Sistemas Inteligentes Aplicados

Carlos Hall

Programa do Curso

- Limpeza/Integração de Dados
- Transformação de Dados
  - Discretização de Variáveis Contínuas
  - Transformação de Variáveis Discretas em Contínuas
  - Transformação de Variáveis Contínuas
- Análise e Seleção de Variáveis (Redução de Dados)

Análise e Seleção de Variáveis

- Parte de uma área chamada de Redução de Dados
- Obtenção de uma representação reduzida em volume mas que produz resultados de análise idênticos ou similares
- Melhora o desempenho dos modelos de aprendizado
- Objetivo: Eliminar atributos redundantes ou irrelevantes

Análise e Seleção de Variáveis

- Métodos Dependentes do Modelo (Wrapper)
- Métodos Independentes do Modelo (Filter)

Análise e Seleção de Variáveis

- Métodos Independentes do Modelo (Filter)
  - Tipo do Atributo de Saída (Tipo de Aplicação)
    - Saída Contínua (Ex.: Previsão, Inferência, etc)
    - Saída Discreta (Ex.: Classificação)
  - Tipo do Atributo de Entrada
    - Entrada Contínua
    - Entrada Discreta

Métodos Independentes do Modelo

- Entrada Contínua / Saída Contínua
  - Correlação Cruzada
  - PCA modificado
  - Least Squares Estimator (LSE)
  - Single Input Effectiveness (SIE)
- Entrada Contínua / Saída Discreta
  - Teste de Student (A)

Métodos Independentes do Modelo

- Entrada Discreta / Saída Contínua
  - Teste de Student (B)
  - Testes para Entrada Contínua / Saída Contínua, após se transformar o atributo de entrada discreto em contínuo
- Entrada Discreta / Saída Discreta
  - Teste do χ^2

Correlação Cruzada

- Método aplicável a entradas contínuas / saída contínua
- Mede relação linear entre variável de entrada e variável de saída
- Caso haja relação fortemente não-linear, não dará bons resultados
- Pode indicar também o atraso (dead-time, delay) entre as variáveis

Correlação Cruzada

- Definições:
  - Matriz de dados de entrada: X $[m \text{ x } k]$
  - j-ésima variável de entrada: x_j $[m \text{ x } 1]$
  - Variável de saída: y $[m \text{ x } 1]$
  - Função de correlação cruzada: c_j $[2m\text{-}1 \text{ x } 1]$

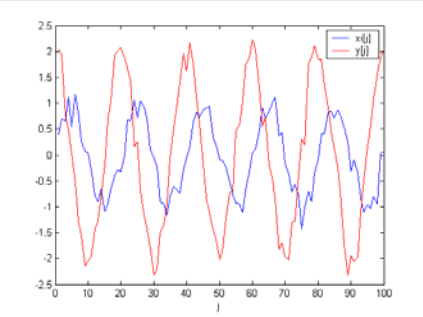
Correlação Cruzada

- Definições:
  - Matriz de dados de entrada: X $[m \text{ x } k]$
  - j-ésima variável de entrada: x_j $[m \text{ x } 1]$
  - Variável de saída: y $[m \text{ x } 1]$
  - Função de correlação cruzada: c_j $[2m\text{-}1 \text{ x } 1]$

$$c_j[\tau] = \frac{\sum_{i=1}^{m} \left(x_j[i-\tau] - \mu_j\right) \cdot \left(y[i] - \mu_y\right)}{m \sigma_x \sigma_y}, \quad \tau = -m \ldots m$$

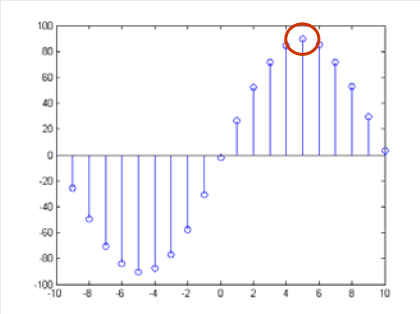
Correlação Cruzada

- Exemplo:



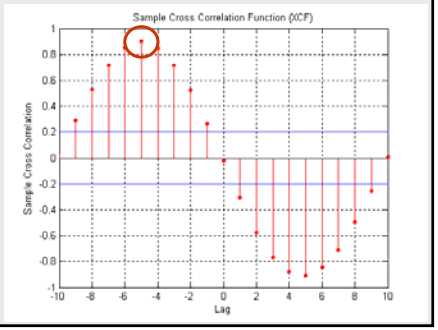
Correlação Cruzada

- No Matlab (1):
  - c = xcorr(xi,y);
  - stem(-99:99,c)
  - xlim([-10 10])



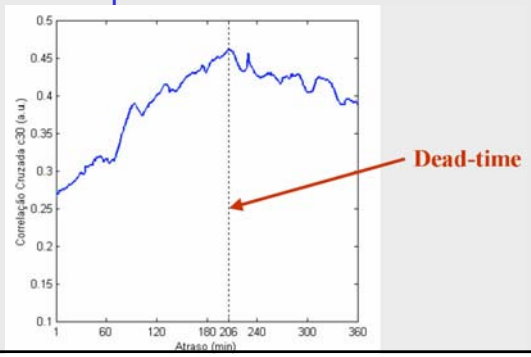
Correlação Cruzada

- No Matlab (2):
  - crosscorr(xi,y);
  - xlim([-10 10])



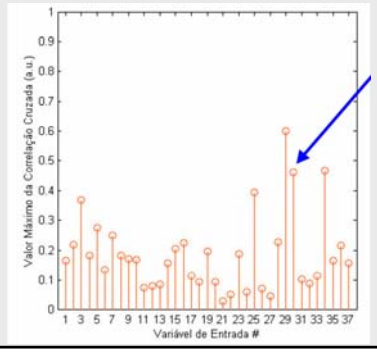
Correlação Cruzada

- Outro exemplo:



Correlação Cruzada

- Para todas as variáveis:



PCA modificado

- Método aplicável a entradas contínuas / saída contínua
- PCA tradicional é uma transformação de coordenadas, que pode ser usado para redução de dados
- Um modificação no algoritmo permite utilizá-lo como método de seleção de variáveis

PCA modificado

- PCA original:
  - Matriz de dados de entrada: $X\ [n \text{ x } k]$
  - Decompõe-se como:

$$\boldsymbol{X} = \boldsymbol{v}_1 \boldsymbol{p}_1^T + \boldsymbol{v}_2 \boldsymbol{p}_2^T + \ldots + \boldsymbol{v}_k \boldsymbol{p}_k^T$$

  - Define-se uma quantidade $L \leq k$

$$\boldsymbol{X} = \boldsymbol{v}_1 \boldsymbol{p}_1^T + \boldsymbol{v}_2 \boldsymbol{p}_2^T + \ldots + \boldsymbol{v}_L \boldsymbol{p}_L^T + \boldsymbol{E}, \quad L \leq k$$

PCA modificado

- PCA original:

$$\boldsymbol{X} = \boldsymbol{v}_1 \boldsymbol{p}_1^T + \boldsymbol{v}_2 \boldsymbol{p}_2^T + \ldots + \boldsymbol{v}_L \boldsymbol{p}_L^T + \boldsymbol{E}, \quad L \leq k$$

  - Vetores de loading: $p_j\ [k \text{ x } 1], j = 1 \ldots L$
  - Vetores de score: $v_j\ [n \text{ x } 1], j = 1 \ldots L$
  - Matriz de loadings: $\boldsymbol{P}\ [k \text{ x } L]$
  - Matriz de scores: $\boldsymbol{V}\ [m \text{ x } L]$

$$\boldsymbol{v}_j = \boldsymbol{X} \cdot \boldsymbol{p}_j \longrightarrow \boldsymbol{V} = \boldsymbol{X} \cdot \boldsymbol{P}$$

PCA modificado

- PCA original:
  - A matriz $\boldsymbol{X}\ [m \text{ x } k]$ é substituída pela matriz $\boldsymbol{V}\ [m \text{ x } L]$
  - Problema: perde-se o sentido físico com as novas variáveis $v_j\ [n \text{ x } 1], j = 1 \ldots L$
  - Isto porque o PCA tradicional é um método de redução de dimensionalidade, e não de seleção de variáveis
  - PCA modificado: implica em um seleção sobre as variáveis originais

PCA modificado

- PCA modificado:
  - Primeiro componente principal: $p_1\ [k \text{ x } 1]$
  - Cada elemento de p_1 indica o peso da variável original x_j na combinação linear que define a variável modificada v_1.
  - Maior elemento em p_1: maior importância

PCA modificado

- PCA modificado:
  1. Calcular PCA para matriz $\boldsymbol{X}$ $[m \text{ x } k]$
  2. Definir o número L de variáveis desejadas
  3. Selecionar a variável original x_j que corresponde ao maior elemento do vetor p_1
  4. Remover a coluna x_j da matriz $\boldsymbol{X}$ $[m \text{ x } k]$, gerando a matriz $\boldsymbol{Z}$ $[m \text{ x } k\text{-}1]$
  5. Regredir a variável x_j na matriz $\boldsymbol{Z}$, e calcular a matriz residual $\boldsymbol{E}$
  6. Redefinir a matriz $\boldsymbol{X} = \boldsymbol{E}$
  7. Retornar ao item 3 até selecionar L variáveis

PCA modificado

- PCA modificado:
  1. Calcular PCA para matriz $\boldsymbol{X}$ $[m \text{ x } k]$
  2. Definir o número L de variáveis desejadas
  3. Selecionar a variável original x_j que corresponde ao maior elemento do vetor p_1
  4. Remover a coluna x_j da matriz $\boldsymbol{X}$ $[m \text{ x } k]$, gerando a matriz $\boldsymbol{Z}$ $[m$
  5. **Regredir** $\boldsymbol{W} = \boldsymbol{Z}^T \cdot x_j / \left(x_j^T \cdot x_j\right)$ $\quad \boldsymbol{E} = \boldsymbol{Z} - x_j \cdot \boldsymbol{W}^T$ $\boldsymbol{Z}$, e calcular a matriz residual $\boldsymbol{E}$
  6. Redefinir a matriz $\boldsymbol{X} = \boldsymbol{E}$
  7. Retornar ao item 3 até selecionar L variáveis

PCA modificado

- PCA modificado:
  - Mantém as variáveis originais
  - Menor compactação que o PCA original
  - Possível colinearidade entre as variáveis originais

Least Squares Estimator (LSE)

- Método aplicável a entradas contínuas / saída contínua
- Não supõe relação linear entre entrada e saída
- Lineariza possíveis relações não-lineares
- Busca expressar o comportamento da variação Δy da variável de saída y em função das variações Δx das diversas variáveis de entrada x

Least Squares Estimator (LSE)

- Descrição resumida:
  - Seja um sistema com n entradas x_i (i=1..n) e uma saída y
  - Δy: vetor que contém as variações da variável y
  - Δx_i: vetor que contém as variações da variável x_i
  - Seja a função F abaixo:

$$F = \Delta y = b_1 \Delta x_1 + b_2 \Delta x_2 + \ldots + b_n \Delta x_n$$

  - Os coeficientes b_i indicam a importância, ou relevância, da variável x_i em relação à saída y, no sentido estatístico
  - Os coeficientes b_i são calculados pelo método dos mínimos quadrados

Least Squares Estimator (LSE)

- Algoritmo:
  - Seja uma função diferenciável y que descreve um sistema de n entradas e uma saída

$$y = f(x_1, x_2, x_3, \ldots, x_n) \qquad [x_1, x_2, x_3, \ldots, x_n]^T \in [0,1]^n$$

  - Suponha que há disponível um conjunto de p "pares" de dados desta função (amostra)

$$[x_1^j, x_2^j, x_3^j, \ldots, x_n^j, y^j]^T, \quad j = 1 \ldots p$$

Least Squares Estimator (LSE)

- Algoritmo:
  - Sejam o j-ésimo e o k-ésimo valores de saída, respectivamente, y_j e y_k
  - Seja um ponto fixo arbitrário: $[X_1, X_2, X_3, \ldots, X_n]^T$
  - Expansão em Série de Taylor:

$$y_j = f(X_1, X_2, X_3, \ldots X_n) + \sum_{i=1}^{n} \left[\left(\frac{\partial f}{\partial x_i} \Big|_{x_i = X_i} \right) \cdot (x_i^j - X_i) \right] + r_j$$

Least Squares Estimator (LSE)

- Algoritmo:
  - Expansões em Série de Taylor:

$$y_j = f(X_1, X_2, X_3, \ldots X_n) + \sum_{i=1}^{n} \left[\left(\frac{\partial f}{\partial x_i} \Big|_{x_i = X_i} \right) \cdot (x_i^j - X_i) \right] + r_j$$

$$y_k = f(X_1, X_2, X_3, \ldots X_n) + \sum_{i=1}^{n} \left[\left(\frac{\partial f}{\partial x_i} \Big|_{x_i = X_i} \right) \cdot (x_i^k - X_i) \right] + r_k$$

Least Squares Estimator (LSE)

- Algoritmo:
  - r_j, r_k: resíduos de alta ordem, podem ser ignorados sem risco de perder muita informação se

$$\left|(x_i^j - X_i)\right| \le 1 \qquad \left|(x_i^k - X_i)\right| \le 1$$

  - Ou seja, os dados têm que estar normalizados pela faixa de variação!

Least Squares Estimator (LSE)

- Algoritmo:
  - Subtraindo as expressões:

$$y_j - y_k = \sum_{i=1}^{n}\left[b_i \cdot (x_i^j - x_i^k)\right]$$

onde

$$b_i = \frac{\partial f}{\partial x_i}\Big|_{x_i = X_i}$$

Least Squares Estimator (LSE)

- Algoritmo:
  - Subtraindo as expressões:

$$y_j - y_k = \sum_{i=1}^{n}\left[b_i \cdot (x_i^j - x_i^k)\right]$$

onde

$$b_i = \frac{\partial f}{\partial x_i}\Big|_{x_i = X_i}$$

Mas quem é a função f ???

Least Squares Estimator (LSE)

- Na prática:
  - Considerando dois índices, j e k, pode-se definir um "vetor variação" como:

$$[x_1^j - x_1^k, x_2^j - x_2^k, x_3^j - x_3^k, \ldots, x_n^j - x_n^k, y^j - y^k]^T$$

  - Base de dados contém p pares de dados
  - Existe portanto uma quantidade de "vetores variação" dada por $m = C_2^p$
  - Essa quantidade pode ser muito grande!

Least Squares Estimator (LSE)

- Na prática:
  - Somente q (<<m) vetores variação são selecionados aleatoriamente
  - Pode-se então reescrever a expressão anterior de forma matricial:

$$\Delta y = \Delta x_1 \cdot b_1 + \Delta x_2 \cdot b_2 + \ldots + \Delta x_n \cdot b_n$$

onde Δy [q x 1]
Δx_i [q x 1], i = 1...n
b_i [1 x 1], i = 1...n

Least Squares Estimator (LSE)

- Na prática:
  - Somente q (<<m) vetores variação são selecionados aleatoriamente
  - Pode-se então reescrever a expressão anterior de forma matricial:

$$\Delta y = \Delta X \cdot b$$

onde Δy [q x 1]
ΔX [q x n]
b [n x 1]

Least Squares Estimator (LSE)

- Na prática:
  - Solução do sistema: $b = \Delta X^{-1} \cdot \Delta y$
  - Problema: se $q > n$, não existe solução exata ou única para b (sistema sobredeterminado)
  - Solução: estimador por Mínimos Quadrados, utilizando a pseudo-inversa

$$b^* = (\Delta X^T \cdot \Delta X)^{-1} \cdot \Delta X^T \cdot \Delta y$$

Least Squares Estimator (LSE)

- Seleção de variáveis:

$$\Delta y = \Delta x_1 \cdot b_1 + \Delta x_2 \cdot b_2 + \ldots + \Delta x_n \cdot b_n$$

  - Cada coeficiente b_i indica o grau de importância da variável correspondente x_i
  - Os valores b_i podem ser positivos ou negativos
  - Define-se então:

$$impo(x_i) = |b_i| \Big/ \sum_{j=1}^{n} |b_j| \quad \longrightarrow \quad \sum_{i=1}^{n} impo(x_i) = 1$$

Single Input Effectiveness (SIE)

- Método aplicável a entradas contínuas / saída contínua
- Tradução: Efetividade de uma Entrada Isolada
- Calcula o grau de efetividade de cada entrada em relação à saída
- Estes graus definem um ranking das entradas

Single Input Effectiveness (SIE)

- Contudo, o método pressupõe uma relação linear entre as entradas e a saída, e então aplica métodos da álgebra linear
- Assim, inicialmente é necessário estimar uma matriz de transferência G, de modo que $y = G \cdot x$
- Caso haja uma relação não-linear (como na maioria dos casos), o método não é aplicável, e/ou seus resultados não são confiáveis

Teste de Student

- Método aplicável para entradas contínuas / saída discreta, ou para entradas discretas / saída contínua
- É um Teste de Hipótese, oriundo da área de Inferência Estatística
- Por simplicidade, apresentaremos o caso de variáveis discretas binárias
- No caso de variáveis discretas com mais categorias, deve-se utilizar o método ANOVA (Analysis of Variance)

Teste de Student

- Pressupõe que a variável discreta (entrada ou saída) divide os valores disponíveis da variável contínua (saída ou entrada) em dois grupos
- Cada grupo contém os valores contínuos que estão associados a um dos valores discretos

Teste de Student

- Exemplo: Base de Dados Meteorológicos

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Teste de Student

- Exemplo: Base de Dados Meteorológicos

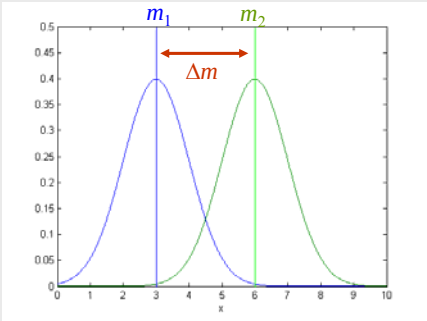
Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Teste de Student

- No exemplo: Base de Dados Meteorológicos
  - Variável de entrada Temperatura:
    - **Saída Não: 85, 80, 65, 72, 71 (média 74,6)**
    - **Saída Sim: 83, 70, 68, 64, 69, 75, 75, 72, 81 (média 73,0)**
  - Variável de entrada Umidade:
    - **Saída Não: 85, 90, 70, 95, 91 (média 86,2)**
    - **Saída Sim: 86, 96, 80, 65, 70, 80, 70, 90, 75 (média 79,1)**
- As diferenças entre as médias são significativas??

Teste de Student

- A diferença entre as médias é significativa?



Teste de Student

- Premissas:
  - Separa-se os valores da variável contínua que são correspondentes às duas categorias da variável discreta, formando assim duas variáveis contínuas distintas, x_1 e x_2
  - Existem n_1 amostras da variável x_1, e n_2 amostras da variável x_2
  - Essas variáveis têm médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, respectivamente
  - As variáveis têm distribuição normal
  - As variâncias são iguais ($\sigma_1^2 = \sigma_2^2$)
    - Caso não sejam, deve-se aplicar o teste de Welch

Teste de Student

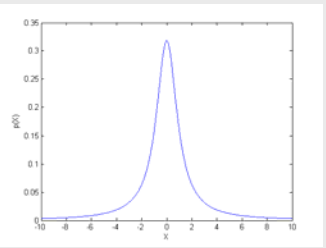
- Hipóteses:
  - H_0: $\mu_1 = \mu_2$ ($\mu_1 - \mu_2 = 0$) Hipótese Nula
  - H_1: $\mu_1 \neq \mu_2$ ($\mu_1 - \mu_2 \neq 0$) Hipótese Alternativa
  - Objetivo do Teste de Hipótese: rejeitar a hipótese nula!
  - Estatística de Teste (Welch): $$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Teste de Student

- Hipóteses:
  - H_0: $\mu_1 = \mu_2$ ($\mu_1 - \mu_2 = 0$) Hipótese Nula
  - H_1: $\mu_1 \neq \mu_2$ ($\mu_1 - \mu_2 \neq 0$) Hipótese Alternativa
  - Objetivo do Teste de Hipótese: rejeitar a hipótese nula!
  - Estatística de Teste (Student): $$t = \frac{|\mu_1 - \mu_2|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

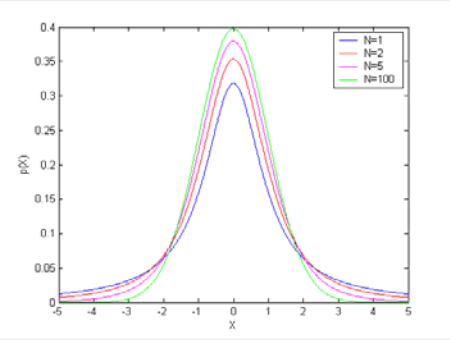
Teste de Student

- Distribuição de Student:
  - Família de distribuições, definidas pelo número de graus de liberdade, N
  - Matlab: p = tpdf(x,N)
    - x = -10:0.01:10;
    - p = tpdf(x,1);
    - plot(x,p)



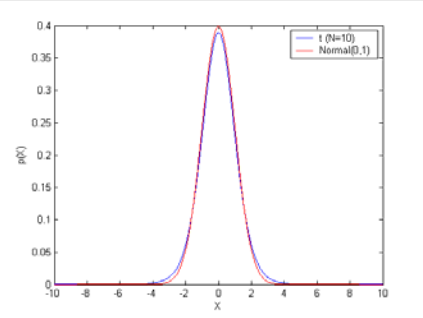
Teste de Student

- Distribuição de Student:



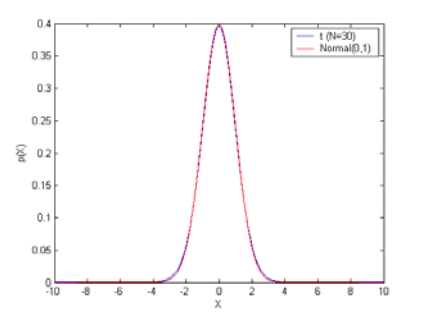
Teste de Student

- Distribuição de Student x Normal:



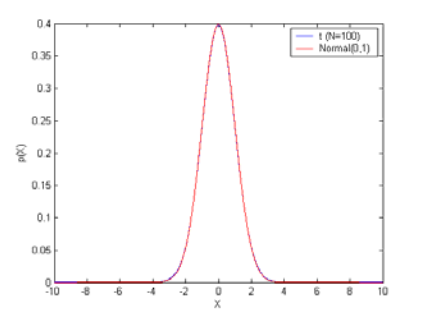
Teste de Student

- Distribuição de Student x Normal:



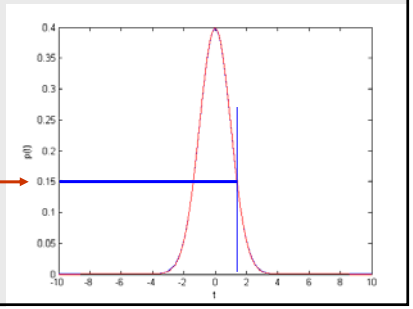
Teste de Student

- Distribuição de Student x Normal:



Teste de Student

- Valor p:
  - Número de graus de liberdade: N = $n_1 + n_2 - 2$
  - p = tpdf(t, N);



Teste de Student

- Interpretação do Valor p:
  - O valor p indica a probabilidade de que a diferença observada entre as médias tenha ocorrido por acaso
  - Quanto menor o valor p, maior a probabilidade de que as médias das variáveis sejam realmente diferentes
  - Normalmente trabalha-se com um limiar em 5%, ou seja, valores p menores que 5% (0,05) indicam significância estatística na diferença observada

Teste de Student

- No exemplo: Base de Dados Meteorológicos
  - Variável de entrada Temperatura:
    - Saída Não: x_1 = {85, 80, 65, 72, 71}
      $n_1 = 5$
      $\mu_1 = 74,6$
      $\sigma_1^2 = 62,3$
    - Saída Sim: x_2 = {83, 70, 68, 64, 69, 75, 75, 72, 81}
      $n_2 = 9$
      $\mu_2 = 73,0$
      $\sigma_2^2 = 38,0$

Teste de Student

- No exemplo: Base de Dados Meteorológicos
  - Variável de entrada Temperatura:
    - Saída Não: x_1 = {85, 80, 65, 72, 71}
      $n_1 = 5$
      $\mu_1 = 74,6$
      $\sigma_1^2 = 62,3$
    - Saída Sim: x_2 = {83,
      $n_2 = 9$
      $\mu_2 = 73,0$
      $\sigma_2^2 = 38,0$

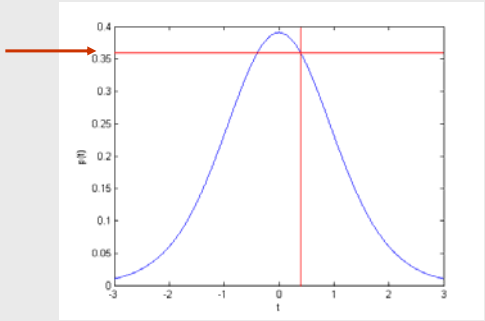
$$t = \frac{|74,6 - 73,0|}{\sqrt{\frac{62,3}{5} + \frac{38}{9}}} = 0,3917$$

$$N = 5 + 9 - 2 = 12$$

$$p = tpdf(0,3917,\ 12) = 0,3598 = 35,98\%$$

Teste de Student

- Variável de entrada Temperatura:



Teste de Student

- No exemplo: Base de Dados Meteorológicos
  - Variável de entrada Umidade:
    - Saída Não: $x_1 = \{85, 90, 70, 95, 91\}$
      $n_1 = 5$
      $\mu_1 = 86{,}2$
      $\sigma_1^2 = 94{,}7$
    - Saída Sim: $x_2 = \{86, 96, 80, 65, 70, 80, 70, 90, 75\}$
      $n_2 = 9$
      $\mu_2 = 79{,}1$
      $\sigma_2^2 = 104{,}4$

Teste de Student

- No exemplo: Base de Dados Meteorológicos
  - Variável de entrada Umidade:
    - Saída Não: $x_1 = \{85, 90, 70, 95, 91\}$
      $n_1 = 5$
      $\mu_1 = 86,$
      $\sigma_1^2 = 94$
    - Saída Sim: $x_2 = \{86$
      $n_2 = 9$
      $\mu_2 = 79{,}1$
      $\sigma_2^2 = 104{,}4$

$$t = \frac{|86{,}2 - 79{,}1|}{\sqrt{\frac{94{,}7}{5} + \frac{104{,}4}{9}}} = 1{,}2848$$

$N = 5 + 9 - 2 = 12$

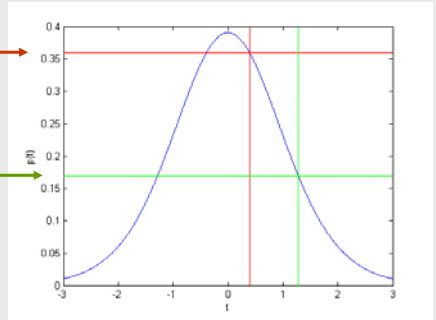
$p = tpdf(1{,}2848,\ 12) = 0{,}1691 = 16{,}91\%$

Teste de Student

- No exemplo: Base de Dados Meteorológicos:

Temperatura

Umidade



Teste de Student

- No exemplo: Base de Dados Meteorológicos:
  - Temperatura: p = 0,3598 = 35,98%
  - Umidade: p = 0,1691 = 16,91%
  - Em relação ao limiar de 5%, nenhuma das variáveis de entrada é significativa (ou relevante)
  - Contudo, pode-se dizer que a variável umidade provavelmente é mais significativa que a variável temperatura

Teste do χ^2

- Método aplicável a entradas discretas / saída discreta
- Também é um Teste de Hipótese
- Baseado na construção de Tabelas de Contingência (também chamadas de Matrizes de Confusão)
- Aplicável a qualquer número de categorias para cada variável discreta

Teste do χ^2

- Exemplo: Base de Dados Meteorológicos

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Teste do χ^2

- Exemplo: Base de Dados Meteorológicos

Valores da Variável Vento para cada Classe da Variável Jogar:

Não Jogar (5 casos): Não, Sim, Sim, Não, Não
- **Vento Não: 3 casos**
- **Vento Sim: 2 casos**

Jogar (9 casos): Não, Não, Não, Sim, Não, Não, Sim, Sim, Não
- **Vento Não: 6 casos**
- **Vento Sim: 3 casos**

Teste do χ^2

- Tabela de Contingência para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	3	6	9
	Sim	2	3	5
	Total	5	9	14

Teste do χ^2

- Hipóteses:
  - H_0: As freqüências das linhas e colunas são independentes
  - H_1: As freqüências das linhas e colunas são dependentes
  - Objetivo do Teste de Hipótese: rejeitar a hipótese nula!
  - Calcula-se inicialmente qual seria a tabela de contingência no caso da hipótese nula ser verdadeira

Teste do χ^2

- Tabela de Contingência Observada para o caso geral com duas categorias:

		Saída		
		Não	Sim	Total
Entrada	Não	O_{11}	O_{12}	TL_1
	Sim	O_{21}	O_{22}	TL_2
	Total	TC_1	TC_2	T

Teste do χ^2

- Tabela de Contingência Esperada para o caso geral com duas categorias:

		Saída		
		Não	Sim	Total
Entrada	Não	E_{11}	E_{12}	TL_1
	Sim	E_{21}	E_{22}	TL_2
	Total	TC_1	TC_2	T

Teste do χ^2

- Tabela de Contingência Esperada para o caso geral com duas categorias:

		Saída		
				Total
Entrada	Não	$E_{ij} = TL_i \times TC_j / T$		TL_1
	Sim	E_{21}	E_{22}	TL_2
	Total	TC_1	TC_2	N

Teste do χ^2

- Tabela de Contingência Observada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	3	6	9
	Sim	2	3	5
	Total	5	9	14

Teste do χ^2

- Tabela de Contingência Observada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	3	6	9
	Sim	2	3	5
	Total	5	9	14

$T = 14$

$TL_1 = 9$
$TL_2 = 5$

$TC_1 = 5$
$TC_2 = 9$

Teste do χ^2

- Tabela de Contingência Esperada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não			9
	Sim			5
	Total	5	9	14

$E_{11} = 9 \times 5 / 14 = 3,2$
$E_{12} = 9 \times 9 / 14 = 5,8$
$E_{21} = 5 \times 5 / 14 = 1,8$
$E_{22} = 5 \times 9 / 14 = 3,2$

Teste do χ^2

- Tabela de Contingência Observada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	3	6	9
	Sim	2	3	5
	Total	5	9	14

Teste do χ^2

- Tabela de Contingência Esperada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	3,2	5,8	9
	Sim	1,8	3,2	5
	Total	5	9	14

Teste do χ^2

- Hipóteses:
  - H_0: As freqüências das L linhas e C colunas são independentes
  - H_1: As freqüências das L linhas e C colunas são dependentes
  - Estatística de Teste (χ^2): $$\chi^2 = \sum_{i=1}^{N_i}\sum_{j=1}^{N_j}\frac{(E_{ij}-O_{ij})^2}{E_{ij}}$$

Teste do χ^2

- Tabela de Contingência Esperada para a variável Vento:

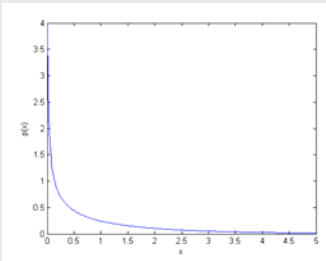
$$\chi^2 = \frac{(3,2-3,0)^2}{3,2}+\frac{(5,8-6,0)^2}{5,8}+\frac{(1,8-2,0)^2}{1,8}+\frac{(3,2-3,0)^2}{3,2}$$

		Jogar?		
		Não	Sim	Total
Vento?	Não	3,2	5,8	9
	Sim	1,8	3,2	5
	Total	5	9	14

$\chi^2 = 0,0541$

Teste do χ^2

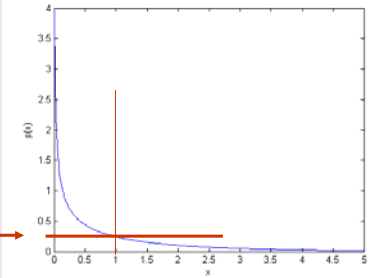
- Distribuição de χ^2:
  - Família de distribuições, definidas pelo número de graus de liberdade, N
  - Matlab: p = chi2pdf(x,N)
    - x = 0:0.01:5;
    - p = chi2pdf(x,1);
    - plot(x,p)



Teste do χ^2

- Valor p:
  - Número de graus de liberdade: N = (L-1)x(C-1)
  - p = chi2pdf(chi2, N);

Valor p →

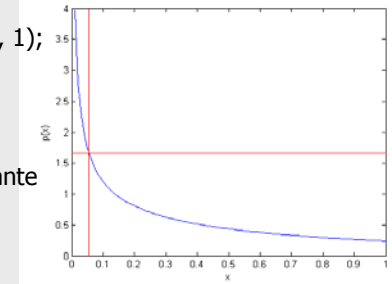


Teste de Student

- Interpretação do Valor p:
  - O valor p indica a probabilidade de que a dependência observada entre as freqüências tenha ocorrido por acaso
  - Quanto menor o valor p, maior a probabilidade de que as freqüências das variáveis sejam realmente dependentes
  - Normalmente trabalha-se com um limiar em 5%, ou seja, valores p menores que 5% (0,05) indicam significância estatística na dependência observada

Teste do χ^2

- No exemplo:
  - Número de graus de liberdade: N = (2-1)x(2-1) = 1
  - χ^2 = 0,0541
  - p = chi2pdf(0,0541, 1);
  - p = 1,67 = 167%
  - Ou seja, a variável Vento NÃO É relevante



Teste do χ^2

- Exemplo alterado:

Valores da Variável Vento para cada Classe da Variável Jogar:

Não Jogar (5 casos): Sim, Sim, Sim, Sim, Não
Vento Não: 1 caso
Vento Sim: 4 casos

Jogar (9 casos): Não, Não, Não, Não, Não, Não, Sim, Sim, Não
Vento Não: 7 casos
Vento Sim: 2 casos

Teste do χ^2

- Tabela de Contingência Observada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	1	7	8
	Sim	4	2	6
	Total	5	9	14

Teste do χ^2

- Tabela de Contingência Observada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não			8
	Sim			6
	Total	5	9	14

$E_{11} = 8 \times 5 / 14 = 3,2$
$E_{12} = 8 \times 9 / 14 = 5,8$
$E_{21} = 6 \times 5 / 14 = 1,8$
$E_{22} = 6 \times 9 / 14 = 3,2$

Teste do χ^2

- Tabela de Contingência Esperada para a variável Vento:

		Jogar?		
		Não	Sim	Total
Vento?	Não	2,9	5,1	8
	Sim	2,1	3,9	6
	Total	5	9	14

Teste do χ^2

- Tabela de Contingência Esperada para a variável Vento:

$$\chi^2 = \frac{(2,9-1,0)^2}{2,9} + \frac{(5,1-7,0)^2}{5,1} + \frac{(2,1-4,0)^2}{2,1} + \frac{(3,9-2,0)^2}{3,9}$$

Vento?	Não	$\chi^2 = 4,60$	5,1	8
				6
				14

p = chi2pdf(0,0541, 1);
p = 0,0187 = 1,87%

Ou seja, a variável Vento É relevante