Estatística Descritiva e Inferencial

Rodrigo de Souza Bulhões
MAC5779 - Engenharia de Software Experimental

IME-USP, 12/09/2011

Resumo

1. Introdução à estatística
2. Análise exploratória de dados
   - Tipos de variáveis e níveis de mensuração
   - Intervalos de classes e histograma
   - Medidas de posição e dispersão
3. Medidas de associação
   - Entre variáveis categóricas

Definição

O que é estatística?

- Conjunto de técnicas que permite, de forma sistemática, recolher, organizar, descrever, analisar, modelar, prever e interpretar *dados* oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.
- Ela estuda fenômenos não deterministas, i.e., cujos resultados ou conseqüências está permanentemente associado a um forte grau de incerteza. Assim, a Teoria das Probabilidades está-lhe naturalmente subjacente.
- Também pode ser abordada como uma disciplina matemática, com o rigor que a qualquer ciência se impõe.

Definição (cont.)

Estatística descritiva

- São métodos que permitem descrever e resumir a informação resultante da observação de uma *amostra*, por intermédio de gráficos e valores característicos.
- De grande utilidade sobretudo ao lidar com grande massa de informação.
- É a primeira etapa a ser realizada na análise de dados.

Definição

População ou universo

Conjunto de informações todos os elementos em estudo que tenha, entre si, uma característica comum, interessando à Estatística a análise das propriedades das populações susceptíveis de representação numérica.

Definição (cont.)

Amostra

- É um subconjunto da população.
- Seu uso é viável se a população é muito grande ou se o processo de pesquisa é destrutivo (e.g., avaliar a percentagem de falhas de uma "população de fósforos").
- Não necessariamente é aleatória, embora deseje-se que ela seja representativa e imparcial — preserve características essenciais da população (e.g., provando o sabor de uma comida — mexe-se e retira-se uma colher).
- A Teoria da Amostragem é o campo da Estatística que estuda os diversos mecanismos de seleção de elementos para compor uma amostra.

Variáveis e escalas

Os tipos de variáveis são:

- Qualitativa nominal.
- Qualitativa ordinal.
- Quantitativa discreta.
- Quantitativa contínua.

Os níveis de mensuração são:

- Escala nominal.
- Escala ordinal.
- Escala intervalar.
- Escala de razão.

Nominal

- Uma variável é dita qualitativa ordinal quando seus possíveis valores representam atributos e/ou qualidades, para a qual não existe nenhuma ordenação nas possíveis realizações.
- Seu nível de mensuração é o mais baixo.
- Consiste num conjunto de categoria de respostas qualitativamente distintivas e mutuamente excludentes.

Nominal

Exemplo: Numa sala há 8 alunos, 5 dos quais do sexo masculino. Convencionando os números 1 e 2 respectivamente para homens e mulheres, tudo o que se pode fazer é escrever

João	$\longrightarrow$	1			
Pedro	$\longrightarrow$	1	Maria	$\longrightarrow$	2
Carlos	$\longrightarrow$	1	Adriana	$\longrightarrow$	2
Alberto	$\longrightarrow$	1	Patrícia	$\longrightarrow$	2
Otávio	$\longrightarrow$	1			

ou dizer que há $5 \times 1 = 5$ homens e $3 \times 2 = 3$ mulheres (?).

Nominal

- O exemplo ilustra que não faz sentido a operação de multiplicação, pois os números 1 e 2 representam atributos.
- Da mesma maneira que não vale a de adição (qual seria o significado?).

Nominal

Exemplo: Dados referentes ao provedor utilizado pelos alunos do exemplo anterior.

Aluno	Nome	Gênero	Provedor
1	João	1	A
2	Pedro	1	A
3	Carlos	1	B
4	Alberto	1	A
5	Otávio	1	B
6	Maria	2	C
7	Adriana	2	C
8	Patrícia	2	C

em que A, B e C são nomes dos provedores.

Nominal

Sejam

- n_j é o número de indivíduos que utilizam o provedor j ou freqüência absoluta da categoria j, $j \in \{A, B, C\}$.
- $n = n_A + n_B + n_C$ é o total de indivíduos na amostra, $n \in \mathbb{N}$.
- $f_j = \dfrac{n_j}{n}$ é a freqüência relativa da categoria j, $j \in \{A, B, C\}$.

Vejamos agora a distribuição da variável Provedor.

Provedor (j)	**Freq. absoluta** (n_j)	**Freq. relativa** (f_j)
A	3	0,375
B	2	0,25
C	3	0,375

Ordinal

- Na variável qualitativa ordinal existe uma ordem nos seus resultados.
- Elas podem ser avaliadas em termos de "mais que" ou "menos que", mas não permitem operações aritméticas.

Exemplo: Considere a distribuição da variável Classe Social.

Classe Social	**Freq. absoluta**	**Freq. relativa**	**Ângulo**
1-Baixa	1	0,125	45°
2-Média	5	0,625	225°
3-Alta	2	0,25	90°

onde o ângulo da classe social j é $\phi_j = \frac{n_j}{n} \times 360°$, $j \in \{1, 2, 3\}$.

Ordinal

Figure: Gráfico de setores (“pizza”) da variável Classe Social.

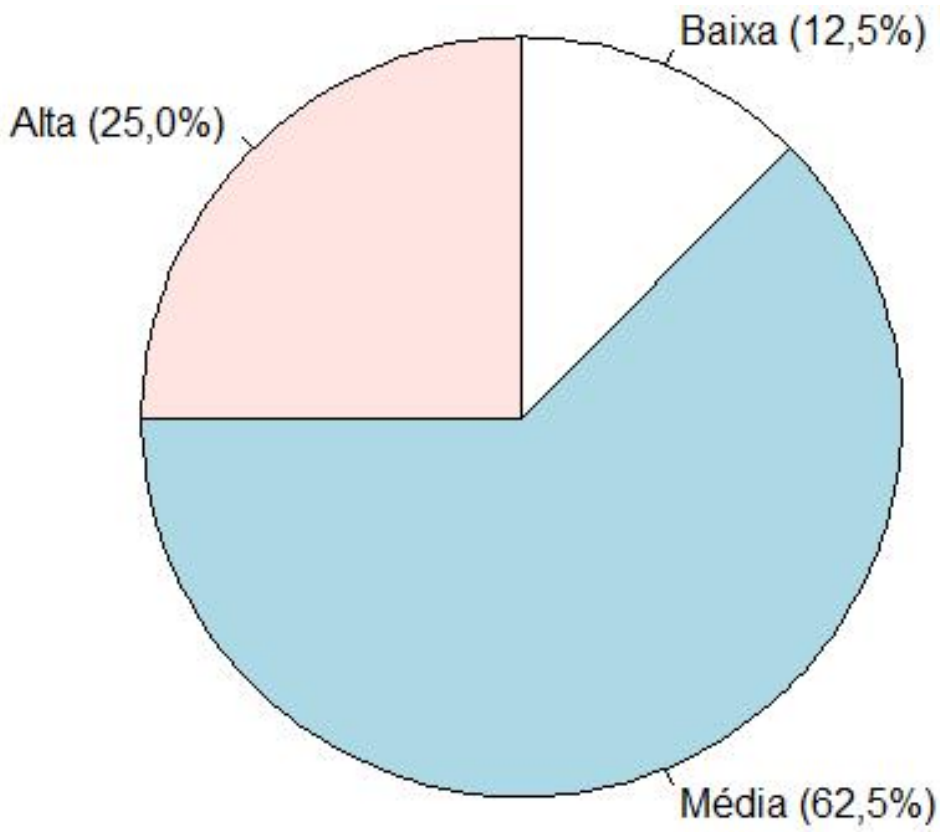
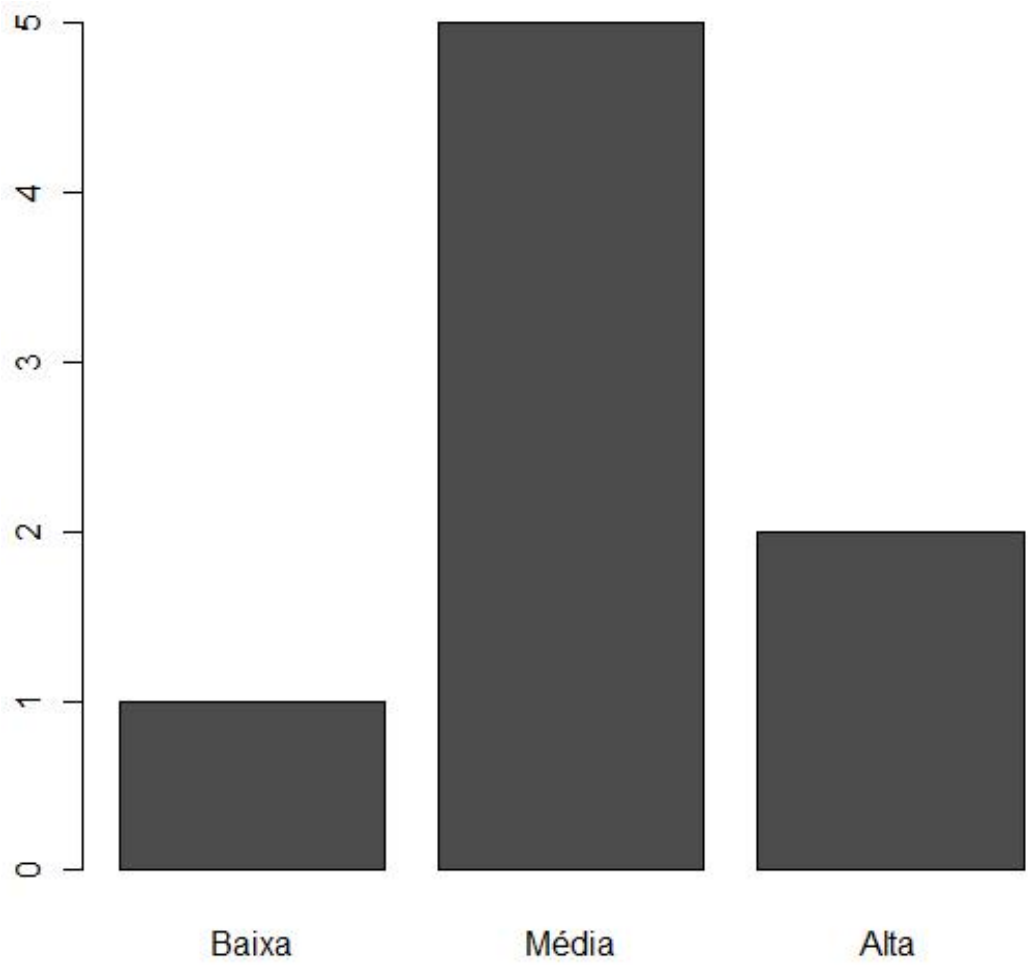


Figure: Gráfico de barras da variável Classe Social.



Variáveis quantitativas

Variáveis discretas

São variáveis cujos possíveis valores formam um conjunto finito ou enumerável de números, as quais resultam freqüentemente de uma contagem (e.g., número de acessos a um site).

Variáveis contínuas

São variáveis cujos possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração (e.g., tempo de resposta do sistema).

Intervalar

- Caracterizada por uma unidade de medida (arbitrária, porém fixa) e um zero relativo (convencional).
- Ela permite as operações de adição (nem sempre) e subtração, mas nunca multiplicação e divisão.
- Também podem ser avaliadas em termos de “mais que” ou “menos que”.

Intervalar

- **Escalas termométricas**: é claro que $40°C = 104°F$ é maior que $10°C = 50°F$, mas $\frac{40°C}{10°C} = 4$ é diferente de $\frac{104°F}{50°F}$ =2,08. Também 0°C não indica ausência de calor.
- **Calendário**: o ano zero existiu - ele não significa "ausência de ano".

- Com suas diferenças, todas as operações podem ser realizadas.

Corpos	$^\circ C$	**Diferenças** ($^\circ C$)	$^\circ F$	**Diferenças** ($^\circ F$)
A	10		50	
		10		18
B	20		68	
		20		36
C	40		104	
		60		108
D	100		212	

- $60^\circ C = 3 \times 20^\circ C = 108^\circ F = 3 \times 36^\circ F$.
- $\sqrt{\frac{60^\circ C}{10^\circ C}} = \sqrt{\frac{108^\circ F}{18^\circ F}}$.

De razão

- Sua diferença para a escala intervalar é que o zero é absoluto - ele significa ausência.
- Todas as operações da aritmética podem ser realizadas (e.g., média do número de filhos, tempo médio de espera do servidor, etc.)

Intervalos de classes e histograma

- É muito comum agrupar as variáveis quantitativas em intervalos de classes, para facilitar a interpretação.
- O número de classes não pode ser nem muito grande, nem muito pequeno.
- **Regra de Sturges**: $L = [1 + \log_2 n]$.
- Calcule $h = \frac{\max\{x_1,\ldots,x_n\} - \min\{x_1,\ldots,x_n\}}{L}$.
- Escreva os intervalos $[l_0, l_1[, [l_1, l_2[, \ldots, [l_{L-1}, l_L[$, em que $l_0 = \min\{x_1, \ldots, x_n\}$, $l_1 = l_0 + h, \ldots, l_L = l_{L-1} + h$.
- Defina f_k como sendo o número de elementos que estão contidos em $[l_{k-1}, l_k[$, com $k \in \{1, \ldots, L\}$.

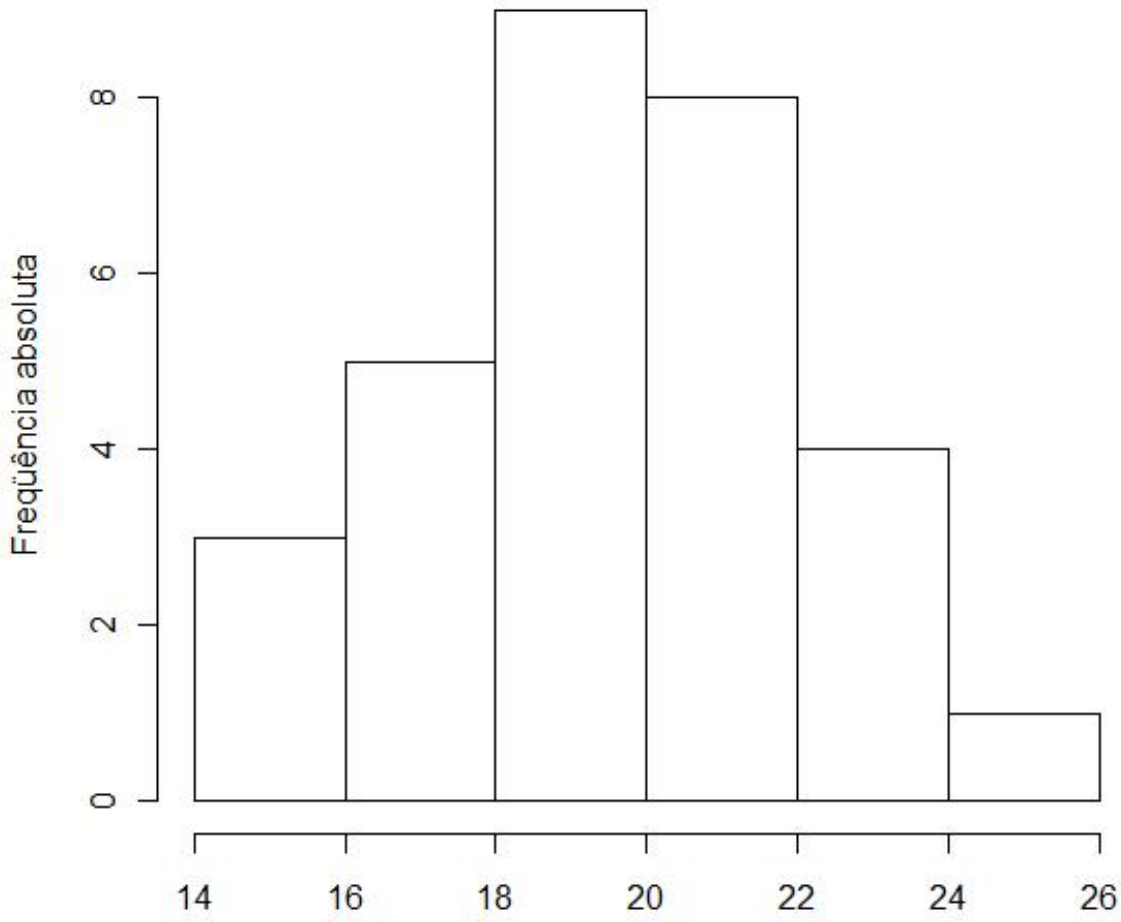
Dado um conjunto com $n = 30$ observações de uma variável contínua, tem-se $L = 6$, com

Classe	Ponto médio	Freqüência
[14, 16[	15	3
[16, 18[	17	5
[18, 20[	19	9
[20, 22[	21	8
[22, 24[	23	4
[24, 26[	25	1

Histograma

Conjunto de retângulos adjacentes com bases de comprimentos $l_k - l_{k-1}$ e alturas f_k, com $k \in \{1, \ldots, L\}$.

Figure: Histograma da tabela anterior.



Algumas medidas de posição:

- A média
- A moda
- Os quantis (extremos e quartis)

Algumas medidas de dispersão:

- A variância e o desvio padrão
- O coeficiente de variação de Pearson
- O coeficiente de simetria
- O coeficiente de curtose

Medidas de posição

Média aritmética

É definida como

$$\overline{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{x_1 + \cdots + x_n}{n}.$$

Moda

É todo valor da variável em que a freqüência atinge um máximo local. A distribuição empírica diz *unimodal*, *bimodal* ou *plurimodal* consoante apresente uma, duas ou mais modas, respectivamente. Chama-se *classe modal* àquela que possuir a maior freqüência.

Medidas de posição (cont.)

Quantis

Sem perda de generalidade, considere $x_{(1)}, \ldots, x_{(n)}$ uma amostra ordenada, i.e., $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Assim:

- $x_{(1)} = \min\{x_1, \ldots, x_n\}$.
- $Q_1 = x_{([\frac{n}{4}])}$
- $Md = Q_2 = \begin{cases} x_{(\frac{n+1}{2})} & , \quad n \text{ ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & , \quad n \text{ par} \end{cases}$
- $Q_3 = x_{([\frac{3n}{4}])}$
- $x_{(n)} = \max\{x_1, \ldots, x_n\}$.

Medidas de variabilidade

Medidas de dispersão

- Variância: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$.
- Desvio padrão: $s = \sqrt{s^2}$.
- Coeficiente de variação: $CV = \frac{s^2}{\overline{x}}$.
- Coef. de assimetria: $b_1 = \frac{m_3^2}{(s^2)^3}$, onde $m_3 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^3$.
- Coef. de achatamento: $g_2 = \frac{m_4}{(s^2)^2}$, onde $m_4 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^4$.

Figure: Variável simétrica em forma de sino

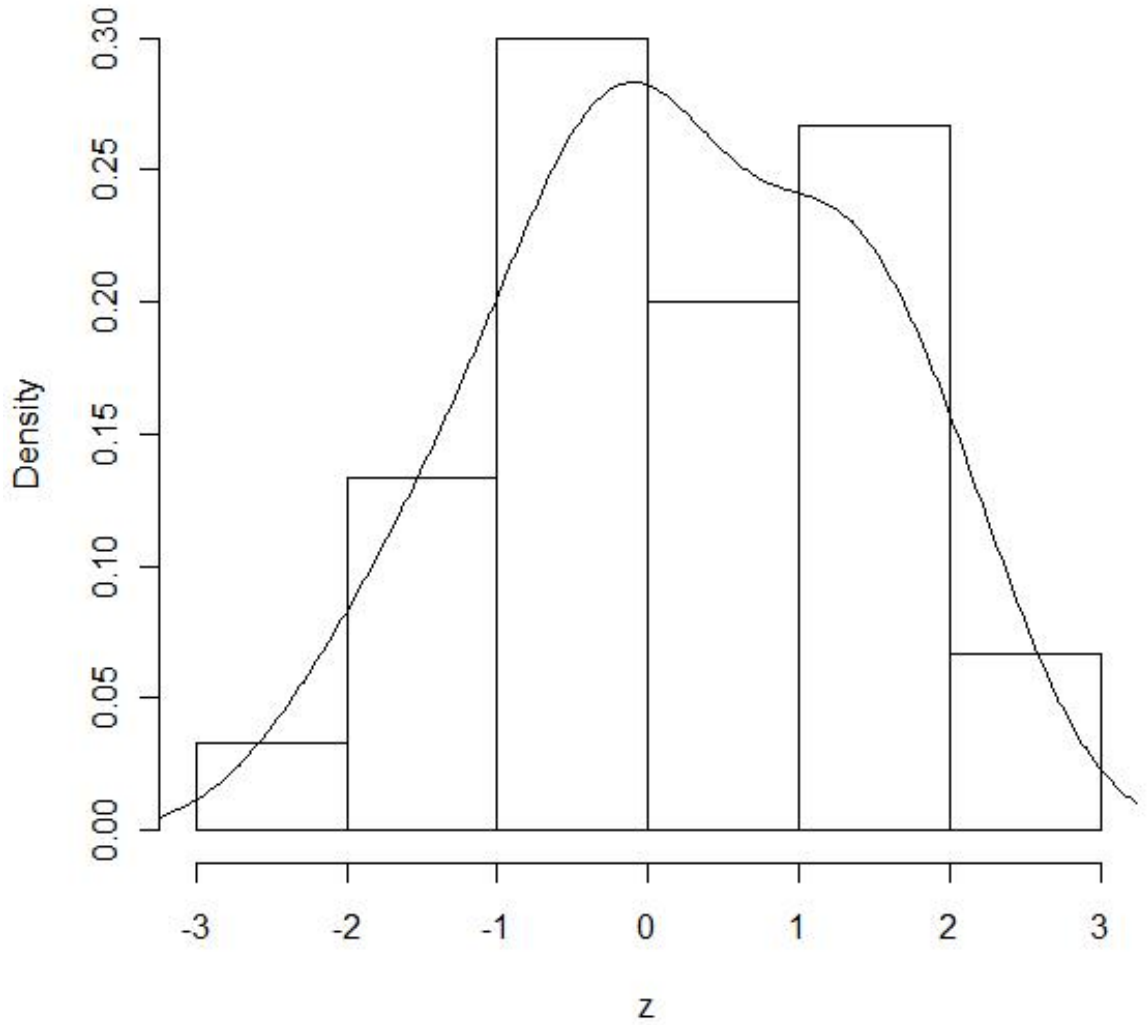


Figure: Variável assimétrica

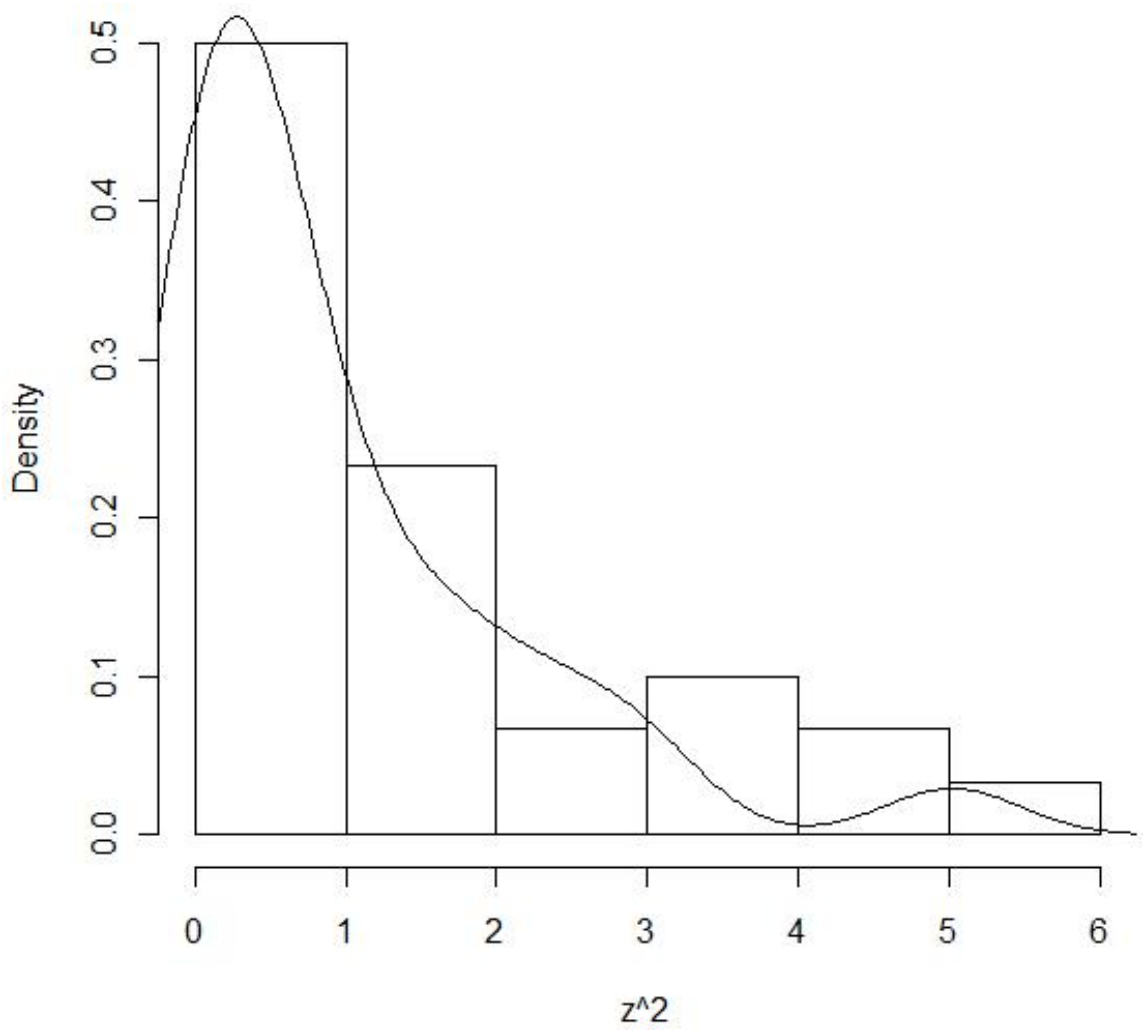
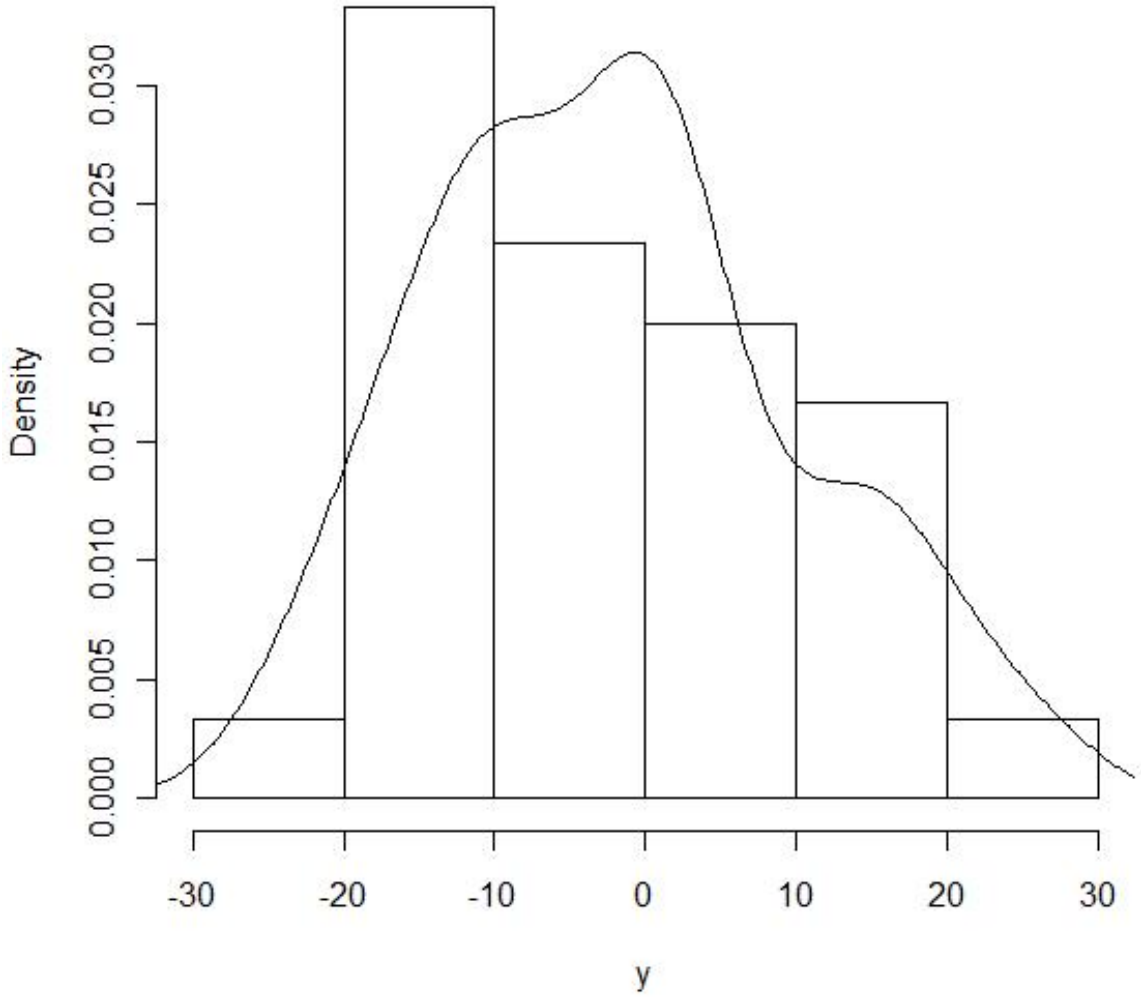


Figure: Variável simétrica e achatada



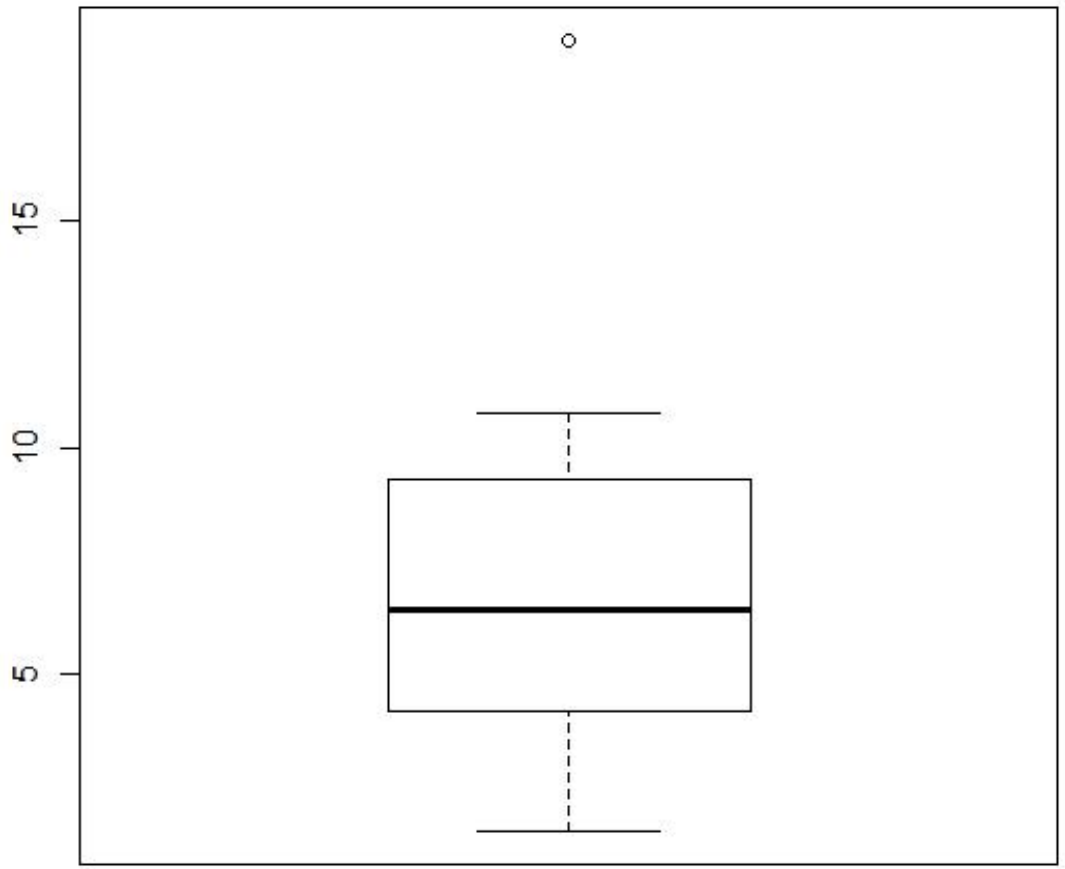
Outliers

Como detectar a presença de valores atípicos?

- Analisando o *boxplot*.
- Ele utiliza os quantis apresentados em sua estrutura.
- Assim como o histograma, dá indícios sobre variabilidade, assimetria e curtose da distribuição amostral.

Outliers

Figure: Boxplot de uma variável “contaminada”.



Associação entre variáveis categóricas

Tabela de contingência

Considere o cruzamento de duas variáveis qualitativas com I e c atributos, conforme a tabela $I \times c$ a seguir.

	Atributo B				
Atributo A	B_1	B_2	$\cdots$	B_c	Totais
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ic}	$n_{I.}$
Totais	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

Medidas de associação

Exemplos de medidas

- Qui-Quadrado de Pearson: $X^2 = \sum_{j=1}^{l} \sum_{k=1}^{c} \frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}}$, onde $n^*_{jk} = \frac{n_{j.} n_{.k}}{n}$.
- Quadrado da contingência: $\Phi^2 = \frac{X^2}{n}$.
- V de Cramer: $V = \sqrt{\frac{\phi^2}{\min\{l-1, c-1\}}}$.

Exemplo

	Tipo de programas		
Gênero	Informação	Esporte	Novela
Masculino	120	300	30
Feminino	150	20	380

- Obtivemos X^2 =542,539, Φ^2 =0,543 e V =0,737.