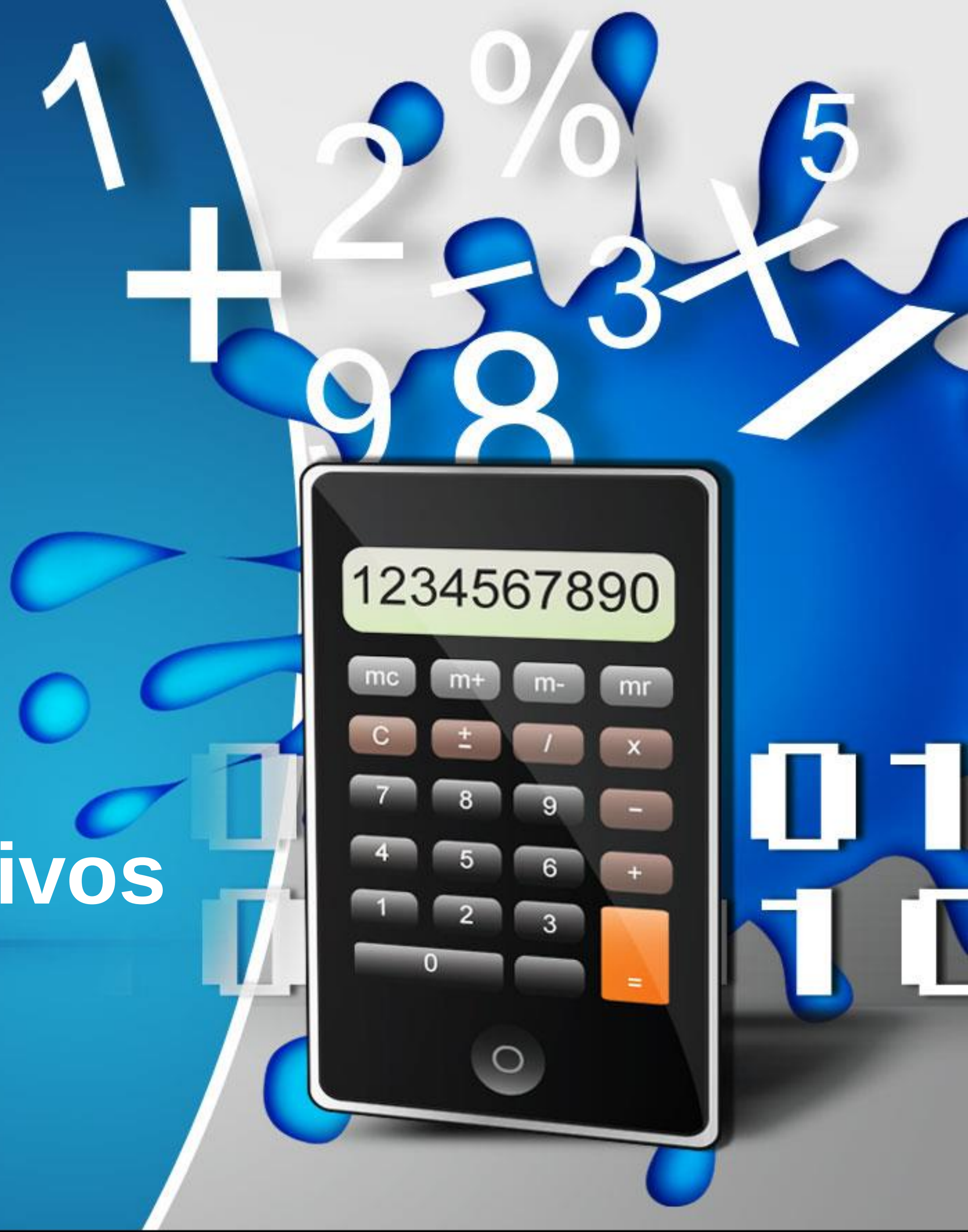
Cálculos Alternativos

Análise Exploratória dos dados de uma amostra.

Reinaldo Madarazo - 2014



Média, Variância e Desvio-Padrão

- As seguintes expressões podem ser usadas para cálculos de média, variância e desvio-padrão para dados não agrupados:

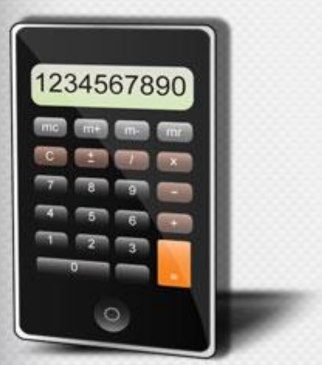
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textit{onde } \mathrm{n} \textit{ é o tamanho da amostra}$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

ou

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

$$s = \sqrt{s^2}$$



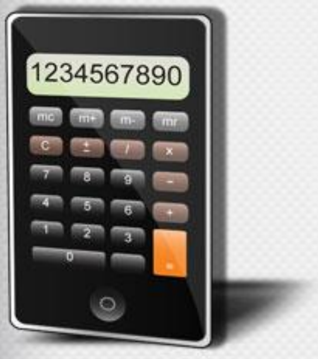
Exemplo

- Considere a amostra formada por 10, 11, 13, 15 e 18. Determinar a média, variância e desvio-padrão usando as fórmulas alternativas vistas.

x_i	x_i^2
10	100
11	121
13	169
15	225
18	324
Soma: 67	Soma: 939

$$n = 5 \Rightarrow \bar{x} = \frac{67}{5} = 13{,}4$$

$$s^2 = \frac{1}{5-1}\left[939 - \frac{(67)^2}{5}\right] = 10{,}3 \Rightarrow s = \sqrt{10{,}3} = 3{,}21$$



Dados sujeitos a repetições

- Cada dado $\boldsymbol{x_i}$ possui uma frequência $\boldsymbol{n_i}$.
- Assim, existem $\boldsymbol{k}$ classes de frequências ou número de agrupamentos.

$$n = \sum_{i=1}^{k} n_i \quad onde\ n\ é\ o\ tamanho\ da\ amostra$$

- Assim, podemos calcular a média, variância e desvio-padrão:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i.n_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$$

$$ou$$

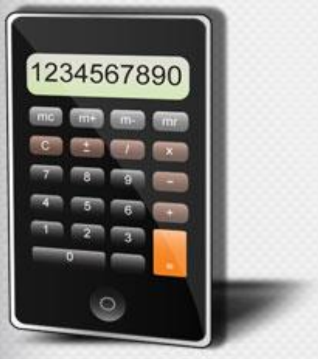
$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2.n_i - \frac{1}{n}\left(\sum_{i=1}^{k} x_i.n_i\right)^2\right] \ e\ \ s = \sqrt{s^2}$$

Exercícios

- Considere a seguinte tabela de dados:

x_i	10	20	30	40	50	60	70
n_i	1	5	22	24	22	5	1

- Determinar a média e o desvio-padrão.



x_i	n_i	$x_i.n_i$	$(x_i-\ddot{x})$	$(x_i-\ddot{x})^2$	$(x_i-\ddot{x})^2.n_i$
10	1				
20	5				
30	22				
40	24				
50	22				
60	5				
70	1				
Soma:					

Resposta

x_i	n_i	$x_i.n_i$	$(x_i-\ddot{x})$	$(x_i-\ddot{x})2$	$(x_i-\ddot{x})^2.n_i$
10	1	10	-30	900	900
20	5	100	-20	400	2000
30	22	660	-10	100	2200
40	24	960	0	0	0
50	22	1100	10	100	2200
60	5	300	20	400	2000
70	1	70	30	900	900
Soma:	80	3200			10200

$$k = 7 \ e \ n = 80 \Rightarrow \bar{x} = \frac{3200}{80} = 40 \Rightarrow \bar{x} = 40$$

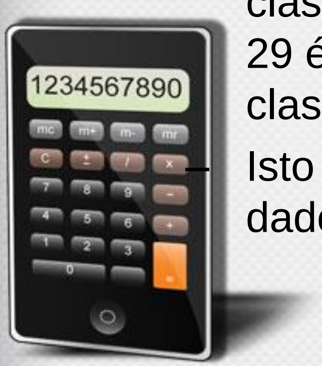
$$s^2 = \frac{1}{80-1}\sum_{i=1}^{7}(x_i - \bar{x})^2.n_i = \frac{10200}{79} = 129{,}11$$

$$s^2 = 129{,}11 \Rightarrow s = \sqrt{129{,}11} = 11{,}36$$

Dados Agrupados

- Limites reais e aparentes: Esta é uma forma antiga de se determinar os limites de uma classe na construção de um histograma.
- Considere as seguintes classes:
  - 10 é o limite aparente inferior da primeira classe.
  - 19 é o limite aparente superior da primeira classe.
  - 20 é o limite aparente inferior da segunda classe.
  - 29 é o limite aparente superiro da segunda classe.
  - Isto pode ser verdadeiro para dados discretos variando de 1 em 1, mas e para dados contínuos? Onde ficaria o 19,5?

Limites da classe
10 a 19
20 a 29
30 a 39
40 a 49
50 a 59



Dados Agrupados

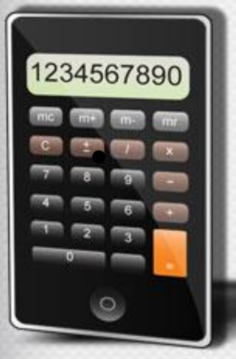
- Para dados contínuos, são usados os limites reais obtidos pela média aritmética simples dos limites diagonais:

$$Limite\ real\ inferior = \frac{Limite\ aparente\ superior\ da\ classe\ i + Limite\ aparente\ inferior\ da\ classe\ i+1}{2}$$

$Assim:$

$$Limite\ real\ inferior\ da\ segunda\ classe = \frac{19+20}{2} = 19{,}5$$

- Assim, o limite inferior da primeira classe seria 9,5 inclusive e o limite superiro seria 19,5 exclusive.



O limite inferior da segunda classe seria 19,5 inclusive e o superior seria 29,5 exclusive (29,5=(29+30)/2).

- Dessa forma os dados contínuos passariam a constar das classes. Essa abordagem não é mais feita atualmente e os limites reais e aparentes não são mais usados.

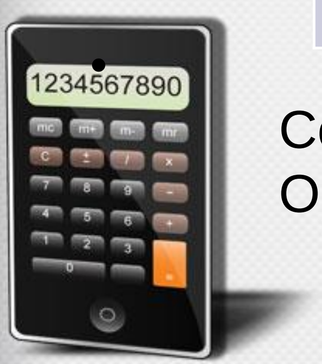
Cálculos Aproximados

- Sejam x_i o ponto médio da classe i; f_i a frequência relativa da classe i; N_i a frequência absoluta acumulada da classe i e F_i a frequência relativa acumulada da classe i.
- Complete a tabela:

Classes	x_i	n_i	N_i	f_i	F_i
90,5 |-- 97,5		3			
97,5 |-- 104,5		8			
104,5 |-- 111,5		10			
111,5 |-- 118,5		8			
118,5 |-- 125,5		11			
125,5 |-- 132,5		6			
132,5 |-- 139,5		2			
139,5 |-- 146,5		2			
Soma:					

Cálculos Aproximados

Classes	x_i	n_i	N_i	f_i	F_i
90,5 |-- 97,5	94	3	3	0,06	0,06
97,5 |-- 104,5	101	8	11	0,16	0,22
104,5 |-- 111,5	108	10	21	0,20	0,42
111,5 |-- 118,5	115	8	29	0,16	0,58
118,5 |-- 125,5	122	11	40	0,22	0,80
125,5 |-- 132,5	129	6	46	0,12	0,92
132,5 |-- 139,5	136	2	48	0,04	0,96
139,5 |-- 146,5	143	2	50	0,04	1,00
Soma:		50		1,00	



Com essa tabela é possível a construção do Histograma e da Ogiva.

Média e Desvio-Padrão

- Usando as seguintes expressões é possível obter a média e o desvio-padrão de dados agrupados:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i . n_i$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2 . n_i - \frac{1}{n}\left(\sum_{i=1}^{k} x_i . n_i\right)^2\right] \; e \;\; s = \sqrt{s^2}$$

Média e Desvio-padrão

- Complete a seguinte tabela e determine a média e o desvio-padrão:

Classes	x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
90,5 |-- 97,5	94	3		
97,5 |-- 104,5	101	8		
104,5 |-- 111,5	108	10		
111,5 |-- 118,5	115	8		
118,5 |-- 125,5	122	11		
125,5 |-- 132,5	129	6		
132,5 |-- 139,5	136	2		
139,5 |-- 146,5	143	2		
Soma:		50		

Média e Desvio-Padrão

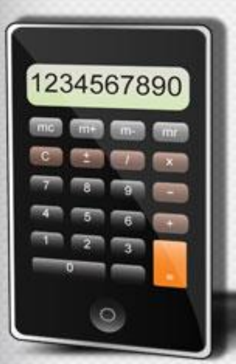
- Solução:

Classes	x_i	n_i	$x_i.n_i$	$x_i^2.n_i$
90,5 |-- 97,5	94	3	282	26508
97,5 |-- 104,5	101	8	808	81608
104,5 |-- 111,5	108	10	1080	116640
111,5 |-- 118,5	115	8	920	105800
118,5 |-- 125,5	122	11	1342	163724
125,5 |-- 132,5	129	6	774	99846
132,5 |-- 139,5	136	2	272	36992
139,5 |-- 146,5	143	2	286	40898
Soma:		50	5764	672016

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i.n_i = \frac{5764}{50} = 115,28 \Rightarrow \vec{x} = 115,28$$

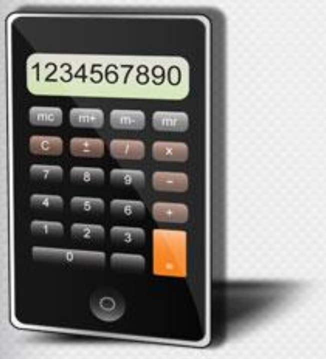
$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2.n_i - \frac{1}{n}\left(\sum_{i=1}^{k} x_i.n_i\right)^2\right] =$$

$$= \frac{1}{49}\left[672016 - \frac{5764^2}{50}\right] = 153,92 \Rightarrow s^2 = 153,92 \Rightarrow s = 12,4065$$



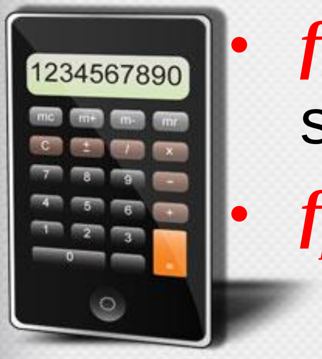
Separatrizes

- São valores que dividem o rol que contém a mesma quantidade de elementos. Ex.:
- **Mediana:** divide o rol em duas metades.
- **Quartis:** divide o rol em quatro partes iguais.



Simbologia

- n = tamanho da amostra.
- l_i = limite inferior da classe separatriz.
- N_a = frequência acumulada da classe anterior à da separatriz.
- h_i = amplitude da classe da medida.
- n_i = frequência absoluta da classe separatriz.
- f_i = frequência relativa da classe separatriz.
- f_a = frequência relativa da classe anterior à da separatriz.
- f_p = frequência relativa posterior da medida.

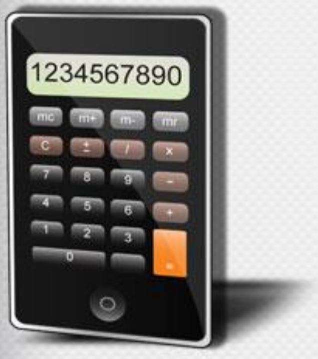


Moda

- A moda pode ser obtida pelo método de King:

$$M_o = l_i + \frac{h_i \cdot f_p}{f_a + f_p}$$

- Ou pelo método de Czuber:

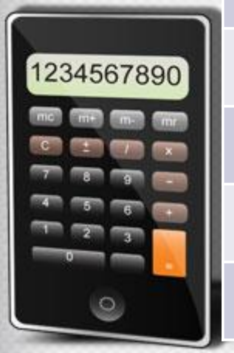


$$M_o = l_i + h_i \frac{f_i - f_a}{2f_i - (f_a + f_p)}$$

Exemplo

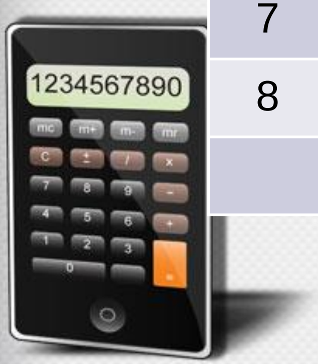
- Usaremos os dados da tabela dos dados agrupados usada nos cálculos da média e desvio padrão.

i	Classes	x_i	n_i	N_i	f_i	F_i
1	90,5 |-- 97,5	94	3			
2	97,5 |-- 104,5	101	8			
3	104,5 |-- 111,5	108	10			
4	111,5 |-- 118,5	115	8			
5	118,5 |-- 125,5	122	11			
6	125,5 |-- 132,5	129	6			
7	132,5 |-- 139,5	136	2			
8	139,5 |-- 146,5	143	2			
	Soma:		50			



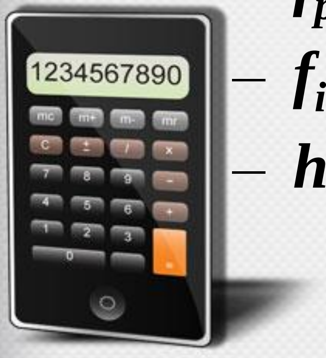
Exemplo - Resposta

i	Classes	x_i	n_i	N_i	f_i	F_i
1	90,5 |-- 97,5	94	3	3	0,06	0,06
2	97,5 |-- 104,5	101	8	11	0,16	0,22
3	104,5 |-- 111,5	108	10	21	0,20	0,42
4	111,5 |-- 118,5	115	8	29	0,16	0,58
5	118,5 |-- 125,5	122	11	40	0,22	0,80
6	125,5 |-- 132,5	129	6	46	0,12	0,92
7	132,5 |-- 139,5	136	2	48	0,04	0,96
8	139,5 |-- 146,5	143	2	50	0,04	1,00
	Soma:		50		1,00	



Exemplo

- A **classe modal** (que contém a moda) é a classe de maior frequência. No nosso caso é a *classe 5* ($n_5=11$).
- Assim, temos:
  - $\boldsymbol{i = 5}$
  - $\boldsymbol{l_i = l_5 = 118{,}5}$
  - $\boldsymbol{f_a = f_4 = 0{,}16}$
  - $\boldsymbol{f_p = f_6 = 0{,}12}$
  - $\boldsymbol{f_i = f_5 = 0{,}22}$
  - $\boldsymbol{h_5 = 125{,}5 - 118{,}5 = 7}$ (valor igual para todas as classes)

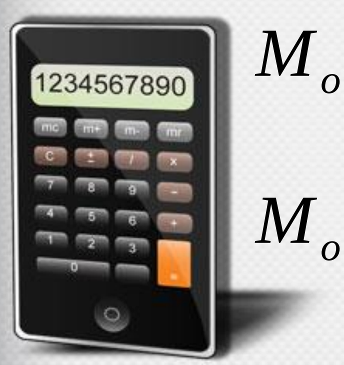


Exemplo - Moda

- King:

$$M_o = l_i + \frac{h_i . f_p}{f_a + f_p} = 118{,}5 + \frac{7.0{,}12}{0{,}16 + 0{,}12} \Rightarrow M_o = 121{,}5$$

- Czuber



$$M_o = l_i + h_i \frac{f_i - f_a}{2 f_i - (f_a + f_p)} = 118{,}5 + 7 \frac{(0{,}22 - 0{,}16)}{2.0{,}22 - (0{,}16 + 0{,}12)}$$

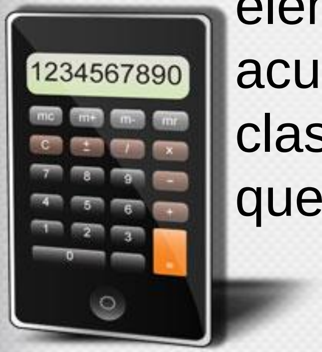
$$M_o = 121{,}125$$

Mediana

- A mediana pode ser obtida por:

$$Md = l_i + \frac{1}{n_i}\left(\frac{n}{2} - N_a\right)h_i$$

- Para determinar a *classe da mediana*, escolhemos a classe que acumula *50%* dos dados. No nosso caso, a amostra tem *50* elementos e *50%* correspondem à *25* dados. A classe que acumula *25* dados é a **classe 4**. Podemos também utilizar a classe de frequências relativas acumuladas, escolhendo a classe que *acumula 0,50*.



Mediana - Exemplo

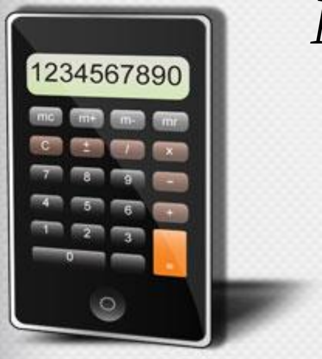
- A mediana se encontra na classe 4 (i = 4).
  - $i = 4$
  - $l_i = l_4 = 111{,}5$
  - $N_a = N_3 = 21$
  - $n_i = n_4 = 8$
  - $h_i = h_4 = 7$

Mediana - Exemplo

$$Md = l_i + \frac{1}{n_i}\left(\frac{n}{2} - N_a\right)h_i = 111{,}5 + \frac{1}{8}\left(\frac{50}{2} - 21\right)7 =$$

$$= 111{,}5 + 7.\frac{(25-21)}{8}$$

$$Md = 115$$



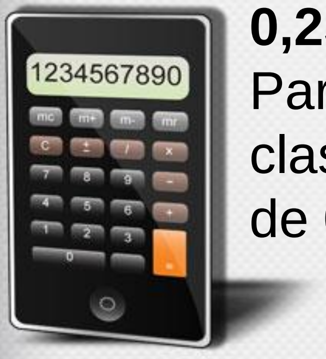
Quartis

- Os quartis podem ser obtidos pelas expressões:

$$Q_1 = l_i + \frac{1}{n_i}\left(\frac{n}{4} - N_a\right)h_i$$

$$Q_3 = l_i + \frac{1}{n_i}\left(3\frac{n}{4} - N_a\right)h_i$$

- Para determinar a *classe do primeiro quartil*, escolhemos a classe que acumula **n/4** dados ou com frequência relativa acumulada de **0,25**. No nosso caso, isso ocorre na **classe 3 (i = 3).**



Para determinar a *classe do terceiro quartil*, escolhemos a classe que acumula **3n/4** dados ou com frequência acumulada de **0,75**. No nosso caso, isso ocorre na **classe 5 (i = 5).**

Exemplos

$Q_1 \rightarrow i = 3$
$l_i = l_3 = 104{,}5$
$N_a = N_2 = 11$
$n_i = n_3 = 10$
$h_i = h_3 = 7$

$Q_3 \rightarrow i = 5$
$l_i = l_5 = 118{,}5$
$N_a = N_4 = 29$
$n_i = n_5 = 11$
$h_i = h_5 = 7$

$$Q_1 = l_i + \frac{1}{n_i}\left(\frac{n}{4} - N_a\right)h_i = 104{,}5 + \frac{1}{10}\left(\frac{50}{4} - 11\right)7 =$$

$$= 104{,}5 + 7\,\frac{12{,}5 - 11}{10} \Rightarrow Q_1 = 105{,}55$$

$$Q_3 = l_i + \frac{1}{n_i}\left(3\frac{n}{4} - N_a\right)h_i = 118{,}5 + \frac{1}{11}\left(3\frac{50}{4} - 29\right)7 =$$

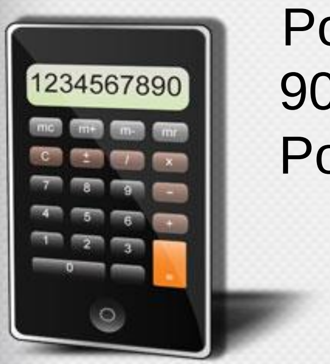
$$= 118{,}5 + 7\,\frac{(37{,}5 - 29)}{11} \Rightarrow Q_3 = 123{,}91$$

Percentis

- Para determinação do percentil P_k usaremos a seguinte expressão:

$$P_k = l_i + \frac{1}{n_i}\left(k\frac{n}{100} - N_a\right)h_i$$

- Para determinar a classe onde se encontra o $\boldsymbol{P_k}$, escolhemos a classe que acumula $\boldsymbol{P_k\%}$ dos dados. Pode-se usar tanto a frequência absoluta como a relativa acumulada.



Por exemplo, queremos obter a *classe do* P_{90}. Como $n=50$, 90% equivale a $90.50/100 = 45$. Escolhemos a **classe 6 (i = 6).** Podemos também escolher a classe que *acumula 0,90*.

•Exemplo – P_{90}

- $\boldsymbol{i = 6}$
- $\boldsymbol{l_i = l_6 = 125{,}5}$
- $\boldsymbol{N_a = N_5 = 40}$
- $\boldsymbol{n_i = n_6 = 6}$
- $\boldsymbol{h_i = h_6 = 7}$

$$P_k = l_i + \frac{1}{n_i}\left(k\frac{n}{100} - N_a\right)h_i$$

$$P_{90} = l_6 + \frac{1}{n_6}\left(90\frac{n}{100} - N_5\right)h_6 =$$

$$= 125{,}5 + \frac{1}{6}(45 - 40).7$$

$$P_{90} = 131{,}33$$

Bibliografia

- Morettin, Luiz Gonzaga. Estatística Básica Volume Único. Probabilidade e Inferência. Pearson.
- Rodrigues, Cláudio Nestor. Apostila On-Line de Pesquisa Operacional (2012) http://pt.slideshare.net/claudionr/distribuio-de-frequencia
- Blog de Estatística do Prof. Alexandre: http://alexandreprofessor.blogspot.com.br/p/medidas-separatrizes.html