Aula 4: Estatísticas Descritivas e Gráficos, Medidas de Resumo

Ricardo Ceneviva
ceneviva@usp.br

Universidade de São Paulo
Departamento de Ciência Política

31 de agosto de 2012

Sobre o que vamos conversar hoje?

1. Estatísticas Descritivas

Sobre o que vamos conversar hoje?

1. Estatísticas Descritivas

2. Medidas de Tendência Central

Sobre o que vamos conversar hoje?

1. Estatísticas Descritivas
2. Medidas de Tendência Central
3. Medidas de Dispersão

Estatísticas Descritivas

- Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.

Estatísticas Descritivas

- Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.

- As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias.

Estatísticas Descritivas

- Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.

- As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias.

- A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

Aspectos Gerais e Forma da Distribuição

- Ao estudarmos a distribuição de freqüências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

Aspectos Gerais e Forma da Distribuição

- Ao estudarmos a distribuição de freqüências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

- Tendência Central

Aspectos Gerais e Forma da Distribuição

- Ao estudarmos a distribuição de freqüências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

- Tendência Central
- Variabilidade

Aspectos Gerais e Forma da Distribuição

- Ao estudarmos a distribuição de freqüências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

- Tendência Central
- Variabilidade
- Forma

Tendência Central

- A tendência central da distribuição de frequências de uma variável é caracterizada pelo valor (ou faixa de valores) típico da variável.

Tendência Central

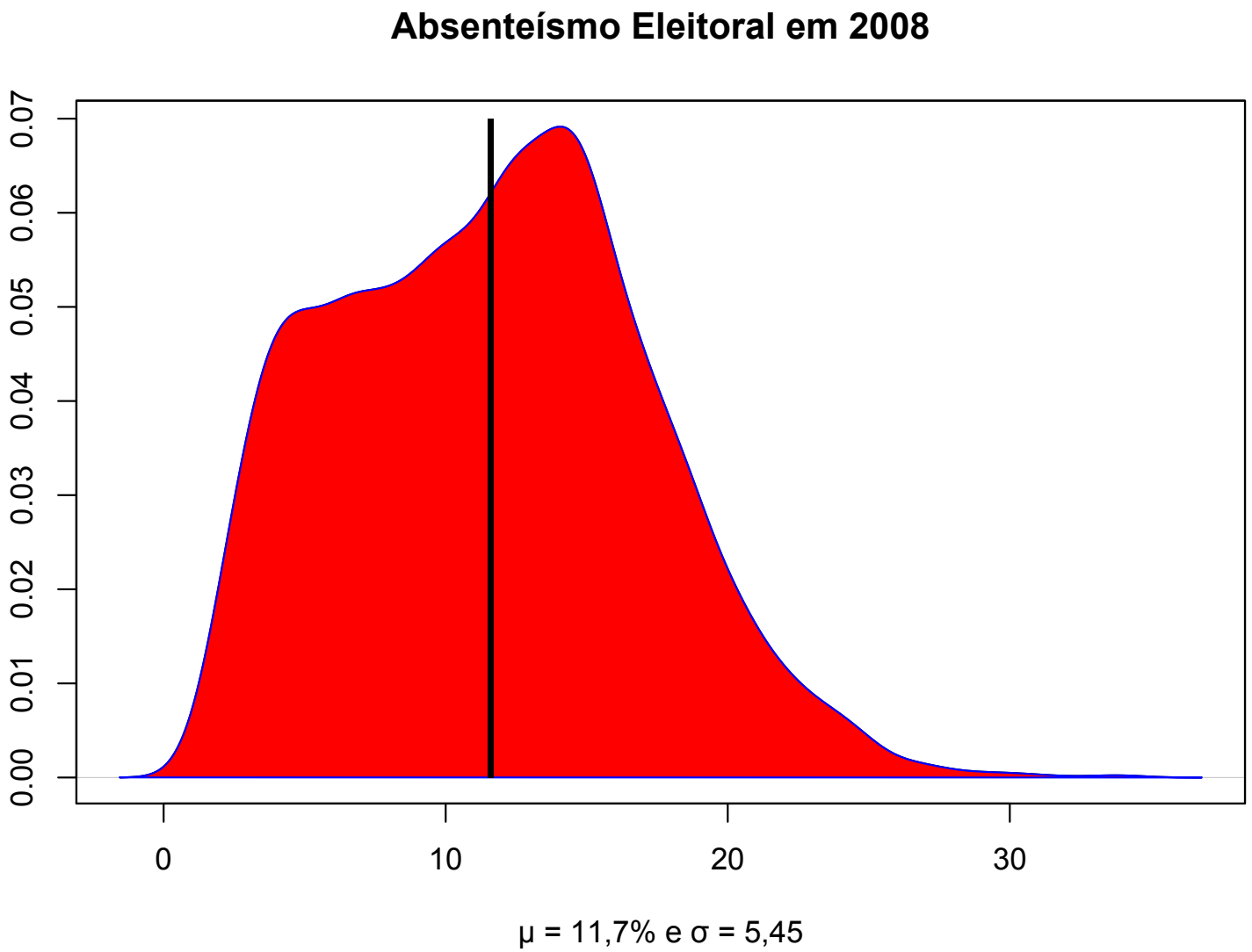
- A tendência central da distribuição de frequências de uma variável é caracterizada pelo valor (ou faixa de valores) típico da variável.

- Podemos representar o que é típico é através do valor mais frequente da variável, chamado de moda, do valor central, chamado de mediana, ou simplesmente pelo valor médio da variável.

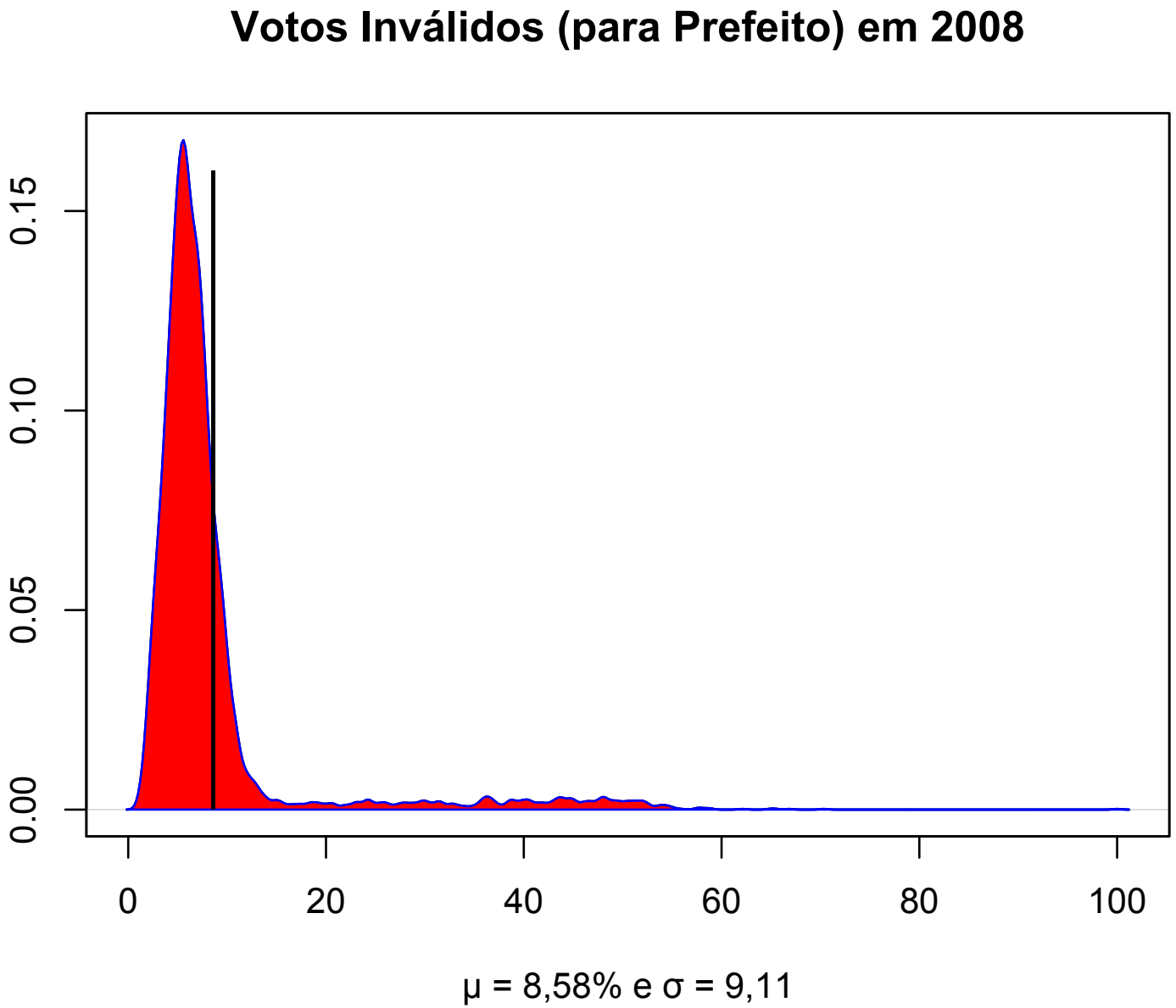
Variabilidade

- Para descrevermos adequadamente a distribuição de frequências de uma variável, além da informação do valor representativo da variável (tendência central), é preciso descrever também o quanto estes valores variam, ou seja, o quão dispersos eles são.

Variabilidade



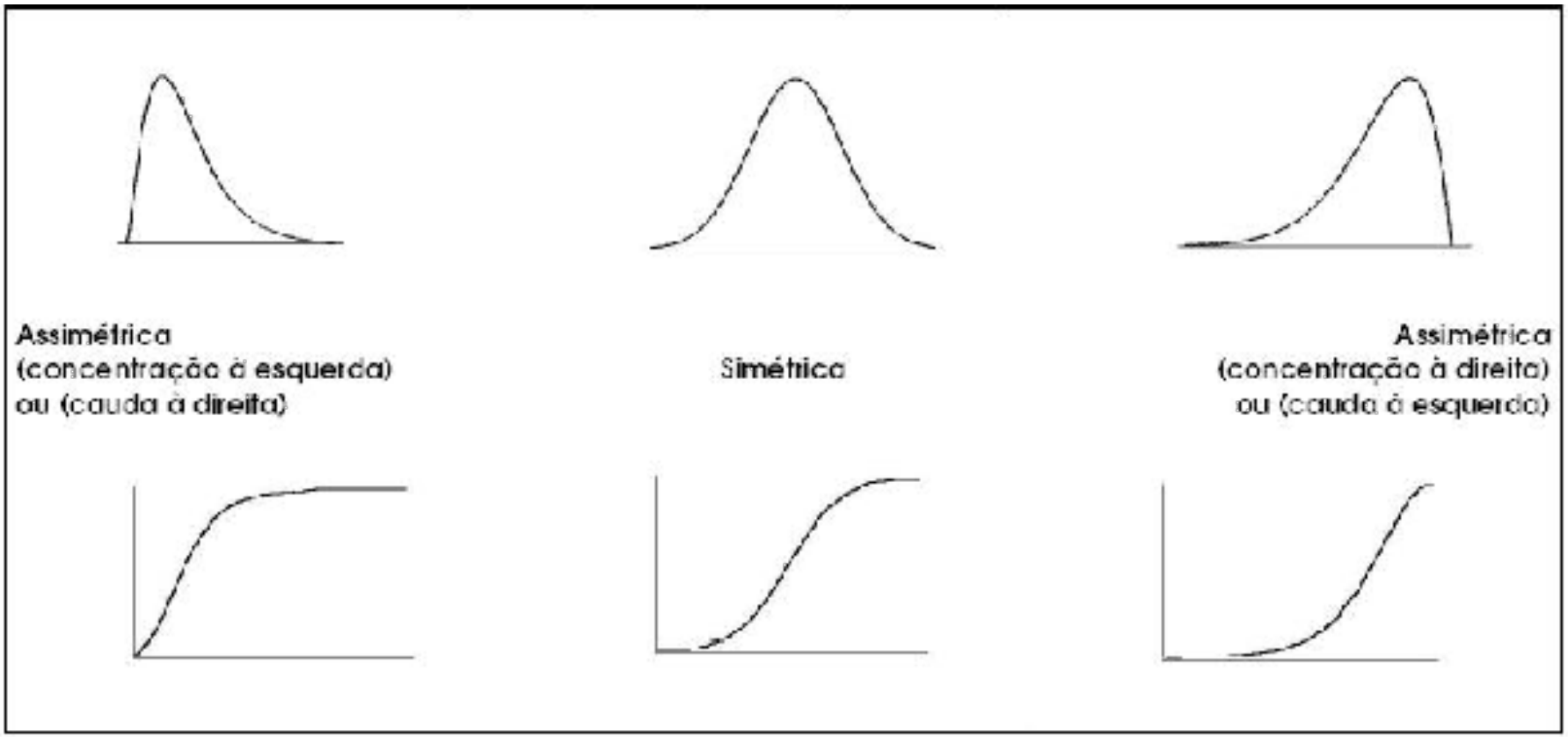
Variabilidade



Forma da Distribuição

- A distribuição de frequências de uma variável pode ter várias formas, mas existem três formas básicas

Variabilidade



Média

- **Conceito familiar:** É a soma das observações dividia pelo número total delas.

Média

- **Conceito familiar:** É a soma das observações dividia pelo número total delas.
- **Conceito formal:** Se x1, ..., xn são os n valores da variável X, a média aritmética de X pode ser descrita:

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

Média

- Se os dados estiverem resumidos em uma tabela de frequências, então a média de X pode ser escrita:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n} \tag{2}$$

Média

- Se os dados estiverem resumidos em uma tabela de frequências, então a média de X pode ser escrita:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n} \tag{2}$$

- Ou, usando a frequência relativa:

$$\bar{x} = \frac{\sum_{i=1}^{n} fr_i x_i}{n} \tag{3}$$

Média para Dados Agrupados:

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

$$\overline{X} = \frac{4*0+5*1+7*2+\ldots+1*5}{20} = 1{,}65$$

$$\overline{X} = 0{,}2*0+0{,}25*1+0{,}35*2+\ldots+0{,}05*5 = 1{,}65$$

Média para Dados Agrupados:

Classe de salários	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

$$\overline{X} \approx \frac{10*6{,}00+12*10{,}00+8*14{,}00+\ldots+1*22{,}00}{36} = 11{,}22$$

$$\overline{X} \approx 0{,}27*6{,}0+0{,}33*10{,}0+0{,}22*14{,}0+\ldots+0{,}03*22{,}0 = 11{,}22$$

Mediana

- A **Mediana** é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente.
- Se o número de observações for ímpar o mediana é a posição central da série:

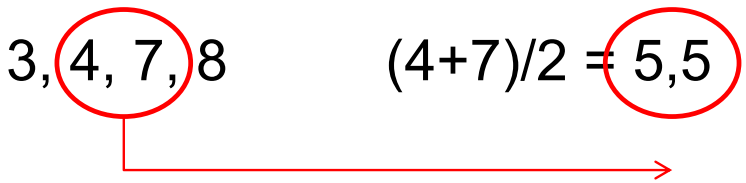
3, 4, 7, 8, 8

Mediana

- A **Mediana** é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente.
- Se o número de observações for ímpar o mediana é a posição central da série:

3, 4, 7, 8, 8

- Se o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais:



Mediana

- Conceito formal:

Considere as estatísticas de ordem $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$

- A **mediana** da variável X pode ser definida como:

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{, se n ímpar;} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{, se n par.} \end{cases}$$

Mediana

0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 5

$$md(X) = \frac{x_{\left(\frac{20}{2}\right)} + x_{\left(\frac{20}{2}+1\right)}}{2} = \frac{x_{(10)} + x_{(11)}}{2} = \frac{2+2}{2} = 2$$

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

Mediana

Classe de salários	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

$$md(X) \approx \frac{Pm_{(18)} + Pm_{(19)}}{2} = \frac{10{,}00 + 10{,}00}{2} = 10{,}00$$

Moda

- A **Moda** é definida como a realização mais frequente do conjunto de valores observados.

Moda

- A **Moda** é definida como a realização mais frequente do conjunto de valores observados.

- Pode haver mais de uma moda, ou seja, a distribuições dos valores pode ser bimodal, trimodal, ..., multimodal.

Moda

0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 5

$$mo(X) = 2$$

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

Moda

Classe de salários	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

$$mo(X) \approx 10{,}00$$

Medidas de Resumo

- Para calcularmos a **Moda** precisamos apenas da tabela de frequências. (Variáveis qualitativas e quantitativas)

Medidas de Resumo

- Para calcularmos a **Moda** precisamos apenas da tabela de frequências. (Variáveis qualitativas e quantitativas)

- Para calcularmos a **Mediana**, precisamos ordenar as realizações da variável. (Variáveis qualitativas ordinal e quantitativas)

Medidas de Resumo

- Para calcularmos a **Moda** precisamos apenas da tabela de frequências. (Variáveis qualitativas e quantitativas)
- Para calcularmos a **Mediana**, precisamos ordenar as realizações da variável. (Variáveis qualitativas ordinal e quantitativas)
- Para calcularmos a **Média**, precisamos que as variáveis sejam mensuráveis. (Variáveis quantitativas)

Medidas de Dispersão

- O resumo de um conjunto de dados por uma única medida de posição central esconde toda a informação sobre a variabilidade do conjunto de observações.

Medidas de Dispersão

- O resumo de um conjunto de dados por uma única medida de posição central esconde toda a informação sobre a variabilidade do conjunto de observações.

- Considere a nota de cinco grupos de alunos:

Grupo A (Variável X): 3,4,5,6,7
Grupo B (Variável Y): 1,3,5,7,9
Grupo C (Variável Z): 5,5,5,5,5
Grupo D (Variável W): 3,5,5,7
Grupo E (Variável V): 3,5,5,6,6

$$\overline{X} = \overline{Y} = \overline{Z} = \overline{W} = \overline{V} = 5$$

Medidas de Dispersão

- A média de cada grupo é igual, e com isso não conseguimos informação sobre sua variabilidade.

Medidas de Dispersão

- A média de cada grupo é igual, e com isso não conseguimos informação sobre sua variabilidade.

- Para resumir a variabilidade de um conjunto de dados utiliza-se a dispersão dos dados em torno da média, e as medidas mais usadas são a **variância** e o **desvio padrão**.

Medidas de Dispersão

- A média de cada grupo é igual, e com isso não conseguimos informação sobre sua variabilidade.
- Para resumir a variabilidade de um conjunto de dados utiliza-se a dispersão dos dados em torno da média, e as medidas mais usadas são a **variância** e o **desvio padrão**.
- Os desvios da média para o grupo A são -2, -1, 0, 1, 2. (Para qualquer conjunto de dados a soma destes desvios é zero!!!)

Medidas de Dispersão

- Considerar o total dos módulos dos desvios:

$$\sum_{i=1}^{n} | \, x_i - \overline{X} \, |$$

Medidas de Dispersão

- Considerar o total dos módulos dos desvios:

$$\sum_{i=1}^{n} | x_i - \overline{X} |$$

- Considerar o total dos quadrados dos desvios

$$\sum_{i=1}^{n} \left(x_i - \overline{X} \right)^2$$

Medidas de Dispersão

- Mas como comparar essas medidas quando os conjuntos de dados tem tamanhos diferentes?

Medidas de Dispersão

- Mas como comparar essas medidas quando os conjuntos de dados tem tamanhos diferentes?

- É melhor exprimir as medidas como médias:

$$dm(X) = \frac{\sum_{i=1}^{n} | x_i - \overline{X} |}{n} \qquad \text{var}(X) = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X} \right)^2}{n}$$

Medidas de Dispersão

- Mas como comparar essas medidas quando os conjuntos de dados tem tamanhos diferentes?

- É melhor exprimir as medidas como médias:

$$dm(X) = \frac{\sum_{i=1}^{n} | x_i - \overline{X} |}{n} \qquad \text{var}(X) = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X} \right)^2}{n}$$

Variância da População vs. Variância da Amostra

O tamanho da amostra é subtraído de 1 devido ao fator de correção de Bessel, que visa uma estimativa mais precisa. No cálculo de variância para toda a população, este corretor é dispensado.

Medidas de Dispersão

- As medidas de variabilidade indicam o quão homogêneo é um conjunto de dados.
- Para os grupos A e E tem-se:

$$dm(X) = \frac{2+1+0+1+2}{5} = 1{,}2$$

$$\text{var}(X) = \frac{4+1+0+1+4}{5} = 2{,}0$$

$$dm(V) = \frac{2+0+0+1+1}{5} = 0{,}8$$

$$\text{var}(V) = \frac{4+0+0+1+1}{5} = 1{,}2$$

Medidas de Dispersão

- A **variância** é uma medida de dispersão igual ao quadrado da dimensão dos dados, pode causar problemas de interpretação. Costuma-se usar, então o **desvio padrão**, que é definido como a raiz quadrada positiva da variância.

$$dp(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{n}}$$

Medidas de Dispersão

- O **Desvio Padrão**, indica, em média, qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (Média).

Medidas de Dispersão

- O **Desvio Padrão**, indica, em média, qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (Média).

- Se os dados estiverem agrupados em tabelas de frequências, as medidas de dispersão são dadas por:

$$\text{ar}(X) = \frac{\sum_{i=1}^{k} f_i \left(x_i - \overline{X}\right)^2}{n} = \sum_{i=1}^{k} fr_i \left(x_i - \overline{X}\right)^2$$

Medidas de Dispersão

- Uma maneira equivalente de calcular a **Variância**, computacionalmente mais eficiente, é:

$$\text{var}(X) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{X}^2$$

$$= \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \overline{X}^2$$

$$= \sum_{i=1}^{k} fr_i x_i^{\,2} - \overline{X}^2$$

Medidas de Dispersão

- Os valores da **média** e do **desvio padrão** isoladamente não são informativos acerca do comportamento da distribuição: se **ASSIMÉTRICO** ou **SIMÉTRICO**.

Medidas de Dispersão

- Os valores da **média** e do **desvio padrão** isoladamente não são informativos acerca do comportamento da distribuição: se **ASSIMÉTRICO** ou **SIMÉTRICO**.

- Podemos definir uma medida, denominada **quantil de ordem p**, indicada por **q(p)**, onde p é uma proporção qualquer, $0<p<1$, tal que 100% das observações sejam menores do que **q(p)**

Medidas de Dispersão

Os **quantis** mais utilizados são:

Medidas de Dispersão

Os **quantis** mais utilizados são:

1. 1° QUARTIL = q(0,25) = 25° PERCENTIL

Medidas de Dispersão

Os **quantis** mais utilizados são:

1. 1° QUARTIL = q(0,25) = 25° PERCENTIL
2. MEDIANA = q(0,50) = 50° PERCENTIL

Medidas de Dispersão

Os **quantis** mais utilizados são:

1. 1º QUARTIL = q(0,25) = 25º PERCENTIL
2. MEDIANA = q(0,50) = 50º PERCENTIL
3. 3º QUARTIL = q(0,75) = 75º PERCENTIL

Medidas de Dispersão

- Formalmente, para as estatística de ordem x(1), ...,x(n). O p-quantil é definido por:

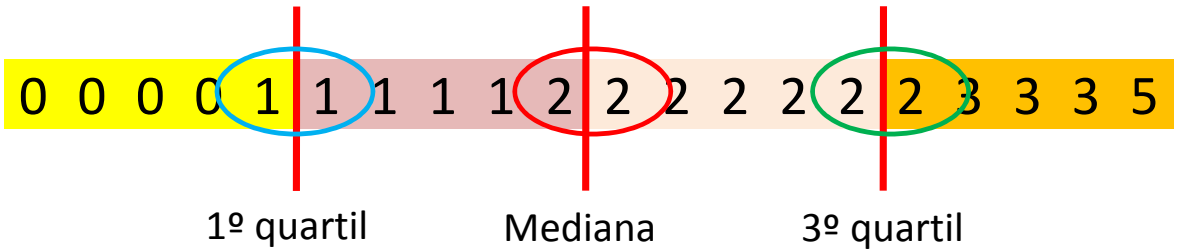
$$q(p) = \begin{cases} x_{(i)}, & se\ p = p_i = \dfrac{i - 0{,}5}{n},\ i = 1,2,...,n \\ (1 - f_i)q(p_i) + f_i q(p_{i+1}), & se\ p_i < p < p_{i+1} \\ x_{(1)}, & se\ p < p_1 \\ x_{(n)}, & se\ p > p_n \end{cases}$$

onde $f_i = \dfrac{(p - p_i)}{(p_{i+1} - p_i)}.$

Medidas de Dispersão

1º Quartil = 1
Mediana = 2
3º Quartil = 2

Exemplo - Par



N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

Medidas de Dispersão

1º Quartil = 1
Mediana = 2
3º Quartil = 2

Exemplo - Ímpar

0 0 0 0 **1** 1 1 1 1 **2** 2 2 2 2 **2** 2 3 3 3

1º quartil Mediana 3º quartil

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.21	0.21	21%
1	5	0.26	0.47	26%
2	7	0.37	0.84	37%
3	3	0.16	1.00	16%
Total	19	1		100%

Medidas de Dispersão

Classe de salários	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	10/36 =0,278	0,278	27,78%
[8,00; 12,00)	12	12/36 =0,333	0,611	33,33%
[12,00; 16,00)	8	8/36 =0,222	0,833	22,22%
[16,00; 20,00)	5	5/36 =0,139	0,972	13,89%
[20,00; 24,00)	1	1/36 =0,029	1,000	2,78%
Total	36	1		100%

1º Quartil = 6,00

Mediana = 10,00

3º Quartil = 14,00

Medidas de Dispersão Alternativa

- Uma medida de dispersão alternativa ao desvio padrão é a **distância interquartílica**, consiste na diferença entre terceiro e o primeiro quartil:

$$d_q = q(0,75) - q(0,25) \tag{4}$$

Resistência

- Os quartis são medidas de posição resistentes.

Resistência

- Os quartis são medidas de posição resistentes.
- Uma medida de posição ou dispersão é resistente quando for pouco afetada por mudanças de uma pequena porção de dados.

Resistência

- Os quartis são medidas de posição resistentes.
- Uma medida de posição ou dispersão é resistente quando for pouco afetada por mudanças de uma pequena porção de dados.
- A mediana é uma medida resistente, a média e o desvio padrão não são medidas resistentes.

Exemplo

- Considere as populações dos 20 municípios mais populosos de Minas Gerais, segundo o censo do IBGE de 2000.

Município	População
Belo Horizonte	2.238.526
Contagem	538.017
Uberlândia	501.214
Juiz de Fora	456.796
Montes Claros	306.947
Betim	306.675
Uberaba	252.051
Governador Valadares	247.131
Ribeirão das Neves	246.846
Ipatinga	212.496

Município	População
Santa Lúcia	184.903
Sete Lagoas	184.871
Divinópolis	183.962
Poços de Caldas	135.627
Ibirité	133.044
Teófilo Otoni	129.429
Patos de Minas	123.881
Sabará	115.352
Barbacena	114.126
Varginha	108.998

Medidas de Resumo

Município	População
Belo Horizonte	2.238.526
Contagem	538.017
Uberlândia	501.214
Juiz de Fora	456.796
Montes Claros	306.947
Betim	306.675
Uberaba	252.051
Governador Valadares	247.131
Ribeirão das Neves	246.846
Ipatinga	212.496

Média=336.044
Desvio padrão=454.389
3º quartil= 306.811
Mediana = 198.700
1º quartil= 131.234

Município	População
Santa Lúcia	184.903
Sete Lagoas	184.871
Divinópolis	183.962
Poços de Caldas	135.627
Ibirité	133.044
Teófilo Otoni	129.429
Patos de Minas	123.881
Sabará	115.352
Barbacena	114.126
Varginha	108.998

Sem BH

Média= 235.914
Desvio padrão=129.667
3º quartil= 306.675
Mediana = 184.903
1º quartil= 129.429

Simetria

- Os cinco valores são importantes para se ter uma boa idéia da assimetria da distribuição dos dados:

Simetria

- Os cinco valores são importantes para se ter uma boa idéia da assimetria da distribuição dos dados:

- $x_1, q_1, Mediana, q_3, x_n$

Simetria

- Os cinco valores são importantes para se ter uma boa idéia da assimetria da distribuição dos dados:
- $x_1, q_1, Mediana, q_3, x_n$
- Para uma distribuição simétrica deveríamos ter:

Simetria

- Os cinco valores são importantes para se ter uma boa idéia da assimetria da distribuição dos dados:

- $x_1, q_1, Mediana, q_3, x_n$

- Para uma distribuição simétrica deveríamos ter:

1. $Mediana - x_1 \approx x_n - Mediana$
2. $Mediana - q_1 \approx q_3 - Mediana$
3. $q_1 - x_1 \approx x_n - q_3$
4. Distâncias entre mediana e q1, q3 menores do que distâncias entre os extremos e q1, q3.

Distribuição Assimétrica

$$89.702 \neq 2.039.826$$

$$67.466 \neq 108.111$$

$$22.236 \neq 1.931.715$$

$$Mediana - q_3 < extremo - q_3$$

Box Plot

- O Box Plot é o gráfico que contém os valores da mediana, 1o e 3o quartis, limite superior e inferior e observações discrepantes.

Box Plot

- O Box Plot é o gráfico que contém os valores da mediana, 1o e 3o quartis, limite superior e inferior e observações discrepantes.

- O limite inferior é obtido por:

$$Li = q_1 - (1,5)dq$$

Box Plot

- O Box Plot é o gráfico que contém os valores da mediana, 1o e 3o quartis, limite superior e inferior e observações discrepantes.

- O limite inferior é obtido por:

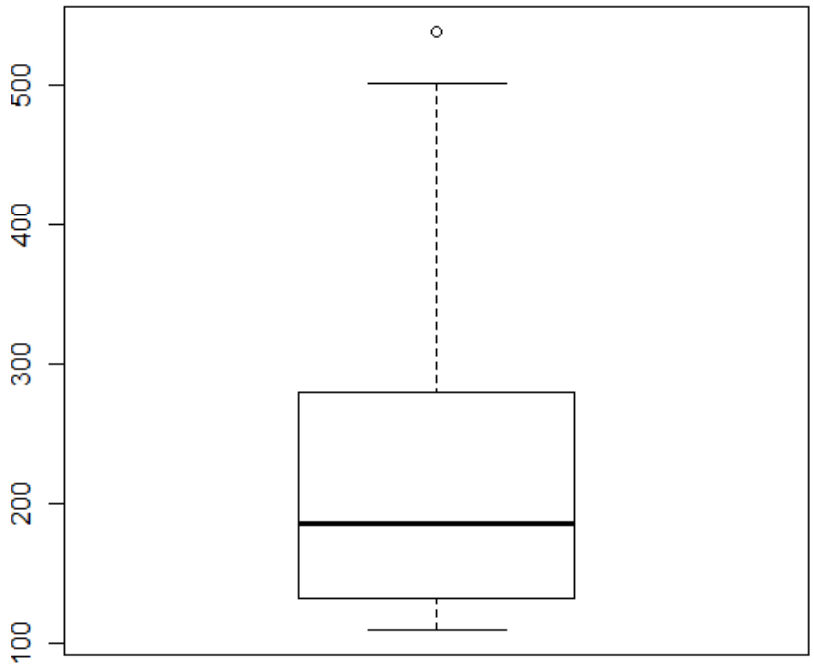
  $Li = q_1 - (1,5)dq$

- O limite superior é obtido por:
  $Li = q_3 + (1,5)dq$

Distribuição Assimétrica

Com BH

Sem BH



Distribuição Simétrica