Notas de aulas de Estatística Econômica

Marcos Minoru Hasegawa

2020-09-24

Sumário

Licença

Sobre o material

A situação especial causada pela pandemia da COVID-19 forçou a muitos professores criarem materiais para facilitar aulas remotas das suas disciplinas. A disciplina SE305 Estatística Econômica e Introdução à Econometria da UFPR não poderia ser diferente. Então, o objetivo deste material é de suprir a falta das bibliografias básicas na sua versão digital com a disponibilização de forma digital e gratuita o que seria o material das notas das aulas da disciplina de Estatística Econômica. Não é o ideal, mas a ideia é melhorar o material com tempo.

Sobre o Autor

Professor do Departamento de Economia da Universidade Federal do Paraná. Engenheiro Agrônomo pela UNESP/Jaboticabal, Mestrado em Economia Agrária pela ESALQ/USP e Doutorado em Economia Aplicada pela ESALQ/USP, é um dos professores responsáveis pelas disciplinas de SE305 Estatística Econômica e Introdução à Econometria e SE308 Econometria ambas do curso de Economia da Universidade Federal do Paraná (UFPR).

Capítulo 1

Estatística Descritiva

1.1 Medidas de posição

Este tópico está baseado no material de Sartoris (2013).

Trata-se de medidas de tendência central ou resumo. Como os nomes dizem, tratam-se de medidas que tratam de resumir a massa de valores e um único número.

1.1.1 Variável Aleatória

- variável aleatória (v.a.) é uma variável que está associada a uma *distribuição de probabilidade*.
- Ou seja, cada valor da v.a. está associada a uma probabilidade.
- O resultado do lançamento de uma dado, que poder ser qualquer número de 1 a 6, está associada a uma probabilidade de 1/6.

1.1.2 Média Aritmética Simples

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.1}$$

onde $i = 1, ..., n$

Exemplo 1

Qual é a média aritmética de um grupo de cinco pessoas cujas idades são em ordem crescente, 21,23,25,28 e 31. Para responder, basta aplicar (1.1).

$$\overline{X} = \frac{21 + 23 + 25 + 28 + 31}{5} = 25,6$$

Exemplo 1 no R

```
X <- c(21, 23, 25, 28, 31)
X
```

```
## [1] 21 23 25 28 31
```

```
mediaX <- mean(X)
mediaX
```

```
## [1] 25,6
```

Exemplo 2

Qual é a média aritmética de três provas realizadas por um aluno, cujas notas foram 4,6 e 8. Para responder, basta aplicar (1.1).

$$\overline{X} = \frac{4 + 6 + 8}{3} = 6$$

Exemplo 2 no R

```
X2 <- c(4, 6, 8)
X2
```

```
## [1] 4 6 8
```

```
mediaX2 <- mean(X2)
mediaX2
```

```
## [1] 6
```

1.1.3 Média Aritmética Ponderada

Na média aritmética ponderada, cada valor pode ter importância diferentes do outros valores considerados no computo. A frequência dos valores é muito comumente usada para para dar maior ou menor importância relativa entre os valores considerados no computo da média aritmética ponderada. Veja como fica a fórmula para o cálculo da média aritmética ponderada em (1.2)

$$\overline{X} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i X_i \tag{1.2}$$

onde w_i é a ponderação ou peso associado a iésimo valor de X.

Podemos escrever na forma de frequência relativa dos valores da variável X:

$$f_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{1.3}$$

Exemplo 3

Qual é a média aritmética de um grupo de vinte alunos, oito com 22 anos, sete de 23 anos, três de 25 anos, um de 28 anos e um de 30 anos. Para responder, basta aplicar (1.2).

$$\overline{X} = \frac{22 \times 8 + 23 \times 7 + 25 \times 3 + 28 \times 1 + 30 \times 1}{20} = 23,5$$

Exemplo 3 no R

```
X3 <- c(22, 23, 25, 28, 30)
X3
```

```
## [1] 22 23 25 28 30
```

```
w3 <- c(8, 7, 3, 1, 1)
w3
```

```
## [1] 8 7 3 1 1
```

```
wX3 <- w3 * X3
mediaX3 <- sum(wX3)/sum(w3)
mediaX3
```

```
## [1] 23,5
```

Exemplo 4

Qual é a média ponderada de três provas realizadas por um aluno, cujas notas foram 4, 6 e 8. A primeira prova tem peso igual a 1, a segunda tem peso igual a 2 e a terceira tem peso igual a 3. Para responder, basta aplicar (1.2).

$$\overline{X} = \frac{4 \times 1 + 6 \times 2 + 8 \times 3}{1 + 2 + 3} \cong 6,7$$

Exemplo 4 no R

```
X4 <- c(4, 6, 8)
X4
```

```
## [1] 4 6 8
```

```
w4 <- c(1, 2, 3)
w4
```

```
## [1] 1 2 3
```

```
wX4 <- w4 * X4
mediaX4 <- sum(wX4)/sum(w4)
round(mediaX4, digits = 1)
```

```
## [1] 6,7
```

1.1.4 Média Geométrica Simples

Na média geométrica simples, a forma de obter uma medida resumo ou de tendência central é multiplicar todos os n valores e tirar a raiz enésima do resultado do produtório. Assim é possível ter duas fórmulas para a média geométrica a (1.4) e (1.5).

$$G = \left(\prod_{i=1}^{n} X_i \right)^{\frac{1}{n}} \tag{1.4}$$

ou

$$G = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n} \tag{1.5}$$

O que acontece se um dos valores de X for igual a zero? E se um dos valores for negativo?

Exemplo 5

Sejam três valores 4, 6 e 8. Calcule a média geométrica simples.

$$\sqrt[3]{4 \times 6 \times 8} \cong 5,7690$$

Exemplo 5 no R

```
X5 <- c(4, 6, 8)
X5
```

```
## [1] 4 6 8
```

```
n <- length(X5)
mediaX5 <- prod(X5)^(1/n)
round(mediaX5, digits = 1)
```

```
## [1] 5,8
```

1.1.5 Média Geométrica Ponderada

Na média geométrica ponderada que podem ser calculadas através de duas fórmulas (1.6) e (1.7), cada valor pode ter uma importância diferente em relação aos outros valores no computo da média geométrica. Muito comumente, esta maior ou menor importância pode estar associada a frequência dos valores considerados no cálculo.

$$G = \left(\prod_{j=1}^{k} X_j^{w_j} \right)^{\frac{1}{n}} \tag{1.6}$$

ou

$$G = \sqrt[n]{X_1^{w_1} \times X_2^{w_2} \times \ldots \times X_k^{w_k}} \tag{1.7}$$

onde a $\sum_{j=1}^{k} w_j = n$

Exemplo 6

tomando os valores do exemplo 5 e ponderando por 1,2 e 3, temos:

$$\sqrt[6]{4^1 \times 6^2 \times 8^3} \cong 6,5$$

O exemplo 6 no R

```
x6 <- c(4, 6, 8)
class(x6)
```

```
## [1] "numeric"
```

```
x6
```

```
## [1] 4 6 8
```

```
w6 <- c(1, 2, 3)
w6
```

```
## [1] 1 2 3
```

```
G2 <- round((prod(x6^w6))^(1/sum(w6)), 1)
G2
```

```
## [1] 6,5
```

1.1.6 Média Harmônica

É o inverso da média dos inversos dos valores da variável que pode ser calculada através das fórmulas (1.8) e (1.9).

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \tag{1.8}$$

$$H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_n}} \tag{1.9}$$

O que acontece se um dos valores de X for igual a zero? Para entender essa situação, use o conceito de limite fazendo o valor tender a zero.

Exemplo 7

Tomando o exemplo das notas, temos:

$$H = \frac{3}{\frac{1}{4} + \frac{1}{6} + \frac{1}{8}} \cong 5,5.$$

1.1.7 Média Harmônica Ponderada

Na média harmônica ponderada, assim como na média aritmética ponderada e na média geométrica ponderada, cada valor pode ter uma importância em relação aos outros valores considerados no seu cálculo. Comumente, a frequência do valor pode associaar uma maior ou menor importância no cálculo da média harmônica ponderada que pode ser calculada através das fórmulas (1.10) e (1.11)

$$H = \frac{n}{\sum_{j=1}^{k} w_j \frac{1}{X_j}} \tag{1.10}$$

ou

$$H = \frac{n}{w_1 \frac{1}{X_1} + w_2 \frac{1}{X_2} + \ldots + w_k \frac{1}{X_k}} \tag{1.11}$$

onde a $\sum_{j=1}^{k} w_j = n$

Exemplo 8

Tomando o exemplo das notas

$$H = \frac{6}{\frac{1}{4} \times 1 + \frac{1}{6} \times 2 + \frac{1}{8} \times 3} \cong 6,3.$$

Observação

Tanto para as médias simples como para as ponderadas, a média aritmética é maior do que a média geométrica e essa, por sua vez, é maior que a harmônica. Isso só não vale quando todos os valores são iguais. Veja de forma esquemática em (1.12)

$$\overline{X} \geq G \geq H \tag{1.12}$$

Exemplo 9

O aluno tira as seguintes notas bimestrais: 3,4,5,7 e 8,5. Determine qual seria sua média final se esta fosse calculada dos três modos, aritmética, geométirca e harmônica, em cada um dos seguintes casos: i) as notas têm o mesmo peso e; ii) as notas têm pesos diferentes.

i) As notas dos bimestres têm os mesmos pesos.

$$\overline{X} = \frac{3 + 4,5 + 7 + 8,5}{4} = 23/4 = 5,75$$

$$G = \sqrt[4]{3 \times 4,5 \times 7 \times 8,5} = \sqrt[4]{803,25} \cong 5,32$$

$$H = \frac{4}{\frac{1}{3} + \frac{1}{4,5} + \frac{1}{7} + \frac{1}{8,5}} \cong 4,90$$

ii) Suponha que agora os pesos para as notas bimestrais sejam, 30%, 25%, 25% e 20%.

$$\overline{X} = 0,3 \times 3 + 0,25 \times 4,5 + 0,25 \times 7 + 0,20 \times 8,5 = 5,475$$

$$G = 3^{0,3} \times 4,5^{0,25} \times 7^{0,25} \times 8,5^{0,2} = \cong 5,05$$

$$H = \frac{1}{0,3\frac{1}{3} + 0,25\frac{1}{4,5} + 0,25\frac{1}{7} + 0,2\frac{1}{8,5}} \cong 4,66$$

1.1.8 Mediana

é o valor que divide um conjunto e dados ordenados ao meio, ou seja, dois grupos de valores de igual tamanho. Com base na definição de mediana, o valor da mediana pode ser obtida através da sua posição que proporciona duas situações: i) o número de valores é impar e ii) o número de valores é par.

i) Quando o número de valores é impar, a posiçãodo valor correspondente a mediana é obtida através de (1.13):

$$PMediana_{impar} = \frac{n+1}{2} \tag{1.13}$$

onde n é o número de valores considerado no cálculo.

ii) Quando o número de valores é par, a posição da mediana é obtida através da média entre os dois valores centrais do conjunto de valores ordenados de menor a maior. O primeiro valor central é definido pela posição obtida através de (1.14)

$$P1Mediana_{par} = \frac{n}{2} \tag{1.14}$$

onde n é o número de valores considerado para o cálculo.

O segundo valor central é definido pelas posição obtida através de (1.15)

$$P2Mediana_{par} = \frac{n}{2} + 1 \tag{1.15}$$

onde n é o número de valores considerado para o cálculo.

Assim, a mediana quando o número de valores é par é obtida através da média aritmética simples dos valores correspondentes as posições obtidas por (1.14) e por (1.15) através de (1.16)

$$Mediana_{par} = \frac{ValorCentral_1 + ValorCentral_2}{2} \tag{1.16}$$

Exemplo numérico de Mediana quando o número de valores é impar

Seja um conjunto de valores 2,-3,1,-2,0,-1,3. Obtenha a mediana.

Primeiramente ordena-se do menor para o maior.

-3,-2,-1,0,1,2,3

Como se trata de número impar de valores o valor central que divide o conjunto de valores em dois subconjuntos de igual tamanho é o valor da mediana. Neste caso é o zero.

Mediana no R

```
w <- c(-3, -2, -1, 0, 1, 2, 3)
mediana1 <- median(w)
print(mediana1)
```

```
## [1] 0
```

Exemplo numérico de Mediana quando o número de valores é par

No exemplo anterior o conjunto de dados era composto por um número ímpar de valores. Neste exemplo o número de valores ordenado de menor a maior é par. Nesse caso, apesar de existir vários critérios, o mais usual é tirar a média aritmética simples entre os dois valores centrais do conjunto de valores ordenados de menor a maior. Uma vez que não existe um valor que separe dois subconjuntos de igual tamanho, a média aritmética simples destes dois valores é o valor da mediana quando o número total de valores não é impar.

Sejam os valores -2,1,3,2,-3,1. Obtenha a mediana.

Primeiramente ordena-se os seis valores.

-3,-2,-1,1,2,3

Note que trata-se de conjunto com um número par de valores.

Dessa forma, toma-se os dois valores centrais que são -1 e 1 e calcula-se a média aritmética simples. Ou seja, a mediana para este conjunto com seis valores é igual a zero.

O exemplo do número par de valores no R

```
v <- c(-3, -2, -1, 1, 2, 3)
mediana2 <- median(v)
print(mediana2)
```

```
## [1] 0
```

1.1.9 Quartis ou Quartiles

são os valores que dividem o conjunto de dados ordenados em quatro subjconjuntos de igual tamanho. Ou seja são valores do conjunto que definem o primeiro quarto dos dados (25%), a metade dos dados (50%) que coincide com a mediana, os três quartos dos dados (75%).

Dessa forma para obter os valores que dividem o conjunto de dados ordenados de menor a maior e quatro subconjuntos de igual tamanho, é necessário definir qual é a posição desses valores. Uma vez definido as suas posições pode-se obter os valores corretamente.

A posição do valor que separa o primeiro do segundo quartil é definido por (1.17).

$$PQ_1 = \frac{(n+1)}{4} \tag{1.17}$$

onde n é o número de valores. A posição do valor que separa o segundo do terceiro quartil é definido por (1.18).

$$PQ_3 = \frac{3(n+1)}{4} \tag{1.18}$$

onde n é o número de valores.

Note que o termo genérico é percentil. Por exemplo, o quintis são os valores que dividem o conjunto de ados ordenados de menor a maior em cinco subconjuntos de igual tamanho.

Quartis no R

No R tem uma função específica para a obtenção dos quartis.

```
p <- c(0:100)
length(p)
```

```
## [1] 101
```

```
quantile(p)
```

```
##   0%  25%  50%  75% 100%
##    0   25   50   75  100
```

```
faixainterquant <- quantile(p, 0.75) - quantile(p,
    0.25)
faixainterquant
```

```
## 75%
##  50
```

Quartis no R

No R tem uma função específica para a obtenção dos quartis.

```
p2 <- c(1:100)
length(p2)
```

```
## [1] 100
```

```
quantile(p2)
```

```
##     0%    25%    50%    75%   100%
##   1,00  25,75  50,50  75,25 100,00
```

```
faixainterquant2 <- quantile(p2, 0.75) - quantile(p2,
    0.25)
faixainterquant2
```

```
##  75%
## 49,5
```

1.1.10 Moda

Moda é o elemento de maior frequência, ou seja, que aparece o maior número de vezes. Pode haver mais de uma moda em um conjunto de valores:

- Unimodal
- Bimodal
- Multimodal
- Amodal

Moda no R Não existe uma função da moda para pronto uso no R. É necessário criar uma função segue abaixo.

```
# criando a função moda no R

getmode <- function(v) {
    uniqv <- unique(v)
    uniqv[which.max(tabulate(match(v, uniqv)))]
}
z <- c(2, 1, 2, 3, 1, 2, 3, 4, 1, 5, 5, 3, 2, 3)
moda1 <- getmode(z)
print(moda1)
```

```
## [1] 2
```

1.2 Medidas de dispersão

Este tópico está baseado nos materiais de Sartoris (2013) e Morettin and Bussab (2013).

Medem como os dados estão *agrupados*, mais ou menos próximos entre si, seja, mais ou menos dispersos.

1.2.1 Variância

A variância é a somatória dos quadrados dos desvios em relação a média, dividido pelo número de observações. Note que a ideia inicial de dispersão foi a distância de cada valor do conjunto dados da variável em relação à media da variável. Mas como trata-se da distância relativa de cada valor em relação a média dos valores da variável, a sua somatória sempre resulta zero. Pois os desvios em relação a médias são compostos de valores positivos e negativos por estarem acima ou abaixo da média e assim a somatória das mesmas resulta zero, sempre. Portanto, a soma dos desvios não tem utilidade como medida de dispersão. Mas se a soma for dos quadrados dos desvios, isso é resolvido. Por isso, a variância é o valor médio dos quadrados dos desvios em relação à media. Ou seja,

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

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

Note que a variância da amostra é um estimador não viesado da variância populacional. A diferença entre variância populacional e variância amostral será apresentada mais adiante. A interpretação intuitiva da diferença entre ambas as variâncias é de que quando se trabalha com amostra, está tendo acesso a parte das informações e isso precisa ser penalizado. Note que esta penalização se dá para amostras pequenas pois se trata de subtrair uma unidade do número de observações. O que acontece com a diferença entre variância populacional e a variância amostral quando i) o número de observações torna-se muito grande, tipo bem maior que 30 e; ii) o número de observações tende ao infinito.

Variância no R

Aproveitando os dados de alturas de 30 pessoas:

```
tail(X)
```

```
## [1] 21 23 25 28 31
```

```
varX <- round(var(X), 4)
varX
```

```
## [1] 15,8
```

Note que a variância no R é a variância amostral, cujo denominador é $(n-1)$.

Em termos práticos, a variância tem uma desvantagem: a unidade do seu resultado é o quadrado da unidade original da variável. Portanto, se a variável

em questão é preço de uma mercadoria em Reais, a sua variância será Reais ao quadrado. Tal fato dificulta a sua interpretação. Por isso é apresentado o desvio padrão como medida de dispersão na sequência.

1.2.2 Desvio Padrão

É a raiz quadrada da variância. No desvio padrão, denotado como $d.p.(X)$ ou σ, não tem o efeito do quadrado.

$$d.p.(X) \cong \sigma = \sqrt{var(X)}$$

Portanto, a sua interpretação é clara e direta por ter a mesma unidade da sua variável original. Desta forma, o desvio padrão facilita a sua análise juntamente com as medidas de posição como a média aritmética simples, por exemplo.

Desvio Padrão no R

```
tail(X)
```

```
## [1] 21 23 25 28 31
```

```
dpX <- round(sd(X), 4)
dpX
```

```
## [1] 3,9749
```

Note que, da mesma forma que a variância no R, o desvio padrão calculado no R tem como base a variância cujo numerador é $(n-1)$.

Fórmula alternativa da Variância

Desenvolvendo a fórmula da definição da variância tem-se:

$$\begin{aligned} var(X) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n}\sum_{i=1}^{n}2X_i\overline{X} + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}n\overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{XX} + \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2. \end{aligned}$$

Em outras palavras

$$var(X) = \text{média dos quadrados} - \text{quadrado da média}.$$

Exemplo de variância e desvio padrão no R

Tomando o exemplo numérico da tabela 2.7 (Sartoris, 2013, p.40) sobre notas do aluno A tem-se:

Aluno A	notas	$notas^2$
Economia	3	9
Contabilidade	2	4
Administração	4	16
Matemática	1	1
Somatória	**10**	**30**
Média	**2,5**	**7,5**

$$var(X) = 7,5 - (2,5)^2 = 1,25$$

$$dp(X) = \sqrt{1,25} = 1,12$$

```
X3 <- c(3, 2, 4, 1)
mediaX3e2 <- sum(X3^2)/length(X3)
mediaX3 <- sum(X3)/length(X3)
varX3 <- mediaX3e2 - mediaX3^2
varX3
```

```
## [1] 1,25
```

```
dpX3 <- round(sqrt(varX3), 4)
dpX3
```

```
## [1] 1,118
```

1.2.3 Histograma

O histograma é uma ferramenta da estatística descritiva para mostrar visualmente, de forma bastante simples, como os valores da variável estão distribuídos. Mas também permite ter uma ideia visual da dispersão do conjunto de valores. Portanto, não se trata de uma medida de dispersão. Mas deveria ser a primeira coisa a se obter das variáveis de interesse em um trabalho de pesquisa.

Considere a altura de 30 pessoas medidas em centímetros.

Tabela 2.1 - Altura de 30 pessoas em cm.

159	168	172	175	181
161	168	173	176	183
162	169	173	177	185
164	170	174	178	190
166	171	174	179	194
167	171	174	180	201

Usando o R para construir o histograma do exemplo numérico

Os dados são inputados na variável X.

```
X <- c(159, 161, 162, 164, 166, 167, 168, 168, 169,
    170, 171, 171, 172, 173, 173, 174, 174, 174, 175,
    176, 177, 178, 179, 180, 181, 183, 185, 190, 194,
    201)
head(X)
```

```
## [1] 159 161 162 164 166 167
```

```
obsX <- length(X)
obsX
```

```
## [1] 30
```

```
faixaX <- range(X)
faixaX
```

```
## [1] 159 201
```

Usando a função hist do R para elaborar o histograma de altura

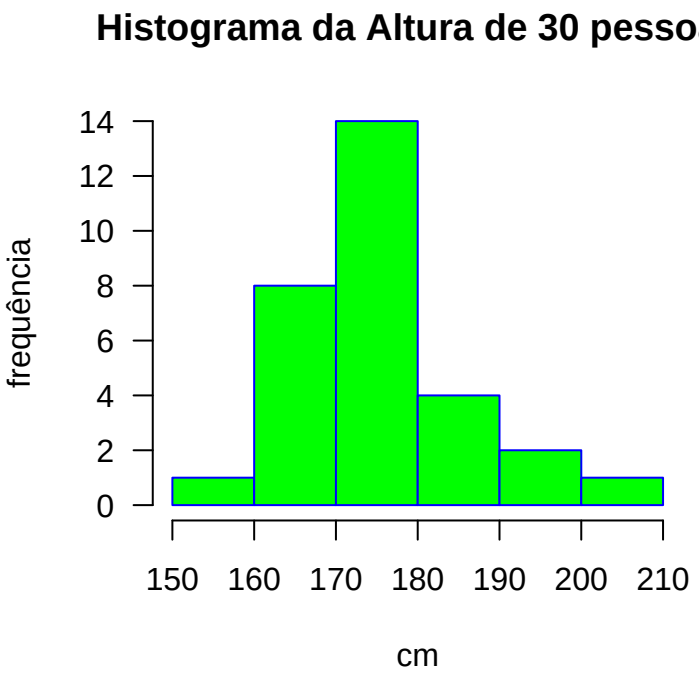
```
grafico1 <- hist(
                  X,
                  main="Histograma da Altura",
                  xlab="cm",
                  ylab="frequência",
                  border="blue",
                  col="green",
                  xlim=c(150,210),
                  las=1,
                  breaks=5,
                  right=FALSE
                )

grafico1
```

onde

- **main=“Histograma da Altura de 30 pessoas”** titulo do histograma
- **xlab=“cm”** rotulo do eixo horizontal
- **ylab=“frequência”** rotulo do eixo vertical
- **border=“blue”** cor do contorno das barras
- **col=“green”** cor das barras
- **xlim=c(150,210)** limite inferior e superior
- **las=1** rotacao do rotulo dos numeros do eixo vertical
- **breaks=5** número de classes
- **right=FALSE** define intervalo do tipo [a,b), se FALSE, e (a,b], se TRUE.

Obtendo o histograma



O pacote **ggplot2** gera gráficos e histogramas melhor elaborados.

Obtendo o histograma usando uma forma alternativa

Agrupando essas pessoas em **classes** de 10 cm temos:

classes	frequência
[150 ; 160[	1
[160 ; 170[	8
[170 ; 180[	14
[180 ; 190[	4
[190 ; 200[	2
[200 ; 210[	1

Fazendo isso no R:

```
nobs <- c(1:30)
dataX <- as.data.frame(cbind(nobs, X))
# transformando em data frame
tail(dataX)
```

```
##    nobs   X
## 25   25 181
## 26   26 183
```

```
## 27   27 185
## 28   28 190
## 29   29 194
## 30   30 201
```

```
# mostrando as seis últimas observações
quebras <- seq(150, 210, by = 10)
# definindo os intervalos
quebras
```

```
## [1] 150 160 170 180 190 200 210
```

```
dataX.cut <- cut(dataX$X, quebras, right = FALSE)
# construindo as classes fechado a esq e aberto a
# direita
dataX.freq <- table(dataX.cut)
# obtendo a frequência para cada classe.
dataXfreq <- cbind(dataX.freq)
# colocando os dados em colunas
dataXfreq
```

```
##           dataX.freq
## [150,160)          1
## [160,170)          8
## [170,180)         14
## [180,190)          4
## [190,200)          2
## [200,210)          1
```

1.2.4 Diagrama de caixa (Boxplot)

O texto sobre o diagrama de caixa foi baseado em Morettin and Bussab (2013).

Boxplot ou caixa de bigode também é uma ferramenta da estatística descritva que permite visualizar a dispersão dos valores da variável em análise. O que define o diagrama de caixa são os quartis. A parte inferior e superior da caixa, são respectivamente o primeiro quartil (Q_1) e o terceiro quartil (Q_3). A linha que corta da caixa é a mediana ou o segundo quartil (Q_2). Os bigodes que são as linhas que se estendem a partir da caixa, são calculado com base na amplitude interquartil (AIQ). A amplitude interquartil é a diferença entre os valores do terceiro e do primeiro quartis. Ou seja,

$$AIQ = Q_3 - Q_1$$

O bigode inferior denominado LI é calculado subtraindo $1,5 \times AIQ$ do valor do primeiro quartil Q_1. Ou seja,

$$LI = Q_1 - 1,5 \times AIQ$$

O bigode superior, denominado LS, é calculado somando $1,5 \times AIQ$ ao valor da terceiro quartil Q_3. Ou seja,

$$LS = Q_1 + 1,5 \times AIQ$$

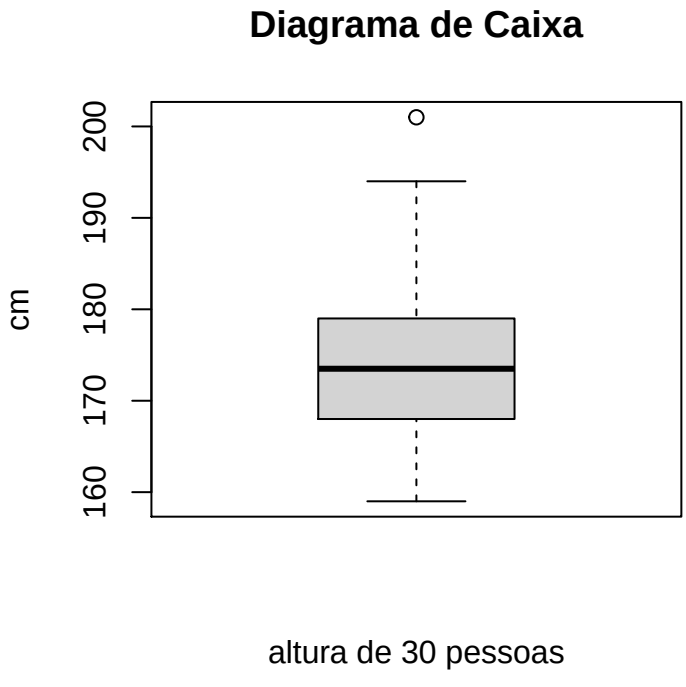
Os valores que forem menor que o LI ou maior que o LS são denominados valores discrepantes oui *outliers*. Os valores discrepantes, quando existentes, são colocados separadamente no diagrama de caixa mantendo a distancia relativa do limite inferior ou do limite superior.

Toma-se o mesmo exemplo da altura de 30 pessoas para apresentar o boxplot.

O código seria:

```
boxplot(X, data = dataX, main = "Diagrama de Caixa",
    ylab = "cm", xlab = "altura de 30 pessoas")
```

e o resultado segue abaixo.



1.3 Medidas de relação linear entre duas variáveis

Este assunto tem como base o material de Sartoris (2013).

Parece um pouco estranho incluir esse tópico logo depois das medidas de dispersão. Mas a variância é um caso especial da covariância que é a primeira medida de relação linear entre duas variáveis.

O coeficiente de correlação utiliza a covariância e o desvio padrão para resolver o problema de interpretação do resultado da covariância.

1.3.1 Covariância

pode ser estendida como uma *variância conjunta* entre duas variáveis. Ou seja,

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Fórmula alternativa da Variância

Também existe a fórmula alternativa da covariância.

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{XY}.$$

Fórmula alternativa da Covariância

Em outras palavras

$$cov(X,Y) = \text{média dos produtos de X e Y} - \text{produto das médias de X e Y.}$$

Covariância no R

Tomando o exemplo de consumo e renda da tabela 2.11 (Sartoris, 2013, p.42) tem-se

Ano	Consumo(X)	Renda(Y)	(XY)
1	600	1.000	600.000
2	700	1.100	770.000
3	800	1.300	1.040.000
4	900	1.400	1.260.000
Somatória	**3.000**	**4.800**	**3.670.000**
Média	**750**	**1.200**	**917.500**

Covariância no R

```
C1 <- c(600, 700, 800, 900)
R1 <- c(1000, 1100, 1300, 1400)
mediaC1 <- sum(C1)/length(C1)
mediaR1 <- sum(R1)/length(R1)
mediaC1R1 <- sum(C1 * R1)/length(C1)
covC1R1 <- mediaC1R1 - mediaC1 * mediaR1
covC1R1
```

```
## [1] 17500
```

```
cov(C1, R1)
```

```
## [1] 23333,33
```

Note que a função covariância no R é calculada dividindo por $(n-1)$ e não por n.

1.3.2 Coeficiente de Correlação

É obtido dividindo a covariância pelos desvios padrões das variáveis, retirando-se o efeito dos valores de cada variável. Como as unidades das variáveis se cancelam matematicamente, o coeficiente de correlação é um número puro que varia entre -1 e +1. Essa característica o torna mais fácil e claro a sua interpretação. Ou seja,

$$corr(X,Y) \cong \rho_{xy} = \frac{cov(X,Y)}{dp(X) \times dp(Y)}$$

onde

$$-1 \leq \rho \leq +1$$

Portanto, quando o coeficiente de correlação é igual a zero ou muito próximo a zero, significa que as duas variáveis analisadas não tem relação do tipo linear entre elas. Quando a o coeficiente de correlação é igual a -1 ou próximo de -1, tal fato indica que a existência de uma relação do tipo linear entre as duas vari áveis analisadas, sendo que as variações ocorrem no setido oposto. Ou seja, quando uma das variáveis aumenta de valor, a outra diminui. Quando o coeficiente de correlação é igual a +1 ou muito próximo de um positivo, tal fato indica que as duas variáveis tem uma relação do tipo linear, sendo que as variações em ambas as variáveis ocorrem no mesmo sentido. Ou seja, quando uma das variáveis aumenta de valor, a outra aumenta também. O que significa

o coeficiente de correlação ser: i) exatamente igual a zero; ii) ser exatamente igual a -1 e; exatamente igual a +1?

Correlação no R

```
medC1 <- sum(C1)/length(C1)
medR1 <- sum(R1)/length(R1)
varC1 <- (sum((C1 - medC1)^2))/length(C1)
varC1
```

```
## [1] 12500
```

```
varR1 <- (sum((R1 - medR1)^2))/length(R1)
varR1
```

```
## [1] 25000
```

```
dpC1 <- abs(sqrt(varC1))
dpR1 <- abs(sqrt(varR1))
corrC1R1 <- round(covC1R1/(dpC1 * dpR1), 4)
corrC1R1
```

```
## [1] 0,9899
```

Ou simplesmente

```
round(cor(C1, R1), 4)
```

```
## [1] 0,9899
```

Capítulo 2

Revisão de Literatura

Aqui o estado da arte mundo afora.

Capítulo 3

Metodologia

We describe our methods in this chapter.

Capítulo 4

Aplicações

Some *significant* applications are demonstrated in this chapter.

4.1 Exemplo um

4.2 Examplo dois

Capítulo 5

Considerações Finais

Terminado um excelente livro digital.

Referências Bibliográficas

Morettin, P. A. and Bussab, W. d. O. (2013). *Estatística Básica*. Saraiva, São Paulo, 8 edition.

Sartoris, A. (2013). *Estatística e Introdução à Econometria*. Saraiva, São Paulo, 2 edition.