Métodos Estatísticos Básicos

Aula 3 - Medidas de tendência central

Prof. Regis Augusto Ely

Departamento de Economia
Universidade Federal de Pelotas (UFPel)

Abril de 2014

- As medidas de tendência central são estatísticas que caracterizam um conjunto de dados, sendo o valor em torno do qual se agrupam as observações.
- **Média aritmética** $(\overline{X})$**:** é o quociente entre a soma dos valores dos nossos dados e o número total de dados, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.
- Quando temos dados não-agrupados pela frequência das observações, calculamos a _média aritmética simples_.
  Ex: 10, 14, 13, 15, 16, 18, 12.
  Logo, $\overline{X} = \frac{(10+14+13+15+16+18+12)}{7} = 14$.
- **Desvio em relação à média:** é a diferença entre um elemento de um conjunto de valores e a média aritmética desse conjunto, ou seja, $d_i = X_i - \overline{X}$.
  Ex: $d_1 = 10 - 14 = -4$; $d_2 = 14 - 14 = 0$, etc.

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

- Se os dados estiverem agrupados em uma tabela de frequência, devemos calcular uma *média aritmética ponderada*. $\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} fi}$.

  Ex:

Dados	Frequência
0	2
1	6
2	10
3	12
4	4
Total	34

$$\overline{X} = \frac{0x2+1x6+2x10+3x12+4x4}{34} = 2,3$$

- Se os dados estiverem agrupados em intervalos de classe, utilizamos a *média aritmética ponderada*, definindo Xi como o ponto médio da classe i. Ex:

Estaturas(cm)	Frequência (fi)	Ponto médio (Xi)	Xi.fi
$50 \vdash 54$	4	52	208
$54 \vdash 58$	9	56	504
$58 \vdash 62$	11	60	660
$62 \vdash 66$	8	64	512
$66 \vdash 70$	5	68	340
$70 \vdash 74$	3	72	216
Total	40		2440

Assim, $\overline{X} = \frac{2440}{40} = 61$

- **Média geométrica** ($\overline{X_g}$)**:** é a raíz n-ésima do produto dos dados, $\bar{X}_g = \sqrt[n]{\prod_{i=1}^{n} X_i}$, onde $\prod_{i=1}^{n} X_i = X_1.X_2....X_n$.
  Ex: 10,60,360. $\overline{X_g} = \sqrt[3]{10.60.360} = 60$.
- Note que aplicando log, temos $log\overline{X_g} = \frac{1}{n}\sum_{i=1}^{n} logX_i$, de modo que o logaritmo da média geométrica é igual à média aritmética dos logaritmos dos valores observados.
- Assim, podemos notar que a média geométrica é uma média aritmética suavizada. Ela é muita utilizada em finanças e engenharia.

- Sempre teremos $\overline{X_g} \leq \overline{X}$, valendo a igualdade apenas se $x_i = x_j$ para todo $i \neq j$, ou seja, se todos os dados são iguais.
- Para provar que $\overline{X_g} \leq \overline{X}$ basta observar que, para o caso de apenas duas observações, X_1 e X_2:
  $\overline{X}^2 - \overline{X_g^2} = (\frac{X_1+X_2}{2})^2 - X_1.X_2 = \frac{X_1^2-2.X_1.X_2+X_2^2}{4} = (\frac{X_1-X_2}{2})^2 \geq 0.$
  Logo, como $\overline{X}^2 - \overline{X_g^2} \geq 0$, temos $\overline{X_g} \leq \overline{X}$.

- Se os dados forem agrupados, calculamos a <u>*média geométrica ponderada*</u>, $\overline{X_g} = \sqrt[\sum_{i=1}^{n} fi]{\prod_{i=1}^{n} X_i^{f_i}} = \sqrt[\sum_{i=1}^{n} fi]{X_1^{f_1}.X_2^{f_2}...X_n^{f_n}}$. Ex:

X_i	f_i
1	2
3	4
9	2
27	1
Total	9

$$\overline{X_g} = \sqrt[9]{1^2.3^4.9^2.27^1} = 3,8296$$

- Se os dados forem agrupados com intervalo de classe o procedimento é o mesmo, porém agora X_i será o ponto médio de cada classe.

- **Média harmônica** ($\overline{X}_h$)**:** É o inverso da média aritmética dos inversos de cada elemento do conjunto de dados, $\overline{X}_h = (\frac{1}{n}.\sum_{i=1}^{n} X_i^{-1})^{-1} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$.
- A média harmônica é bastante utilizada na física, quando trabalhamos com grandezas que variam inversamente. Ex: velocidade e tempo.
- Sempre teremos $\overline{X}_h \leq \overline{X}_g \leq \overline{X}$, valendo a igualdade apenas se todos os dados forem iguais. Podemos ver isso para o caso de apenas duas observações, X_1 e X_2:
$\overline{X}_h = [\frac{1}{2}.(\frac{1}{X_1} + \frac{1}{X_2})]^{-1} = \frac{2X_1.X_2}{X_1+X_2} = \frac{2.\overline{X}_g^2}{2.\overline{X}} = \frac{\overline{X}_g.\overline{X}_g}{\overline{X}}$.
Como $\overline{X}_g \leq \overline{X}$, temos $0 \leq \frac{\overline{X}_g}{\overline{X}} \leq 1$. Logo, $\overline{X}_h \leq \overline{X}_g$.

- Se os dados forem agrupados, calculamos a *média harmônica ponderada*, $\overline{X}_h = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{X_i}}$. Ex:

Classes	f_i	X_i	$\frac{f_i}{X_i}$
$1 \vdash 3$	2	2	2/2=1
$3 \vdash 5$	4	4	4/4=1
$5 \vdash 7$	8	6	8/6=1,33
$7 \vdash 9$	4	8	4/8=0,5
$9 \vdash 11$	2	10	2/10=0,2
total	20		4,03

$$\overline{X}_h = \frac{20}{4{,}03} = 4,96$$

- Se os dados forem agrupados sem intervalo de classe o procedimento é o mesmo, porém agora X_i será o valor de cada elemento do conjunto de dados.

- **Obs 1:** a média harmônica não aceita valores iguais a zero como dados de uma série.
- **Obs 2:** quando os valores da variável não forem muito diferentes, verifica-se a seguinte relação, $\overline{X_g} = \frac{(\overline{X}+\overline{X_h})}{2}$.
  Ex: {10,1; 10,1; 10,2; 10,4; 10,5}
  $\overline{X} = \frac{51{,}3}{5} = 10,2600$
  $\overline{X_h} = \frac{5}{0{,}4874} = 10,2574$
  $\overline{X_g} = \frac{10{,}26+10{,}2574}{2} = 10,2587$

- **Moda** (M_o)**:** É o valor que ocorre com maior frequência em uma série de dados.
- Se os dados estiverem não agrupados, devemos procurar o valor que mais se repete.
  Ex: {7, 8, 9, 10, 10, 10, 11, 12}. Temos Mo=10.
- Se nenhum valor aparece mais vezes do que outro, chamamos a série de amodal.
  Ex: {3, 5, 8, 10, 12} não apresenta moda. É *amodal*.
- Se dois ou mais valores repetem o mesmo número de vezes, a série tem mais de um valor modal.
  Ex: {2, 3, 4, 4, 4, 5, 6, 7, 7, 7, 8, 9} tem duas modas, é *bimodal*. Suas modas são 4 e 7.
- O emprego da moda é utilizado apenas em alguns casos específicos, pois a média aritmética possui maior estabilidade. Ex: salários.

- **Sem intervalos de classe:** com os dados agrupados, é possível determinar a moda apenas olhando o dado com a maior frequência.

Temperatura	Frequência
0ºC	3
1ºC	9
2ºC	12
3ºC	6

$$M_o = 2^{\underline{o}}\text{C}$$

- **Com intervalos de classe:** a classe com a maior frequência é a classe modal. A moda será um valor compreendido entre os limites da classe moda.

- **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.
- **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. Ex:

Classes	Frequências
$54 \vdash 58$	9
$58 \vdash 62$	11
$62 \vdash 66$	8
$66 \vdash 70$	5

A classe modal é $58 \vdash 62$, logo $M_o = \frac{58+62}{2} = 60$; e
$M_c = 58 + \frac{(11-9)}{(11-9)+(11-8)}.4 = 59,6$

- **Mediana** (M_e)**:** é o valor que separa um conjunto de dados (dispostos em ordem crescente ou decrescente) em dois subconjuntos de 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 $\frac{n}{2}+1$.
- Ex. 1: $\{1, 3, 0, 0, 2, 4, 1, 2, 5\}$.
  1º colocamos a série em ordem crescente $\{0, 0, 1, 1, 2, 2, 3, 4, 5\}$.
  2º como existem 9 elementos, a mediana será o elemento de número $\frac{n+1}{2} = \frac{10}{2} = 5$. Assim, $M_e = 2$.

- Ex. 2: $\{1, 3, 0, 0, 2, 4, 1, 3, 5, 6\}$.
  1º colocamos a série em ordem crescente $\{0, 0, 1, 1, 2, 3, 3, 4, 5, 6\}$.
  2º como existem dez elementos, a mediana será a média dos elementos de número $\frac{n}{2} = \frac{10}{2} = 5$ e $\frac{n}{2} + 1 = 6$, logo
  $M_e = \frac{2+3}{2} = 2,5$.
- A mediana depende da posição dos valores da série. Ela não se deixa influenciar por valores extremos, como é o caso da média. Já a moda, depende da frequência. Estes três valores em geral são diferentes.
- Ex. 3: $\{5, 7, 10, 10, 18\}$. $\overline{X} = 10$, $M_o = 10$, $M_e = 10$.
- Ex. 4: $\{5, 5, 10, 13, 67\}$. $\overline{X} = 20$, $M_o = 5$, $M_e = 10$.

- Se o somatório das frequências for ímpar, a mediana será o elemento $\frac{\sum_{i=1}^{n} fi+1}{2}$.
- Identificamos facilmente esse elemento através da frequência acumulada. Ex:

Xi	fi	Fi
0	2	2
1	6	8
2	9	17
3	13	30
4	5	35
total	35	

Temos $\frac{\sum fi+1}{2} = \frac{36}{2} = 18$. Logo, $M_e = 3$.

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

Xi	fi	Fi
12	1	1
14	2	3
15	1	4
16	2	6
17	1	7
20	1	8
total	8	

Temos $\frac{\sum_{i=1}^{n} f_i}{2} = \frac{8}{2} = 4$ e $\frac{\sum_{i=1}^{n} f_i}{2} + 1 = \frac{8}{2} + 1 = 5$. Assim, $M_e = \frac{15+16}{2} = 15,5$.

- Para calcularmos a mediana de dados agrupados com intervalo de classe, seguimos as seguintes etapas:
  1º Determinamos as frequências acumuladas.
  2º Calculamos $\frac{\sum_{i=1}^{n} fi}{2}$.
  3º Marcamos a classe correspondente a frequência acumulada imediatamente superior a $\frac{\sum_{i=1}^{n} fi}{2}$. Essa será a classe mediana.
  4º Temos que$M_e = l^* + \frac{[(\frac{\sum_{i=1}^{n} f_i}{2} - FAA).h*]}{f*}$, 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.

- Ex:

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	

$$\frac{\sum fi}{2} = 20 \rightarrow \text{Classe mediana: } 58 \vdash 62.$$

$$l* = 58,\ FAA = 13,\ f* = 11,\ h* = 4.$$

$$Md = 58 + \frac{[(20-13)x4]}{11} = 58 + \frac{28}{11} = 60,54.$$

- Quando desejamos obter o ponto que divide a distribuição em duas partes iguais.
- Quando há valores extremos que afetam demais a média.
- Em variáveis como salário.

- Existem outras medidas de posição que não são medidas de tendência central, como os quartis, decis e percentis, conhecidas genericamente por separatrizes.
- **Quartil:** são os valores que dividem a série em quatro partes iguais. Precisamos 3 quartis para dividir a série em quatro partes.
- Note que o segundo quartil (Q_2) será sempre igual a mediana.

- Ex. 1: $\{5, 2, 6, 9, 10, 13, 15\}$.
  1º ordenamos a série $\{2, 5, 6, 9, 10, 13, 15\}$.
  2º calculamos a mediana, que será o segundo quartil $M_e = Q2 = 9$.
  3º dividimos a série em dois grupos $\{2, 5, 6\}$ e $\{10, 13, 15\}$.
  4º calculamos os outros quartis como sendo as medianas desses dois grupos $Q_1 = 5$ e $Q_2 = 13$.
- Ex. 2: $\{1, 1, 2, 3, 5, 5, 6, 7, 9, 9, 10, 13\}$.
  $M_e = Q_2 = \frac{5+6}{2} = 5,5$. Logo, temos $\{1, 1, 2, 3, 5, 5\}$ com $Q_1 = 2,5$, e $\{6, 7, 9, 9, 10, 13\}$ com $Q_3 = 9$.

- 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=$l^* + \frac{[(3\frac{\sum fi}{4} - FAA).h^*]}{f*}$

- Ex:

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	

$\frac{\sum fi}{2} = 20 \rightarrow$ classe mediana: $58 \vdash 62.\ l* = 58,\ FAA = 13,\ f* = 11,\ h* = 4$

$$M_e = Q_2 = 58 + \frac{[(20-13)x4)}{11} = 60,54.$$

$\frac{\sum fi}{4} = 10 \rightarrow$ classe mediana do 1º grupo: $54 \vdash 58.$

$$Q1 = 54 + \frac{[(10-4)x4]}{9} = 56,66.$$

$\frac{3.\sum fi}{4} = 30 \rightarrow$classe mediana do 3º grupo: $62 \vdash 66.$

$$Q3 = 62 + \frac{[(30-24)x4]}{8} = 65.$$

- **Decil:** são os valores que dividem a série em dez partes iguais. Precisamos 9 decis para dividir a série em 10 partes.
- O procedimento é análogo, porém agora o 5º decil será igual ao 2º quartil, que será igual à mediana.
- Ex: calcule o 3º decil da tabela anterior.
  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 (ou centil)

- Serão os 99 valores que separam a série em 100 partes iguais, de modo que $P_{50} = M_e$, $P_{25} = Q_1$ e $P_{75} = Q_3$. O cálculo é análogo, mas utilizando$k.\frac{\sum fi}{100}$.