Medidas de Tendência Central e Dispersão

Guilherme de Alencar Barreto

gbarreto@ufc.br

Grupo de Aprendizado de Máquinas – GRAMA
Programa de Pós-Graduação em Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
www.researchgate.net/profile/Guilherme_Barreto2/

Conteúdo da Apresentação

1. Objetivo Geral
2. Medidas de Tendência Central e Dispersão
3. Exemplos no Matlab/Octave

Objetivo Geral

Objetivo Geral da Aula

Introduzir conceitos de medidas de tendência central e de dispersão visando o uso como descritoras quantitativas de um conjunto de observações.

Parte I

Medidas de Tendência Central e Dispersão

Introdução

- São necessários dois tipos de medida para descrever adequadamente um conjunto de dados.
- Medidas de tendência central fornecem informação quanto ao “meio” de um conjunto de números.
- A média é uma medida de tendência central.
- Medidas de dispersão indicam se os valores estão relativamente próximos uns dos outros ou separados.
- A variância ou o desvio-padrão são medidas de dispersão.
- Porém existem várias outras medidas, cujas as mais importantes iremos detalhar a seguir.

Medidas de Tendência Central

- Medidas de tendência central são usadas para indicar um valor que tende a **tipificar**, ou a representar melhor, um conjunto de números.
- As três medidas mais usadas são a **média**, **mediana** e a **moda**.
- Mas existem ainda várias outras de interesse para o analista de dados, tais como média podada, média ponderada, média harmônica e média geométrica.

Média Aritmética (MA ou $\bar{x}$)

- A média de um conjunto de números (amostra ou dados amostrais), $\{x_1, x_2, ..., x_n\}$, é representada pelo símbolo $\bar{x}$ (leia-se "x barra"), sendo calculada pela seguinte expressão

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}} \tag{1}$$

- A Eq. (1) é um caso especial da média de uma variável aleatória discreta já discutida na Eq. (**??**)s.
- A média aritmética é a idéia que ocorre à maioria das pessoas quando se fala em "média".
- A média aritmética possui certas propriedades matemáticas convenientes, sendo por isso a mais importante das medidas de tendência centra que estudaremos.

Propriedades da Média Aritmética

1. A média de um conjunto de números pode sempre ser calculada e é única.
2. A média é afetada por todos os valores do conjunto: se um valor se modifica, a média também se modifica.
3. Somando-se (ou subtraindo-se, ou multiplicando-se, ou dividindo-se) uma constante a cada valor do conjunto, a média ficará aumentada do valor dessa constante.
4. A *soma dos desvios* dos números de um conjunto em relação à média é zero:

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x}) = 0}$$

Média Ponderada

- A fórmula anterior para calcular a média aritmética supõe que cada observação tenha a mesma importância. Embora este seja o caso mais geral, há exceções. Nestes casos, usamos a fórmula da média ponderada:

$$MP = \boxed{\text{média ponderada} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}}, \qquad (2)$$

em que $w_i > 0$ é o peso do i-ésimo valor.

Média Ponderada (exemplos)

- Suponha que em uma turma de alunos haverá 3 avaliações no semestre. As duas primeiras têm peso 0,3 (i.e. valem 30% do total de pontos do curso) e a última tem peso 0,4.
- Um aluno com as seguintes notas: 80 (primeiro exame), 90 (segundo exame) e 96 (exame final), terá a seguinte nota final:

$$MP = \frac{0,30(80) + 0,30(90) + 0,40(96)}{0,30 + 0,30 + 0,40} = 89,4$$

- Em outro curso há um exame no meio do semestre e um exame final, este com o dobro do peso daquele. Um aluno com notas 95 (primeiro exame) e 89 (exame final) terá a seguinte média final:

$$MP = \frac{1(95) + 2(89)}{1 + 2} = 91,0$$

Mediana

- A **mediana** de um conjunto de números é o valor que é maior que uma metade dos valores e menor que a outra metade.
- O algoritmo para determinar a mediana é o seguinte:
  1. Ordenar os valores.
  2. Verificar se há um número ímpar ou par de valores.
  3. Para um número ímpar de valores,a mediana é o valor do meio. Para um número par de valores, a mediana é a média dos dois valores do meio.
- Por exemplo, a mediana do conjunto $\{5, 6, 8\}$ é 6, pois é valor do meio. Para o conjunto $\{7, 8, 9, 10\}$ é $(8 + 9)/2 =$ 8,5.

Mediana versus Média Aritmética

- A média é influenciada por $cada$ valor do conjunto, em especial pelos valores extremos (i.e. valores muito baixos ou muito altos).
- Por outro lado, a mediana é relativamente **insensível** aos valores extremos.
- A média possui certas propriedades matemáticas que a tornam atraente.
- A etapa de ordenação dos dados no cálculo da mediana requer custo computacional adicional, podendo torna-se alta para grande quantidade de dados.
- O cálculo da mediana não dá para ser feito com um máquina de calcular simples, ao contrário do que ocorre com a média aritmética.

Moda

- A **moda** é o valor que ocorre com mais freqüência em um conjunto de números.
- Por exemplo, para o conjunto $\{10, 10, 8, 6, 10\}$, a moda é igual a 10, pois o mesmo ocorre 3 vezes.
- Comparando com a média e a mediana, a moda é a menos útil das medidas de tendência central, porque não se presta à análise matemática, ao contrário do que ocorre com as outras duas medidas.
- De um ponto de vista puramente descritivo, a moda indica o valor "típico" em termos da maior ocorrência.
- Quando todos ou quase todos os valores ocorrem aproximadamente com a mesma freqüência,a moda nada acrescenta em termos de descrição dos dados.

Média Truncada

- A **média truncada** (ou podada) é a média aritmética dos valores restantes após o descarte de uma certa quantidade dos maiores e menores valores do conjunto.
- Esta definição de média é útil porque sabe-se que a média aritmética é muito influenciada por valores extremos.
- Por exemplo, para o conjunto $\{2, 5, 3, 3, 4, 7, 12\}$, a média aritmética é igual a 5,1429, enquanto a média truncada é igual 4,4, após o descarte do menor valor (2) e do maior valor (12).

Média Geométrica

- A média geométrica (MG) de um conjunto de n números é obtida por meio da seguinte fórmula:

$$\boxed{MG = \sqrt[n]{\prod_{i=1}^{n} x_i} = (x_1 \cdot x_2 \cdots x_n)^{\frac{1}{n}}}. \tag{3}$$

- A média geométrica de um conjunto de números é sempre menor ou igual à média aritmética dos membros desse conjunto.
- As duas médias são iguais se e somente se todos os membros do conjunto são iguais.

Média Harmônica

- A média harmônica (MH) de um conjunto de n números é obtida por meio da seguinte fórmula:

$$\boxed{MH = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}}, \qquad (4)$$

  para $x_i > 0$ para todo x_i
- A MH nunca é maior do que a MG ou do que a MA (i.e. $MH \leq MG \leq MA$).
- Utiliza-se a MH quando se esté tratando de observações de grandezas inversamente proporcionais, tais como velocidade e tempo.

Exemplo 1: Média Harmônica ou Aritmética?

Determinar a velocidade média de um automóvel que viajou metade da **distância** a 40 Km\h e a outra metade a 60 Km\h.

Solução:

Pode-se mostrar que a solução desse problema é dada pela média harmônica das velocidades em cada metade do percurso, ou seja:

$$\bar{v} = \frac{2}{\frac{1}{40} + \frac{1}{60}} = \frac{2 \cdot 40 \cdot 60}{40 + 60} = 48 \text{ Km}\backslash \text{ h} \tag{5}$$

Exemplo 2: Média Harmônica ou Aritmética?

Determinar a velocidade média de um automóvel que viajou metade do **tempo** do percurso a 40 Km\h e a outra metade a 60 Km\h.

Solução:

Neste caso, pode-se mostrar que a solução desse problema é dada pela média aritmética das velocidades em cada metade do tempo, ou seja:

$$\bar{v} = \frac{40 + 60}{2} = 50 \text{ Km}\backslash \text{ h} \tag{6}$$

Intervalo

- O **intervalo** de um grupo de números é a medida de dispersão mais simples de calcular e entender, pois focaliza o maior (x_{max}) e o menor (x_{min}) valor no conjunto .
- O intervalo pode ser expresso de duas maneiras:
  1. A *diferença* entre o maior e o menor valor: $x_{max} - x_{min}$.
  2. O maior e o menor valor no conjunto: $[x_{min}, x_{max}]$.
  3. Por exemplo, para o conjunto $\{1, 10, 25\}$, a diferença entre o maior e o menor valor é 25-1=24. Alternativamente, pode-se dizer que o intervalo de valores vai de 1 a 25.
  4. Note que o mero conhecimento de que um intervalo de um conjunto de números é 44 não nos diz nada a respeito dos números em si.
  5. Porém, dizer que o intervalo vai de 300 a 344 já nos dá uma informação adicional sobre a magnitude dos números.

Desvio Médio Absoluto (DMA)

- O **desvio médio absoluto** (DMA) mede o desvio médio dos valores em relação à média do grupo, ignorando o sinal do desvio, sendo calculado pela seguinte fórmula:

$$\boxed{DMA = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}}, \qquad (7)$$

  em que $|u|$ é o valor absoluto (módulo) de u.
- Esta medida só tem significado devido à presença da função *valor absoluto*, caso contrário seu valor será sempre zero.

Existem funções prontas no Matlab/Octave que calculam todas as medidas de tendência central e dispersão vistas.

Medida	**Comando**
Média aritmética	`mean`
Média geométrica	`geomean`
Média harmônica	`harmmean`
Média truncada	`trimmean`
Mediana	`median`
Moda	`mode`
Variância	`var`
Desvio-padrão	`std`
Intervalo	`min` e `max`
Desvio médio absoluto	`mad`

Exemplo Numérico no Matlab/Octave

>> X = [2 5 3 3 4 7 12];

Comando	**Valor**
`mean(X)`	5.1429
`geomean(X)`	4.3660
`harmmean(X)`	3.7984
`trimmean(X,20)`	4.4000
`median(X)`	4
`mode(X)`	3
`var(X,1)`	10.1224
`std(X,1)`	3.1816
`min(X), max(X)`	2,12
`mad(X)`	2.4898