Revisão de Estatística e Probabilidade

Magnos Martinello

Universidade Federal do Espírito Santo - UFES
Departamento de Informática – DI
Laboratório de Pesquisas em Redes Multimidia – LPRM

tatística descritiva X inferência estatística

- Estatística descritiva : métodos que envolvem a coleta, a apresentação e a caracterização de um conjunto de dados de modo a descrever apropriadamente as várias características deste conjunto;
- Inferência estatística : métodos que possibilitam a estimativa de uma característica de uma população ou a tomada de decisão referente à população com base somente em resultados de amostra;

População x Amostra

- População (ou universo) : conjunto de indivíduos sobre os quais se quer obter informações
  - Totalidade dos itens ou objetos.
- Exemplo: Todas as empresas do Brasil, censo do IBGE
- Amostra: parte selecionada da população.
- Exemplo: empresas brasileiras de capital aberto
  - Informações imprecisas devido à variação amostral
  - Imprecisão compensada pela viabilidade
    Ex: pesquisas eleitorais, análise de ativos (retorno X risco)
- Amostra: deve ser representativa da população

Estatística Descritiva

- Objetivos

1. Explicar propriedades numéricas dos dados

2. Descrever medidas de :
   - Tendência central
   - Variação
   - Forma

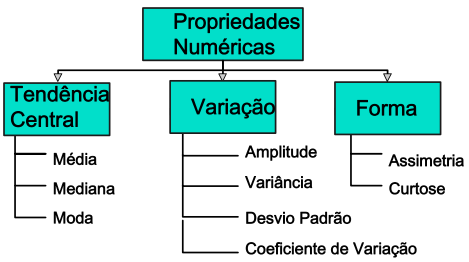
Caracterizando um conjunto de dados

- Quando temos um conjunto de observações, em geral há 2 tipos de medidas que nos interessam: medidas de posição e medidas de dispersão ou variação.

- Entre as medidas de posição, temos especial interesse nas medidas de tendência central: média, mediana e moda.

- Entre as medidas de dispersão, as mais importantes são a variância e o desvio-padrão.

Propriedades e Medidas



Medidas de Tendência Central

- Média
  - Medida mais comum
  - Atua como um 'Centro de Gravidade'
  - Suponha uma amostra de n observações X_1, X_2,..., X_n. A média aritmética, é definida como:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Prof. Magnos Martinello – UFES

Medidas de Tendência Central

- Exemplo : A cotação diária do dólar para venda ao longo da semana foi: 2,11 ; 2,09 ; 2.11 ; 2.07 ; 2.16.

A) Qual é a média ?

B) Se a única informação que você tem é que uma pessoa vendeu 10 dólares nesta semana, qual o melhor palpite que você pode formar quanto ao retorno em reais desta venda?

É afetada por valores extremos ('Outliers')

Prof. Magnos Martinello – UFES

Medidas de Tendência Central

- Mediana
  - Valor do meio da seqüência ordenada
    - *Assim, 50% das observações estão abaixo da mediana, enquanto 50% estão acima.*
    - *n é ímpar, a mediana é o valor numérico de posição (n+1)/2, na amostra ordenada.*
    - *n é par, a mediana é a média das duas observações do meio da amostra ordenada.*

Prof. Magnos Martinello – UFES

Medidas de Tendência Central

- Exemplo : A cotação diária do dólar para venda ao longo da semana foi: 2,11 ; 2,09 ; 2.11 ; 2.07 ; 2.16.

A) Qual é a mediana?

B) Considere os seguintes valores: 2,4,6,8,80. Se você quer uma medida que melhor reflita estes dados, você escolheria a média ou a mediana?

C) E se os valores forem 4,4,6,40,40?

Mediana não é afetada por valores extremos ('Outliers')

Medidas de Tendência Central

- Moda
  - Valor que ocorre com maior frequência em um conjunto de dados
  - Não é afetada por valores extremos
  - Pode existir mais de uma moda

Medidas de Tendência Central

- Exemplo : A cotação diária do dólar para venda ao longo da semana foi: 2,11 ; 2,09 ; 2.11 ; 2.07 ; 2.16.

A) Qual é a moda?

B) Considere os seguintes valores: 2,4,6,8,80.

Qual é a moda?

C) E se tivermos os seguintes valores observados 4,4,6,40,40?

Qual é a moda ?

- Moda é afetada por valores extremos ?

Resumo: medidas de tendência central

Medida	Equação	Descrição
Média	$\Sigma X_i / n$	Centro de Gravidade
Mediana	$\frac{(n+1)}{2}$ Posição	Valor do meio quando ordenado
Moda	max f(X)	Mais Freqüente

Necessidade de medidas de dispersão

- Considere uma empresa exportadora de frangos. Esta empresa esta preocupada com as variações na cotação do dólar e deseja obter uma medida que quantificasse estas variações.

  1. O cálculo de medidas de tendência central vai ajudar esta firma neste objetivo?

  Amostra 1 - {3,90; 3,89; 3,88; 3,91; 3,89; 3,87; 3,90; 3,88; 3,92}
  Amostra 2 - {3,88; 3,84; 3,90; 3,90; 3,93; 3,97; 3,86; 3,81; 3,95}

Necessidade de medidas de dispersão

- Ambas as amostras têm a mesma média, mediana e moda. Mas note que a amostra 2 é mais espalhada que a amostra 1.
- É, portanto, importante desenvolvermos medidas de dispersão.

Medidas de Dispersão ou Variação

- Medidas de dispersão indicam como valores estão distribuídos em torno de um determinado ponto
- As medidas que vamos estudar aqui, medem como as variáveis estão distribuídas em torno da média: variância e desvio padrão.

Qual a importância das medidas de dispersão?

- Quanto maior o desvio padrão e/ou a variância, maior vai ser o intervalo onde os valores de uma variável aleatória pode cair.
- Neste caso, significa que a média não tem muita informação acerca dos valores de uma realização da variável aleatória.

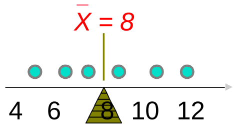
Exemplo: Lucro e Risco

- Pense no caso do lucro. Se o lucro de uma empresa tiver muita dispersão, significa que ele pode ser muito alto ou muito baixo.
- Isto significa dizer que, por exemplo, um investimento nas ações desta empresa pode trazer um ganho elevado ou uma perda elevada.

Medidas de Dispersão

- Indicam a variação em torno da média

$$Var(X) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \qquad S = \sqrt{Var(X)}$$



Necessidade de medidas de dispersão

- Qual é a variância e o desvio padrão de cada amostra ?

  Amostra 1 - {3,90; 3,89; 3,88; 3,91; 3,89; 3,87; 3,90; 3,88; 3,92}

  Amostra 2 - {3,88; 3,84; 3,90; 3,90; 3,93; 3,97; 3,86; 3,81; 3,95}

Voltando ao Exemplo

- Variância: Amostra 1 = 0,000222

  Amostra 2 = 0,0024

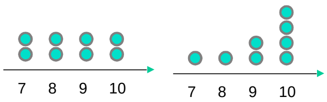
  Difícil de interpretar porque, se a amostra é medida em reais, a variância é medida em reais ao quadrado.
- Desvio-Padrão: Amostra 1 = 0,0149

  Amostra 2 = 0,0490

  As oscilações da amostra em torno da média são de, aproximadamente, 1 centavo e meio na amostra 1 e cinco centavos na amostra 2.

Amplitude

1. Diferença entre a menor e a maior observação
2. Mede a dispersão total no conjunto de dados
3. Ignora a maneira como os dados são distribuídos



Quantis

- Tanto a média quanto o desvio padrão são afetados, de forma exagerada, por valores extremos;
- Podemos definir uma medida, chamada quantil de ordem p denotada por q(p), onde p é uma proporção qualquer, 0< p<1, tal que 100p% das observações sejam menores do que q(p)
- Exemplos
  - q(0.25) = primeiro quartil = 25 percentil
  - q(0.5) = mediana = 5 decil = 50 percentil

Coeficiente de Variação

- Medida de dispersão **relativa**
- Sempre em %
- Mostra a variação relativa à média
- Ao contrário da variância ou do desvio-padrão, ele não depende da unidade de medida.

$$CV = \frac{S}{\bar{X}} \times 100\%$$

- Evidentemente, se a média aritmética for zero ou muito próxima de zero, o coeficiente de variação deixa de ser uma medida útil.

Resumo de medidas de dispersão

Medida	Equação	Descrição
Amplitude	$X_{maior} - X_{menor}$	Dispersão total
Desvio Padrão	$\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n}}$	Dispersão em torno da média
Variância	$\dfrac{\Sigma(X_i - \bar{X})^2}{n}$	Dispersão quadrada em torno da média
Coef. de Variação	$(S / \bar{X}).100\%$	Variação Relativa

Propriedades e Medidas

- Propriedades Numéricas
  - Tendência Central
    - Média
    - Mediana
    - Moda
  - Variação
    - Amplitude
    - Variância
    - Desvio Padrão
    - Coeficiente de Variação
  - Forma
    - Assimetria
    - Curtose

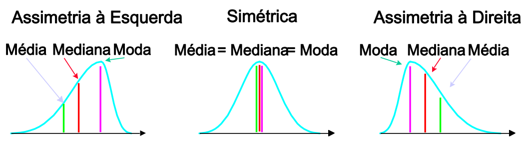
Medidas de Forma

Assimetria

- Descreve como os dados estão distribuídos em cada lado da média
- O coeficiente de assimetria de uma distribuição simétrica em relação à média é nulo.

$$a_3 = \frac{m_3}{\sqrt{m_2^3}}$$

$$m_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k}{n}$$

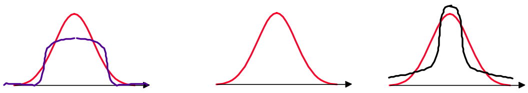


Coeficiente de Curtose

- Mede o nível de achatamento em relação a uma distribuição Normal
- A curtose de uma Normal vale 3: distribuição **mesocúrtica**
- Menor que 3: distribuição **platicúrtica**
- maior que 3: distribuição **leptocúrtica**

$$\boxed{a_4 = \frac{m_4}{m_2^2}}$$

platicúrtica mesocúrtica leptocúrtica



Prof. Magnos Martinello – UFES