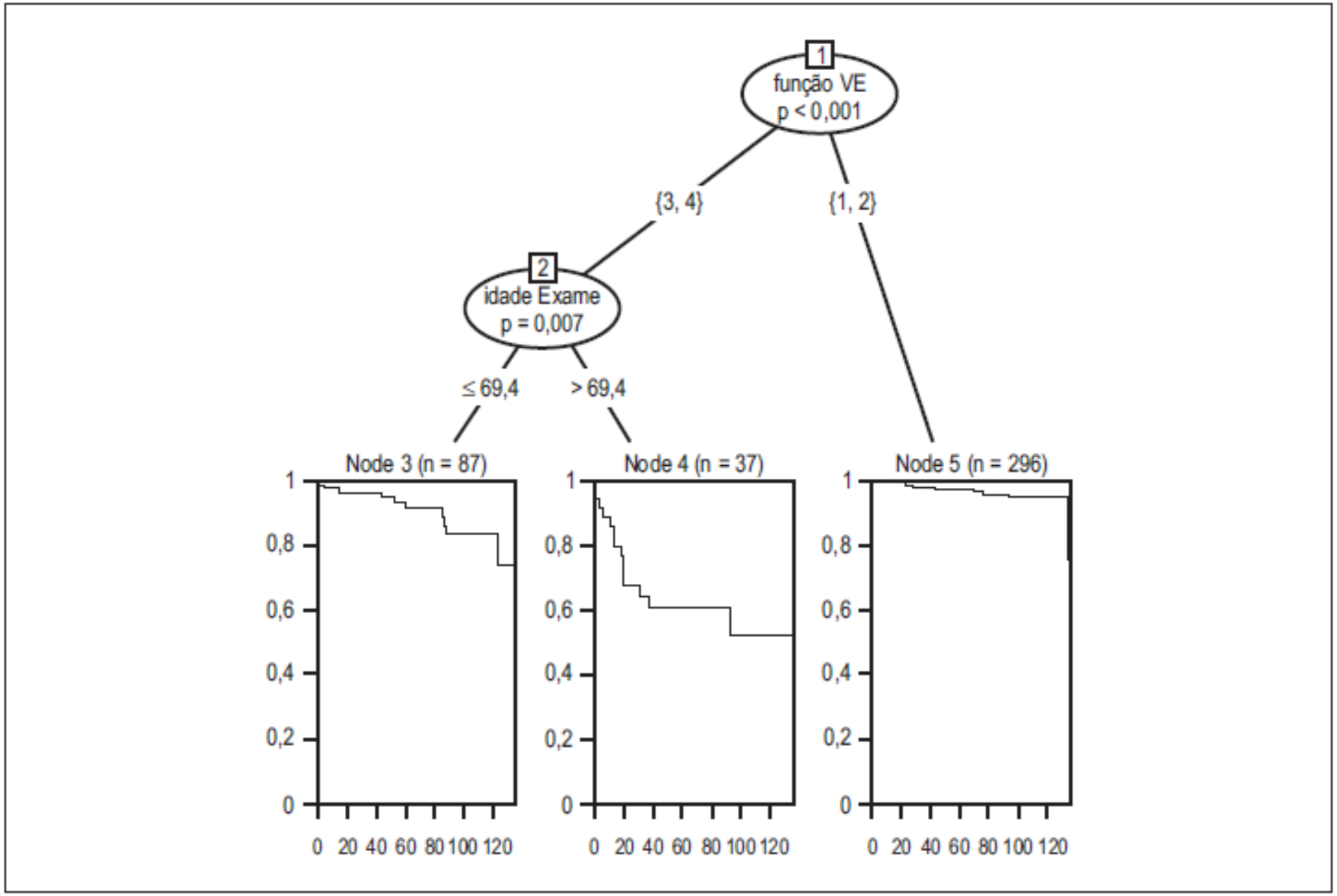


Figura 1 – *Árvore de sobrevida para o desfecho mortalidade por doenças do aparelho circulatório. (1) Função do VE normal; (2) disfunção do VE leve; (3) disfunção do VE moderada; (4) disfunção do VE grave. VE: ventrículo esquerdo.*

9 – Miscelânea

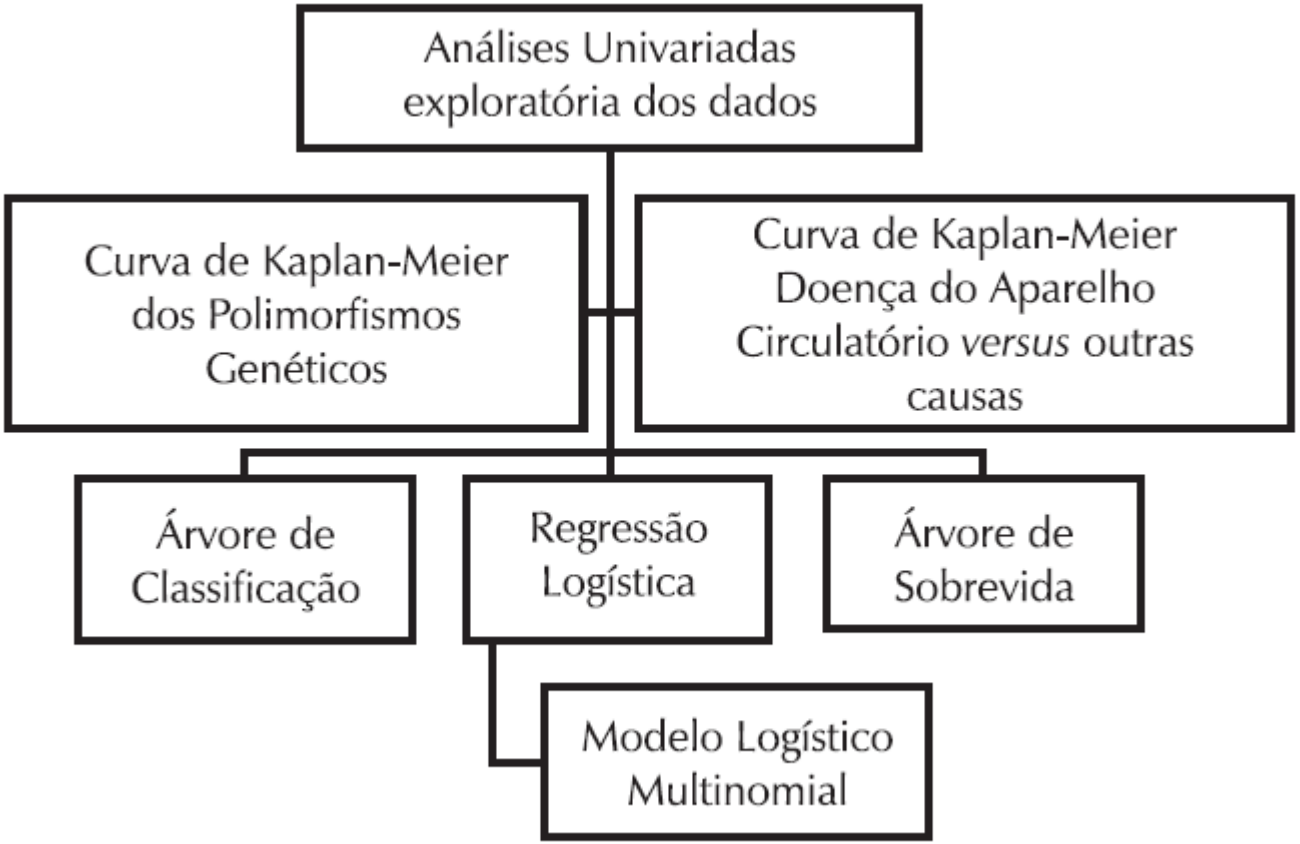
Alves et al., 2013

- Objetivo: Avaliar a sobrevida em hemodialisados e sua associação com polimorfismo dos genes do sistema reninaangiotensina: deleção/inserção do gene que codifica a enzima conversora da angiotensina I e o M235T do angiotensinogênio.
- Métodos: Estudo observacional desenhado para ver o papel dos genes do sistema renina-angiotensina.

  Foram analisados 473 pacientes tratados com hemodiálise crônica em quatro unidades de diálise do Estado do Rio de Janeiro.

  As taxas de sobrevida foram calculadas pelo método de Kaplan-Meier e as diferenças entre as curvas avaliadas pelos testes de: Tarone-Ware, Peto-Prentice e Log-rank.

  Foram utilizados também modelos de regressão logística e multinomial.


Análises Univariadas
exploratória dos dados
Curva de Kaplan-Meier
dos Polimorfismos
Genéticos
Curva de Kaplan-Meier
Doença do Aparelho
Circulatório *versus* outras
causas
Árvore de
Classificação
Regressão
Logística
Árvore de
Sobrevida
Modelo Logístico
Multinomial

10 – Redes Neurais Artificiais – Multilayer Perceptron

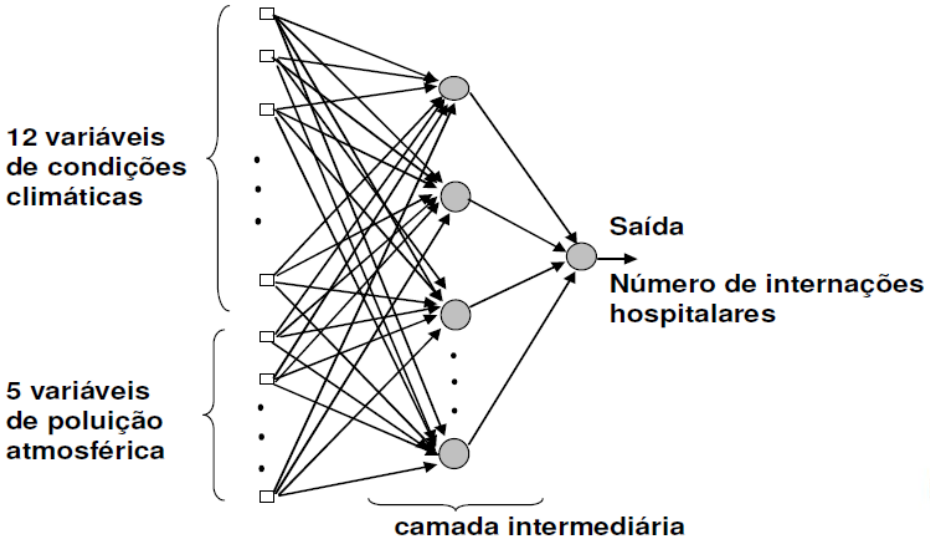
MSc Emilia M. Nascimento (Pesquisa Operacional)

Objetivo

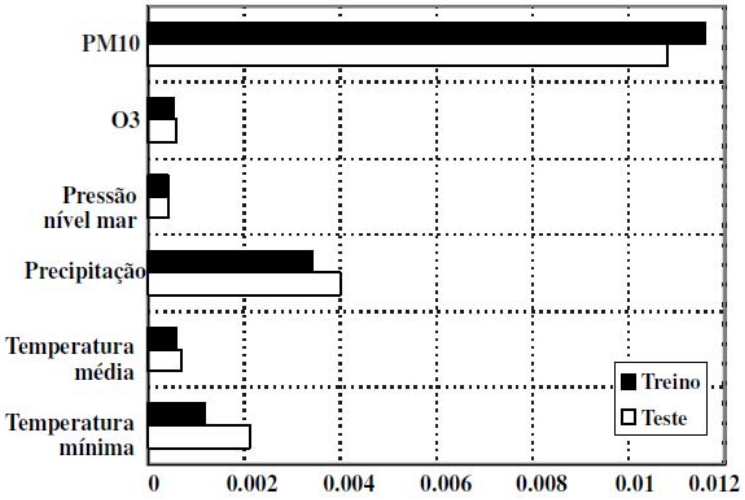
- Investigar a associação entre a poluição atmosférica e condições climáticas no número de internações hospitalares, por motivo de bronquiolite infantil.
- Uso das redes neurais MLP para reproduzir a análise de WILLEMS et al. (2005, apud Nascimento, 2006), que utilizaram os modelos aditivos generalizados.

Método

- 419 pacientes acompanhados de 1977 a 2000, em 34 hospitais de Paris
- 12 variáveis climáticas e 5 de poluição atmosférica



Arquitetura do modelo neural.



Análise de relevância

11 – Redes Neurais Artificiais – Self-Organizing Map

Pereira et al., 2010b

Objetivo

Reconhecimento de padrões em dados de emissão de raios gama.

Método

- Análise de cluster usando o SOM
- 422 amostras completas $\rightarrow$ 17 variáveis
- Topologias: 25 nós (5x5)
  49 nós (7x7)
  100 nós (10x10)
  225 nós (15x15)

variable	*ID*	*Description*
1	T_{50}	*time* measure *representing* the *arrival* of 50% of the flux
2	T_{90}	*time* measure *representing* the *arrival* of 90% of the flux
3	F_1	time-integrated fluence in spectral channels 20-50 keV
4	F_2	time-integrated fluence in spectral channels 50-100 keV
5	F_3	time-integrated fluence in spectral channels 100-300 keV
6	F_4	time-integrated fluence in spectral channels over 300 keV
7	P_{64}	peak flux measured in 64ms bins
8	P_{256}	peak flux measured in 256ms bins
9	P_{1024}	peak flux measured in 1024ms bins
10	T_{64}	trigger threshold, i.e, number of counts in 64 ms required to trigger the second most brightly illuminated detector
11	T_{256}	trigger threshold on the 256 ms timescale
12	T_{1024}	trigger threshold on the 1024 ms timescale
13	*Lat*	galactic latitude
14	*Lon*	galactic longitude
15	F_T	sum of the four fluencies (F1 + F2 + F3 + F4)
16	H_{32}	spectral hardness, obtained from fluence relation F3/F2
17	T_{321}	spectral hardness, obtained from relation F3/(F1+ F2)

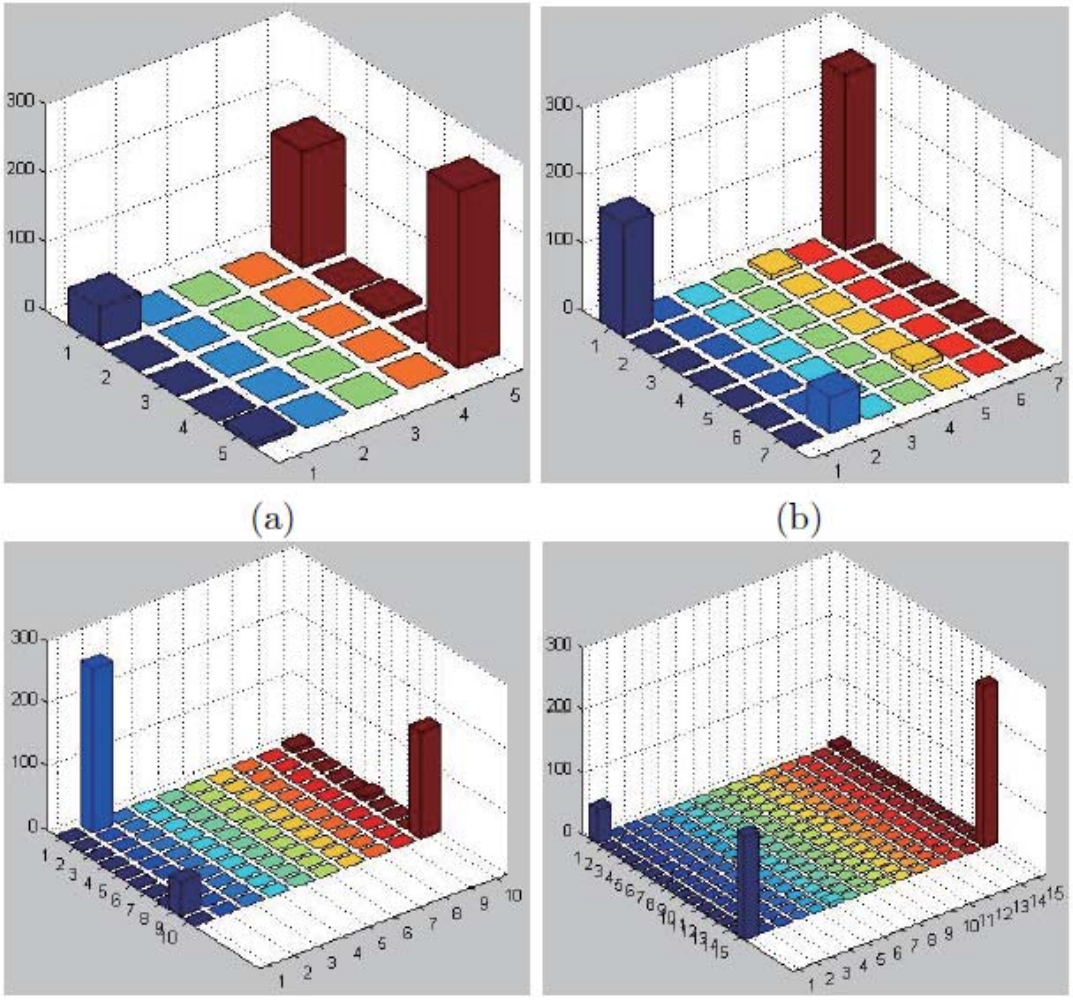
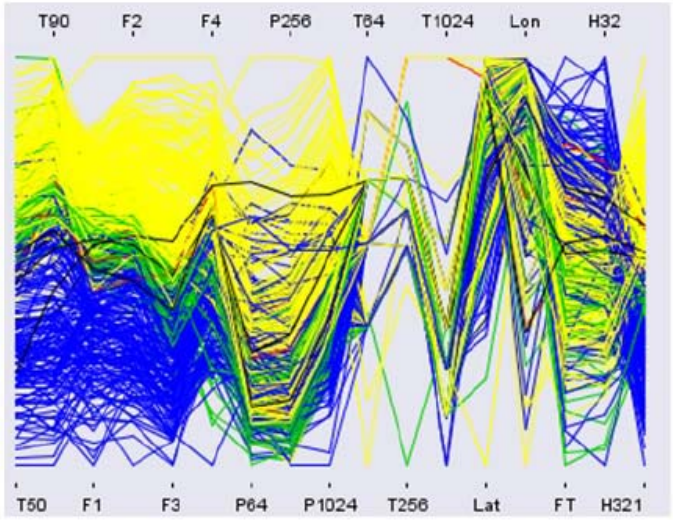
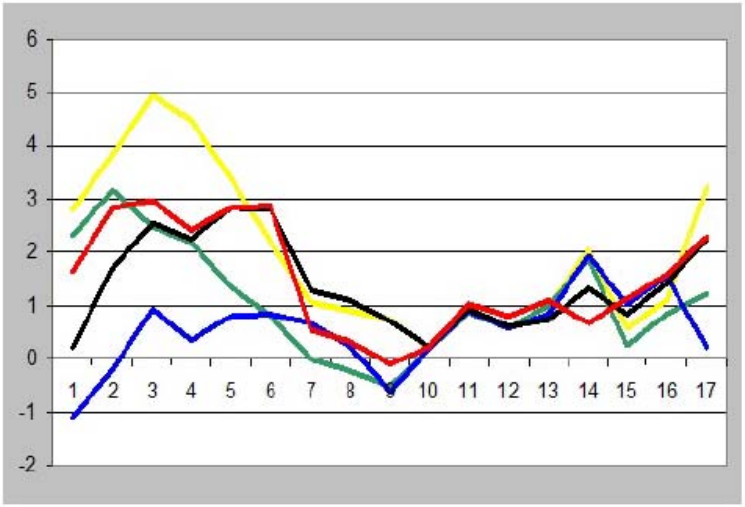


Figure 1: Clustering using Kohonen's maps of 5×5 nodes (a), 7×7 nodes (b), 10×10 nodes (c), and 15×15 nodes (d)





Coordenadas paralelas (perfis)

Valores médios

Class 1 Class 2 Class 3 Class 4 Class 5

12 – Máquinas de vetores suporte

MSc Rodrigo A. Collazo (Pesquisa Operacional)

Objetivo

- Predição do risco de morte por síndrome coronariana aguda usando a ferramenta SVM, que consiste de uma máquina de aprendizado criada a partir de três corpos teóricos:
    - teoria de generalização,
    - núcleo,
    - e teoria de otimização.
- Proposição de critérios alternativos para a seleção das variáveis.

Método

- Dados provenientes da Tese de Doutorado de REIS (2007);
- Os dados foram coletados a partir de um estudo de coorte prospectivo, com pacientes internados com SCA, no município de Niterói, RJ;
- A coleta das informações foi feita no período entre julho/agosto de 2004 e junho/julho de 2005;
- Pacientes internados em cinco hospitais, sendo três públicos e dois privados;
- Condições: idade superior a 20 anos e não apresentar: sinais de doenças neoplásicas em fase terminal, politraumatismos e demência.

- Amostra: 25 variáveis agrupadas em 6 categorias:
  - **variáveis antropométricas, sociais e hábitos de vida**: idade, índice de massa corporal (IMC), sexo, escolaridade, atividade física (AF), tabagismo;
  - **variáveis de história prévia cardiovascular**: infarto do miocárdio prévio (IMP), qualquer revascularização prévia (QRP), história familiar de doença arterial coronariana (DAC);
  - **variáveis clínicas e laboratoriais na admissão hospitalar**: tipo de síndrome coronariana aguda (SCA), tempo para 1º atendimento médico (1ºAM), frequência cardíaca (FC), Classe Killip, creatinina;
  - **variáveis de diagnóstico**: hipertensão arterial sistólica (HAS), colesterol elevado, triglicerídios elevados, colesterol-HDL baixo;
  - **variáveis genéticas**: 7 alelos de 3 polimorfismos;
  - **variável de desfecho**.

- Quatro critérios para seleção de variáveis: o critério desenvolvido em 2008 por CHEN et al. e três modificações de Collazo (2009):
  - Critério CZCL;
  - Critério CZCL Adaptado;
  - Critério CZCL Dual;
  - Critério CZCL Adaptado Dual.

Seleção de variáveis: comparação entre os critérios

Ordem	CZCL	CZCL Adaptado	CZCL Dual	CZCL Adap. Dual	MIFS-U
1	QRP	QRP	Creatinina	Creatinina	Idade
2	Alelo E3	HAS	FC	FC	QRP
3	1°AM	1°AM	Idade	Killip	Creatinina
4	Alelo E2	Alelo I	HDL	Idade	IMC
5	Infarto Prévio	Infarto Prévio	Tabagismo	Alelo E3	Genótipo DD
6	HAS	AF	Escolaridade	Alelo D	Genótipo E4E4

Seleção de Variáveis e Previsão do Desfecho

Variáveis	**v**	**escala**	**a** (%)	**e** (%)	**s** (%)
Creatinina, QRP, Idade	0.16	0.3	97.5	98.3	87.5
Creatinina, QRP, Idade, HAS	0.16	0.6	97.7	98.6	87.5
Creatinina, QRP, Idade, FC	0.16	0.1	96.1	97.6	79.3
Creatinina, QRP, Idade, HAS, FC	0.16	0.3	96.4	97.9	79.3
Creatinina, QRP, Idade, 1ºAM	0.17	0.5	70.4	70	75
Creatinina, QRP, Idade, HAS, 1ºAM	0.15	0.6	79.8	81.1	64.3
Creatinina, QRP, Idade (Reis, 2007)	-	-	70.7	69.8	78.3

- O conjunto de variáveis que apresentou o melhor desempenho foi o mesmo encontrado por REIS (2007), que usou redes neurais e um critério de informação na escolha de variáveis.

- O resultado foi a construção de um classificador que superou o desempenho da RNA *feedforward* construída sobre o mesmo banco de dados e apresentada em REIS (2007)

Referências

Altman, D. G. e Bland J. M. – 1991- Improving doctor's understanding of statistics. (with discussion). Journal of the Royal Statistical Society A, 154, 223-267.

ALVES, M. ; Souza e SILVA, NA ; Salis LHA ; PEREIRA, B. de B. ; GODOY, P. H. ; Nascimento EM ; Oliveira, J.M.F. Survival and Predictive Factors of Lethality in Hemodyalisis: D/I Polymorphism of The Angiotensin I-Converting Enzyme and of the Angiotensinogen M235T Genes. Arquivos Brasileiros de Cardiologia (Impresso), v. 103, p. 209-218, 2014.

Altschüller, M.B.C.M.-2006-Prevalência de Anticorpos IgG com Ação Agonista Muscarínica em Pacientes Chagásicos Crônicos Portadores de Disfunção do Nódulo Sinusal com e sem Disfunção Ventricular. Dissertação de Mestrado em Medicina (Cardiologia), FM/UFRJ.

Box, G. E. P.- 1979- Robustness in the strategy of scientific model building. In R.L. Launer and G.N. Wilkinson. (eds.) Robustness in Statistics, Academic Press.

Breiman ,L-2001-Statistical Modeling: The two cultures (with discussion), Statistical Sciences ,16(3),199-231

Bussab, W.- 2004- Entrevista ao Boletim da ABE 58, AnoXX,14-20.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. -1983 - Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.

Clayton D, Hills,M-1993-Statistical Models in Epidemiology. Oxford Univ. Press.

Collazo, R.A -2009- Aplicação de "Support Vector Machines" à Classificação do Risco de Morte de Pacientes com Síndrome Coronariana Aguda. Dissertação de Mestrado (Pesquisa Operacional), COPPE/UFRJ

Cox,D.R.– 2004 – The accidental statistician. Significance ,1, 27-29.

Diamond GA, Forrester JS-1983- Clinical trials and statistical verdicts: probable grounds for appeal. ANN. Inter. Med. ,98, 385-394

Dr. Fisher – 2004- Dr Fisher casebook, Significance,1, 26.(editoriais de Significance)

Freeman PR- 1993-The role of p-value in analyzing trials results (with discussion) Stat Med ,12,1443-1458

Healy, M.J.R.- 1979- Does medical Statistics exist? Bulletin Applied Statistics, 6, 137-183

Hoffman,P. e Grinstein, G. – 1997- Visualizations for high dimensional data mining- table visualizations (http://home.concast.net/~patrick.hoffman/viz/MIV-data mining.pdf)

Kanji, J.G.-2006 - 100 Statistical Tests. Sage Publications.

Landesmann, M. C. P., Fonseca, L. M. B., Pereira, B.B., Nascimento, E.M., Rosado-de-Castro, P. H., Souza, S. A. L., Lima, R. S. L., Pedrosa, R.C. -2011- Iodine-123 Metaiodobenzylguanidine Cardiac Imaging as a Method to Detect Early Sympathetic Neuronal Dysfunction in Chagasic Patients With Normal or Borderline Electrocardiogram and Preserved Ventricular Function. Clinical Nuclear Medicine. , v.36, p.757 - 761.

Lindsey,JK-1995- Introductory Statistics : A Modelling Approach. Oxford Sc.Publ.

Mannarino, VL-2009- Avaliação dos cuidados paliativos sobre sintomas angustiantes, capacidade funcional e qualidade de vida nos pacientes com câncer primário de pulmão, Dissertação de Mestrado em Medicina (Clínica Médica), FM/UFRJ.

Nascimento, EM-2006-Redes Neurais Artificiais: Uma Aplicação no Estudo da Poluição Atmosférica e seus Efeitos Adversos à Saúde, Dissertação de Mestrado (Pesquisa Operacional), COPPE/UFRJ.

Nascimento, EM-2010- Tópicos em Aprendizado Estatístico na Pesquisa Clinica. Tese de Doutorado (Pesquisa Operacional). COPPE/UFRJ

Nascimento, E.M., Pereira, B.B., Seixas, J. M. -2009- Redes Neurais Artificiais: Uma Aplicação no Estudo da Poluição Atmosférica e Seus Efeitos Adversos à Saúde. Revista Brasileira de Biometria. , v.27, p.37 - 50.

Nascimento, E.M., Pereira, B.B., Basto, S.T., Ribeiro Filho, J. -2012- Survival Tree and Meld to Predict Long Term Survival in Liver Transplantation Waiting List. Journal of Medical Systems. , v.36, p.73 - 78.

Pereira,BB-1995- Estatistica em medicina: p-varicação. Revista da SOCERJ,8 (3),73-78

Pereira,BB-1997-Estatística a tecnologia da ciência. Boletim da ABE ,13(37),27-35.

Pereira,BB-1997-Estatística a tecnologia da ciência 2. Boletim da ABE,16(47),37-38

Pereira,BB- 2001-Estatistica em psiquiatria Rev. Bras. Psiquiatr. 23 (3),168-170

Pereira, B.B., Nascimento, E.M., Felix, F., Tomita, S. -2010a- A non conventional use of survival curves to identify factors for gustatory alterations in patients with chronic otitis media. Revista Brasileira de Biometria. , v.28, p.104 - 111.

Pereira, B.B., Rao, C.R., Oliveira, R.L., Nascimento, E.M. -2010b- Combining Unsupervised and Supervised Neural Networks in Cluster Analysis of Gamma-Ray Burst. Journal of Data Science, v.8, p.327 - 338.

Pereira,BB, Rao,CR e Rao,MB -2013- Data Mining Using Neural Networks: A Guide for Statisticians , Chapman Hall , Londres

Pereira,BB-2014-Procedures for Discriminating Separate or Non-nested Families of Models, Springer,London.

Petriz, J. L. ; Gomes, B. F. O. ; Rua, B. S. ; Azevedo Filho, C. F. ; Hadlich, M. S. ; Mussi, H. T. P. ; Taets, G. C. ; Nascimento, E. M. ; Pereira, B. de B. ; Souza e Silva, NA . Assessment of Myocardial Infarction by Cardiac Magnetic Resonance Imaging and Long-Term Mortality. Arquivos Brasileiros de Cardiologia (Impresso), v. 103, p. 1-20, 2014.

Piantadosi, S.- 1997- Clinical Trials – A Methodological Pespective. Wiley Intercience.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reese,A.–2004–Does significance matters? Significance, 1 , 39-40..

Reis, A. F.- 2007- Modelo Preditivo de Mortalidade na Síndrome Coronariana Aguda Utilizando Redes Neurais Artificiais com Base em Variáveis Clínicas e Genéticas.Tese de Doutorado em Medicina (Clínica Médica/ Pesquisa Clínica) FM/UFRJ

Sackett, D.L. – 2001- Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand). Canadian Medical Association Journal, 165(9), 1226-1237.

Soares, J. F., Siqueira, A. L. -1999- Introdução à Estatística Médica, UFMG.

Terzi, F.V.O., Pedrosa, R.C., Siqueira Filho, A.G., Nascimento, E.M., Pereira, B.B. -2010- Alterações contráteis segmentares e sua associação com arritmias ventriculares complexas, em pacientes chagásicos com eletrocardiograma normal ou "bordeline". Revista da Sociedade Brasileira de Medicina Tropical , v.43, p.557 - 561.

Torsten, H., Kurt H. and Achim Z. -2006- Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651–674.

Tura,B.R. -2001- Aplicações do Data Mining em Medicina. Dissertação de Mestrado (Bioestatística), NESC/UFRJ

Vidal, R.I.O.-2012- Fatores Prognósticos para Sobrevivência e Recorrência do Carcinoma Hepato-Celular Pós-Transplante Hepático em Pacientes Portadores de Infecção pelo Vírus da Hepatite C: Modelo Multinomial com LASSO e de Riscos Competitivos. Dissertação de Mestrado em Medicina (Ciências Cirúrgicas), FM/UFRJ.

Bibliografia Recomendada

Malley,J.D; Malley,K.G. ;Pajevic,S -2011- Statistical Learning for Biomedical Data, Cambridge University Press

Manly, B.J,F,-2008- Métodos Estatísticos Multivariados , 3ª Ed. Artmed Bookman.

Mathews, D.E.; Farewell, V.T.- 2007- Using and Understanding Medical Statistics , 4 th Ed. Krager Pub.