Notas de aulas de Estatística Econômica

Marcos Minoru Hasegawa

2020-10-19

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 Hoffmann (2006), Morettin and Bussab (2013) e Sartoris (2013).

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

1.2.1 Amplitude

A amplitude de um conjunto de valores é a diferença entre o maior elemento e o menor elemento desse conjunto.

1.2.2 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.

1.2.3 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
```

1.2.4 Desvio Absoluto Médio

1.2.5 Diferença Média

1.2.6 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

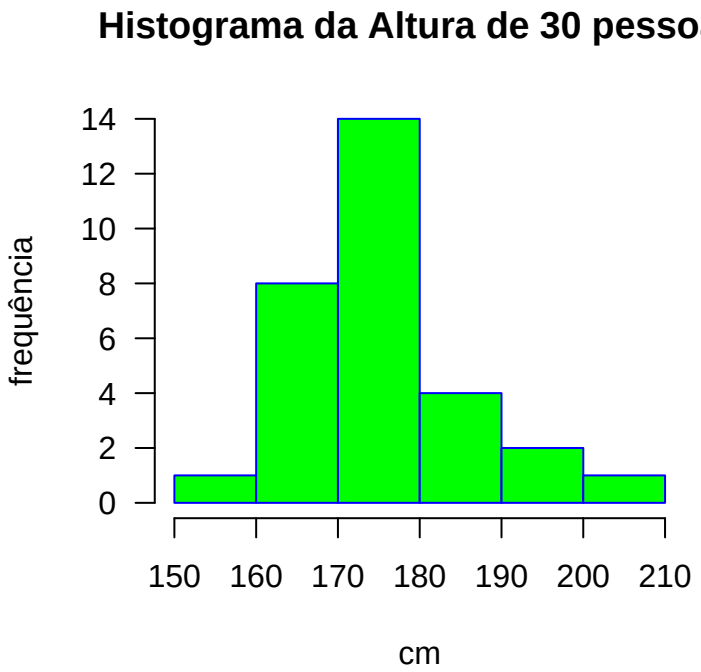
```
grafico1 <- hist(
                  X,
                  main="Histograma da Altura",
                  xlab="cm",
                  ylab="frequência",
                  border="blue",
                  col="green",
                  xlim=c(150,210),
                  las=1,
                  breaks=5,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

1.2.7 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,

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

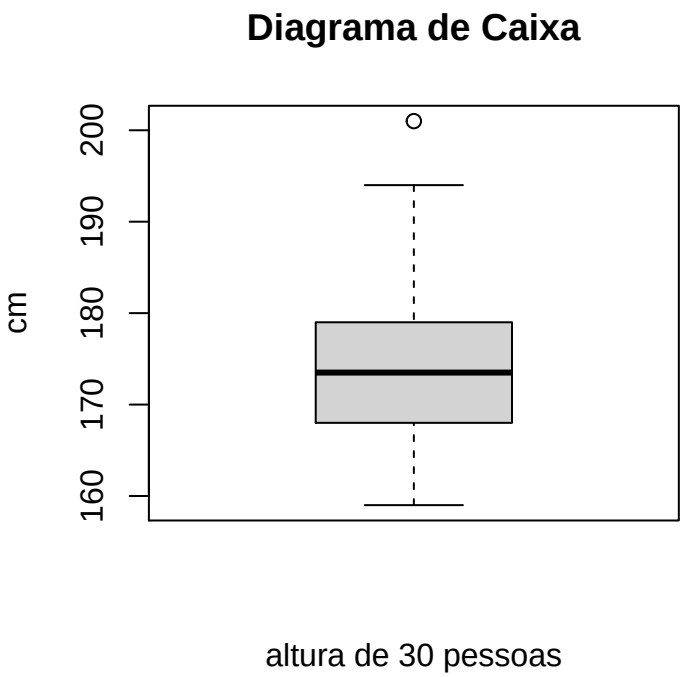
Os valores que forem menor que o LI ou maior que o LS são denominados valores discrepantes oui *outliers*. Os valores discrepantes, quando existentes, são colocados separadamente no diagrama de caixa mantendo a distancia relativa do limite inferior ou do limite superior.

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

$cov(X,Y)$ = média dos produtos de X e Y − 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

Medidas de desigualdade

O assunto sobre medidas de desigualdade está baseada na sua totalidade no capítulo 17 de Hoffmann (2006)

2.1 Príncipio de Pigou-Dalton

A condição de Pigou-Dalton define que as medidas de desigualdades devem ter seus valores aumentados quando há transferência regressivas de renda. Para entender a condicção de Pigou-Dalton, considere uma população com apenas duas pessoas cujas rendas são X_1 e X_2. Então, $\mu = \frac{X_1+X_2}{2}$. No caso de perfeita igualdade, $X_1 = X_2 = \mu$. No caso de uma distribuição com $X_1 \neq X_2$, uma transferência de renda do mais pobre para o mais rico, mantendo a renda média constante, aumenta o grau de desigualdade.

2.2 Transferência Regressiva

Essa tranferência de renda do mais pobre para o mais rico, mantida a renda média constante, é denominada como **transferência regressiva** de renda. Portanto, uma **transferência progressiva** é a transferência de renda do mais rico para o mais pobre.

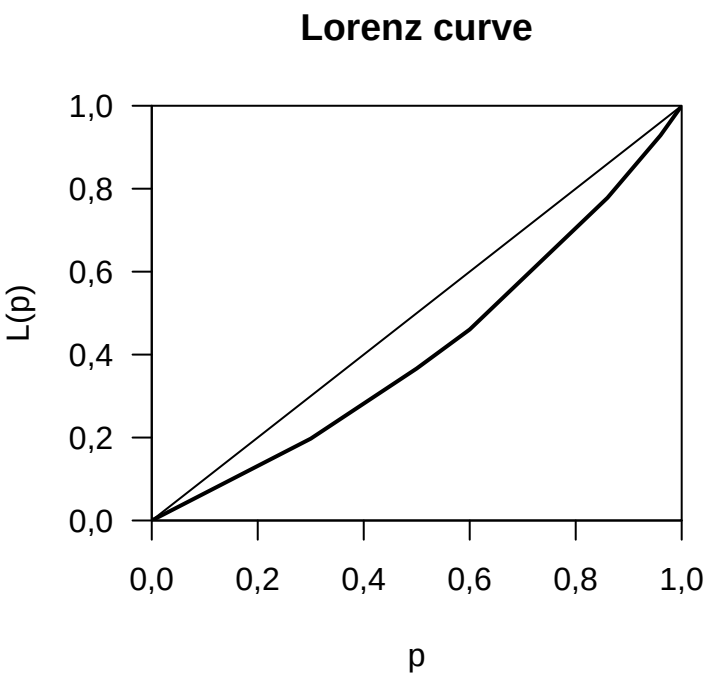
2.3 Curva de Lorenz

Tabela 2.1: Distribuição de pessoas ocupadas conforme renda obtida na atividade exercida no Brasil, de acordo com a PNAD 2003

estrato	% no estrato da população (%)	% no estrato da renda (%)	% acumulada da população ($100p$)	% acumulada da renda (100Φ)
I	30	7	30	7
II	20	9	50	16
III	20	13	70	29
IV	10	10	80	39
V	10	16	90	55
VI	5	13	95	68
VII	4	19	99	87
VIII	1	13	100	100

Considere os dados da tabela 2.1. Na coluna de porcentagem acumulada podemos observar que 70% da população possui 29% da renda. Os percentuais acumlados da população p e da renda Φ formam um plano cartesinao (p, Φ) originando a Cuirva de Lorenz.

```
library(ineq)
# usando os valores do exemplo em porcentagem mesmo
p <- c(30, 20, 20, 10, 10, 5, 4, 1)
r <- c(7, 9, 13, 10, 16, 13, 19, 13)

# calcula o mínimo da curva de Lorenz
Lc.min <- Lc(r, n = p)
# Desenha a curva de Lorenz em um gráfico
plot(Lc.min)
```



Considerando a curva de Lorenz, figura 2.1, que é basicamente a obtida pelo R, figura **??**, mas com algumas indicações, é possível obter algumas definições.

```
knitr::include_graphics("lorenz3.png")
```

A área que corresponde a letra a é denominada área de desigualdade. o seguimento de retas $\overline{AB}$ é chamado de *linha de perfeita igualdade* onde $p = \Phi$ e a área de de desigualdade é zero.

Analisando o casos de máxima desigualdade:

- excluindo-se o fato de renda negativa, considere que apenas um de n indíviduos receba toda a renda e os demais $n - 1$ indivíduos recebam zero de renda.
- Neste caso a pocentagem de renda é zero até o ponto $\dfrac{n-1}{n}$ no eixo horizontal, tornando-se $\Phi = 1$ ao se incluir o último indíviduo.
- Neste caso, a Curva de Lorenz é dada pela poligonal $\widehat{ABC}$ e a área de desigualdade máxima é o triângulo ABC.

2.4 Índice Gini

Considere os dados da tabela 2.1. Seja p o valor da proporção acumulada da população até certo estrato e seja Φ o valor da correspondente proporção

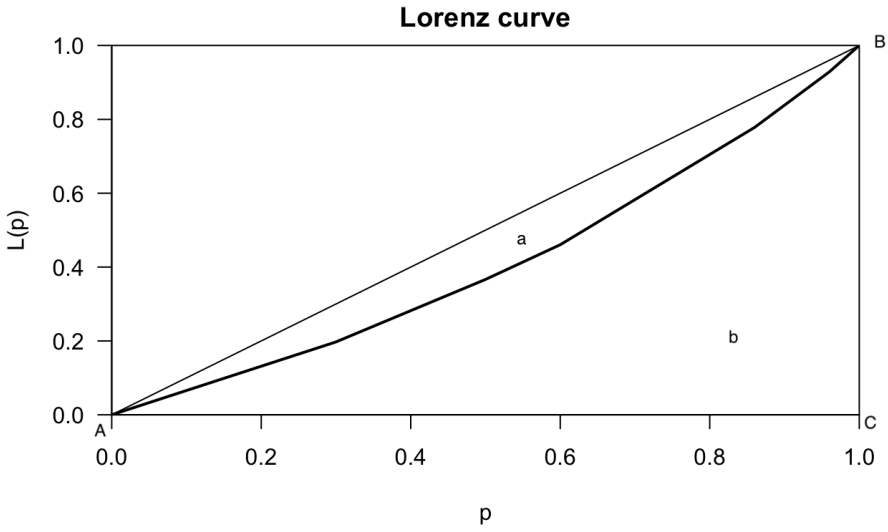


Figura 2.1: A curva de Lorenz com algumas indicações

acumulada da renda. Os pares de valores (p, Φ), para os diversos estratos, definem pontos em um sistema de eixos cartesianos como aparece na figura 2.1. Estes pontos estão sobre a curva de Lonrez, que mostra como a porporção acumulada da renda (Φ) varia em função da proporção acumulada da população (p), com as pessoas ordenadas de acordo com valores crescentes da renda. A área correspondente a a que está entre a reta AB e a curva de Lorenz na figura 2.1, é denominada **área de desigualdade**.

Para entender como ocorre a variação desta área de desigualdade, a área a, primeiro considere uma situação de distribuição de renda com perfeita igualdade, ou seja, uma população em que todos recebem a mesma renda.Nesta situação, a uma população p da população corresponde uma igual proporção Φ da renda total, ou seja, $\Phi = p$. Portanto, a curva de Lorenz dessa distribuição coincide com a reta AB da figura 2.1, denominado, por isso, de **linha de perfeita igualdade**. Neste caso a área de desigualdade é igual a zero.

Considere agora uma outra situação, uma distribuição de renda com o máximo de desigualdade. Considerando que **não** existe a possibilidade de renda negativa, esse seria o caso de uma população com n pessoas, em que uma delas recebe toda a renda e as $n - 1$ restante receba zero de renda. Nesta situação, a proporção acumulada da renda é igual a zero até o ponto do eixo horizontal (abcissa) $\frac{(n-1)}{n}$, tornando-se $\Phi = 1$ quando se se inclui a pessoa que recebe toda a renda. Neste caso, a curva de Lorenz passa a ser a poligonal ABC da figura 2.1. Que é numericamente igual a 0,5 (Por quê?).

Por definição, o **índice de Gini (G)** é uma relação entre a área de desigualdade,

indicada por a que passar a ser denominada de α, e a área do triângulo ABC que é numericamente igual a 0,5, ou seja,

$$G = \frac{\alpha}{0,5} = 2\alpha$$

A fórmula (2.4) é uma das fórmulas de Gini que tem utilidade do ponto de vista teórico.Uma vez que

$$0 \leq \alpha \leq 0,5$$

tem se que

$$0 \leq G \leq 1$$

Ou seja de que o índice de Gini varia entre entre zero, ausência de desigualdade, e um, máxima desigualdade.Adicionalmente, o índice de Gini é um número adimensional.

Uma fórmula alternativa e mais prática do ponto de vista do cálculo do índice de Gini pode ser obitda considerando-se uma distribuição discreta.

Seja uma variável aleatória discreta X_i para $i = 1, \ldots, n$, cujos valores estão em ordem crescente

$$X_1 \leq X_2 \leq \ldots \leq X_{n-1} \leq X_n$$

admitindo-se que os n valores são igualmente prováveis.

a proporção acumulada do número de elementos,m até o i-ésimo elemento, é

$$p_i = \frac{i}{n}, \text{ para } i = 1, \ldots, n$$

A correspondente proporção acumulada de X, até o i-ésimo elemento é

$$\Phi_i = \frac{\sum_{j=1}^{i} X_j}{\sum_{j=1}^{n} X_j} = \frac{1}{n\mu} \sum_{j=1}^{i} X_j$$

onde

$$\mu = \frac{1}{n} \sum_{j=1}^{n} X_j$$

Se X representa a renda individual e se $X_j < X_{j+1}$, Φ_i representa a fração da renda total apropriada pelas pessoas com renda inferior ou igual a X_i. As expressões (2.4) e (2.4) definem as coordenadas (p_i, Φ_i) com $i = 1, \ldots, n$ de n

pontos da curva de Lorenz. A rigor não existe, nesse caso, uma curva, mas uma poligonal cujos vértices são a origemdos exios e os pontos de coordenadas (p_i, Φ_i).

Na sequência é apresentada de forma resumida como se calcula o índice de Gini a partir dos valores de X_i para $i = 1, \ldots, n$ da variável. Na figura 2.1 a soma das áreas a e b totaliza a área do polígono ABC que numericamente é igual a 0,5. Portanto, $a = 0,5 - b$. Ou seja, colocando na notação mais elegante,

$$\alpha = 0,5 - \beta.$$

Substituindo (2.4) em (2.4) obtém-se

$$G = \frac{0,5 - \beta}{0,5} = 1 - 2\beta.$$

Note que a área abaixo da "curva" de Lorenz pode ser representada, de forma aproximada, como a soma das áreas de n trapézios um do lado do outro. Desta forma, a área b da figura 2.1, compreendida entre a poligonal de Lorenz e o eixo das abscissas, é obtida somando-se a área dos n trapézios. Ou seja, a área do i-ésimo trapézio é

$$S_i = \frac{\Phi_{i-1} + \Phi_i}{2} \times \frac{1}{n}$$

onde Φ_{i-1} é a base menor do i-ésimo trapézio; Φ_i é a base maior do i-ésimo trapézio e; $1/n$ é a altura do trapézio que corresponde a pessoa da população composta por n pessoas.

Note que o valor de $\Phi_0 = 0$, ou seja, o valor de Φ_{i-1} para $i = 1$. Com base na fórmula (2.4) é possível obter a área corresponde a b na figura 2.1 ou β nas notações matemáticas no texto

$$\beta = \sum_{i=1}^{n} S_i = \frac{1}{2n} \sum_{i=1}^{n} (\Phi_{i-1} + \Phi_i)$$

Substituindo (2.4) em (2.4), obtém-se

$$G = 1 - \frac{1}{n} \sum_{i=1}^{n} (\Phi_{i-1} + \Phi_i)$$

Considerando a fórmula (2.4) e que $\Phi_0 = 0$, se obtém o índice de Gini em termos da variável X_i. Ou seja,

$$G = 1 - \frac{1}{n^2 \mu}[(2n-1)X_1 + (2n-3)X_2 + \cdots + 3X_{n-1} + 1X_n]$$

Na parte sobre Estatística Descritiva foi apresentada a medida de dispersão chamada Diferença Absoluta Média que é dada por

$$\Delta = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|$$

Trabalhando algebricamente a fórmula (2.4)

$$\Delta = 2\mu - \frac{2}{n^2}[(2n-1)X_1 + (2n-3)X_2 + \cdots + 3X_{n-1} + 1X_n]$$

Se dividir (2.4) por 2μ obtém a fórmula do índice de Gini em termos da medida de dispersão Diferença Absoluta Média

$$G = \frac{\Delta}{2\mu}$$

Tratando-se da distribuição da renda em uma população, a relação (2.4) mostra que o índice de Gini, como medida de do grau de desigualdade, apresenta a vantagem de medir diretamente as diferenças de rendal, levando em consideração diferenças entre as rendas de **todos** os pares de pessoas.

Como Δ é uma medida de dispersão da distribuição, a relação (2.4) mostra que o índice de Gini é uma medida de dispersão relativa. Assim, o conceito de desigualdade de uma distribuição se confunde com o conceito de dispersão relativa.

Com um desenvolvimento algébrico de Δ é possível transformar a fórmula (2.4) em

$$G = \frac{2}{n^2\mu}\sum_{i=1}^{n} iX_i - \frac{1}{n} - 1$$

A relação (2.4) mostra que, no cálculo do índice de Gini, cada valor de X_i da variável aparece poderado por i. Ou seja, X_i aparece poderada pelo respectivo número de ordem na sequência dos valores ordenados.

Exemplo numérico

Para aplicar a fórmula do índice de Gini, utiliza-se os dados apresentados na tabela abaixo, obtidos de Hoffmann (2006).

Tabela 2.2: Valores de X_i, p_i, Φ_i e $\Phi_{i-1} + \Phi_i$ para a população hipotética de 8 elementos

i	p_i	X_i	$\sum_{j=1}^{n} X_j$	Φ_i	$\Phi_{i-1} + \Phi_i$
1	0,125	1	1	0,02	0,02
2	0,250	1	2	0,04	0,06
3	0,375	1	3	0,06	0,10

i	p_i	X_i	$\sum_{j=1}^{n} X_j$	Φ_i	$\Phi_{i-1} + \Phi_i$
4	0,500	2	5	0,10	0,16
5	0,625	4	9	0,18	0,28
6	0,750	8	17	0,34	0,52
7	0,875	13	30	0,60	0,94
8	1,000	20	50	1,00	1,60

Com essas informações é possível calcular o índice de Gini, através de (2.4)

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}(\Phi_{i-1} + \Phi_i);$$

através de (2.4)

$$G = \frac{\Delta}{2\mu},$$

onde

$$\Delta = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|$$

e

$$\mu = \sum_{i=1}^{n} X_i;$$

e através de (2.4)

$$G = \frac{2}{n^2\mu}\sum_{i=1}^{n} iX_i - \frac{1}{n} - 1$$

Usando (2.4), é necessário totalizar a coluna $\Phi_{i-1} + \Phi_i$ na tabela 2.2. Ou seja,

```
options(OutDec = ",")
somaphis <- c(0.02, 0.06, 0.1, 0.16, 0.28, 0.52, 0.94,
    1.6)
somasomaphis <- sum(somaphis)
somasomaphis
```

```
## [1] 3,68
```

$$\sum_{i+1}^{8}(\Phi_{i-1} + \Phi_i) = 3{,}68$$

Portanto

```
giniphis <- 1 - 1/length(somaphis) * somasomaphis
giniphis
```

```
## [1] 0,54
```

$$G = 1 - \frac{1}{8} \times 3{,}68 = 0{,}54$$

Para aplicar a fórmula (2.4) que é a fórmula do índice de Gini em termos de diferença absoluta média, Δ, é necessário calcular a diferença absoluta média com base nos dados de X_i da tabela 2.2.

```
xi <- c(1, 1, 1, 2, 4, 8, 13, 20)

XC <- matrix(xi, nrow = length(xi), ncol = length(xi),
    byrow = FALSE)
XC
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1    1    1    1    1    1    1    1
## [2,]    1    1    1    1    1    1    1    1
## [3,]    1    1    1    1    1    1    1    1
## [4,]    2    2    2    2    2    2    2    2
## [5,]    4    4    4    4    4    4    4    4
## [6,]    8    8    8    8    8    8    8    8
## [7,]   13   13   13   13   13   13   13   13
## [8,]   20   20   20   20   20   20   20   20
```

```
XL <- matrix(xi, nrow = length(xi), ncol = length(xi),
    byrow = TRUE)
XL
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1    1    1    2    4    8   13   20
## [2,]    1    1    1    2    4    8   13   20
## [3,]    1    1    1    2    4    8   13   20
## [4,]    1    1    1    2    4    8   13   20
## [5,]    1    1    1    2    4    8   13   20
## [6,]    1    1    1    2    4    8   13   20
## [7,]    1    1    1    2    4    8   13   20
## [8,]    1    1    1    2    4    8   13   20
```

```
DIFX <- XC - XL
DIFX
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    0    0    0   -1   -3   -7  -12  -19
## [2,]    0    0    0   -1   -3   -7  -12  -19
## [3,]    0    0    0   -1   -3   -7  -12  -19
## [4,]    1    1    1    0   -2   -6  -11  -18
## [5,]    3    3    3    2    0   -4   -9  -16
## [6,]    7    7    7    6    4    0   -5  -12
## [7,]   12   12   12   11    9    5    0   -7
## [8,]   19   19   19   18   16   12    7    0
```

```
ABSDIFX <- abs(DIFX)
ABSDIFX
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    0    0    0    1    3    7   12   19
## [2,]    0    0    0    1    3    7   12   19
## [3,]    0    0    0    1    3    7   12   19
## [4,]    1    1    1    0    2    6   11   18
## [5,]    3    3    3    2    0    4    9   16
## [6,]    7    7    7    6    4    0    5   12
## [7,]   12   12   12   11    9    5    0    7
## [8,]   19   19   19   18   16   12    7    0
```

```
iota <- matrix(1, nrow = length(xi), ncol = 1, byrow = TRUE)
iota
```

```
##      [,1]
## [1,]    1
## [2,]    1
## [3,]    1
## [4,]    1
## [5,]    1
## [6,]    1
## [7,]    1
## [8,]    1
```

```
somacolunaABSDIFX <- t(iota) %*% ABSDIFX
somacolunaABSDIFX
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]   42   42   42   40   40   48   68  110
```

```
somalinhasomacolunaABSDIFX <- somacolunaABSDIFX %*%
    iota
somalinhasomacolunaABSDIFX
```

```
##      [,1]
## [1,]  432
```

```
obs <- length(xi)

delta <- obs^(-2) * somalinhasomacolunaABSDIFX
delta
```

```
##      [,1]
## [1,] 6,75
```

Ou seja,

$$\Delta = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j| = \frac{1}{(8)^2} \times 432 = 6{,}75$$

Com a diferença absoluta média de X_i devidamente calculada, aplica-se a fórmula (2.4)

```
ximedio <- sum(xi)/length(xi)
ximedio
```

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

```
ginidelta <- delta/(2 * ximedio)
ginidelta
```

```
##      [,1]
## [1,] 0,54
```

$$G = \frac{\Delta}{2\mu} = \frac{6{,}75}{2 \times 6{,}25} = 0{,}54$$

Para aplicar a fórmula (2.4) na obtenção do índice de Gini é necessário ponderar cada valor de X_i pela sua respectiva ordem i e soma todos os respectivos produtos

$$\sum_{i=1}^{n} iX_i.$$

usando os dados da tabela 2.2

```
is <- matrix(1:length(xi), nrow = length(xi), ncol = 1,
    byrow = TRUE)
is
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
## [4,]    4
## [5,]    5
## [6,]    6
## [7,]    7
## [8,]    8
```

```
ixi <- t(is) * xi
ixi
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1    2    3    8   20   48   91  160
```

```
somaixi <- sum(ixi)
somaixi
```

```
## [1] 333
```

```
ginifinal <- 2/(obs^2 * ximedio) * somaixi - obs^(-1) -
    1
ginifinal
```

```
## [1] 0,54
```

obtém

$$G = \frac{2}{n^2\mu}\sum_{i=1}^{n} iX_i - \frac{1}{n} - 1 = \frac{2}{8^2 \times 6{,}25} \times 333 - \frac{1}{8} - 1 = 0{,}54$$

2.5 Discrepância Máxima

Discrepância Máxima é a maior distância entre entre a linha AB e a curva de Lorenz da figura 2.1. Portanto, a discrepância máxima é a diferença máxima entre a relação da porcentagem acumulada da população e a sua respectiva porcentagem acumulada da renda numa situação de exata igualdade dada pela reta AB e a relação da porcentagem acumulada da população e a sua respectiva porcentagem acumulada da renda numa situação de desigualdade entre as pessoas dessa população que na figura corresponde a poligonal da curva de Lorenz. De acordo com Hoffmann (2006)

$$D = p_h - \Phi_h$$

Portanto, o cálculo de D através de (2.5) depende de h, um número inteiro positivo. Encontrando-se $i = h$ encontra-se a discrepância máxima D.

Seja uma sequência de valores ordenados em ordem crescente de uma variável discreta X_i

$$X_1 \leq X_2 \leq \ldots \leq X_n$$

sendo válida pelos menos uma desigualdade.

Para o cálculo da discrepância máxima, é importante entender que a mesma ocorre quando a inclinação do segmento da poligonal da curva de Lorenz passa de uma valor menor que um para um valor maior que um. De acordo com Hoffmann (2006), a inclinação do segmento da poligonal é dada por

$$d_i = \frac{X_i}{\mu}$$

Essa mudança é identificada quando o valor de X_i ordenada em ordem crescente passa de uma valor menor que a média μ para um valor maior que a média μ, ou seja,

$$X_i < \mu \;\; para \; 1 \leq i \leq h$$

e

$$X_i \geq \mu \;\; para \; h < i \leq n$$

Nesaas condições percorre-se a sequência de valores em ordem crescente, o valor de $p_i - \Phi_i$ aumenta até a inclusão do h-ésimo elemento.que corresponde a (2.5).

Considerando (2.5), (2.4) e (2.4) chaga-se em

$$D = \frac{h}{n} - \frac{1}{n\mu}\sum_{i=1}^{h} X_i$$

que depois de algumas manobras algébricas torna-se

$$D = \frac{1}{n\mu}\sum_{i=1}^{h}(\mu - X_i)$$

Na parte de estatística descritiva foi apresentado a medida de dispersão desvio absoluto médio

$$\delta = \frac{1}{n}\sum_{i=1}^{n}|X_i - \mu|$$

e considerando que

$$\delta = \frac{1}{n}\left[-\sum_{i=1}^{h}(X_i - \mu) + \sum_{i=h+1}^{n}(X_i - \mu)\right]$$

e que

$$\sum_{i=h+1}^{n}(X_i - \mu) = -\sum_{i=1}^{h}(X_i - \mu).$$

Portanto

$$\delta = \frac{2}{n}\sum_{i=h+1}^{n}(X_i - \mu) = \frac{2}{n}\sum_{i=1}^{h}(\mu - X_i).$$

Comparando (2.5) e (2.5) obtém-se

$$D = \frac{\delta}{2\mu}.$$

Se o Desvio Absoluto Médio, δ, é uma medida de dispersão da distribuição, a fórmula (2.5) mostra que discrepância máxima, da mesma forma que o índice de Gini, é uma medida de dispersão relativa. Retomando o exemplo numérico da tabela 2.2, é possível obter o valor da sua discrepância máxima através de (2.5).

$$D = p_h - \Phi_h = 0,625 - 0,180 = 0,445.$$

O mesmo resultado poderia ser obtido através da fórmula (2.5). Para isso é necessário calcular o desvio absoluto médio, δ, para os valores de X_i do exemplo numérico. Ou seja,

```
somadesvioabsolutomedioxi <- sum(abs(xi - ximedio))
somadesvioabsolutomedioxi
```

```
## [1] 44,5
```

```
desvioabsolutomedioxi <- (length(xi))^(-1) * sum(abs(xi -
    ximedio))
desvioabsolutomedioxi
```

```
## [1] 5,5625
```

$$\delta = \frac{1}{n}\sum_{i=1}^{n}|X_i - \mu| = \frac{1}{8} \times 44{,}5 = 5{,}5625.$$

Portanto,

```
discrepanciamaximaxi <- desvioabsolutomedioxi/(2 *
    ximedio)
discrepanciamaximaxi
```

```
## [1] 0,445
```

$$D = \frac{\delta}{2\mu} = \frac{5{,}5625}{2 \times 6{,}25} = 0{,}445.$$

2.6 Redundância e Índice de Theil

2.6.1 Teoria da Informação

Para entender melhor as medidas de desigualdades de Theil, é necessário introduzir alguns conceitos da teoria da informação.

Seja x, a probabilidadde de ocorrer o evento E.

- Para $x = 1$, a mensagem **evento E ocorreu** não tem nenhum conteúdo informativo.
- Para $x \to 0$, ou seja, para valores muito pequenos de x, a mensagem **evento E ocorreu** tem alto valor informativo.

A segunda situação seria, por exemplo, o caso de uma notícia que nos causas surpresa ou de um *furo* de imprensa. Quando x tende a zero, o conteúdo informativo da mensagem **evento E ocorreu** tende a infinito.

Matematicamente, o conteúdo informativo da mensagem que afirma que determinado evento ocorreu é dado por

$$h(x) = \log \frac{1}{x} = \log x^{-1} = -\log x$$

De acordo com Hoffmann (2006), a escolha da função logarítmica é devido a propriedade de atividade do conteúdo informativo no caso de eventos independentes. Portanto, se E_1 e E_2 são dois eventos independentes com probabilidades x_1 e x_2, respectivamente, a probabilidade de que ambos ocorram é x_1x_2. O conteúdo informativo da mensagem de que ambos os eventos ocorreram é

$$h(x_1x_2) = \log \frac{1}{x_1x_2} = \log \frac{1}{x_1} + \log \frac{1}{x_2} = h(x_1) + h(x_2)$$

Em teoria da informação, normalmente se utiliza logarítmos na base 2 ou logarítmos naturais. Desta forma:

- Logaritmos na base 2: o conteúdo informativo é medido em **bits**.
- Logaritmos naturais: o conteúdo informativo é medido em **nits**.
- 1 bit = 0,693 nit.
- 1 nit = 1,443 bit.

Generalizando o conceito de informação, é apresentando, na sequência, como se mede o conteúdo informativo de uma **mensagem sujeita a erro**, ou **mensagem incerta**.

Para isso, admita-se que a a probabilidade de chover em um determinado dia, em certo local, estabelecida com base em séries históricas, seja $x_1 = 0,5$. Nesse caso o conteúdo da informação **chove** é de

$$h(x_1) = \log \frac{1}{0,5} = \log 2^1 = 1 \text{ bit}$$

Suponha agora que uma previsão de tempo estabeleceu que iria chover. Suponha, também, que, com base nos resultados anteriores de tais previsões, probabilidade de que realmente shova passa a ser $y_1 = 0,68$. De acordo com as novas suposições, o conteúdo da informação **chove** é

$$h(y_1) = \log \frac{1}{o,68} + 0,5564 \text{ bit}$$

O conteúdo informativo da previsão é

$$h(x_1) - h(y_1) = \log \frac{1}{x_1} - \log \frac{1}{y_1} = 1 - 0,5564 = 0,4436 \text{ bit}$$

ou

$$h(x_1) - h(y_1) = \log \frac{1}{x_1} - \log \frac{1}{y_1} = \log \frac{1}{x_1} + \log \left(\frac{1}{y_1}\right)^{-1} = \log \frac{y_1}{x_1} = \log \left(\frac{0,50}{0,68}\right) = 0,4436 \text{ bit.}$$

Ou seja, o conteúdo informativo **chove**, com base na probabilidade x_i nos dados históricos e na probabilidade y_i do histórico de previsões, é de 0,4436 bit.

Generalizando, o coanteúdo informativo de uma mensagem sujeita a erro ou mensagem incerta, como é o caso da previsão, é dado por

$$\log \frac{y}{x}$$

onde

- x é a probabilidade *ex-ante* ou a probabilidade de que o evento ocorra antes de recebida a mensagem;
- y é a probabilidade *ex-post* ou a probabilidade de que oevento ocorra uma vez recebida a mensagem.

Na sequência é apreseantado o conceito de *entropia*.

Entropia de uma distribuição $H(x)$

Seja o universo de n possíveis eventos E_i, para $i = 1, \ldots, n$, mutuamente exclusivos aos quais associa-se as probabilidades x_i. Sabe-se que

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

A informação esperada de uma mensagem certa, ou seja, a esperança matemática do conteúdo informativo da mensagem **ocorreu** E_i, também denominada entropia da distribuição, é

$$H(x) = E[h(x_i)] = \sum_{i=1}^{n} x_i h(x_i) = \sum_{i=1}^{n} x_i \log \frac{1}{x_i} = -\sum_{i=1}^{n} x_i \log x_i$$

Para o caso particular de $x_i = 0$, adota-se a definição

$$x \log x = 0, \quad \text{se } x = 0$$

uma vez que

$$\lim_{x\to 0}(x\log x) = 0$$

Para $0 < x_i \leq 1$ se tem

$$\frac{1}{x_i} \geq 1$$

e

$$\log \frac{1}{x_i} \geq 0.$$

Conclui-se que

$$H(x) = \sum_{i=1}^{n} x_i \log \frac{1}{x_i} = -\sum_{i=1}^{n} x_i \log x_i \geq 0$$

Valor mínimo de $H(x)$

O valor mínimo de $H(x)$ ocorre quando uma das probabilidades é 1 e as demais, consequentemente, são nulas. Nesse caso $H(x) = 0$. Ou seja, na somatória há um único $x_i = 1$ e o restante $x_i = 0$. Portanto,

- quando $x = 0$

  $$x\log x = 0$$

  de acordo com (2.6.1);

- quando $x = 1$ se tem $\log 1 = 0$ e também

  $$x\log x = 0.$$

Assim o valor mínimo do valor esperado do conteúdo informativo $H(x)$ é

$$H(x) = -\sum_{i=1}^{n} x_i \log x_i = 0$$

Valor Máximo de $H(x)$

Para encontrar o valor máximo de $H(x)$ sujeito a condição de que $\sum x_i = 1$, utiliza-se o método do multiplicador de Lagrange, escrevendo a seguinte função

$$\max H(x) = -\sum_{i=1}^{n} x_i \log x_i$$

sujeito a

$$\sum_{i=1}^{n} x_i = 1$$

então

$$\mathcal{L} = -\sum_{i=1}^{n} x_i \log x_i - \lambda \left(\sum_{i=1}^{n} x_i - 1 \right)$$

Igualando a zero as derivadas parciais de (2.6.1) em relação a x_i e admitindo-se que se usa os logaritmos naturais, se tem:

$$\log x_i = -(1 + \lambda), \quad para\ i = 1, \ldots, n$$

sendo que

$$x_i = e^{-(1+\lambda)} = \frac{1}{e^{(1+\lambda)}}.$$

O valor máximo de $H(x)$ acontece quando todos os valores de x_i, ou seja, todos as probabilidades são iguais entre si e, portanto, igual a $\frac{1}{n}$. Nesse caso,

$$H(x) = \sum_{i=1}^{n} x_i \log \frac{1}{x_i} = \sum_{i=1}^{n} \frac{1}{n} \log n = n \frac{1}{n} \log n = \log n$$

Resumindo, o valor esperado da informação ou a entropia da distribuição $H(x)$ varia entre 0 e $\log n$. Ou seja,

$$0 \leq H(x) \leq \log n.$$

A entropia da distribuíção é máxima, ou seja, há um máximo de incerteza a respeito do que pode ocorrer, quando todos os possíveis eventos são igualmente prováveis, ou seja, quando há um máximo de *desordem* no sistema.

Informação de uma mensagem incerta

Finalmente é apresentado o conceito de informação de uma mensagem incerta. Dado o universo de n possíveis eventos E_i, mutuamente exclusivos, com probabilidades x_i, para $i = 1, \ldots, n$, considera-se uma mensagem incerta que poderia ser uma previsão ou uma mensagem duvidosa, que transforma as probabilidades *a priori* x_i em probabilidade *a posteori* y_i, onde y_i é a probabilidade de ocorrência do evento E_i depois de recebido a mensagem. Lembrando (2.6.1), verifica-se que a esperança matemática do conteúdo informativo da mensagem é

$$I(y : x) = \sum_{i=1}^{n} y_i \log \frac{y_i}{x_i}$$

A definição (2.6.1), do conteúdo informativo de uma mensagem certa, é somente um caso especial de (2.6.1), eem que uma probabilidade *a posteriori* é igual a um e todas as outras são iguais a zero, ou seja, $y_j = 1$ e $y_i = 0$ para todo $i \neq j$.

2.6.2 Índice T de Theil

Seja uma população com n pessoas em que cada uma recxebe uma fração não negativa da renda total,

$$y_i \geq 0, \quad com\ i = 1, \ldots, n.$$

Se a renda média é μ e X_i é a renda i-ésima pessoa,

$$y_i = \frac{X_i}{n\mu}.$$

Obviamente,

$$\sum_{i=1}^{n} y_i = 1.$$

Os valores de y_i tem as mesmas propriedades que as probabilidades x_i associadas a um universo de eventos E_i da teoria da informação. Assim sendo, pdoe-se, com base em (2.6.1), definir a **entropia** da distribuição de renda considerada como sendo

$$H(y) = \sum_{i=1}^{n} y_i \log \frac{1}{y_i}.$$

De acordo com (2.6.1), se tem

$$0 \leq H(y) \leq \log n.$$

Assim é possível definir as duas situações extremas:

- o caso de perfeita igualdade na distribuição da renda,

  $$y_i = \frac{1}{n} \quad para\ i = 1, \ldots, n,$$

  se tem $H(y) = \log n$;

- o caso de perfeita desigualdade na distribuição de renda,

  $$y_i = 1, \quad para\ i = 1, \ldots, n,$$

  se tem $H(y) = 0$.

A **entropia** é, portanto, uma medida do grau de igualdade da distribuição. Mas como o objeto de análise é desigualdade, é ,muito mais interessante uma medida de desigualdade. Para isto basta subtrair a entropia do seu valor próprio máximo, $\log n$. Essa medida, denominada **Índice T de Theil** da distribuição é dada por

$$T = \log n - H(y) = \sum_{i=1}^{n}$$

Para o cálculo do índice T de Theil, pode-se usar os logarítmos naturais ou os logarítmos na base 2, obtendo-se o valor de T en *nits* ou *bits*, respectivamente. Na prática, utiliza-se mais o logarítmo natural.

Note que

$$0 \leq T \leq \log n$$

sendo que:

- $T = 0$ corresponde ao caso de uma distribuição da renda com perfeita igualdade e;
- $T = \log n$ corresponde ao caso de uma distribuição da renda com perfeita desigualdade.

De (2.6.2) se tem

$$T = \sum_{i=1}^{n} y_i \log \frac{y_i}{\frac{1}{n}}.$$

Comparando essa equação com (2.6.1), verifica-se que o **índice T de Theil** corresponde à esperança do valor informativo de uma mensagem incerta, em que as probabilidades *a posteriori* são as frações da renda total y_i apropriadas pelas pessoas, e as probabilidades *a priori* são iguais a $1/n$, ou seja, iguais à fração da população correspondente a cada pessoa.

2.6.3 Índice de L de Theil

A outra medida de desigualdade proposta por Theil, o **índice L de Theil**, corresponde à esperança do valor informativo de uma mensagem incerta, em que as probabilidades 8a posteriori* são as frações da população $1/n$ e as probabilidades *a priori* são as frações da renda y_i. Ou seja,

$$L = \sum_{i=1}^{n} \frac{1}{n} \log \frac{\frac{1}{n}}{y_i} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{ny_i}.$$

Verifica-se que o **índice L de Theil** é igual a zero no caso de pefeita igualdade

$$y_i = \frac{1}{n}$$

para todo i.

Basta uma das renda aproximar-se de zero para que o valor de L tenda a infinito, fazendo que o índice L seja inútil quando se trata de comparar distribuições de renda que incluem valores nulos.

Uma vantagem importante das medidas de desigualdades de Theil na análise da distribuição de renda ou da riqueza é que, quando os dados podem ser agrupados com base em um critério qualquer, por exemplo regiões, os valores de T e L podem ser decompostos em uma medida de desigualdade **entre** grupos, por exemplo inter-regional, e uma média poderada das medidas de desigualdades **dentro** de grupos, por exemplo dentro das regiões.

2.7 Variância dos Logaritmos

A variância dos logaritmos das rendas é frequentemente utilizada como medida da desigualdade da distribuição da renda em uma população. Para uma população com n pessoas, em que a renda da i-ésima pessoas aé indicada por X_i para $i = 1, \ldots, n$, a variância dos logaritmos das rendas é dada por

$$V(Z) = \frac{1}{n} \sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2$$

onde

$$Z_i = \log X_i$$

e

$$\overline{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

Nota-se que $V(Z)$ só e definida para $X_i \geq 0$ para $i = 1, \ldots, n$.

Indicando-se por X^* a média geométrica dos X_i , se tem

$$V(Z) = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{X_i}{\overline{X^*}}\right)^2$$

A variância dos logaritmos, da mesmaforma que as medias T e L de Theill, é uma medida de desigualdade que, quando os dados podem ser agrupados segundo um critério qualquer, pode ser decomposta em um componente que corresponde à desigualdade entre os grupos e uma média podenrada das variâncias dos logarítmos dentro dos grupos.

Capítulo 3

Números-Índices

Capítulo 4

Variável Aleatória e Distribuição

Capítulo 5

Considerações Finais

Terminado um excelente livro digital.

Referências Bibliográficas

Hoffmann, R. (2006). *Estatística para Economistas*. Cengage Learning, São Paulo, 4 edition.

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.