# CIÊNCIA DE DADOS - 04

Prof. Júlio Cesar Nievola

PPGIa – PUCPR

25/maio/2019

# O que é Exploração de Dados?

**Uma exploração preliminar dos dados para compreender melhor suas características.**

- Motivações-chave da exploração de dados incluem
  - Ajudar na seleção da técnica correta para pré-processamento ou análise
  - Fazer uso das habilidades humanas de reconhecimento de padrões
    - Pessoas podem reconhecer padrões não capturados pelas técnicas de análise de dados

- Relacionado à área de Análise Exploratória de Dados (EDA)
  - Criada pelo estatístico John Tukey
  - Livro seminal: Exploratory Data Analysis escrito por Tukey
  - Uma boa introdução online pode ser encontrada no capítulo 1 do NIST Engineering Statistics Handbook

http://www.itl.nist.gov/div898/handbook/index.htm

# Técnicas Usadas na Exploração de Dados

- Em EDA, como originalmente definido por Tukey
  - O foco está na visualização
  - Agrupamento e detecção de anomalias eram vistos como técnicas exploratórias
  - Em mineração de dados, agrupamento e detecção de anomalias são grandes áreas de interesse, e não são vistas apenas como exploração
- Nesta discussão de exploração de dados, o foco está em
  - Estatística Sumária
  - Visualização
  - Online Analytical Processing (OLAP)

# Conjunto de Dados Iris

- Muitas das técnicas de exploração de dados são ilustradas com o conjunto de dados da planta Iris.
  - Pode ser obtido do UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - Criada pelo estatístico Douglas Fisher
  - Três tipos de flores (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Quatro atributos (não-classes)
    - Sepal width e length
    - Petal width e length

**Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.**

# Tipos de tabelas

- Tabelas de frequências:
  - Absolutas
  - Relativas
  - Acumuladas

| Quantidade de carros por dia | | | |
|---|---|---|---|
| Data | Quantidade | % | Acumulada |
| 01/jan | 517 | 8,51 | 8,51 |
| 02/jan | 615 | 10,12 | 18,63 |
| 03/jan | 334 | 5,50 | 24,13 |
| 04/jan | 815 | 13,42 | 37,55 |
| 05/jan | 730 | 12,02 | 49,56 |
| 06/jan | 482 | 7,93 | 57,50 |
| 07/jan | 651 | 10,72 | 68,21 |
| 08/jan | 981 | 16,15 | 84,36 |
| 09/jan | 553 | 9,10 | 93,47 |
| 10/jan | 397 | 6,53 | 100,00 |
| TOTAL | 6075 | 100 | 100 |

# Estatística Sumária

- Estatística Sumária são números que resumem as propriedades dos dados
  - Propriedades sumarizadas incluem frequência, posição e dispersão
    - Exemplos:
      - Posição – média
      - Dispersão – desvio padrão
  - A maioria das estatísticas sumárias pode ser calculada em um único passo através dos dados

# Frequência e Moda

- A frequência do valor de um atributo é a percentagem do tempo em que o valor ocorre no conjunto de dados
  - Por exemplo, dado o atributo ‘gênero’ e uma população representativa de pessoas, o gênero ‘feminino’ ocorre cerca de 50% do tempo
- A moda de um atributo é o valor mais frequente do atributo
- As noções de frequência e moda são usadas tipicamente com dados categóricos

# Percentil

- Para dados contínuos, a noção de percentil é mais útil.

- Dados um atributo contínuo ou ordinal $x$ e um número $p$ entre 0 e 100, o *p-ésimo* percentil é um valor $x_p$ de $x$ tal que $p$% dos valores observados de x são menores que $x_p$.

- Por exemplo, o percentil 50 é o valor $x_{50\%}$ tal que 50% de todos os valores de $x$ são menores que $x_{50\%}$.

# Medidas de posição: Média e Mediana

- A média é a medida mais comum de posição de um conjunto de pontos.
- Entretanto, a média é muito sensitiva a outliers.
- Então, a mediana ou uma mediana ajustada é usada comumente.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

# Propriedades da média

- A soma dos desvios de uma amostra é sempre igual a zero

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

- A soma dos quadrados dos desvios em relação à própria média de uma variável é sempre igual a um valor mínimo

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \Rightarrow mínimo$$

# Características da moda

| Vantagens | Desvantagens |
|---|---|
| Fácil de calcular | Pode estar afastada do centro dos dados |
| Não é afetada pelos dados dos extremos da amostra | Difícil de incluir em funções matemáticas |
| Pode ser aplicada em qualquer escala: nominal, ordinal, intervalar e proporcional | Não utiliza todos os dados da amostra |
| | A amostra pode ter mais de uma moda |
| | Algumas amostras podem não ter moda |

# Características da mediana

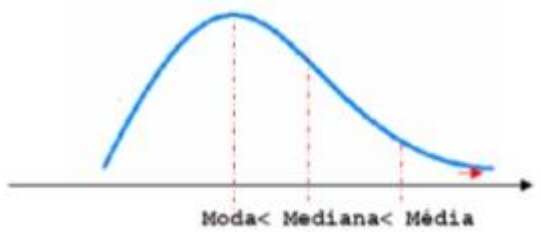
| Vantagens | Desvantagens |
| --- | --- |
| Fácil de calcular | Difícil de incluir em funções matemáticas |
| Não é afeta pelos dados extremos da amostra | Não utiliza todos os dados da amostra |
| É um valor único | |
| Pode ser aplicada nas escalas: ordinal, intervalar e proporcional | |

# Características da média

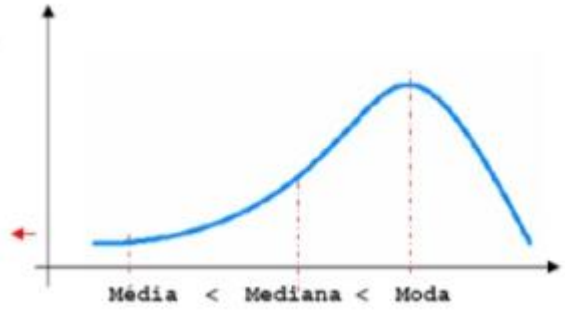
| Vantagens | Desvantagens |
| --- | --- |
| Fácil de compreender e aplicar | É afetada pelos dados extremos da amostra |
| Utiliza todos os dados da amostra | É necessário conhecer todos os dados da amostra |
| É um valor único | |
| Fácil de incluir em funções matemáticas | |
| Pode ser aplicada nas escalas: intervalar e proporcional | |

# Relação entre moda, mediana e média

- Para uma distribuição simétrica, os três valores são iguais
- Se os dados tem distribuição para a direita (ou positiva), a moda representa o pico e a mediana (dividindo a distribuição em duas áreas iguais) está a sua direita e tem-se a relação:
  - Média > Mediana > Moda (embora nem sempre haja moda)




- Se os dados tem distribuição para a esquerda (ou negativa), a moda representa o pico e a mediana (dividindo a distribuição em duas áreas iguais) está a sua esquerda e tem-se a relação:
  - Média < Mediana < Moda (embora nem sempre haja moda)



# Medidas de Dispersão: Faixa e Variância

- Faixa é a diferença entre o máximo e mínimo
- A variância ou desvio padrão é a medida mais comum de desvio de um conjunto de pontos.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Entretanto, elas são sensitivas a outliers, e outras medidas são freqüentemente utilizadas.

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

# Teorema de Chebyshev

- Para qualquer conjunto de dados de uma amostra ou população, a proporção mínima de valores que se encontram dentro de *k* desvios padrão ao redor da média é pelo menos igual a $1 - \frac{1}{k^2}$, sendo *k* uma constante maior que 1.

# Regra prática - Chebyshev

- Em uma distribuição simétrica em forma de sino, a porcentagem de dados contidos no intervalo de um desvio padrão ao redor da média é de 68%. Para uma distribuição assimétrica com acentuada inclinação para um lado, essa porcentagem se aproxima de 90%.
- Em uma distribuição simétrica em forma de sino, a porcentagem de dados contidos no intervalo de dois desvios padrão ao redor da média é de 95%. Para uma distribuição assimétrica com acentuada inclinação para um lado, essa porcentagem se aproxima de 100%.
- Para todas as distribuições, a porcentagem de dados contidos no intervalo de três desvios padrão ao redor da média é próximo de 100%.

# CONCEITOS LIGADOS A TESTES DE HIPÓTESE

Prof. Júlio Cesar Nievola

PPGIa – PUCPR

# Teorema Central do Limite

- Parâmetro: é uma medida numérica que descreve uma população.

- Estatística: é uma medida numérica que descreve uma amostra.

- Teorema Central do Limite:
  - Se de uma população com parâmetros $(\mu, \sigma)$ for retirada uma amostra de tamanho $n$ suficientemente grande, a distribuição de $\bar{X}$ será aproximadamente normal $N\left(\mu, {}^{\sigma}/_{\sqrt{n}}\right)$, qualquer que seja a forma da distribuição da população.

# Hipóteses

- A *hipótese nula*, chamada de $H_0$, é a hipótese sobre a qual devem ser obtidas evidências para poder rejeitá-la.
- A *hipótese alternativa, chamada de* $H_1$, é a hipótese sobre a qual devem ser obtidas evidências para poder aceitá-la.
- A *hipótese nula* e a *hipótese alternativa* descrevem dois possíveis estados mutuamente excludentes, pois as duas hipóteses não podem ser aceitas ou rejeitadas ao mesmo tempo:
  - A *hipótese nula* $H_0$ é o valor correntemente aceito até que se tenham evidências de que esse valor não é mais correto. A hipótese nula é uma afirmação ou ponto de partida do teste de hipóteses;
  - A *hipótese alternativa* $H_1$ será somente aceita se surgirem evidências de que o valor da hipótese nula não é mais correto.

# Teste de Hipótese com uma ou duas caudas

- Um teste de hipótese em uma cauda da distribuição é um teste no qual a hipótese alternativa $H_1$ define a mudança em alguma direção da hipótese nula $H_0$, incluindo na especificação um dos símbolos "$\leq$" ou "$\geq$".

- Um teste de hipótese em duas caudas da distribuição é um teste no qual a hipótese alternativa $H_1$ define uma mudança da hipótese nula $H_0$ sem especificar nenhuma direção, incluindo na especificação o símbolo "$\neq$".

# Intervalo de confiança e *p*-valor

- *Intervalo de confiança* é o intervalo de valores que contém a média da população com uma determinada probabilidade de acerto. O intervalo de valores é construído de uma amostra aleatória retirada da população.

- O *p*-valor é definido como a probabilidade de qualquer média da amostra ser mais extrema do que a média da amostra $\bar{X}$ extraída para o teste, sem rejeitar a hipótese nula, ou seja, tem-se que:
  - O *p*-valor é o nível de significância observado;
  - Se o *p*-valor for maior ou igual a $\alpha$, então a hipótese nula não será rejeitada;
  - Se o *p*-valor menor ou igual a $\alpha$, então a hipótese nula será rejeitada;
  - Quanto menor for o *p*-valor, mais forte deverá ser a evidência para poder rejeitar a hipótese nula.

# Erros nos Testes de Hipótese

- Se $H_0$ for rejeitada, o teste de hipóteses não afirma que $H_0$ seja falsa e sim sugere que há evidências de que $H_0$ seja falsa.

- Nesta situação, o que se pode afirmar é que para o nível de significância $\alpha$ espera-se que, se o teste fosse repetido um grande número de vezes, a proporção de acertos seja (1- $\alpha$) das vezes.

| | $\boldsymbol{H_0}$ verdadeira | $\boldsymbol{H_0}$ falsa |
|---|---|---|
| Não rejeita $\boldsymbol{H_0}$ | Decisão correta | Erro tipo II |
| Rejeita $\boldsymbol{H_0}$ | Erro tipo I | Decisão correta |

# Tipos de Erros

- Erro do tipo I: Acontece quando a hipótese nula é rejeitada, embora seja realmente verdadeira. Este tipo de erro pode ser reduzido diminuindo o nível de significância $\alpha$.
- Erro do tipo II: Acontece quando a hipótese nula não é rejeitada, embora ela seja falsa. Aumentando-se o nível de significância $\alpha$ aumenta-se a chance de cometer um erro do tipo II.

| | Quando $\boldsymbol{H_0}$ for verdadeira | Quando $\boldsymbol{H_0}$ for falsa |
|---|---|---|
| Probabilidade de não rejeitar $\boldsymbol{H_0}$ | $1 - \alpha$ | $\beta$ |
| Probabilidade de rejeitar $\boldsymbol{H_0}$ | $\alpha$ | $1 - \beta$ |

# Exigências dos Testes Paramétricos

- Para que os testes estatísticos paramétricos (tais como correlação, índices z, testes *t*) possam ser aplicados, é necessário que um conjunto de condições sejam obedecidas pelas amostras, isto é, as amostras devem:
  - Ser obtidas aleatoriamente de uma população com distribuição normal;
  - Consistir de observações independentes, exceto para valores pareados;
  - Consistir de valores em uma escala de medidas de tipo intervalar ou de razão;
  - Ter populações com distribuição aproximadamente iguais;
  - Ser relativamente grandes (em geral 30 amostras);
  - Ter um comportamento aproximadamente normal.

# Passos para Teste de Hipóteses

1. Estabelecer a hipótese nula e a hipótese de pesquisa;
2. Definir o nível de risco (ou nível de significância) associado à hipótese nula;
3. Escolher a estatística de teste apropriada;
4. Calcular a estatística de teste;
5. Determinar o valor necessário para rejeição da hipótese nula usando a tabela de valores críticos apropriada para a estatística escolhida;
6. Comparar o valor obtido com o valor crítico;
7. Interpretar os resultados;
8. Relatar os resultados.

# Tipo de teste em função das características das amostras

| Tipo de Análise | Teste Não-Paramétrico | Teste Paramétrico |
|---|---|---|
| Comparar duas amostras relacionadas | Teste de Postos Sinalizados de Wilcoxon | Teste *t* para amostras dependentes |
| Comparar duas amostras não relacionadas | Teste U de Mann-Whitney | Teste *t* para amostras independentes |
| Comparar três ou mais amostras relacionadas | Teste de Friedman | Repeated Measures Analysis of Variance (ANOVA) |
| Comparar três ou mais amostras não relacionadas | Teste H de Kruskal-Wallis | One-way Analysis of Variance (ANOVA) |
| Comparar dados categóricos | Teste Chi-Quadrado / Teste Exato de Fisher | NÃO HÁ |
| Comparar duas variáveis ordenadas por postos | Correlação posição-ordem de Spearman | Correlação produto-momento de Pearson |
| Comparar duas variáveis em que uma delas é discreta dicotômica | Correlação ponto-bi-serial | Correlação produto-momento de Pearson |
| Comparar duas variáveis em que uma delas é contínua dicotômica | Correlação bi-serial | Correlação produto-momento de Pearson |
| Avaliar a aleatoriedade de uma amostra | Testes de execução | NÃO HÁ |