Estatística Básica

Primeiros Conceitos

Sumário

Introdução

Estatística Descritiva Básica

Está apresentação tem o obejetivo de apresentar os conceitos básicos da estatística descritiva: termos, números resumos, tipos de frequências e gráficos.

Termos

Termos

Definicão (Estatística Descritiva)

Consiste em resumir os dados observados em forma de tabelas, gráficos e números resumo.

Definicão (Estatística Inferencial)

É o processo que permite generalizar um conhecimento obtido a partir de resultados particulares para a totalidade dos dados.

Termos

Definicão (População)

É o conjunto de todos os elementos relativos a um determinado fenômeno que possuem pelo menos uma característica em comum.

Definicão (Amostra)

É um subconjunto da população e deverá ser considerada finita.

Termos

Definicão (Censo)

É a coleta exaustiva de informações das "N"unidades populacionais.

Definicão (Amostragem)

É o processo de retirada de informações dos "n"elementos amostrais, no qual deve seguir um método criterioso e adequado (tipos de amostragem).

Temos

Definicão (Dados estatístico)

É qualquer característica que possa ser observada ou medida por um processo.

Definicão (Variável)

É a característica que se deseja observar para se tirar algum tipo de conclusão.

Termos

Classificação das variáveis

- Variável
  - Qualitativa
    - Nominal
    - Ordinal
  - Quantitativa
    - Discreta
    - Contínua

Termos

Classificação das variáveis

- Qualitativas (ou atributos): São características de uma população que não pode ser medidas.
  - Nominal : são utilizados símbolos, ou números, para representar determinado tipo de dados, mostrando, assim, a qual grupo ou categoria eles pertencem.
  - Ordinal ou por postos: quando uma classificação for dividida em categorias ordenadas em graus convencionados, havendo uma relação entre as categorias do tipo “maior do que”,“menor do que”, “igual a”, os dados por postos consistem de valores relativos atribuídos para denotar a ordem de primeiro, segundo, terceiro e, assim, sucessivamente.

Termos

Classificação das variáveis

- Quantitativas: São características populacionais que podem ser quantificadas, sendo classificadas em discretas e contínuas.
  - Discretas: são aquelas variáveis que pode assumir somente valores inteiros num conjunto de valores. É gerada pelo processo de contagem.
  - Contínuas: são aquelas variáveis que podem assumir um valor dentro de um intervalo de valores. É gerada pelo processo de medição.

Números Resumos

Números Resumos

- Estatísticas
  - Tendência Central
    - Moda
    - Média
    - Mediana
  - Separatrizes
    - Valor máximo e Valor mínimo
    - Quartil
    - Percentis
  - Dispersão
    - Amplitude
    - Variâcia
    - Coeficiente de Variação
  - Assimetria e Curtose

Números Resumos

Definicão (Médidas de Têndência Central)

As medidas de tendência central, ou promédias, são valores entorno dos quais os dados observados tendem a se agruparem.

A média aritmética, a mediana e a moda são as mais conhecidas

Números Resumos

Definicão (Média)

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

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

Em palavras, a média a razão entre a soma e a quantidade de valores observados.

Numeros Resumos

Para os dados $x_1, x_2, \ldots, x_n$, com média $\bar{x}$, e uma constante $c \in \mathbb{R}$ são válidas as seguintes propriedades:

1. A soma dos desvios $(x_i - \bar{x})$, para todo $i = 1, \ldots, n$, é zero.
2. Somar (subtrair) uma constante c a todo valor x_i, soma (subtrai) a constante c da média $\bar{x}$.
3. Multiplicar (dividir) uma constante c a todo valor x_i, multiplica (dividir) por c a média $\bar{x}$.
4. A soma dos quadrados dos desvios em relação a média é menor ou igual a soma dos quadrados dos desvios em relação a qualquer valor de c, isto é,

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

Numeros Resumos

Definicão (Rol)

É uma lista de dados quantitativos ordenados em ordem crescente ou decrescente.

Definicão (Mediana)

É o valor M_d que divide o rol de dados em dois grupos com o mesmo número de elementos.

Numeros Resumos

Para os dados $x_1, x_2, \ldots, x_n$, com mediana M_d, e uma constante $c \in \mathbb{R}$ são válidas as seguintes propriedades:

1. Somar (subtrair) uma constante c a todo valor x_i, soma (subtrai) a constante c a mediana M_d.
2. Multiplicar (dividir) uma constante c a todo valor x_i, multiplica (dividir) por c a mediana M_d.
3. A mediana é robusta à presença de valores extremos ou *outliers*.
4. A mediana não é robusta a alterações no número de elementos n.

Definicão (Moda)

É o valor M_o mais frequente em um conjunto de dados.

Para dados agrupados em classe, há três métodos para determinar a moda M_o: Moda Bruta, King e Czuber. Os três baseiam-se na classe de dados com a maior frequência.

Números Resumos

Definicão (Classe modal)

É a classe de dados com maior frequência.

Definicão (Moda Bruta)

É o ponto médio da classe modal.

Definicão (Método de King)

É baseado na influência da frequência da classe anterior e da posterior sobre a classe modal. É dada pela fórmula.

$$M_o = l + c\frac{f_{post}}{f_{ant} + f_{post}},$$

onde

- *l é o limite inferior da classe modal.*
- *c é a amplitude da classe modal.*
- f_{ant} *é a frequência da classe anterior a classe modal.*
- f_{post} *é a frequência da classe posterior a classe modal.*

Definicão (Método de King)

É baseado na influência da frequência da classe anterior e da posterior sobre a classe modal. É dada pela fórmula.

$$M_o = l + c\frac{f_{post}}{f_{ant} + f_{post}},$$

onde

- *l é o limite inferior da classe modal.*
- *c é a amplitude da classe modal.*
- *f_{ant} é a frequência da classe anterior a classe modal.*
- *f_{post} é a frequência da classe posterior a classe modal.*

Definicão (Método de Czuber)

É baseado nas diferença entre frequência da classe modal e da classe anterior e da classe modal e a classe posterior. É dada pela fórmula.

$$M_o = l + c\frac{f_{mo} - f_{ant}}{2f_{mo} - (f_{ant} + f_{post})},$$

onde

- *l é o limite inferior da classe modal.*
- *c é a amplitude da classe modal.*
- f_{mo} *é a frequência da classe modal.*
- f_{ant} *é a frequência da classe anterior a classe modal.*
- f_{post} *é a frequência da classe posterior a classe modal.*

Números Resumos

Definicão (Separatrizes)

As separatrizes são valores que dividem os dados em conjuntos com o mesmo número de elementos aproximadamente.

As separtrizes mais conhecidas são: os quartis e os percentis

- quartis divide os dados em quatro partes com o mesmo número de elementos.
- percentis divide os dados em cem partes com o mesmo número de elementos.

Números Resumos

Definicão (Quatis)

São os valores $Q_1 \leq Q_2 \leq Q_3$ que dividem os dados em quatro partes com, aproximadamente, o mesmo número de elementos.

O valor Q_1 é chamado de 1° quartil, o Q_2 é o 2° quartil e Q_3 é o 3° quartil.

Pode-se caracterizar um quartil pela quantidade de valores maiores e menores que ele.

1. Q_1 - 25% são menores e 75% são maiores.
2. Q_2 - 50% são menores e 50% são maiores.
3. Q_3 - 75% são menores e 25% são maiores.

Números Resumos

Definicão (Percentis)

São os valores $P_1 \leq P_2 \leq \cdots \leq P_{99}$ que dividem os dados em 100 partes com, aproximadamente, o mesmo número de elementos.

O valor P_1 é o 1º percentil, P_2 é o 2º percentil, P_i é o $i - esimo$ percentil e P_{99} é o 99º percentil.

Pode-se caracterizar um quartil pela quantidade de valores maiores e menores que ele.

1. P_1 - 1% são menores e 99% são maiores.
2. P_{50} - 50% são menores e 50% são maiores.
3. P_i - i% são menores e $(1 - i)$% são maiores.

Definicão (Medidas de Dispersão)

As medidas de dispersão indicam a proximidade dos dados em relação a um valor.

As medidas de dispersão mais conhecidas são: variância, desvio padrão, amplitude total, desvio médio absoluto, amplitude semi-interquartílica e coeficiente de variação.

Definicão (Amplitude Total)

*Para os dados $x_1, \ldots, x_n$ com valor máximo (x_{max}) e valor mínimo *(x_{min}), a amplitude total é dada por*

$$A_t = x_{max} - x_{min}$$

A_t é a diferença entre o maior e o menor valor observado em um conjuto de dados, considera apenas os valores extremos e é de fácil obtenção.

Definicão (Variância populacional)

Para uma população de n dados $x_1, \ldots, x_n$, com média aritmética $\bar{x}$, é definida por

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

Assim, a variância é a média aritmética do quadrado da diferença entre cada valor x_i e a média $\bar{x}$ dos dados.

Definicão (Variância amostral)

Para uma amostra populacional de n dados $x_1, \ldots, x_n$, com média aritmética $\bar{x}$, é definida por

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

A teoria de probabilidades justifica a relação entre a variância amostral s_x^2 e a populacional s^2 dado por

$$s^2 = \frac{n}{n-1} s_x^2,$$

que justifica a troca dos denominadores de n por $(n-1)$ na fórmula da variância populacional para obter a amostral. Entretanto, na prática, os seus valores diferem pouco.

Números Resumos

Sejam um conjunto de n dados $x_i, \ldots, x_n$ e uma constante $k \in \mathbb{R}$. A variância, populacional ou amostral, apresentam as seguintes propriedades:

1. Adicionar ou subtrair o valor k a todos os dados não altera a sua variância.
2. Multiplicar pelo valor k todos dados multiplica a sua variância por k^2.

Definicão (Desvio padrão populacional)

Para uma população de n dados $x_1, \ldots, x_n$, com média aritmética $\bar{x}$, é definida por

$$s_x^2 = \sqrt[2]{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

Assim, o desvio padrão é a raiz quadrada da média aritmética do quadrado da diferença entre cada valor x_i e a média $\bar{x}$ dos dados.

Definicão (Desvio parão amostral)

Para uma amostra populacional de n dados $x_1, \ldots, x_n$, com média aritmética $\bar{x}$, é definida por

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

Números Resumos

Sejam um conjunto de n dados $x_i, \ldots, x_n$ e uma constante $k \in \mathbb{R}$. O desvio padrão, populacional ou amostral, apresenta as seguintes propriedades:

1. Adicionar ou subtrair o valor k a todos os dados não altera o seu desvio padrão.
2. Multiplicar pelo valor k todos dados multiplica o seu desvio padrão por k.
3. O desvio padrão de um conjunto de dados é a raiz quadrada positiva da variância deles.

Definicão (Coeficiente de Variação)

Para os dados $x_1, \ldots, x_n$ com desvio padrão s_x e média $\bar{x}$, o coeficiente de variação é dado por

$$C_v = \frac{s_x}{\bar{x}} \text{ ou por } C_v = \frac{s_x}{\bar{x}} \times 100,$$

nessa última forma, para ser expresso em porcentagem.

O coeficiente de variação deve ser interpretado como a variabilidade dos dados em relação à média. Assim, quanto menor for o coeficiente mais homogêneo é o conjunto de dados.

Tipos de Frequência

Tipos de Frequências

Definicão (Distribuição de Frequência)

É uma serie estatística na qual permanece constante o fato, o local e a época. Onde o interesse se concentra nas repetições dos valores.

Geralmente a distribuição de frequência é representada graficamente em forma de tabela, que faz corresponder os valores observados da variável em estudo e as suas respectivas frequências.

Tipos de Frequências

Seja x_j, a representação de um valor ou de uma classe de valores, a sua frequência pode ser apresentada como no esquema abaixo:

- Frequência
  - Simples
    - Absoluta
    - Relativa
  - Acumulada
    - Absoluta
      - Acima de
      - Abaixo de
    - Relativa
      - Acima de
      - Abaixo de

Tipos de Frequências

- **Frequência simples absoluta** (f) é o número de repetições de x_j ou de valores na classe j.
- **Frequência simples relativa** (fr) é a razão entre o número de repetições de x_j ou de valores na classe j e o número total de observações.
- **Frequência acumulada absoluta do tipo abaixo de** (F_{abaixo}) é a soma das frequências simples dos valores ou classes menores ou iguais a x_j
- **Frequência acumulada relativa do tipo abaixo de** (Fr_{abaixo}) é a razão entre a soma das frequências relativa dos valores ou classes menores ou iguais a x_j e o total de observações.
- **Frequência acumulada absoluta do tipo acima de** (F_{acima}) é a soma das frequências simples dos valores ou classes maiores ou iguais a x_j
- **Frequência acumulada relativa do tipo abaixo de** (Fr_{abaixo}) é a razão entre a soma das frequências relativa dos valores ou classes menores ou iguais a x_j e o total de observações.

Tipos de Frequências

Tabela 1: Frequências do número de acertos de 50 alunos em uma prova.

Acertos	f	$f_r(\%)$	F_{abaixo}	Fr_{abaixo} (%)
2	1	2	1	2
3	4	8	5	10
4	17	34	22	44
5	14	18	36	72
6	7	14	43	86
7	3	6	46	92
8	3	6	49	98
2	1	2	50	100

Tipos de Frequência

Tabela 2: Frequência simples (f) da massa, em gramas, de 140 grãos de feijão.

Classe	Massa	f
1	$0,10 \vdash 0,12$	1
2	$0,12 \vdash 0,14$	4
3	$0,14 \vdash 0,16$	11
4	$0,16 \vdash 0,18$	24
5	$0,18 \vdash 0,20$	32
6	$0,20 \vdash 0,22$	27
7	$0,22 \vdash 0,24$	17
8	$0,24 \vdash 0,26$	15
9	$0,26 \vdash 0,28$	7
10	$0,28 \vdash 0,30$	1
11	$0,30 \vdash 0,32$	1

Gráficos

Definicão (Histograma)

Representação gráfica para um conjunto de dados quantitativos por meio de retângulos justapostos, cujas áreas são proporcionais às frequências das classes que representam.

Gráficos

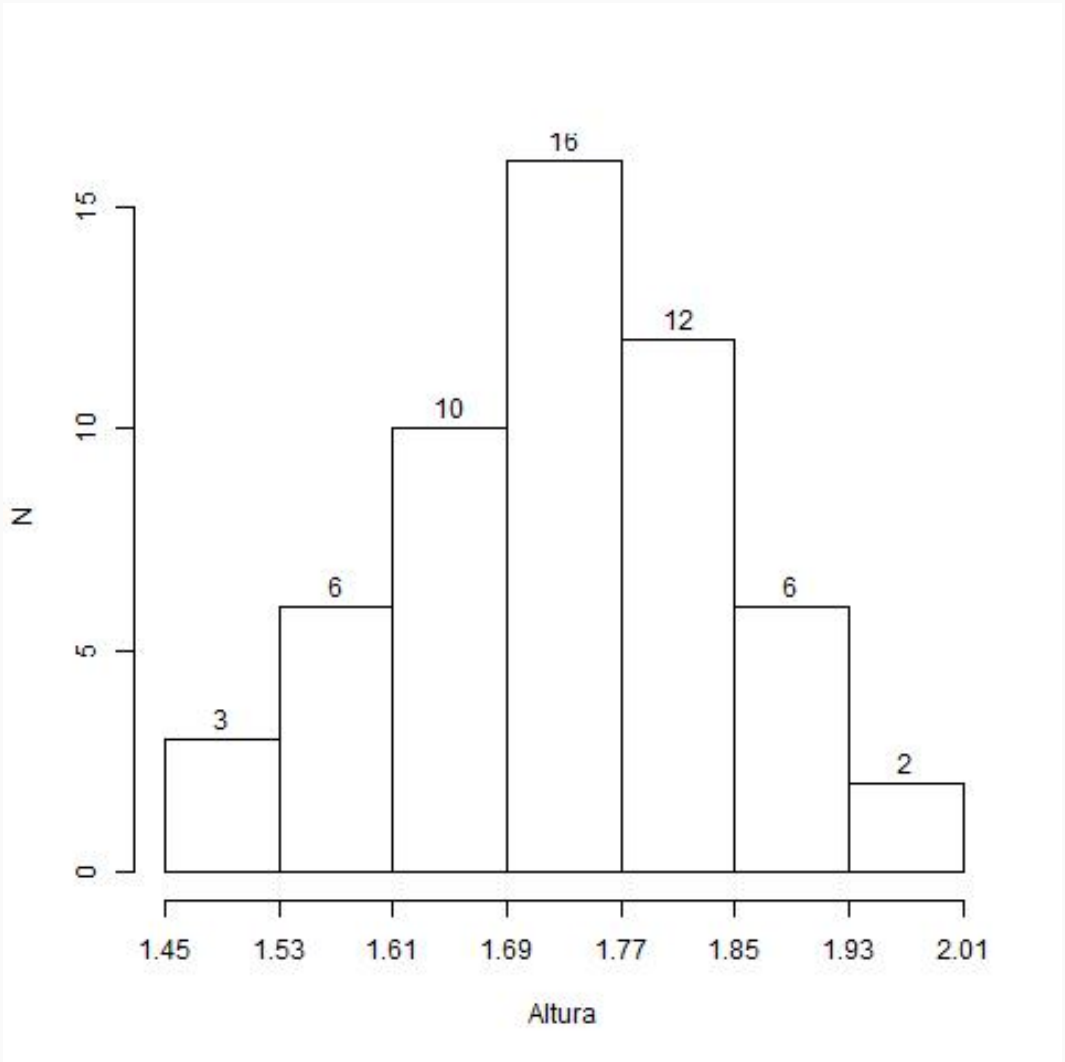

Figura 1: Alturas(m) de 55 homens para a frequência simples.

Gráficos

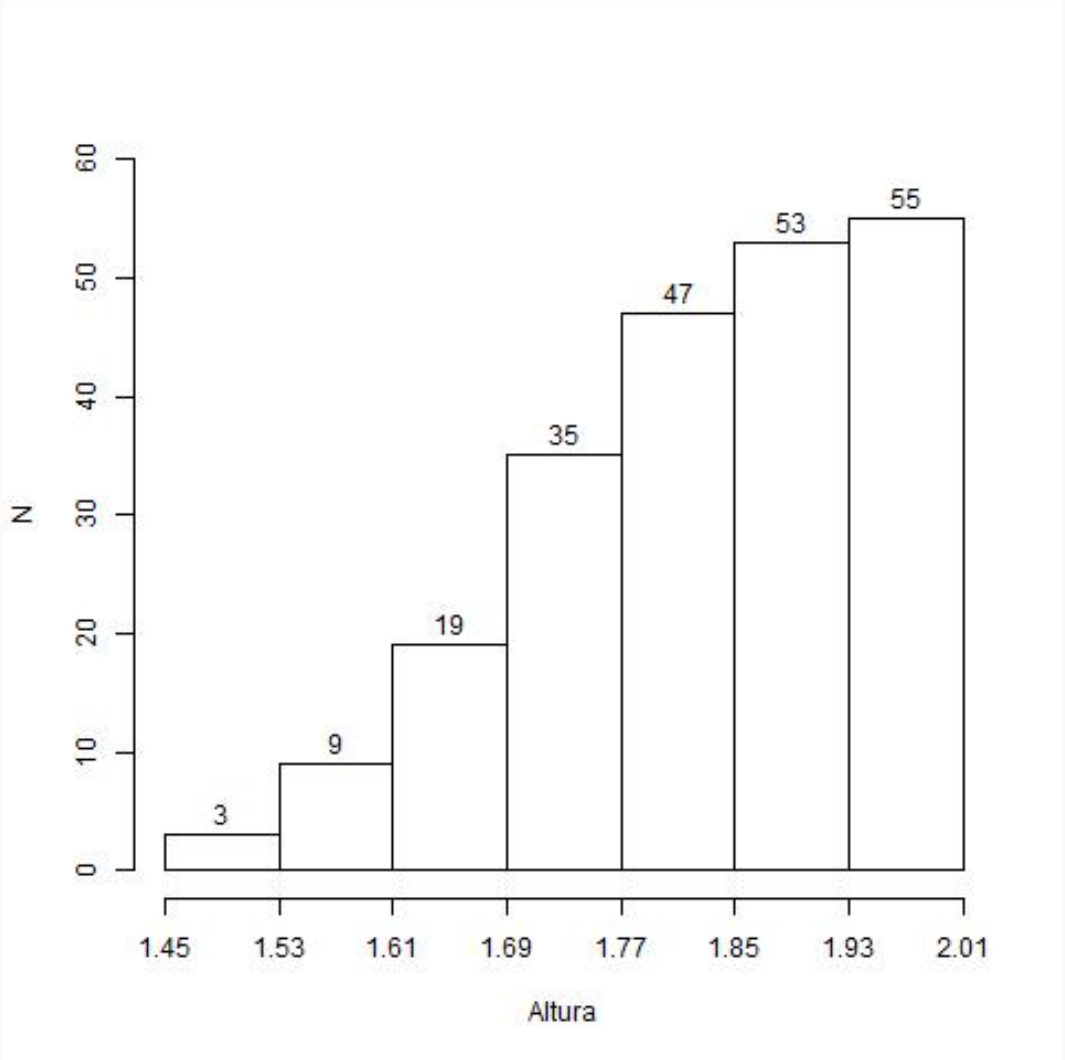

Figura 2: Alturas(m) de 55 homens para a frequência acumulada.

Definicão (Polígono de frequência)

É a linha formada pela união dos pontos médios dos topos das colunas de um histograma.

Para "aterrisar" o polígono de frequências é usual ligar o seu primeiro porto ao limite inferior da primeira classe e o último ao limite superior da última classe.

Definicão (Ogiva de Galton)

É a linha formada pela união dos pontos médios dos topos das colunas de um histograma elaborado a partir de frequências acumuladas.

A Ogiva de Galton é um caso particular de polígono de frequência e ele aproxima a forma da função de distribuição de probabilidade que pode estar associada ao fenômeno estudado.

Gráficos

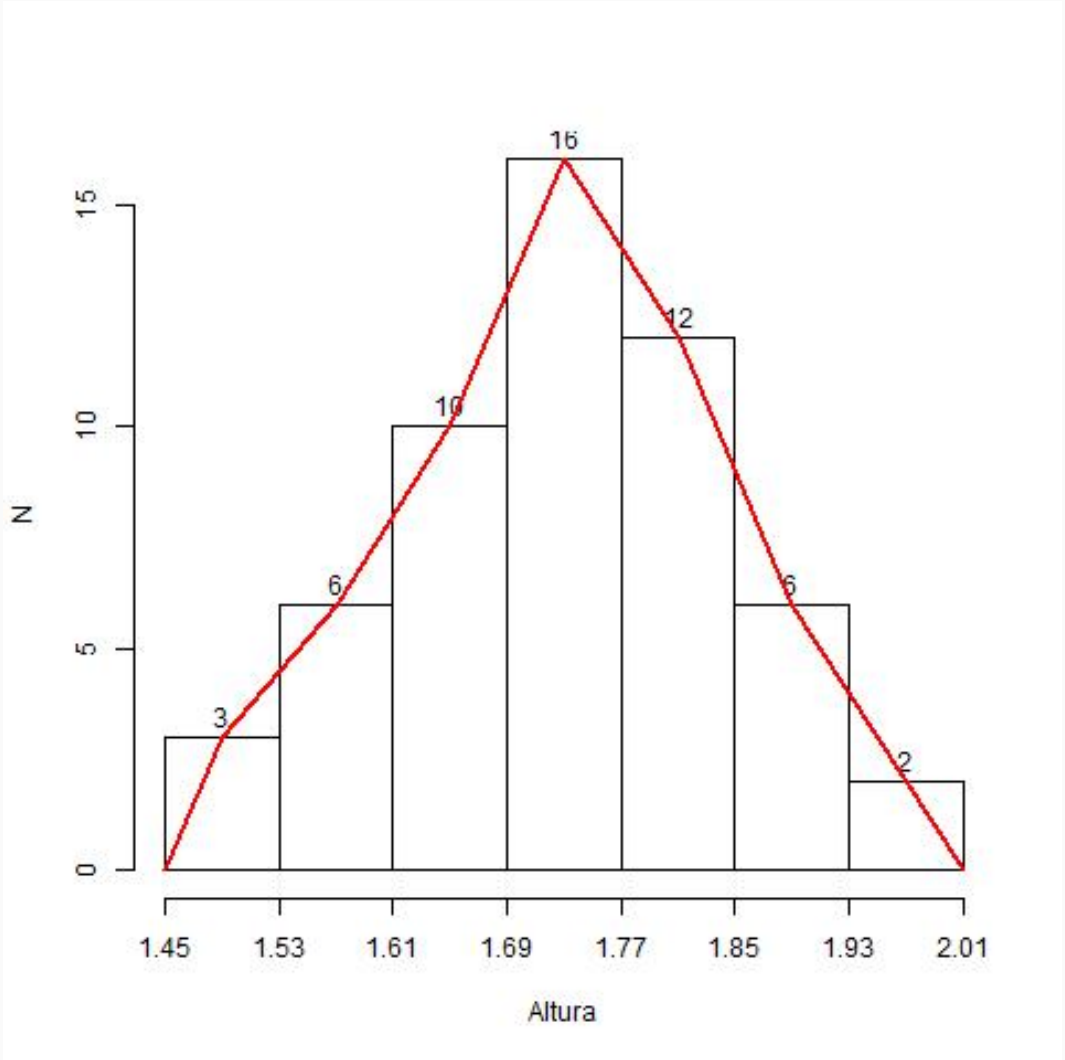

Figura 3: Poligono de frequências das alturas(m) de 55 homens.

Gráficos

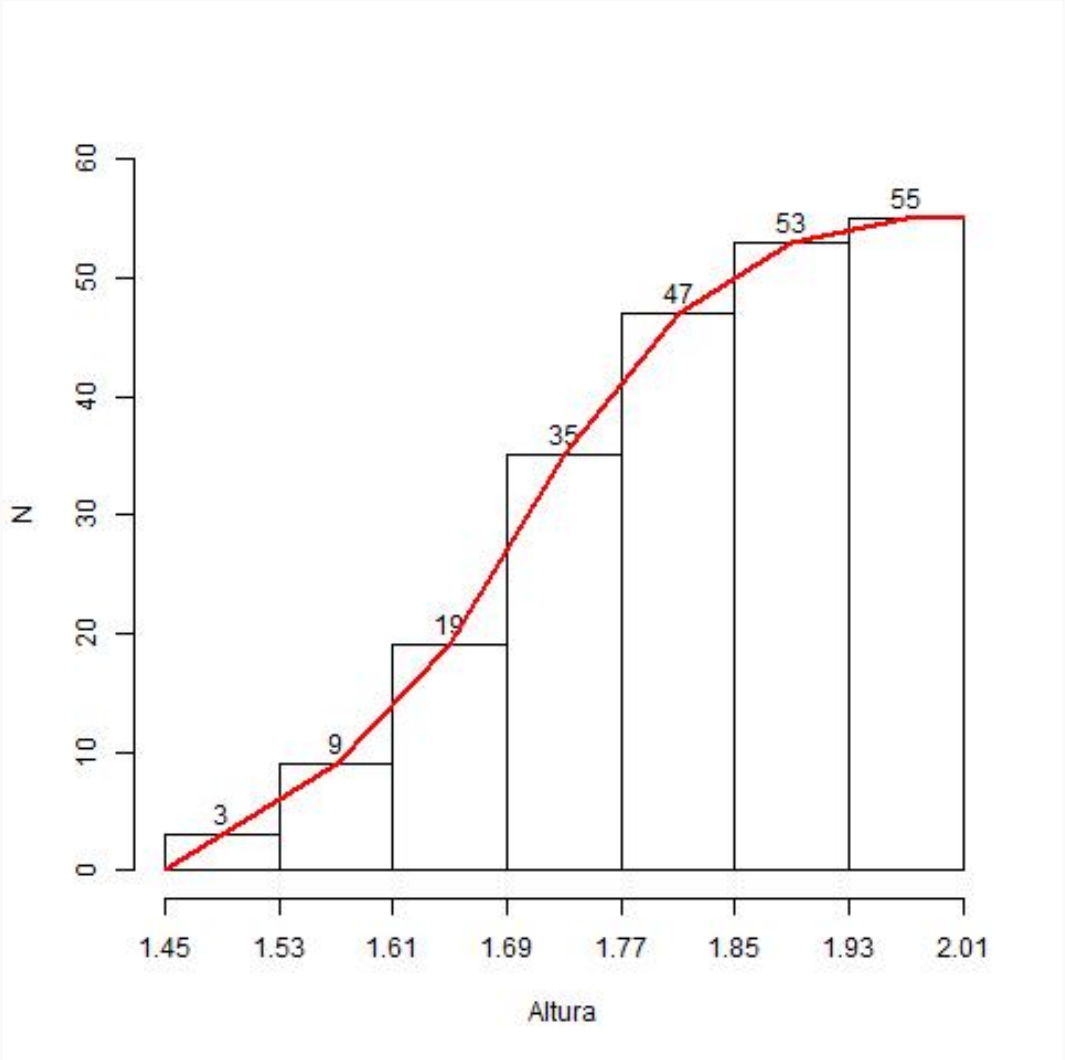

Figura 4: Ogiva de Galton das Alturas(m) de 55 homens.

Definicão (Gráfico de haste)

Representação gráfica para um conjunto de dados quantitativos discretos por meio linhas, não justapostas, cujas alturas são proporcionais às frequências do valores representam.

Gráficos

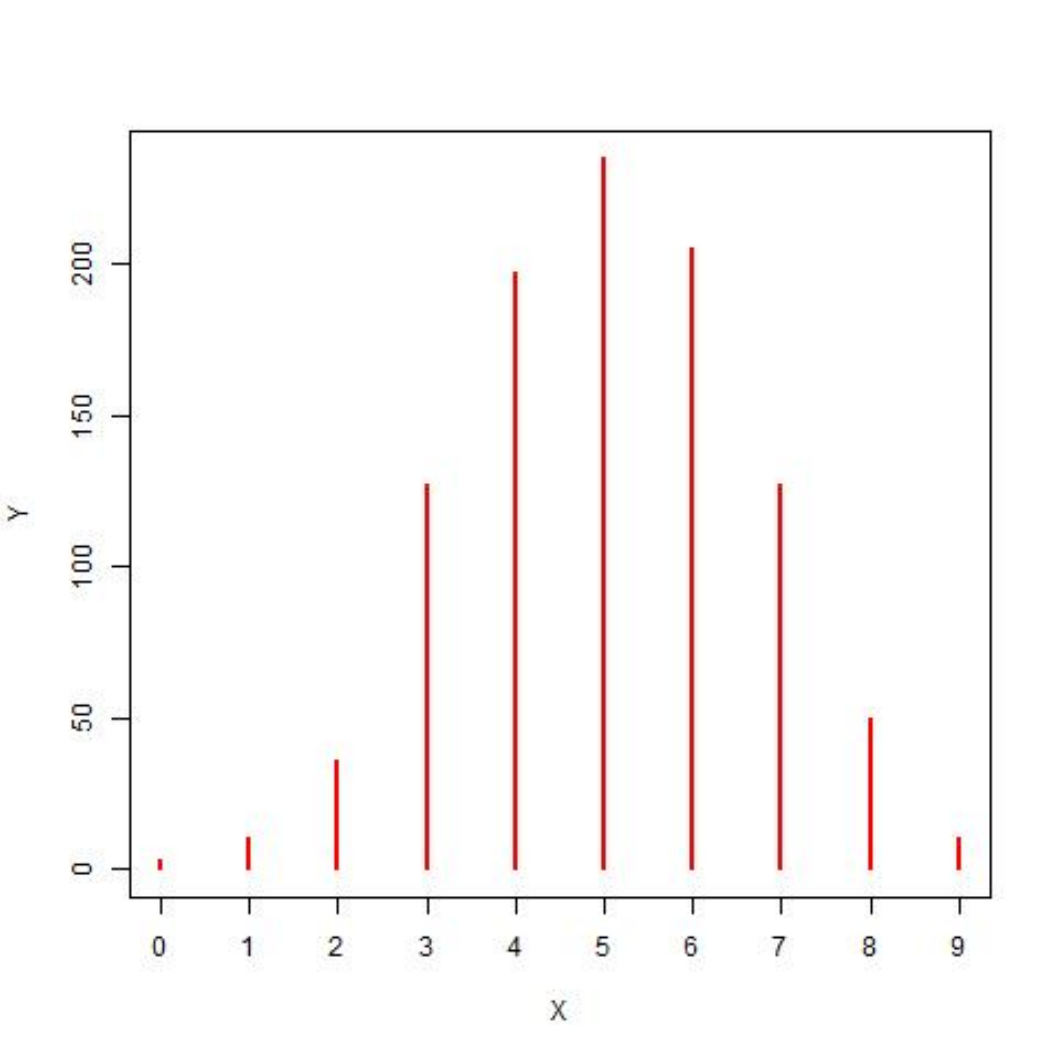

Figura 5: Gráfico de hastes para dados fictícios.

Definicão (Grafico de Barras)

O gráfico de barras é formado por barras retangulares com comprimento proporcional aos valores representados.

Gráficos

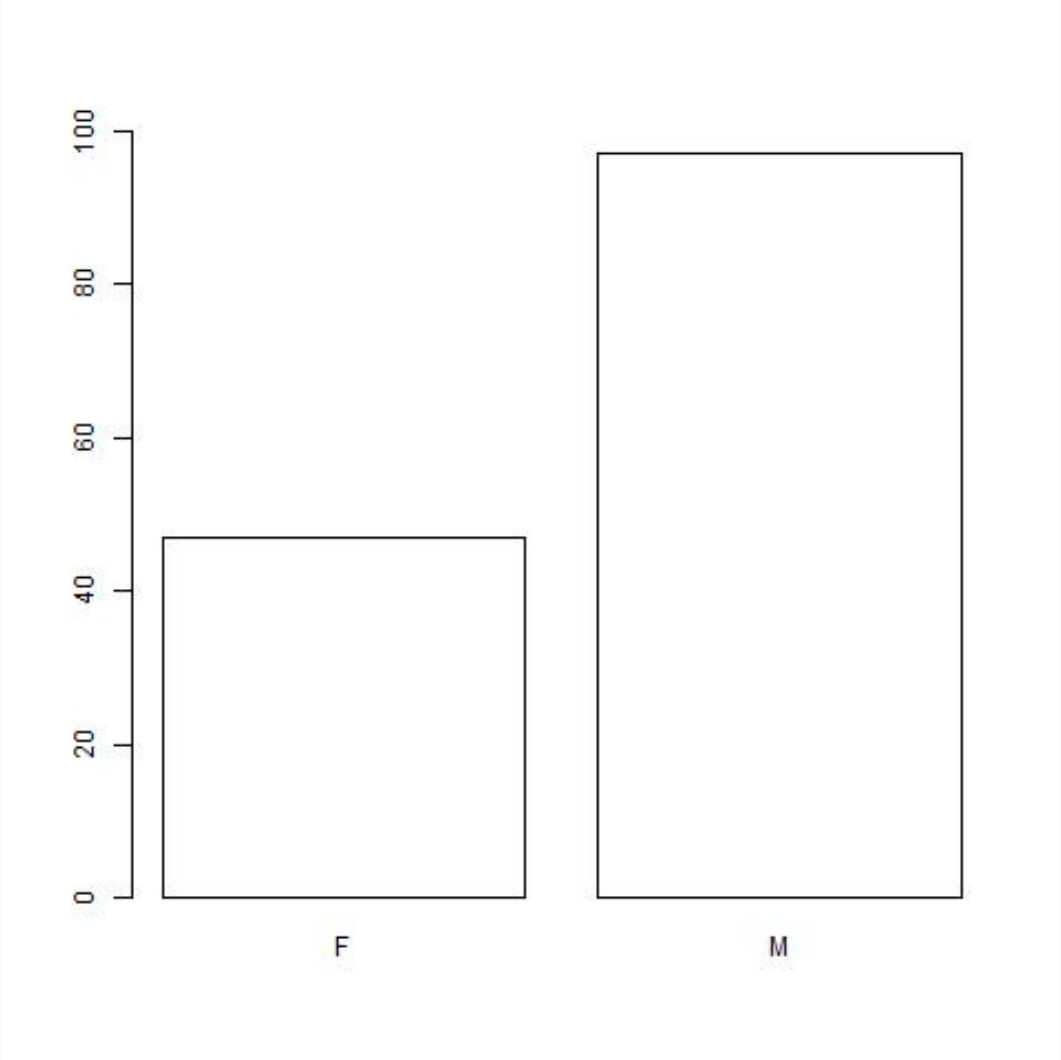

Figura 6: Gráfico de barras para dados fictícios.

Definicão (Gráfico de Dispersão)

É um diagrama matemático que representa duas varíveis em um Plano Cartesiano. Uma variável no eixo X e outra no eixo Y.

Gráficos

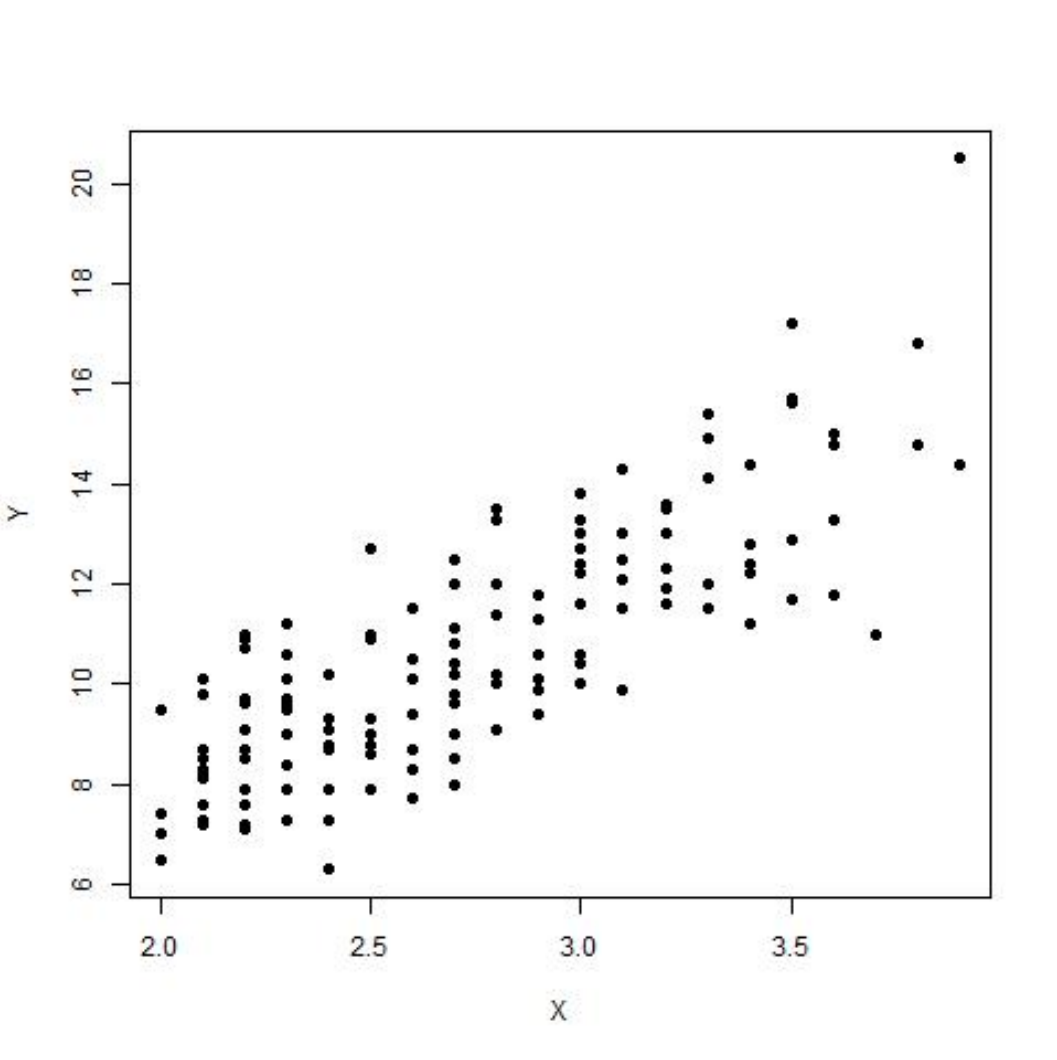

Figura 7: Gráfico de Dispersão para dados fictícios.

Definicão (Box-plot)

É uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes (outliers) dos dados analiasos, construída a partir das seguintes estátísticas: mínimo, máximo, primeiro quartil, mediana e o terceiro quartil.

Gráficos

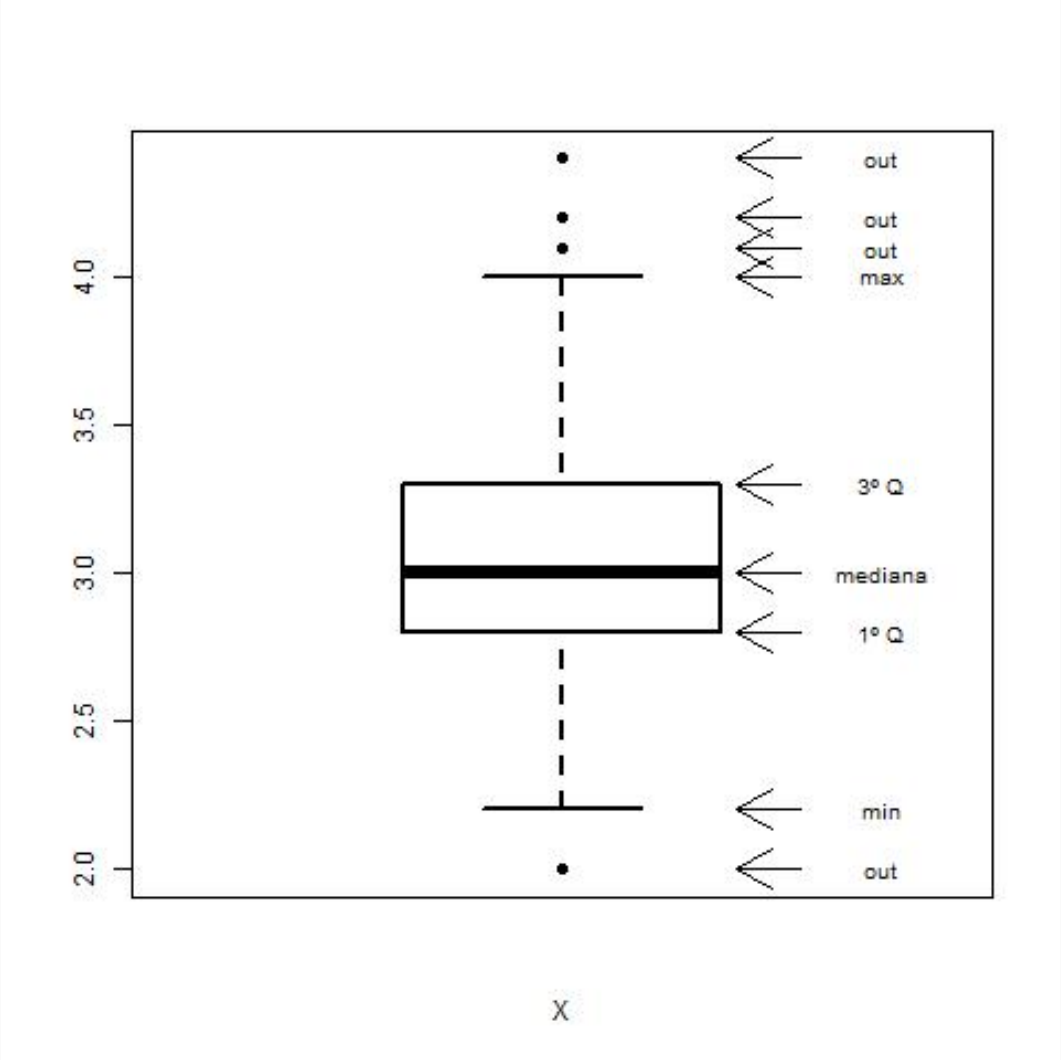

Figura 8: Esquema de um boxplot.

Gráficos

- Esse gráfico coloca 50% dos pontos dentro “caixa”.
- A base da caixa é o 1º quartil, a mediana é indicada por uma linha intermediária e o 3º quartil é a base superior.
- Tem linha chamadas bigodes que se estendem das bases da caixa +/- 1,5 vezes os a altura da caixa (3º quartil – 1º quartil).
- As extemidades dos bigodes representam os valores mínimo é máximo, excluíndo os outliers.
- Os pontos além dos bigodes são considerados outliers.