MAD211 - Estatística para Administração

Estatística Descritiva: Medidas resumo

Prof. Carlos Trucíos
carlos.trucios@facc.ufrj.br
ctruciosm.github.io

Faculdade de Administração e Ciências Contábeis,
Universidade Federal do Rio de Janeiro

Aula 5

Introdução

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.
- Imagine agora que temos a variável `salário`. Como você resumiria os dados dessa variável?

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.
- Imagine agora que temos a variável `salário`. Como você resumiria os dados dessa variável?
- Hoje aprenderemos como sintetizar/resumir dados provenientes de variáveis quantitativas.

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.
- Imagine agora que temos a variável `salário`. Como você resumiria os dados dessa variável?
- Hoje aprenderemos como sintetizar/resumir dados provenientes de variáveis quantitativas.
  - medidas de posição

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.
- Imagine agora que temos a variável `salário`. Como você resumiria os dados dessa variável?
- Hoje aprenderemos como sintetizar/resumir dados provenientes de variáveis quantitativas.
    - medidas de posição
    - medidas de dispersão

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.
- Imagine agora que temos a variável `salário`. Como você resumiria os dados dessa variável?
- Hoje aprenderemos como sintetizar/resumir dados provenientes de variáveis quantitativas.
  - medidas de posição
  - medidas de dispersão
  - medidas de associação

Introdução

- Na última aula vimos tabelas e gráficos para sintetizar/resumir dados qualitativos.
- Imagine agora que temos a variável `salário`. Como você resumiria os dados dessa variável?
- Hoje aprenderemos como sintetizar/resumir dados provenientes de variáveis quantitativas.
  - medidas de posição
  - medidas de dispersão
  - medidas de associação
  - Gráficos: Histograma, Boxplot, scatterplot

Medidas de posição

Medidas de posição: Média

- É a medida de posição mais conhecida.

Medidas de posição: Média

- É a medida de posição mais conhecida.
- Constitui uma medida da posição central dos dados.

Medidas de posição: Média

- É a medida de posição mais conhecida.
- Constitui uma medida da posição central dos dados.

Medidas de posição: Média

- É a medida de posição mais conhecida.
- Constitui uma medida da posição central dos dados.

Média amostral

Sejam as observações $x_1, x_2, \ldots, x_n$, a média é dada por

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

Nota: geralmente $\bar{x}$ é utilizado para denotar a média amostral e μ para denotar a média populacional.

Medidas de posição: Média

Exemplo

A seguinte tabela apresenta as notas finais de 18 alunos de MAD211 da FACC/UFRJ.

2.1	7	6.7	7	6.5	8.1	9.1	9.3	7.8
5.6	8	9.0	6	7.2	8.8	6.3	9.6	7.7

Medidas de posição: Média

Exemplo

A seguinte tabela apresenta as notas finais de 18 alunos de MAD211 da FACC/UFRJ.

2.1	7	6.7	7	6.5	8.1	9.1	9.3	7.8
5.6	8	9.0	6	7.2	8.8	6.3	9.6	7.7

Vamos calcular $\bar{x}$

Medidas de posição: Média

Exemplo

A seguinte tabela apresenta as notas finais de 18 alunos de MAD211 da FACC/UFRJ.

2.1	7	6.7	7	6.5	8.1	9.1	9.3	7.8
5.6	8	9.0	6	7.2	8.8	6.3	9.6	7.7

Vamos calcular $\bar{x}$

$$\bar{x} = \frac{2.1 + 5.6 + 7 + 8 + 6.7 + 9 \cdots + 7.7}{18} = \frac{131.8}{18} = 7.322222$$

Medidas de posição: Mediana

- Outra medida de posição central.

Medidas de posição: Mediana

- Outra medida de posição central.
- É o valor *do meio* quando os valores estão ordenados

Medidas de posição: Mediana

- Outra medida de posição central.
- É o valor *do meio* quando os valores estão ordenados
- Para obter a mediana os valores devem ser ordenados de menor a maior: : $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, (onde $x_{(i)}$ é a i-ésima observação ordenada)

Medidas de posição: Mediana

- Outra medida de posição central.
- É o valor *do meio* quando os valores estão ordenados
- Para obter a mediana os valores devem ser ordenados de menor a maior: : $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, (onde $x_{(i)}$ é a i-ésima observação ordenada)
- Robusta a observações atípicas.

Medidas de posição: Mediana

- Outra medida de posição central.
- É o valor *do meio* quando os valores estão ordenados
- Para obter a mediana os valores devem ser ordenados de menor a maior: : $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, (onde $x_{(i)}$ é a i-ésima observação ordenada)
- Robusta a observações atípicas.

Medidas de posição: Mediana

- Outra medida de posição central.
- É o valor *do meio* quando os valores estão ordenados
- Para obter a mediana os valores devem ser ordenados de menor a maior: : $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, (onde $x_{(i)}$ é a i-ésima observação ordenada)
- Robusta a observações atípicas.

Mediana

$$Mediana(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ for impar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ for par.} \end{cases}$$

Medidas de posição: Mediana

Exemplo

No conjunto de dados anterior:

2.1	7	6.7	7	6.5	8.1	9.1	9.3	7.8
5.6	8	9.0	6	7.2	8.8	6.3	9.6	7.7

Medidas de posição: Mediana

Exemplo

No conjunto de dados anterior:

2.1	7	6.7	7	6.5	8.1	9.1	9.3	7.8
5.6	8	9.0	6	7.2	8.8	6.3	9.6	7.7

Primeiro, ordenamos os dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Medidas de posição: Mediana

Exemplo

No conjunto de dados anterior:

2.1	7	6.7	7	6.5	8.1	9.1	9.3	7.8
5.6	8	9.0	6	7.2	8.8	6.3	9.6	7.7

Primeiro, ordenamos os dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Qual o valor *do meio*?

Medidas de posição: Mediana

(...continuação) Exemplo

Como $n = 18$ (par), a mediana é $Mediana(x) = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$.

Medidas de posição: Mediana

(...continuação) Exemplo

Como $n = 18$ (par), a mediana é $Mediana(x) = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$.

No nosso caso:

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

$$Mediana(x) = \frac{x_{\left(\frac{18}{2}\right)} + x_{\left(\frac{18}{2}+1\right)}}{2} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{7.2 + 7.7}{2} = 7.45$$

Medidas de posição: Moda

- Outra medida de posição central

Medidas de posição: Moda

- Outra medida de posição central
- É o valor que ocorre com maior frequência

Medidas de posição: Moda

- Outra medida de posição central
- É o valor que ocorre com maior frequência
- Podem existir várias modas (nesse caso dizemos que os dados são multimodais)

Medidas de posição: Moda

- Outra medida de posição central
- É o valor que ocorre com maior frequência
- Podem existir várias modas (nesse caso dizemos que os dados são multimodais)
- Útil também quando trabalhamos com variaveis qualitativas.

Medidas de posição: Moda

- Outra medida de posição central
- É o valor que ocorre com maior frequência
- Podem existir várias modas (nesse caso dizemos que os dados são multimodais)
- Útil também quando trabalhamos com variaveis qualitativas.

Medidas de posição: Moda

- Outra medida de posição central
- É o valor que ocorre com maior frequência
- Podem existir várias modas (nesse caso dizemos que os dados são multimodais)
- Útil também quando trabalhamos com variaveis qualitativas.

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

temos que o número 7 aparece duas vezes, e todos os outros valores aparecem apenas 1 vez, logo $Moda(x) = 7$

Medidas de posição: Moda

Nos dados do Titanic,

Tabela 6: Distribuição de Frequências das classe da passagem dos passagéiros do Titanic.

	Freq. absoluta
1st	323
2nd	277
3rd	709

A moda é 3rd (terceira classe)

Outras medidas de posição

Percentil

O k-ésimo percentil (P_k) é um valor tal que *pelo menos* $k\%$ das observações são **menores ou iguais** a esse valor e *pelo menos* $(100-k)\%$ das observações são **maiores ou iguais** a esse valor.

Outras medidas de posição

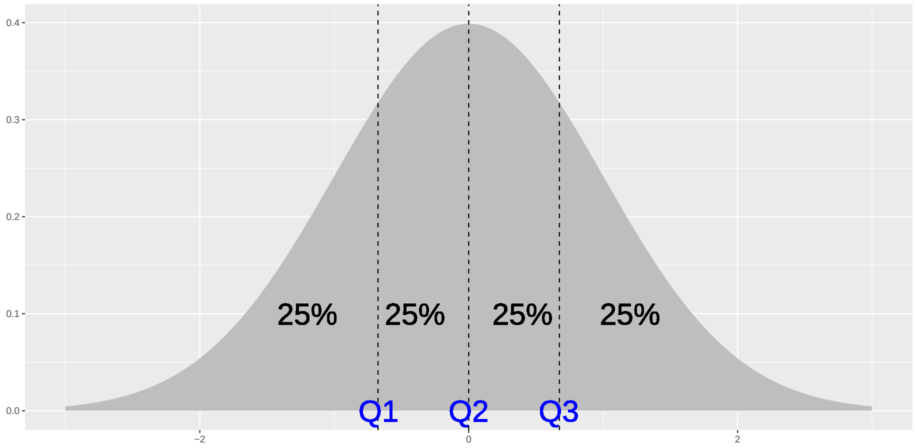
Percentil

O k-ésimo percentil (P_k) é um valor tal que *pelo menos* $k\%$ das observações são **menores ou iguais** a esse valor e *pelo menos* $(100 - k)\%$ das observações são **maiores ou iguais** a esse valor.

Quartil

- Às vezes é interessante dividir os dados em quatro partes, de forma que cada parte tenha aproximadamente 25% das observações.
- Um quartil é um caso particular de um percentil e temos três quartis em total: $Q_1 = P_{25}$, $Q_2 = P_{50}$ (ou mediana) e $Q_3 = P_{75}$

Outras medidas de posição



Outras medidas de posição

Como calcula o k-ésimo percentil

1. Ordene os dados de menor a maior: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
2. Calcule o índice i,
$$i = \left(\frac{k}{100}\right) \times n$$
em que k é o percentil desejado e n é o número e observações
3. Calcular o k-ésimo percentil:

$$P_k = \begin{cases} x_{(\lfloor i \rfloor + 1)}, & \text{se } i \text{ não for inteiro} \\ \dfrac{x_{(i)} + x_{(i+1)}}{2}, & \text{se } i \text{ for inteiro.} \end{cases}$$

Outras medidas de posição

Como calcula o k-ésimo percentil

1. Ordene os dados de menor a maior: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
2. Calcule o índice i,

$$i = \left(\frac{k}{100}\right) \times n$$

   em que k é o percentil desejado e n é o número e observações
3. Calcular o k-ésimo percentil:

$$P_k = \begin{cases} x_{(\lfloor i \rfloor + 1)}, & \text{se } i \text{ não for inteiro} \\ \dfrac{x_{(i)} + x_{(i+1)}}{2}, & \text{se } i \text{ for inteiro.} \end{cases}$$

Provavelmente, você encontrará nos livros ou na internet formas diferentes de calcular os percentis. Não precisa se preocupar, existem vários formas de calcular percentis, só na função `quantile()` do *R* existem 9 formas diferentes!

Como calcula o p-ésimo percentil

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Vamos calcular $Q_1 = P_{25}$, $Q_2 = P_{50}$ e $Q_3 = P_{75}$

Como calcula o p-ésimo percentil

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Vamos calcular $Q_1 = P_{25}$, $Q_2 = P_{50}$ e $Q_3 = P_{75}$

- $i_1 = \left(\frac{25}{100}\right) \times 18 = 4.5$, então $Q_1 = P_{25} = x_{(4+1)} = 6.5$

Como calcula o p-ésimo percentil

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Vamos calcular $Q_1 = P_{25}$, $Q_2 = P_{50}$ e $Q_3 = P_{75}$

- $i_1 = \left(\frac{25}{100}\right) \times 18 = 4.5$, então $Q_1 = P_{25} = x_{(4+1)} = 6.5$
- $i_2 = \left(\frac{50}{100}\right) \times 18 = 9$, e então $Q_2 = P_{50} = \dfrac{x_{(9)} + x_{(10)}}{2} = 7.45$

Como calcula o p-ésimo percentil

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Vamos calcular $Q_1 = P_{25}$, $Q_2 = P_{50}$ e $Q_3 = P_{75}$

- $i_1 = \left(\frac{25}{100}\right) \times 18 = 4.5$, então $Q_1 = P_{25} = x_{(4+1)} = 6.5$
- $i_2 = \left(\frac{50}{100}\right) \times 18 = 9$, e então $Q_2 = P_{50} = \dfrac{x_{(9)} + x_{(10)}}{2} = 7.45$
- $i_3 = \left(\frac{75}{100}\right) \times 18 = 13.5$, então $Q_3 = P_{75} = x_{(13+1)} = 8.8$

Medidas de dispersão

Medidas de dispersão

- As medidas de posição nada nos dizem sobre a variabilidade (dispersão) dos dados

Medidas de dispersão

- As medidas de posição nada nos dizem sobre a variabilidade (dispersão) dos dados
- As medidas de dispersão são um complemento às medidas de posição e juntas nos ajudarão a entender melhor como se comportam nossos dados.

Medidas de dispersão: Amplitude e Amplitude Interquartil

Amplitude

É a medida de dispersão mais simples,

$$\text{Amplitude} = \underbrace{x_{(n)}}_{\text{Máximo}} - \underbrace{x_{(1)}}_{\text{Mínimo}}$$

Medidas de dispersão: Amplitude e Amplitude Interquartil

Amplitude

É a medida de dispersão mais simples,

$$\text{Amplitude} = \underbrace{x_{(n)}}_{\text{Máximo}} - \underbrace{x_{(1)}}_{\text{Mínimo}}$$

Sua vantagem é o simples cálculo mas sua desvantagem é que depende apenas dos 2 valores mais extremos.

Medidas de dispersão: Amplitude e Amplitude Interquartil

Amplitude

É a medida de dispersão mais simples,

$$\text{Amplitude} = \underbrace{x_{(n)}}_{\text{Máximo}} - \underbrace{x_{(1)}}_{\text{Mínimo}}$$

Sua vantagem é o simples cálculo mas sua desvantagem é que depende apenas dos 2 valores mais extremos.

Observações extremas (outliers) afetarão a amplitude!

Medidas de dispersão: Amplitude e Amplitude Interquartil

Amplitude Interquartil (AIQ)

É a diferença entre o terceiro e primeiro quartil,

$$AIQ = Q_3 - Q_1$$

Medidas de dispersão: Amplitude e Amplitude Interquartil

Amplitude Interquartil (AIQ)

É a diferença entre o terceiro e primeiro quartil,

$$AIQ = Q_3 - Q_1$$

Não temos mais os problema da amplitude, mas nada sabemos dos outros 50% das observações.

Medidas de dispersão: Amplitude e Amplitude Interquartil

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Temos que $Q_3 = 8.8$ e $Q_1 = 6.5$. Então

Medidas de dispersão: Amplitude e Amplitude Interquartil

Exemplo

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

Temos que $Q_3 = 8.8$ e $Q_1 = 6.5$. Então

- Amplitude $= x(n) - x(1) = 9.6 - 2.1 = 7.5$
- $AIQ = Q_3 - Q_1 = 8.8 - 6.5 = 2.3$

Medidas de dispersão: Variância

- É uma mas das medidas de dispersão mais conhecidas e utilizadas

Medidas de dispersão: Variância

- É uma mas das medidas de dispersão mais conhecidas e utilizadas
- Seu cálculo utiliza todas as observações

Medidas de dispersão: Variância

- É uma mas das medidas de dispersão mais conhecidas e utilizadas
- Seu cálculo utiliza todas as observações
- Baseia-se na diferença (ao quadrado) dos valores observados e sua média.

Medidas de dispersão: Variância

- É uma mas das medidas de dispersão mais conhecidas e utilizadas
- Seu cálculo utiliza todas as observações
- Baseia-se na diferença (ao quadrado) dos valores observados e sua média.

Medidas de dispersão: Variância

- É uma mas das medidas de dispersão mais conhecidas e utilizadas
- Seu cálculo utiliza todas as observações
- Baseia-se na diferença (ao quadrado) dos valores observados e sua média.

Até agora, não temos feito diferença entre população e amostra. Isto, pois as formulas apresentadas anteriormente são as mesmas independente se as observações são da população ou da amostra.

Medidas de dispersão: Variância

- É uma mas das medidas de dispersão mais conhecidas e utilizadas
- Seu cálculo utiliza todas as observações
- Baseia-se na diferença (ao quadrado) dos valores observados e sua média.

Até agora, não temos feito diferença entre população e amostra. Isto, pois as formulas apresentadas anteriormente são as mesmas independente se as observações são da população ou da amostra.

Basicamente, onde tinhamos $x_1, x_2, \ldots, x_n$ com n sendo o tamaho da amostra, teremos $x_1, x_2, \ldots, x_N$ com N sendo o tamanho da população.

Medidas de dispersão: Variância

Aqui vamos diferenciar entre a variância populacional (σ^2) e a variância amostral (s^2)

Variância populacional

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

em que $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ é a média populacional e N é o tamanho (número de elementos) da população.

Medidas de dispersão: Variância

Aqui vamos diferenciar entre a variância populacional (σ^2) e a variância amostral (s^2)

Variância populacional

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

em que $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ é a média populacional e N é o tamanho (número de elementos) da população.

Na prática, dificilmente calculamos a variância populacional.

Medidas de dispersão: Variância

Variância amostral

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

em que $\bar{x}$ é a media amostral e n é o tamanho da amostra.

Medidas de dispersão: Variância

Variância amostral

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

em que $\bar{x}$ é a media amostral e n é o tamanho da amostra.

Na prática, utilizamo s^2 para estimar o σ^2.

Medidas de dispersão: Variância

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

tinhamos que $\bar{x} = 7.322222$. Vamos calcular s^2.

Medidas de dispersão: Variância

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

tinhamos que $\bar{x} = 7.322222$. Vamos calcular s^2.

$$s^2 = \frac{(2.1 - \bar{x})^2 + (5.6 - \bar{x})^2 + \cdots + (9.6 - \bar{x})^2}{18 - 1} = \frac{53.01111}{17} = 3.118301$$

Medidas de dispersão: Desvio Padrão

- A variância não preserva a mesma unidade dos dados originais (lembre-se, elevamos ao quadrado.)

Medidas de dispersão: Desvio Padrão

- A variância não preserva a mesma unidade dos dados originais (lembre-se, elevamos ao quadrado.)
- Para facilitar a compreensão e interpretação, uma medida de dispersão que preserve a mesma unidade dos dados originais é desejada.

Medidas de dispersão: Desvio Padrão

- A variância não preserva a mesma unidade dos dados originais (lembre-se, elevamos ao quadrado.)
- Para facilitar a compreensão e interpretação, uma medida de dispersão que preserve a mesma unidade dos dados originais é desejada.
- Isto é obtido com a raíz quadrada da variância, essa medida de dispersão recebe o nome de **Desvio Padrão**

Medidas de dispersão: Desvio Padrão

- A variância não preserva a mesma unidade dos dados originais (lembre-se, elevamos ao quadrado.)
- Para facilitar a compreensão e interpretação, uma medida de dispersão que preserve a mesma unidade dos dados originais é desejada.
- Isto é obtido com a raíz quadrada da variância, essa medida de dispersão recebe o nome de **Desvio Padrão**

Medidas de dispersão: Desvio Padrão

- A variância não preserva a mesma unidade dos dados originais (lembre-se, elevamos ao quadrado.)
- Para facilitar a compreensão e interpretação, uma medida de dispersão que preserve a mesma unidade dos dados originais é desejada.
- Isto é obtido com a raíz quadrada da variância, essa medida de dispersão recebe o nome de **Desvio Padrão**

Desvio Padrão

- Desvio padrão da população: $\sigma = \sqrt{\sigma^2}$
- Desvio padrão da amostra: $s = \sqrt{s^2}$

Outras medidas de disperão

Desvio absoluto médio (DAM)

$$DAM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Coeficiente de variação (CV)

$$CV = \Big(\frac{\text{Desvio Padrão}}{\text{Média}} \times 100\Big)\%$$

O CV é interessante pois ele nos diz qual o tamanho do desvio padrão em relação à média.

Outras medidas de disperão

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

tinhamos que $\bar{x} = 7.322222$ e $s^2 = 3.118301$. Vamos calcular o CV.

Outras medidas de disperão

No nosso conjunto de dados

2.1	5.6	6	6.3	6.5	6.7	7.0	7.0	7.2
7.7	7.8	8	8.1	8.8	9.0	9.1	9.3	9.6

tinhamos que $\bar{x} = 7.322222$ e $s^2 = 3.118301$. Vamos calcular o CV.

$$CV = \Big(\frac{\text{Desvio Padrão}}{\text{Média}} \times 100\Big)\% = \Big(\frac{\sqrt{3.118301}}{7.322222} \times 100\Big)\% = 24.1166\%$$

O desvio padrão é $\approx 24.11\%$ do valor da média.

Gráficos para variáveis quantitativas

Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.

Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.
- É contruido utilizando 5 valores:

Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.
- É contruido utilizando 5 valores:
  - Medina (Q_2)

Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.
- É contruido utilizando 5 valores:
    - Medina (Q_2)
    - Quartil 1 (Q_1)

Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.
- É contruido utilizando 5 valores:
  - Medina (Q_2)
  - Quartil 1 (Q_1)
  - Quartil 3 (Q_3)

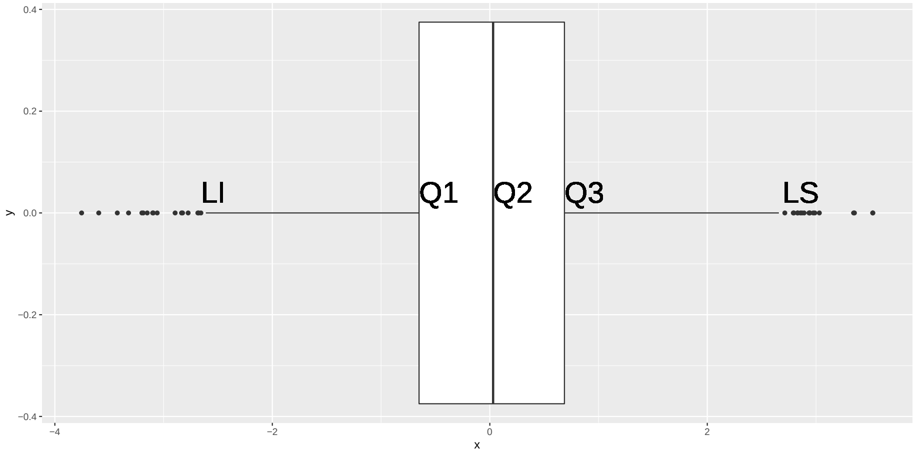
Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.
- É contruido utilizando 5 valores:
  - Medina (Q_2)
  - Quartil 1 (Q_1)
  - Quartil 3 (Q_3)
  - LS $= Q_3 + 1.5 \; AIQ$

Boxplot

- Traz informação do valor central, variabilidade, observações extremas e simetria.
- É contruido utilizando 5 valores:
  - Medina (Q_2)
  - Quartil 1 (Q_1)
  - Quartil 3 (Q_3)
  - LS $= Q_3 + 1.5\ AIQ$
  - LI $= Q_1$ - $1.5 AIQ$

Boxplot



Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).

Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.

Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.
- Em geral, nos permite conhecer como os dados estão distribuidos

Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.
- Em geral, nos permite conhecer como os dados estão distribuidos

Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.
- Em geral, nos permite conhecer como os dados estão distribuidos

Como calcular?

- Procedemos da mesma forma em que construimos as tabelas de frequência para variáveis continuas (Aula 3). Precisaremos formas as classes, definir a amplitude de classe, os limites da classe e a frequência da classe.

Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.
- Em geral, nos permite conhecer como os dados estão distribuidos

Como calcular?

- Procedemos da mesma forma em que construimos as tabelas de frequência para variáveis continuas (Aula 3). Precisaremos formas as classes, definir a amplitude de classe, os limites da classe e a frequência da classe.
- Algumas regras praticas para escolher o número de classes k são:

Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.
- Em geral, nos permite conhecer como os dados estão distribuidos

Como calcular?

- Procedemos da mesma forma em que construimos as tabelas de frequência para variáveis continuas (Aula 3). Precisaremos formas as classes, definir a amplitude de classe, os limites da classe e a frequência da classe.
- Algumas regras praticas para escolher o número de classes k são:
  - Sturges: $k = 1 + 3.322\log(n)$

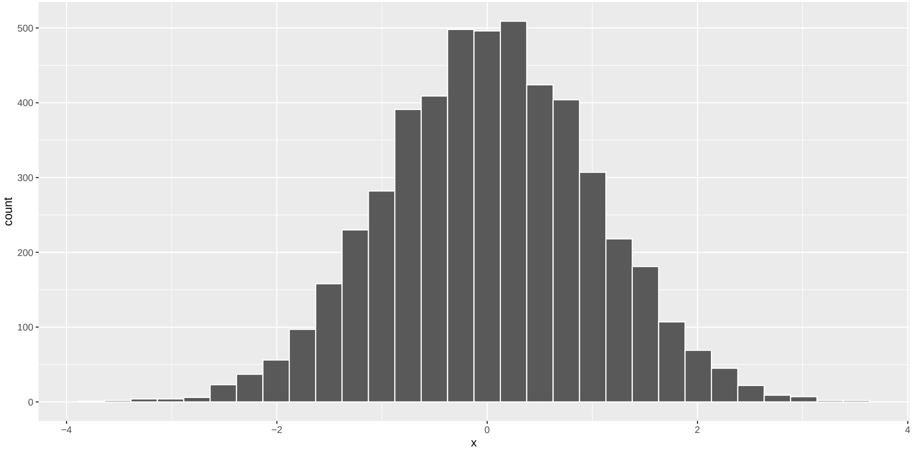
Histograma

- O histograma é um gráfico formado por barras que indicam a frequência dos dados (previamente agrupados em clases).
- Nos permite ter uma ideia da variabilidade e simetria dos dados.
- Em geral, nos permite conhecer como os dados estão distribuidos

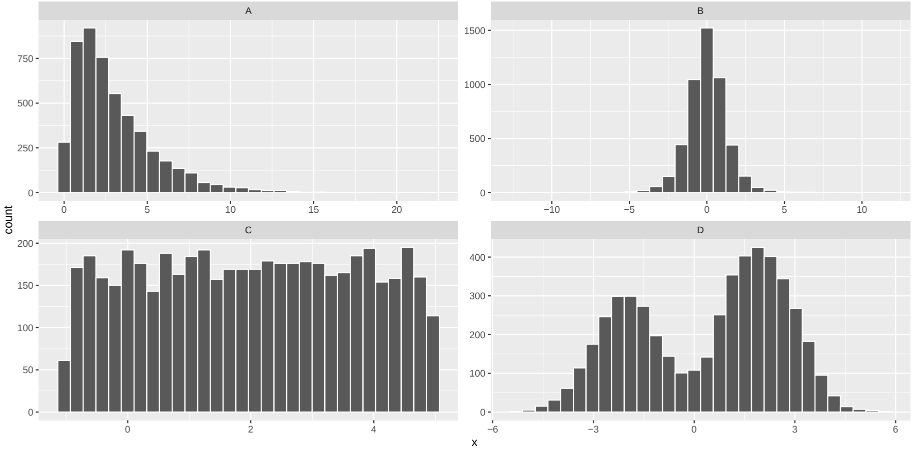
Como calcular?

- Procedemos da mesma forma em que construimos as tabelas de frequência para variáveis continuas (Aula 3). Precisaremos formas as classes, definir a amplitude de classe, os limites da classe e a frequência da classe.
- Algumas regras praticas para escolher o número de classes k são:
  - Sturges: $k = 1 + 3.322 \log(n)$
  - $k = \sqrt{n}$

Histograma



Histograma


A
B
C
D
count
x
0
250
500
750
1500
1000
−10
−5
5
10
15
20
50
100
150
200
2
4
300
400
−6
−3
3
6

Medidas de associação entre duas variáveis

Medidas de associação entre duas variáveis.

- Frequentemente estamos interessados na relação de associação entre duas variáveis.

Medidas de associação entre duas variáveis.

- Frequentemente estamos interessados na relação de associação entre duas variáveis.
- Nesta seção aprenderemos sobre o gráfico de dispersão e duas medias de associação amplamente utilizadas: covariância e correlação.

Medidas de associação entre duas variáveis.

- Frequentemente estamos interessados na relação de associação entre duas variáveis.
- Nesta seção aprenderemos sobre o gráfico de dispersão e duas medias de associação amplamente utilizadas: covariância e correlação.

Medidas de associação entre duas variáveis.

- Frequentemente estamos interessados na relação de associação entre duas variáveis.
- Nesta seção aprenderemos sobre o gráfico de dispersão e duas medias de associação amplamente utilizadas: covariância e correlação.

Gráfico de dispersão Bi-dimensional

- Também conhecido como *Nuvem de pontos* ou *scatter plot*.
- É uma representação gráfica no plano cartesiano dos pares (x, y) em que x e y são os valores observados das duas variáveis em análise.

Medidas de associação entre duas variáveis.

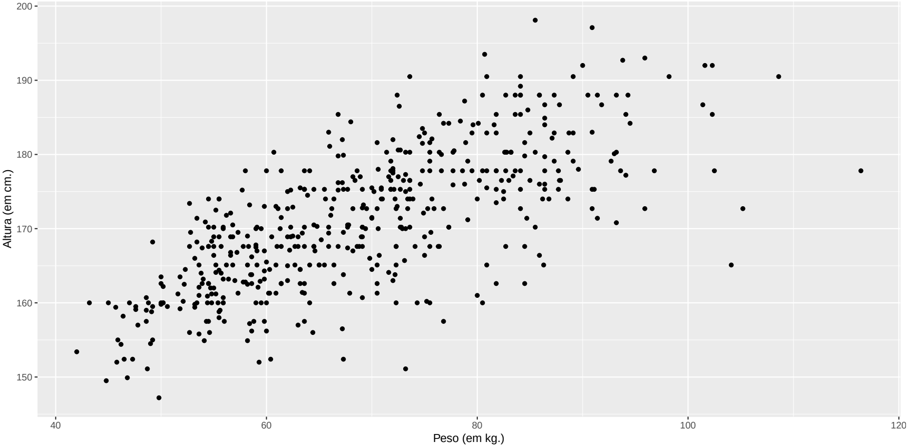


Figura 1: Gráfico de dispersão (peso X altura) de 507 individuos

Você acha que existe alguma relação de associação entre *altura* e *peso*?

Medidas de associação entre duas variáveis.

Sejam X e Y duas variáveis de interesse com $(x_1, y_1), \ldots, (x_n, y_n)$ e os valores observados de X e Y em uma amostra de tamanho n.

Covariância amostral

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Medidas de associação entre duas variáveis.

Sejam X e Y duas variáveis de interesse com $(x_1, y_1), \ldots, (x_n, y_n)$ e os valores observados de X e Y em uma amostra de tamanho n.

Covariância amostral

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

No conjunto de dados utilizado no gráfico de dispersão temos que $s_{xy} \approx 90.05$. Como interpretar esse valor?

Medidas de associação entre duas variáveis.

Sejam X e Y duas variáveis de interesse com $(x_1, y_1), \ldots, (x_n, y_n)$ e os valores observados de X e Y em uma amostra de tamanho n.

Covariância amostral

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

No conjunto de dados utilizado no gráfico de dispersão temos que $s_{xy} \approx 90.05$. Como interpretar esse valor?

- valores positivos indicam uma relação linear direta (ou positiva)

Medidas de associação entre duas variáveis.

Sejam X e Y duas variáveis de interesse com $(x_1, y_1), \ldots, (x_n, y_n)$ e os valores observados de X e Y em uma amostra de tamanho n.

Covariância amostral

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

No conjunto de dados utilizado no gráfico de dispersão temos que $s_{xy} \approx 90.05$. Como interpretar esse valor?

- valores positivos indicam uma relação linear direta (ou positiva)
- valores negativos indicam uma relação linear inversa (ou negativa)

Medidas de associação entre duas variáveis.

Sejam X e Y duas variáveis de interesse com $(x_1, y_1), \ldots, (x_n, y_n)$ e os valores observados de X e Y em uma amostra de tamanho n.

Covariância amostral

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

No conjunto de dados utilizado no gráfico de dispersão temos que $s_{xy} \approx 90.05$. Como interpretar esse valor?

- valores positivos indicam uma relação linear direta (ou positiva)
- valores negativos indicam uma relação linear inversa (ou negativa)
- valores muito proximos de zero indicam que não há nenhuma associação linear entre as variáveis

Medidas de associação entre duas variáveis.

Sejam X e Y duas variáveis de interesse com $(x_1, y_1), \ldots, (x_n, y_n)$ e os valores observados de X e Y em uma amostra de tamanho n.

Covariância amostral

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

No conjunto de dados utilizado no gráfico de dispersão temos que $s_{xy} \approx 90.05$. Como interpretar esse valor?

- valores positivos indicam uma relação linear direta (ou positiva)
- valores negativos indicam uma relação linear inversa (ou negativa)
- valores muito proximos de zero indicam que não há nenhuma associação linear entre as variáveis
- $s_{xy} = s_{yx}$

Medidas de associação entre duas variáveis.

- No exemplo anterior vimos que $s_{xy} \approx 90.05$ o que implica uma relação positiva, mas *quão forte é essa relação?*

Medidas de associação entre duas variáveis.

- No exemplo anterior vimos que $s_{xy} \approx 90.05$ o que implica uma relação positiva, mas *quão forte é essa relação?*
- Para responder essa pergunta precisamos de algum valor de referência para saber se a relação é forte ou não.

Medidas de associação entre duas variáveis.

- No exemplo anterior vimos que $s_{xy} \approx 90.05$ o que implica uma relação positiva, mas *quão forte é essa relação?*
- Para responder essa pergunta precisamos de algum valor de referência para saber se a relação é forte ou não.
- Além disso, o valor da covariância depende das unidades de medida (por exemplo, se utizarmos a altura em metros e não em centímetros teremos que $s_{xy} \approx 0.9$)

Medidas de associação entre duas variáveis.

Coeficiente de correlação de Pearson

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

em que

- s_{xy} é a covariância amostral entre x e y,
- s_x é o desvio padrão de x e
- s_y é o desvio padrão de y.

Medidas de associação entre duas variáveis.

Coeficiente de correlação de Pearson

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

em que

- s_{xy} é a covariância amostral entre x e y,
- s_x é o desvio padrão de x e
- s_y é o desvio padrão de y.

Propriedades

- $r_{xy} = r_{yx}$

Medidas de associação entre duas variáveis.

Coeficiente de correlação de Pearson

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

em que

- s_{xy} é a covariância amostral entre x e y,
- s_x é o desvio padrão de x e
- s_y é o desvio padrão de y.

Propriedades

- $r_{xy} = r_{yx}$
- $-1 \leq r_{xy} \leq 1$

Coeficiente de correlação de Pearson

Exemplo

- No conjunto de dados utilizado no gráfico de dispersão temos que $r_{xy} \approx 0.72$.

Coeficiente de correlação de Pearson

Exemplo

- No conjunto de dados utilizado no gráfico de dispersão temos que $r_{xy} \approx 0.72$.
- Como 0.72 é positivo e proximo de 1, dizemos que a relação entre é x e y é positiva (ou direta) e que esta relação é forte

Coeficiente de correlação de Pearson

Exemplo

- No conjunto de dados utilizado no gráfico de dispersão temos que $r_{xy} \approx 0.72$.
- Como 0.72 é positivo e proximo de 1, dizemos que a relação entre é x e y é positiva (ou direta) e que esta relação é forte

Coeficiente de correlação de Pearson

Exemplo

- No conjunto de dados utilizado no gráfico de dispersão temos que $r_{xy} \approx 0.72$.
- Como 0.72 é positivo e proximo de 1, dizemos que a relação entre é x e y é positiva (ou direta) e que esta relação é forte

Na próxima aula utilizaremos um conjunto de dados real e faremos, com ajuda do *R*, um pouco de análise exploratória de dados utilizando o visto até aqui.

Outros coeficientes de correlação

- O coeficiente de correlação de Pearson é util quando as duas variáveis de interesse são continuas.
- Contudo, às vezes queremos calcular a correlação entre outros tipos de variáveis. Para isto, existem outros coeficientes de correlação.
  - Coeficiente de correlação de Spearman (recomendado quando os dados estão em escala ordinal)
  - Coeficiente de correlação de Kendall (recomendado quando os dados estão em escala ordinal)
  - Coeficiente de contingência (se usa quando as duas variáveis estão em escala nomial)
  - Etc.

Leituras recomendadas

- Anderson, D. R; Sweeney, D. J.; e Williams, T. A. (2008). *Estatística Aplicada à Administração e Economia*. 2ed. Cengage Learning. **Cap 3**
- Freund, J. E.; Perles, B. M. (2014). *Modern elementary statistics*. 12ed. Pearson College Division. **Chapter 2 - 3**
- Morettin, P. A; Bussab, W. O (2004). *Estatística Básica*. 5ed. Saraiva. **Cap 3**