Métodos Estatísticos Básicos

Aula 3 - Medidas de Tendência Central e de Posição

Regis A. Ely

Departamento de Economia
Universidade Federal de Pelotas

06 de julho de 2020

Conteúdo

Medidas de tendência central

Medidas de tendência central são estatísticas que representam o ponto central de um conjunto de dados

- **Média**: os tipos mais comuns são a média *aritmética*, *harmônica* e *geométrica*
- **Mediana**: valor que separa a metade maior e a metade menor de um conjunto de dados
- **Moda**: valor mais frequente de um conjunto de dados

Média aritmética simples

Média aritmética: razão entre a soma dos valores e o número de observações de um conjunto de dados

- Para dados brutos, calculamos a *média aritmética simples*
  - $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, onde n é o número de observações

Exemplo no R: média aritmética de um conjunto com 7 valores

```
x <- c(10, 14, 13, 15, 16, 18, 12)
mean(x)
```

```
[1] 14
```

Desvio em relação à média

Desvio em relação à média: é a diferença entre um valor específico dos dados e a média aritmética do conjunto de todos os dados

- $d_i = X_i - \overline{X}$

Exemplo no R: desvios em relação à média do exemplo anterior

```
x - mean(x)
```

```
[1] -4  0 -1  1  2  4 -2
```

Média aritmética ponderada

Média aritmética ponderada: cada observação entra com um peso diferente no cálculo da média

- Para dados agrupados em uma tabela de frequência, calculamos a *média aritmética ponderada*[1]
- $\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i}$, onde f_i é o peso da observação i

Exemplo no R: dados utilizados na Aula 1 estão agrupados por classes

```r
library(HistData)
DrinksWages
```

[1] Se os dados estiverem agrupados em intervalos de classe, X_i sempre será o ponto médio da classe.

Média aritmética ponderada

A média do salário semanal de todos os indivíduos é obtida ponderando o salário (`wage`) pelo número de pessoas em cada profissão (`n`)

- A função `weighted.mean` faz isso no `R`[2]:

```
library(tidyverse)
DrinksWages %>%
  summarise("Salário Médio" = weighted.mean(wage, n))
```

```
  Salário Médio
1      24.59782
```

[2] Para esse cálculo utilizaremos o operador %>%, que é carregado com o pacote `tidyverse`. O livro escrito pelo criador deste pacote pode ser obtido `aqui`.

Propriedades da média aritmética

A média aritmética sempre respeita as seguintes propriedades[3]:

1. A soma algébrica dos desvios em relação à média é nula:
   - $\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (Xi - \overline{X}) = \sum_{i=1}^{n} Xi - n.\overline{X} = 0$
2. Somando ou subtraindo uma constante c a todos os elementos do conjunto de dados, a média também aumentará em c:
   - $\frac{\sum_{i=1}^{n} (Xi + c)}{n} = \frac{\sum_{i=1}^{n} Xi}{n} + \frac{n.c}{n} = \overline{X} + c$
3. Multiplicando (ou dividindo) todos os valores por uma constante c, a média será multiplicada (ou dividida) por c:
   - $\frac{\sum_{i=1}^{n} (c.Xi)}{n} = \frac{c \sum_{i=1}^{n} Xi}{n} = c.\overline{X}$

[3] A média aritmética ponderada também respeita estas propriedades, porém deve-se sempre multiplicar as expressões pelos pesos f_i.

Propriedades da média aritmética

Exemplo no `R`: testando as propriedades da média aritmética

```r
## Soma dos desvios em relação à média
sum(mtcars$hp - mean(mtcars$hp))
```

```
[1] 0
```

```r
## Média de hp+10 é igual à média de hp mais 10?
all.equal(mean(mtcars$hp + 10), mean(mtcars$hp) + 10)
```

```
[1] TRUE
```

```r
## Média de hp*10 é igual à média de hp vezes 10?
all.equal(mean(mtcars$hp * 10), mean(mtcars$hp) * 10)
```

```
[1] TRUE
```

Média geométrica simples

Média geométrica: é a raíz n-ésima do produtório dos dados

- $\overline{X}_g = \sqrt[n]{\prod_{i=1}^{n} X_i}$, onde $\prod_{i=1}^{n} X_i = X_1 \times X_2 \times \cdots \times X_n$

O logaritmo da média geométrica é igual à média aritmética dos logaritmos:

- $log\overline{X_g} = \frac{1}{n}\sum_{i=1}^{n} logX_i$

Logo, a média geométrica é uma média aritmética suavizada:

- É muita utilizada em finanças, para calcular médias de taxas de juros

Média geométrica ponderada

Média geométrica ponderada: cada observação tem um peso diferente no cálculo (útil para dados em tabela de frequência)

- $\overline{X}_g = \sqrt[\sum_{i=1}^{n} f_i]{\prod_{i=1}^{n} X_i^{f_i}} = \sqrt[\sum_{i=1}^{n} f_i]{X_1^{f_1} \times X_2^{f_2} \times \cdots \times X_n^{f_n}}$

Exemplo no R: média geométrica simples da taxa Selic em 2020[4]

```r
library(lubridate) # Para lidar com datas
library(psych) # Para calcular média geométrica
url_bcb <- "http://api.bcb.gov.br/dados/serie/bcdata.sgs.1178/dados?formato=csv"
selic <- read_csv2(url_bcb) %>% # Ler dados da taxa Selic anualizada
  mutate(data = dmy(data)) %>% # Criar variável data com padrão dia/mês/ano
  filter(data >= "2020-01-01") # Filtrar ano de 2020 em diante
geometric.mean(selic$valor) # Calcular média geométrica simples da taxa Selic
```

```
[1] 3.514823
```

[4] Ao invés de usar a função `geometric.mean`, pode-se obter o mesmo resultado com `prod(x)^(1/length(x))`

Propriedades da média geométrica

A média geométrica sempre respeita as seguintes propriedades:

1. A média geométrica é sempre menor ou igual a média aritmética, sendo igual apenas no caso em que todas as observações tem o mesmo valor
2. O produtório dos dados permanece inalterado se todas as observações forem substituídas pelo valor da média geométrica
3. A média geométrica da razão das observações em duas séries é igual à razão de suas médias geométricas
4. A média geométrica da multiplicação de duas séries é igual a multiplicação de suas respectivas médias geométricas

Propriedades da média geométrica

Exemplo no `R`: propriedades da média geométrica para dados fictícios

```
x <- c(1, 4, 9, 12, 14, 19)
y <- c(2, 4, 7, 11, 20, 32)
geometric.mean(x) < mean(x)
```

```
[1] TRUE
```

```
all.equal(prod(x), prod(rep(geometric.mean(x), length(x))))
```

```
[1] TRUE
```

```
all.equal(geometric.mean(x/y), geometric.mean(x)/geometric.mean(y))
```

```
[1] TRUE
```

```
all.equal(geometric.mean(x*y), geometric.mean(x)*geometric.mean(y))
```

```
[1] TRUE
```

Média harmônica simples

Média harmônica: É o inverso da média aritmética dos inversos de cada observação

- $\overline{X}_h = (\frac{1}{n} . \sum_{i=1}^{n} X_i^{-1})^{-1} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$
- Algumas utilizações comuns da média harmônica incluem:
  1. Na física, com dados de velocidade e tempo
  2. No cálculo de scores padronizados de concursos
  3. Em finanças, para calcular médias de razões contábeis

Exemplo no `R`: média harmônica simples da variável `x`, criada no exemplo anterior

```
harmonic.mean(x) # ou 1/mean(1/x)
```

```
[1] 3.8253
```

Média harmônica ponderada e propriedades

Se os dados estiverem agrupados por distribuição de frequência, deve-se utilizar a *média harmônica ponderada*:

- $\overline{X}_h = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{X_i}}$

A média harmónica é sempre menor ou igual a geométrica, valendo a igualdade apenas se todas observações forem iguais. Podemos testar isso no R:

```r
harmonic.mean(x) < geometric.mean(x)
```

```
[1] TRUE
```

Moda simples

Moda: valor que ocorre com maior frequência em uma série de dados

- Se nenhum valor ocorre mais vezes do que outro, a série é amodal
- Se há valores mais frequentes que se repetem o mesmo número de vezes, a série tem mais de um valor modal
- A média aritmética possui maior estabilidade que a moda

Dados agrupados: com dados agrupados em frequência simples, é possível determinar a moda apenas olhando o dado com a maior frequência. Para dados agrupados em intervalos de classe, teremos mais de uma alternativa

Moda simples

Exemplo no `R`: Na base de dados `mtcars`, os dados não estão agrupados em frequências, mas para calcular a moda do número de carburadores, podemos utilizar a função `table` para agrupar os dados e então verificar os valores com maior frequência

```
table(mtcars$carb)
```

```
 1  2  3  4  6  8
 7 10  3 10  1  1
```

Pode-se ver que esta é uma série *bimodal*, sendo que o número de carburadores que aparece com mais frequência é 2 e 4

Moda simples

Exemplo no `R`: Na base de dados `DrinksWages`, os dados já estão agrupados em frequências, de acordo com as profissões dos indivíduos. Para calcular a moda dos salários, basta encontrar a linha com o maior valor da variável `n`

```
DrinksWages %>% filter(n == max(DrinksWages$n))
```

```
  class                trade sober drinks wage   n
1     A general labourer    71     85 18.5 156
```

A profissão mais frequente é a de *general labourer*, com salário semanal de 18.5 xelins

Moda com intervalos de classe

Quando os dados estão agrupados em intervalos de classe, a classe com a maior frequência é a classe modal. A moda será um valor compreendido entre os limites da classe modal. Veremos duas alternativas para calcular a moda:

1. **Moda bruta**: $M_o = (\frac{l^*+L^*}{2})$, onde l^* é o limite inferior da classe modal e L^* o limite superior da classe modal
2. **Moda de Czuber**: $M_c = l^* + (\frac{d_1}{d_1+d_2}).h^*$, onde d_1 é a frequência da classe modal menos a frequência da classe anterior a modal; d_2 é a frequência da classe modal menos a frequência da classe posterior a modal; e h^* é a amplitude da classe modal

Moda com intervalos de classe

Vamos calcular a moda bruta e a moda de Czuber utilizando o exemplo da Aula 2:

Idade	Frequência
21 $\vert\!-$ 23	4
23 $\vert\!-$ 25	7
25 $\vert\!-$ 27	5

A classe modal, que ocorre com maior frequência, é 23 $\vert\!-$ 25. Assim:

- $M_o = \frac{23+25}{2} = 24$
- $M_c = 23 + \frac{(7-4)}{(7-4)+(7-5)} \times 2 = 24,2$

Mediana simples

Mediana: é o valor que separa uma série (disposta em ordem crescente ou decrescente) em duas partes com o mesmo número de elementos

- Se a série tiver número ímpar de termos, a mediana será o elemento $\frac{n+1}{2}$
- Se a série tiver numero par de termos, a mediana será a média dos elementos $\frac{n}{2}$ e $\frac{n}{2}+1$
- Útil quando há valores extremos nos dados. *Ex: salário*

Considere os dados: $\{1, 3, 0, 0, 2, 4, 1, 2, 5\}$. Devemos:

1. Colocar a série em ordem crescente $\{0, 0, 1, 1, 2, 2, 3, 4, 5\}$
2. Como existem 9 elementos, a mediana será o elemento de número $\frac{n+1}{2} = \frac{10}{2} = 5$. Assim, $M_e = 2$

Mediana simples

Para dados agrupados em frequência simples, em ordem crescente:

- Se o somatório das frequências for ímpar, a mediana será o elemento $\frac{\sum_{i=1}^{n} fi+1}{2}$
- Se o somatório das frequências for par, a mediana será a média dos termos $\frac{\sum_{i=1}^{n} f_i}{2}$ e $\frac{\sum_{i=1}^{n} f_i}{2} + 1$

Exemplo no R: mediana simples do número de carburadores

```
median(mtcars$carb)
```

```
[1] 2
```

Mediana com intervalos de classe

Para calcularmos a mediana de dados agrupados em intervalos de classe, devemos:

1. Determinar as frequências acumuladas,
2. Calcular $\sum\limits_{i=1}^{n} fi/2$,
3. Marcar a classe correspondente a frequência acumulada imediatamente superior a $\sum\limits_{i=1}^{n} fi/2$. Essa será a classe mediana,
4. Calcular $M_e = l^* + \frac{[(\sum_{i=1}^{n} f_i/2 - FAA).h*]}{f*}$, onde $l*$ é o limite inferior da classe mediana, FAA é a frequência acumulada da classe anterior à classe mediana, $f*$ é a frequência simples da classe mediana, e $h*$ é a amplitude do intervalo da classe mediana

Mediana com intervalos de classe

Considere os seguintes dados agrupados em intervalos de classe:

Classes	fi	Fi
$50 \vdash 54$	4	4
$54 \vdash 58$	9	13
$58 \vdash 62$	11	24
$62 \vdash 66$	8	32
$66 \vdash 70$	5	37
$70 \vdash 74$	3	40
Total	40	

Mediana com intervalos de classe

Vamos aplicar os passos 1 a 4 no exemplo anterior:

1. As frequências acumuladas são dadas na coluna F_i
2. $\frac{\sum fi}{2} = 20$
3. *Classe mediana*: $58 \vdash 62$
4. Temos $l^* = 58$, $FAA = 13$, $f^* = 11$, $h^* = 4$, logo, $M_e = 58 + \frac{[(20-13)x4]}{11} = 58 + \frac{28}{11} = 60,54$

- Assim, a mediana é 60,54

Separatrizes

Medidas de posição são estatísticas que representam a tendência de concentração de um conjunto de dados ao redor de certos pontos

As medidas de posição mais usuais são chamadas *separatrizes* e incluem[5]:

- **Quartil**: são os valores que dividem os dados em quatro partes
  - São necessários 3 quartis para dividir os dados em quatro partes
- **Decil**: são os valores que dividem a série em dez partes
  - São necessários 9 decis para dividir os dados em dez partes
- **Percentil**: são os valores que separam a série em cem partes
  - São necessários 99 percentis para dividir os dados em cem partes

[5] Há uma equivalência entre as medidas de posição: $P_{50} = D_5 = Q_2 = M_e$

Quartil simples

Considere o seguinte conjunto de dados: $\{5, 2, 6, 9, 10, 13, 15\}$

- Para calcular os quartis, devemos:
    1. Ordenar a série $\{2, 5, 6, 9, 10, 13, 15\}$
    2. Calcular a mediana, que será o segundo quartil, $M_e = Q2 = 9$
    3. Dividir a série em dois grupos $\{2, 5, 6\}$ e $\{10, 13, 15\}$[6]
    4. Calcular os outros quartis como sendo as medianas desses dois grupos, $Q_1 = 5$ e $Q_2 = 13$

Exemplo no R: quartis da potência dos veículos na base `mtcars`

```r
quantile(mtcars$hp)
```

```
   0%   25%   50%   75%  100%
 52.0  96.5 123.0 180.0 335.0
```

[6] A divisão é sempre feita com base no valor da mediana

Quartil com dados agrupados

- Se os dados forem *agrupados sem intervalos de classe*, utilizamos $\frac{\sum_{i=1}^{n} f_i}{2}$ e $\frac{\sum_{i=1}^{n} f_i}{2} + 1$ para calcular as posições dos quartis
- Se os dados forem *agrupados com intervalos de classe*, utilizamos a mesma fórmula da mediana para calcular os quartis, entretanto substituímos $\frac{\sum_{i=1}^{n} f_i}{2}$ por $k\frac{\sum_{i=1}^{n} f_i}{4}$, sendo k o número do quartil:
  - $Q1 = l^* + \frac{[(\frac{\sum fi}{4} - FAA).h^*]}{f^*}$
  - $Q2 = l^* + \frac{[(2\frac{\sum fi}{4} - FAA).h^*]}{f^*}$
  - $Q3 = l^* + \frac{[(3\frac{\sum fi}{4} - FAA).h^*]}{f^*}$

Quartil com intervalos de classe

Seguindo os passos no exemplo utilizado para a mediana:

- $\frac{\sum fi}{2} = 20 \rightarrow$ *Classe mediana*: $58 \vdash 62$
- Temos $l* = 58$, $FAA = 13$, $f* = 11$, $h* = 4$, logo $M_e = Q_2 = 58 + \frac{[(20-13)x4)}{11} = 60,54$
- $\frac{\sum fi}{4} = 10 \rightarrow$ *Classe mediana do 1º grupo*: $54 \vdash 58$
- Logo, $Q1 = 54 + \frac{[(10-4)x4]}{9} = 56,66$
- $\frac{3.\sum fi}{4} = 30 \rightarrow$ *Classe mediana do 3º grupo*: $62 \vdash 66$
- Logo, $Q3 = 62 + \frac{[(30-24)x4]}{8} = 65$

Decil simples

O procedimento de cálculo dos decis é análogo aos quartis, porém agora o 5º decil será igual ao 2º quartil, que será igual à mediana
Exemplo no R: calcular o segundo, terceiro e oitavo decis da potência dos veículos na base `mtcars`

```
quantile(mtcars$hp, c(.2, .3, .8))
```

```
  20%   30%   80% 
 93.4 106.2 200.0 
```

Decil com intervalos de classe

O procedimento de cálculo dos decis para dados com intervalos de classe é o mesmo utilizado para os quartis, basta alterar o valor de K pelo número do decil e dividir $\sum fi$ por 10
Exemplo: calcular o 3º decil da tabela com intervalos de classe utilizada nos exemplos da mediana e quartil

- Como K=3, temos $3.\frac{\sum fi}{10} = 3.\frac{40}{10} = 12$, e a classe mediana é $54 \vdash 58$
- Logo, $D_3 = 54 + \frac{[(12-4)x4]}{9} = 57,55$

Percentil

No caso dos percentis, teremos $P_{50} = M_e$, $P_{25} = Q_1$ e $P_{75} = Q_3$

- O cálculo é análogo ao do quartil e decil, mas utilizando $k.\frac{\sum fi}{100}$

Exemplo no R: calcular os percentis 34, 67 e 95 da potência dos veículos na base de dados `mtcars`

```
quantile(mtcars$hp, c(0.34, 0.67, 0.95))
```

```
   34%    67%    95% 
109.54 175.00 253.55 
```

Exemplo no R

Podemos resumir todas as medidas de tendência central para a potência dos veículos na base `mtcars` usando um comando do `R`:

```
mtcars %>%
  summarise(
    `Média aritmética` = mean(hp),
    `Média geométrica` = geometric.mean(hp),
    `Média harmônica` = harmonic.mean(hp),
    `Moda` = which.max(tabulate(mtcars$hp)),
    `Mediana` = median(hp),
    `Quartil 1` = quantile(hp)[2],
    `Quartil 3` = quantile(hp)[4],
    `Mínimo` = min(hp),
    `Máximo` = max(hp)
    )
```

Potência de veículos e medidas de posição

O resultado será:

```
                          [,1]
Média aritmética 146.6875
Média geométrica 131.8837
Média harmônica  118.2289
Moda             110.0000
Mediana          123.0000
Quartil 1         96.5000
Quartil 3        180.0000
Mínimo            52.0000
Máximo           335.0000
```