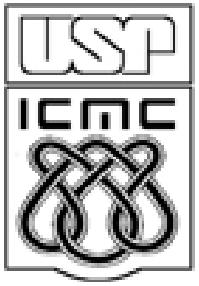


Aula Prática 2

2012

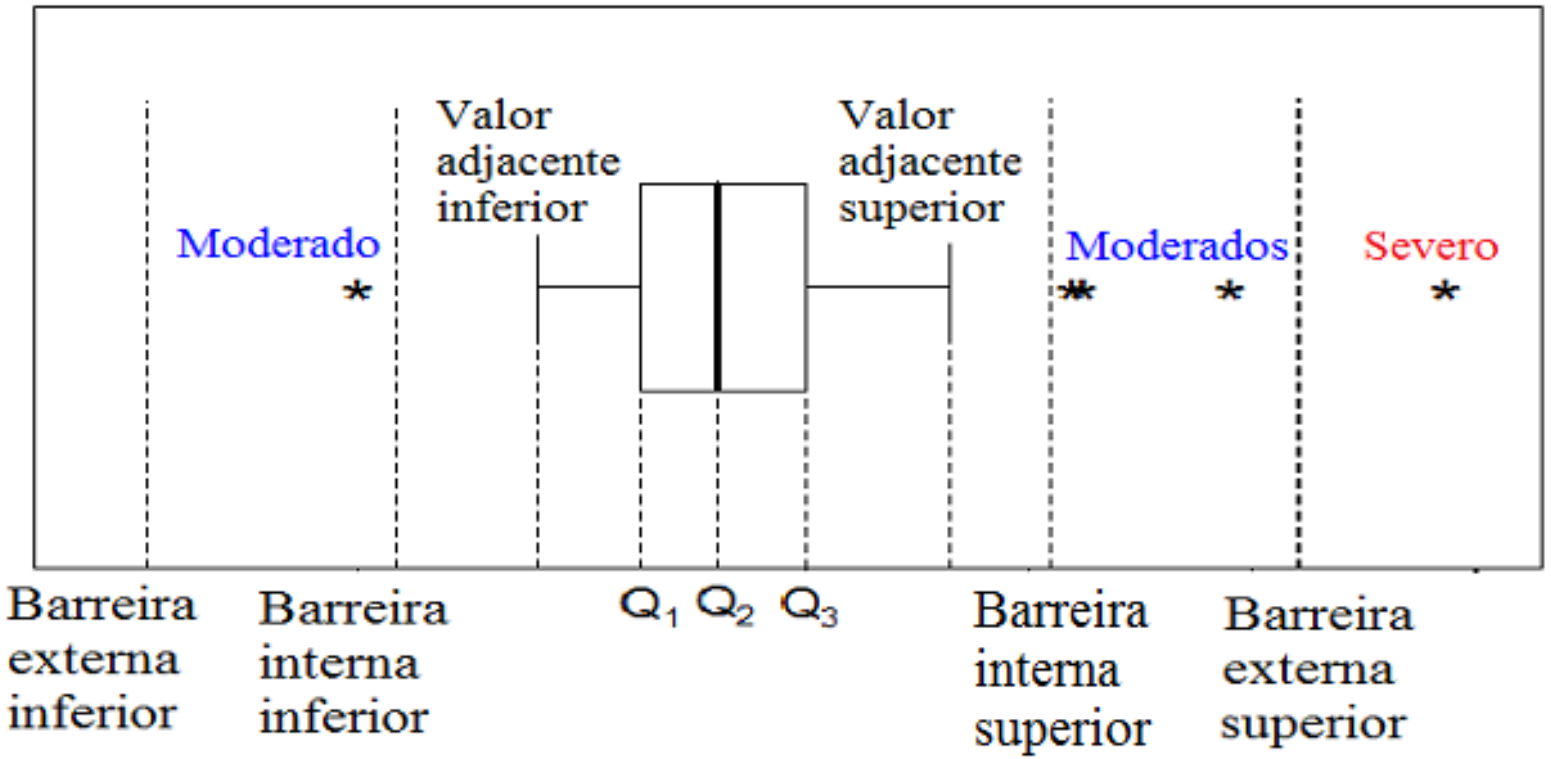
1. Dispersão

2. Assimetria e Curtose

3. Concentração e desigualdade

1. Dispersão

Gráfico de caixas (box plot, caixa-de-bigodes, *box-and-whisker plot*)



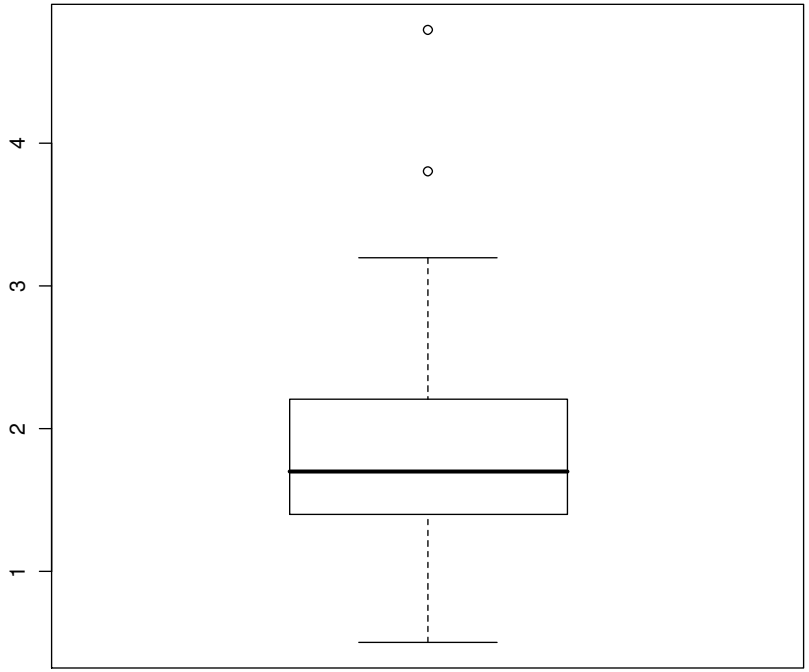
Valor adjacente inferior: menor valor no conjunto de dados que não é extremo (pode ser igual a $x_{(1)}$).

Valor adjacente superior: maior valor no conjunto de dados que não é extremo pode ser igual a $x_{(n)}$.

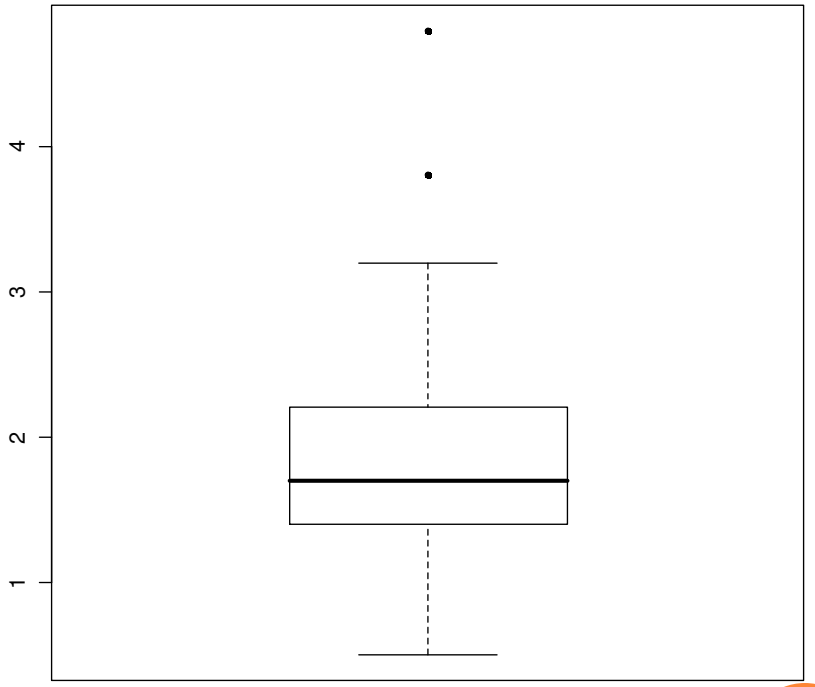
Box plot

```
> x = c(1.5, 1.9, 1.7, 1.6, 3.8, 1.3, 2.2, 1.8, 1.3, 0.5, 1.6, 1.4, 1.7, 1.7,
1.9, 0.7, 2.2, 2.3, 2.4, 2.3, 1.8, 2.7, 1.3, 1.7, 2.0, 1.1, 2.1, 1.6, 1.3,
2.2, 1.5, 2.3, 1.1, 1.8, 1.2, 2.0, 1.5, 1.5, 2.6, 1.6, 1.4, 2.2, 1.5, 1.2,
2.0, 1.3, 2.6, 1.9, 1.3, 2.4, 3.2, 1.9, 4.8)
```

```
> boxplot(x)
```



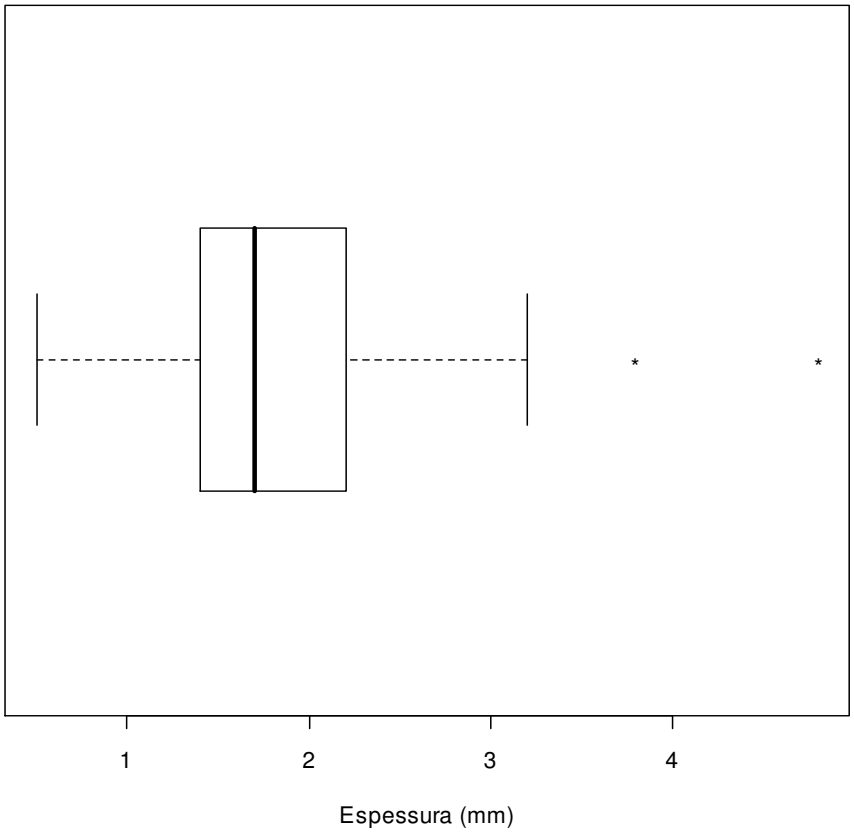

```
> boxplot(x, pch = 20)
```



Obs. Na construção do gráfico de caixas, quanto maior for n, melhor.

Gráfico de caixa

```
> boxplot(x, pch = "*",
horizontal = TRUE, xlab =
"Espessura (mm)")
```




```
> bx = boxplot(x, plot = FALSE)
> names(bx)
```

```
[1] "stats" "n"     "conf"
"out"   "group" "names"
```

`bx$stats`: valor adjacente inferior, Q_1, Q_2, Q_3 e valor adjacente superior.

`bx$n`: número de observações.

`bx$out`: observações extremas.

```
> bx$stats
     [,1]
[1,]  0.5
[2,]  1.4
[3,]  1.7
[4,]  2.2
[5,]  3.2
```

Gráfico de caixa

```
> boxplot(x, pch = "*", horizontal = TRUE, xlab = "Espessura (mm)")
> identify(bx$out, rep(1, length(bx$out)), match(bx$out, x))
```

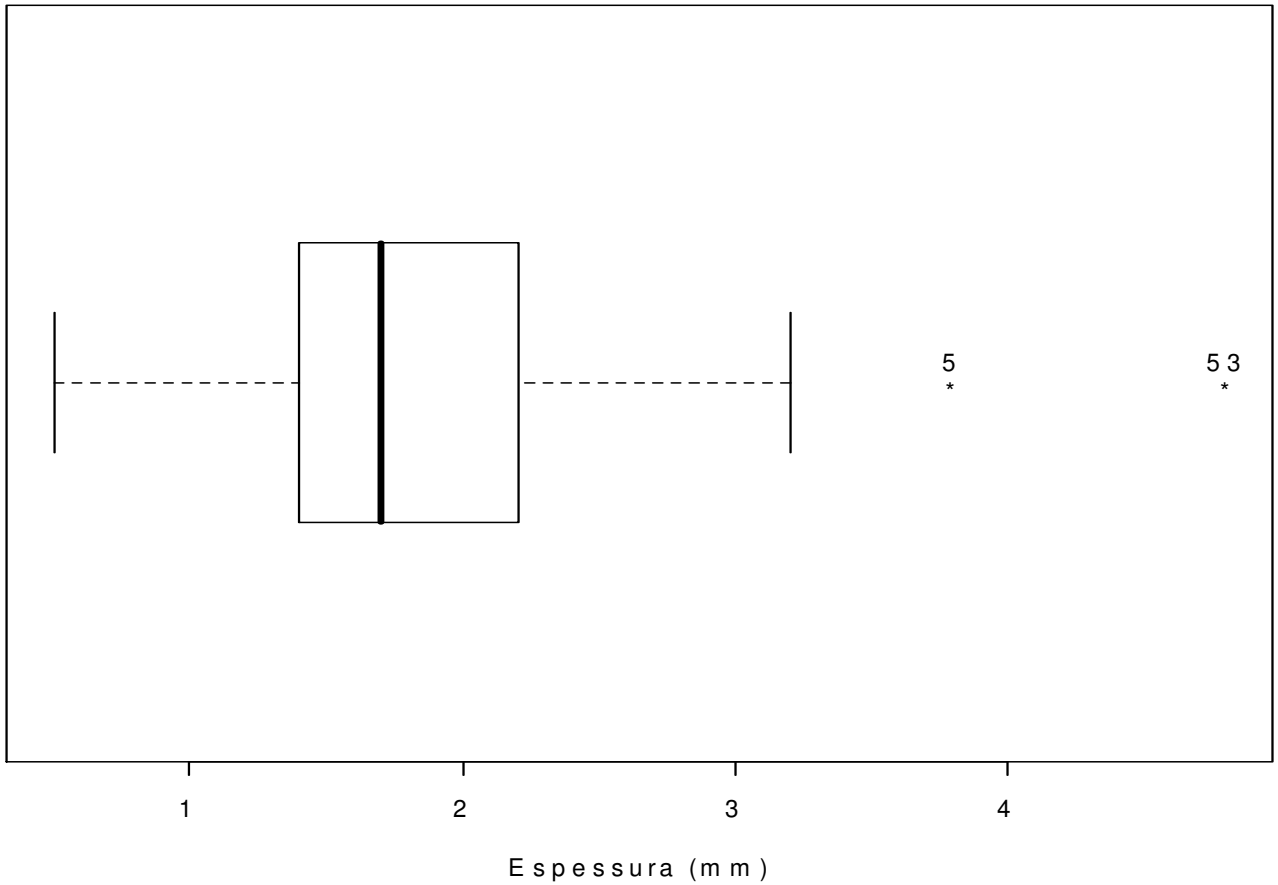
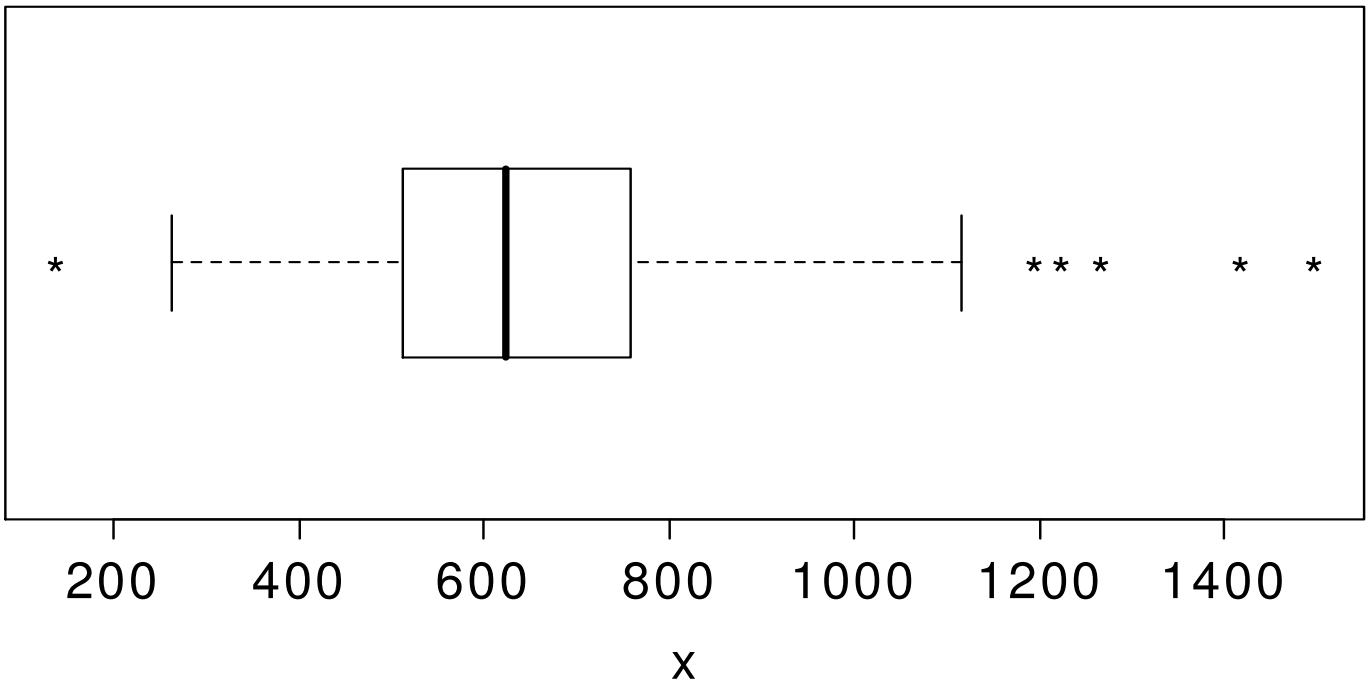


Gráfico de caixa

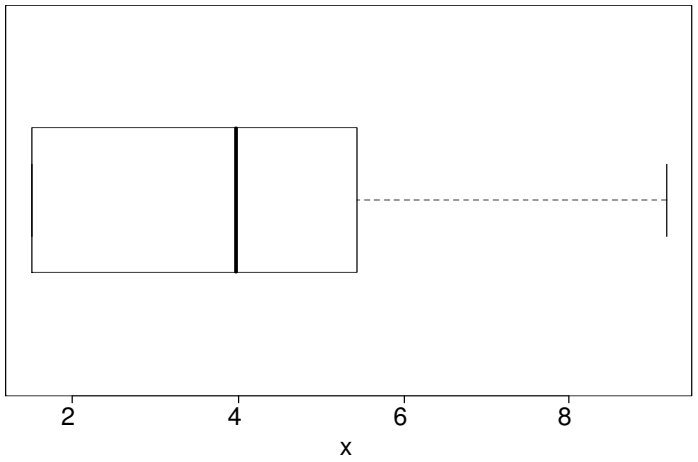
O que é possível observar em um gráfico de caixa?

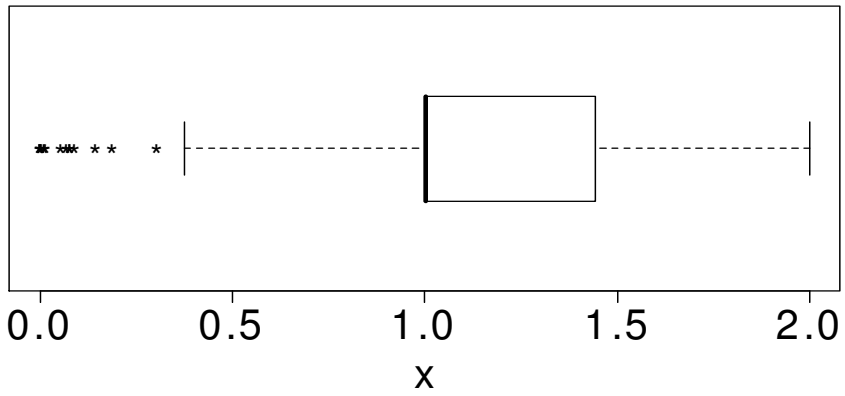


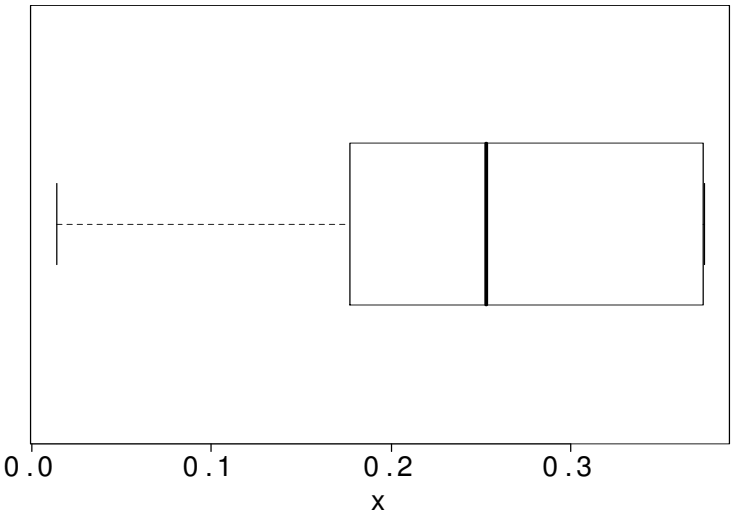
Medida de posição ($M = Q_2$). Medida de dispersão ($d_q = Q_3 - Q_1$).

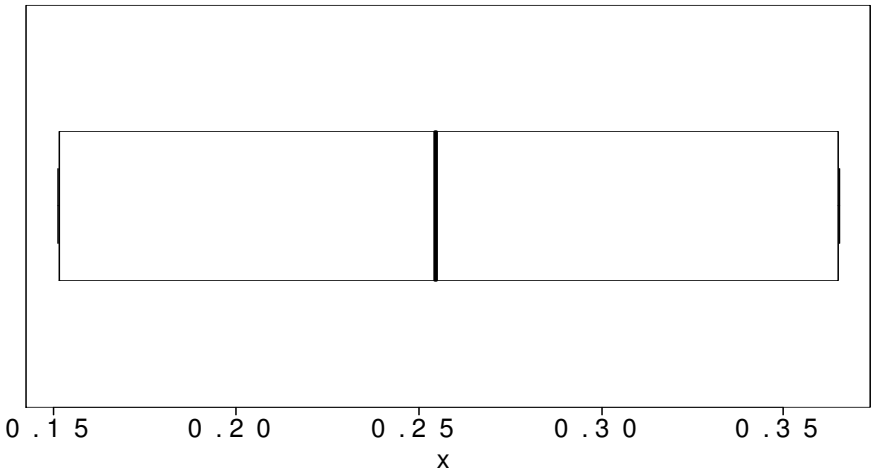
Simetria. Valores extremos.

Boxplot







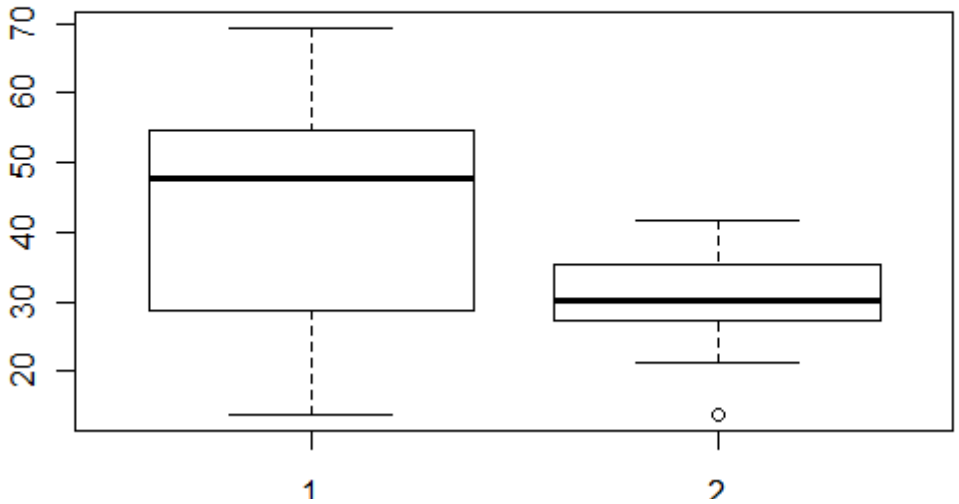


Exercício. Descreva conjuntos de dados correspondentes a cada um dos gráficos.

Exercício 3, Lista 2

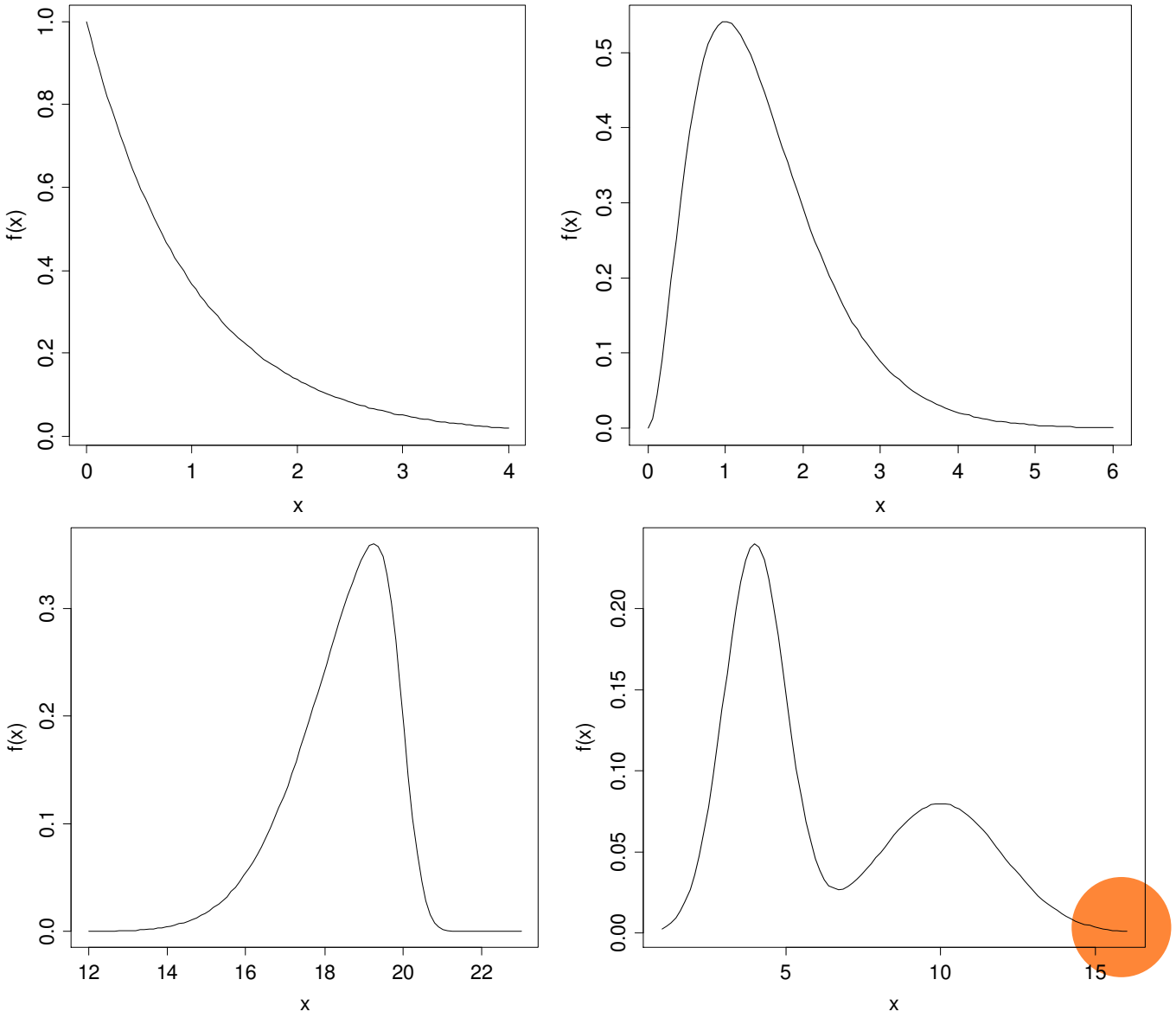
Um estudo dos efeitos do tabagismo nos padrõees de sono é conduzido. A medida observada é o tempo, em minutos, que se leva para dormir, em fumantes e não fumantes. Comente o tempo de impacto que o fumo aparenta ter no tempo que se leva para dormir.

```
> x1 <- c(69.3,56.0,22.1,47.6,53.2,48.1,52.7,34.4,60.2,43.8,
     23.2,13.8)
> x2 <-c(28.6,25.1,26.4,34.9,29.8,28.4,38.5,30.2,30.6,31.8,
    41.6,21.1,36.0,37.9,13.9)
> boxplot(x1,x2)
```



2. Assimetria e Curtose

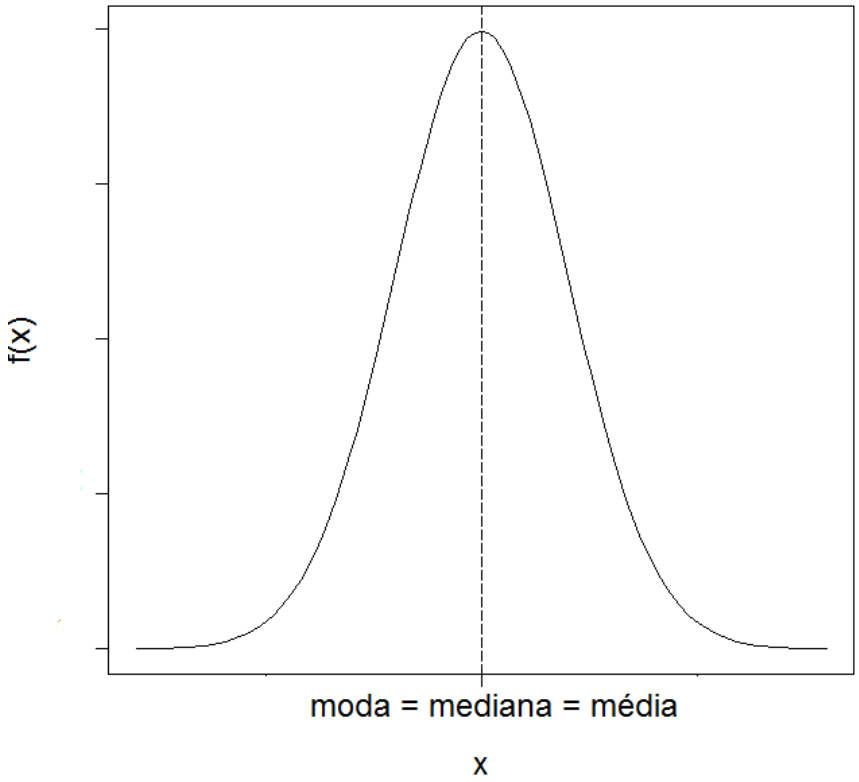
Distribuições assimétricas:

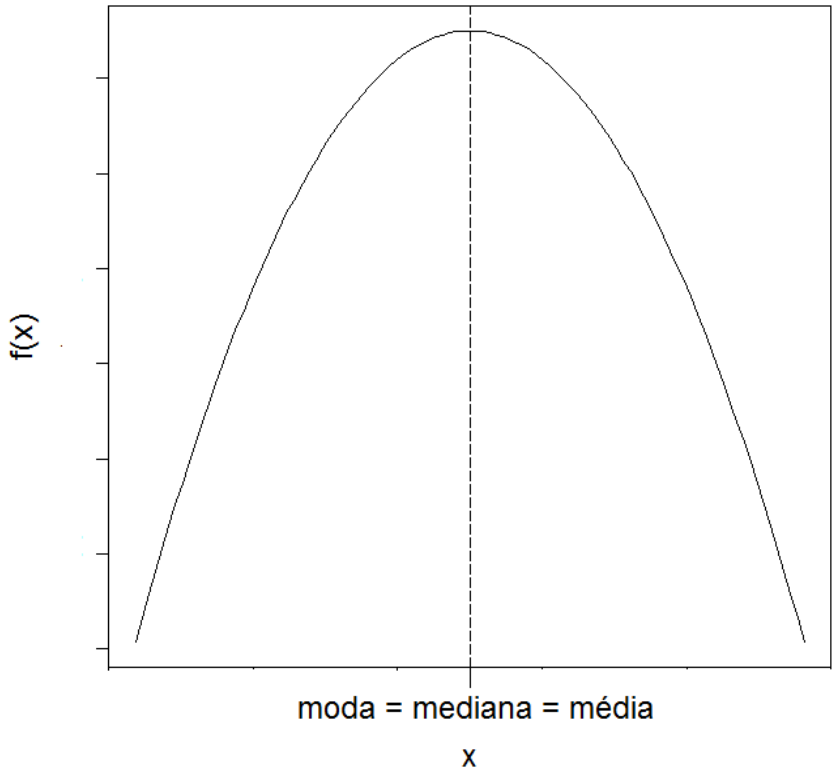


Relação entre moda, mediana e média

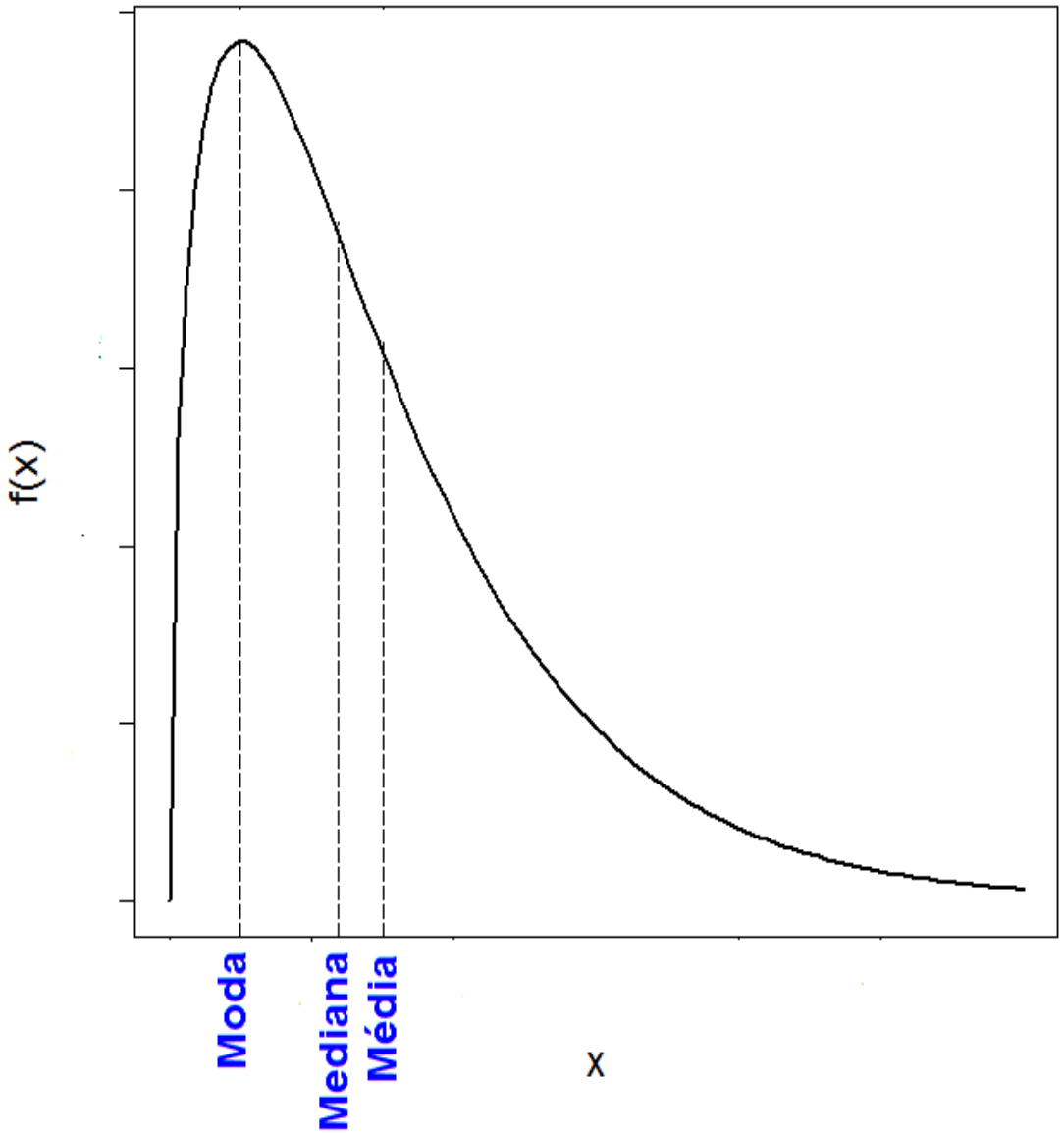
Supomos que a distribuição é unimodal e que a média existe.

Distribuição simétrica: moda = mediana = média.





Relação entre moda, mediana e média

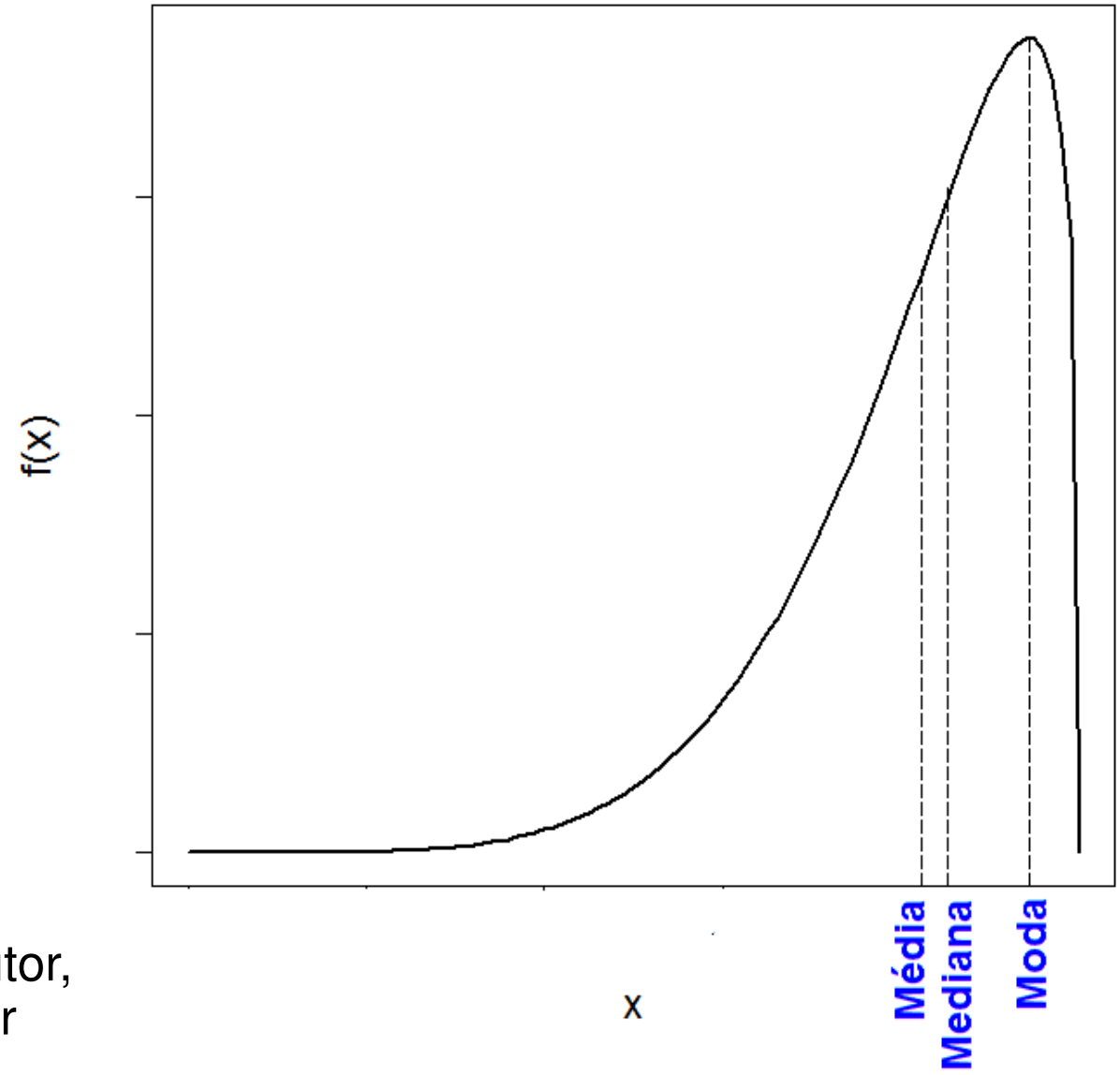


Distribuição assimétrica à direita ou assimétrica positiva (*right skewed* ou *positive skewed*): média > mediana > moda.

Cauda direita (*right tail*) é mais longa.

Relação entre moda, mediana e média

Distribuição assimétrica à esquerda ou assimétrica negativa (*left skewed* ou *negative skewed*): média < mediana < moda.

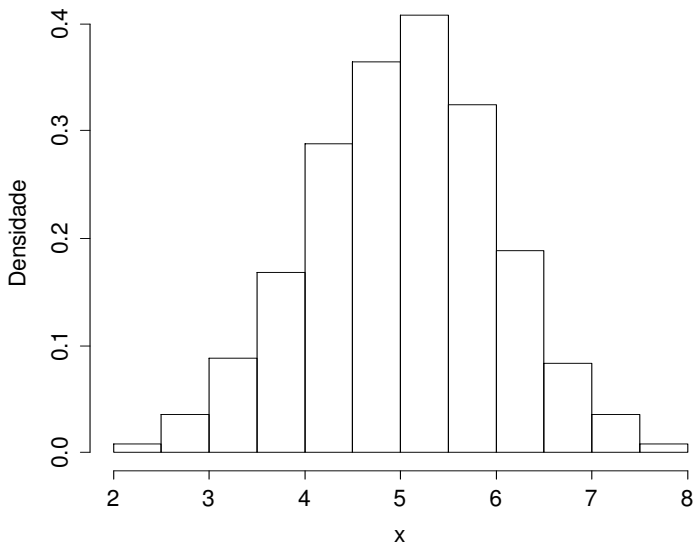


Cauda esquerda (*left tail*) é mais longa.

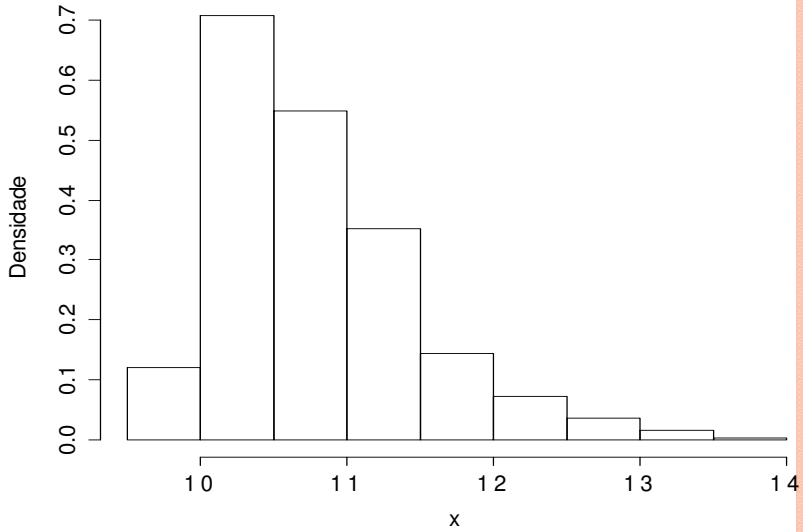
Obs. Dependendo do autor, há troca de esquerda por direita.

Assimetria em conjuntos de dados

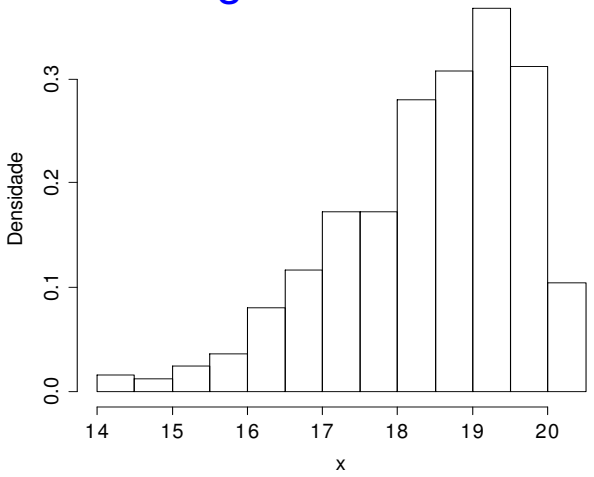
Aproximadamente simétrico

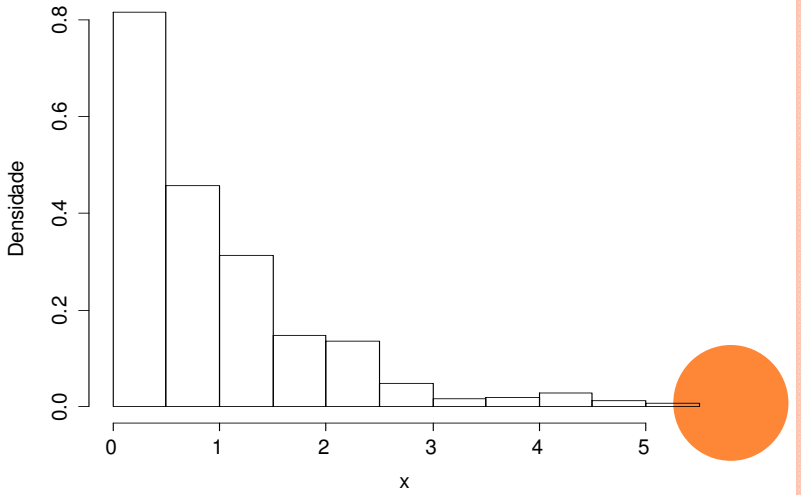


Assimetria positiva

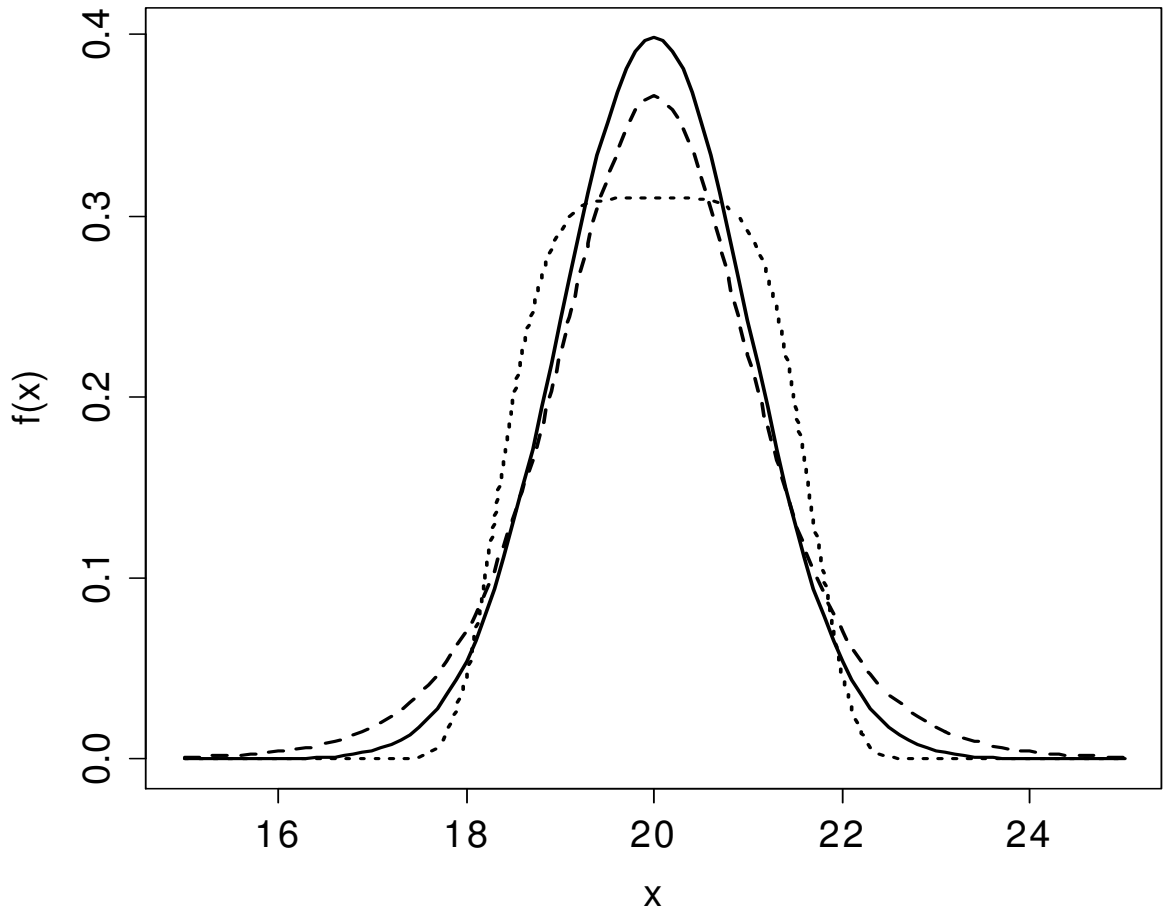


Assimetria negativa





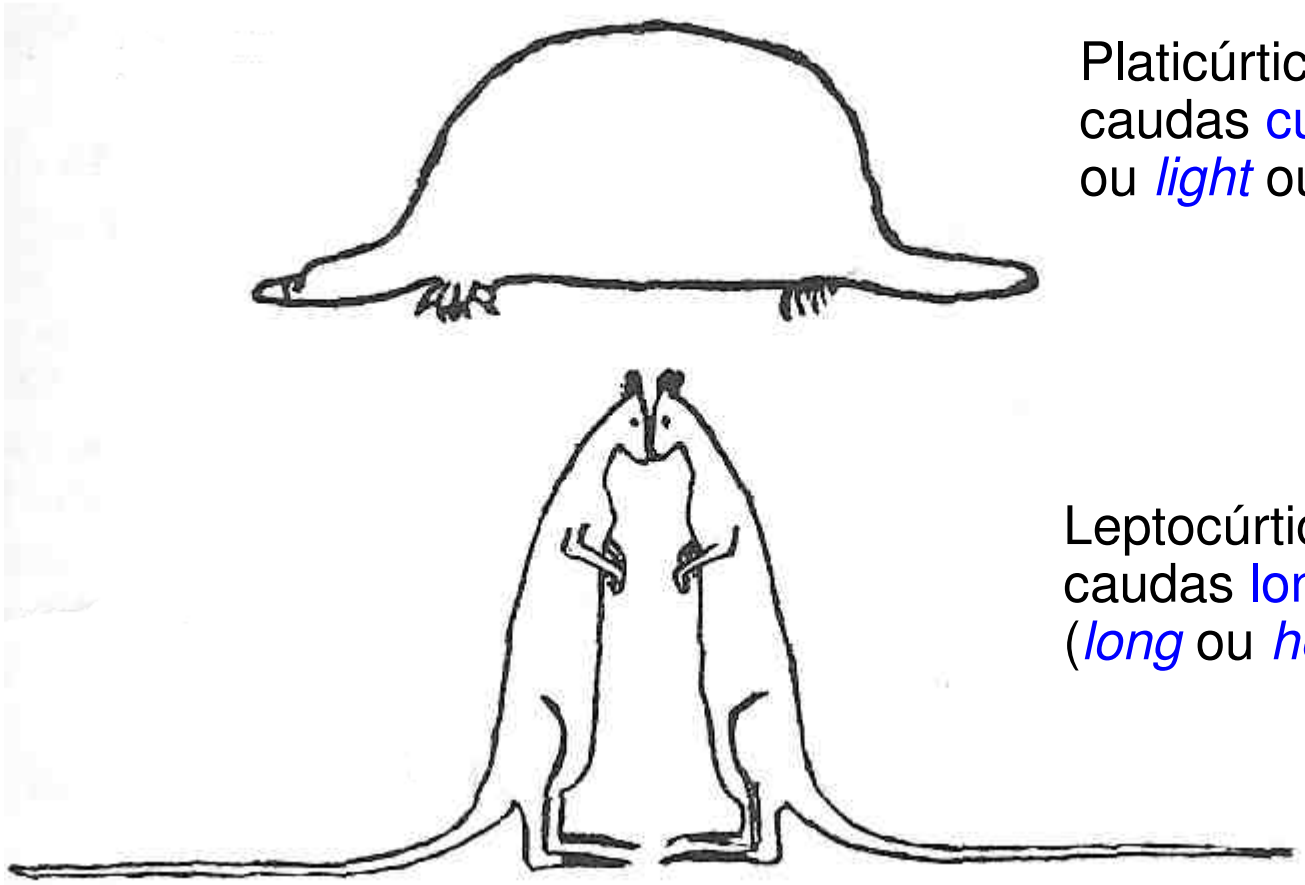
Distribuições simétricas com médias e variâncias iguais:



Diferenças quanto ao afastamento em relação à média, mas que não são captadas pela variância.

Este fato sugere utilizarmos o 4º momento central para quantificar estas diferenças.

Distribuições platicúrticas, mesocúrticas e leptocúrticas



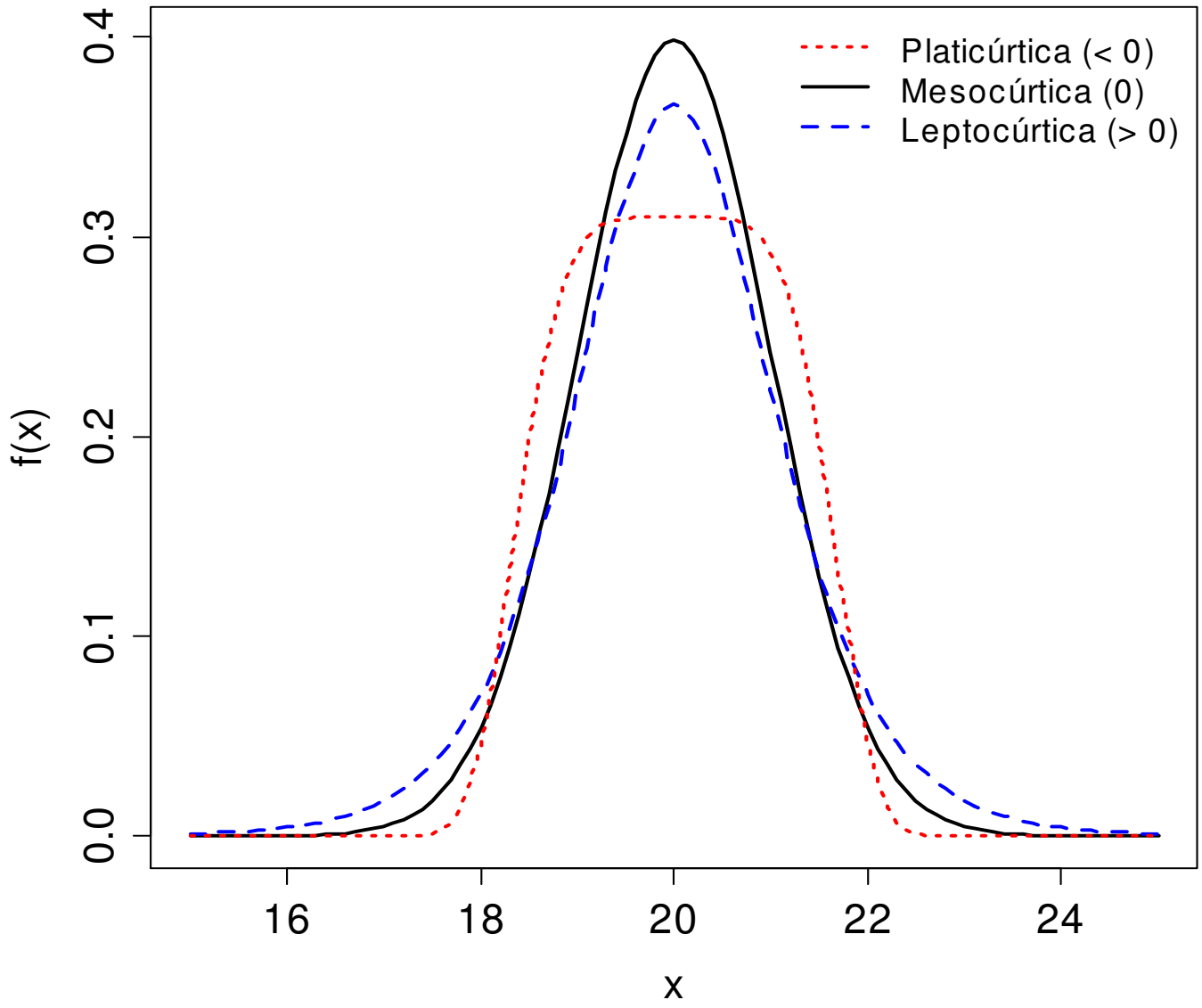
Platicúrtica (*platykurtic*): caudas curtas ou leves (*short* ou *light* ou *thin*).

Leptocúrtica (*leptokurtic*): caudas longas ou pesadas (*long* ou *heavy* ou *thick* ou *fat*).

Fonte. Bulmer, M. G. (1979), *Principles of Statistics*, Dover: New York.

Mesocúrtica (*mesokurtic*): caudas neutras (nem curtas e nem longas).

Distribuições simétricas com médias e variâncias iguais:



Momentos em R

Pacote `moments`

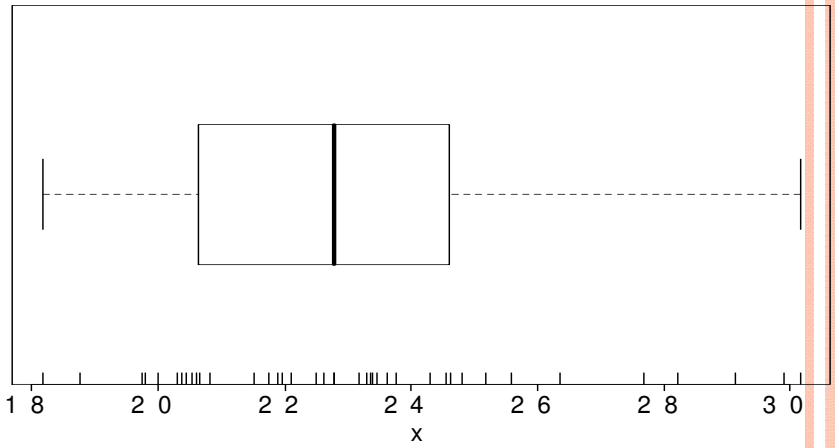
```
> library(moments)
```

40 observações

```
x = c(21.88, 22.61, 23.28, 25.19, 19.73, 22.50, 26.35, 22.09, 24.63,
19.79, 23.61, 29.91, 29.12, 23.17, 23.38, 21.75, 18.77, 20.38,
20.45, 28.23, 18.17, 30.15, 20.60, 19.99, 24.56, 25.59, 20.66,
23.76, 23.35, 22.77, 21.52, 20.54, 20.30, 23.45, 27.69, 24.82,
24.30, 21.97, 20.82, 22.78)
```

```
> boxplot(x, horizontal = TRUE,
xlab = "X")

> rug(x)
```



Momentos centrais até ordem 4:

```
> (mom4 = all.moments(x, order.max = 4, central = TRUE))
```

```
[1] 1.000000e+00 4.440892e-16 8.521595e+00 1.843078e+01 2.206282e+02
```

k = 0 1 2 3 4

Assimetria e curtose em R

Pacote `moments`

```
> skewness(x)
[1] 0.7409046
```

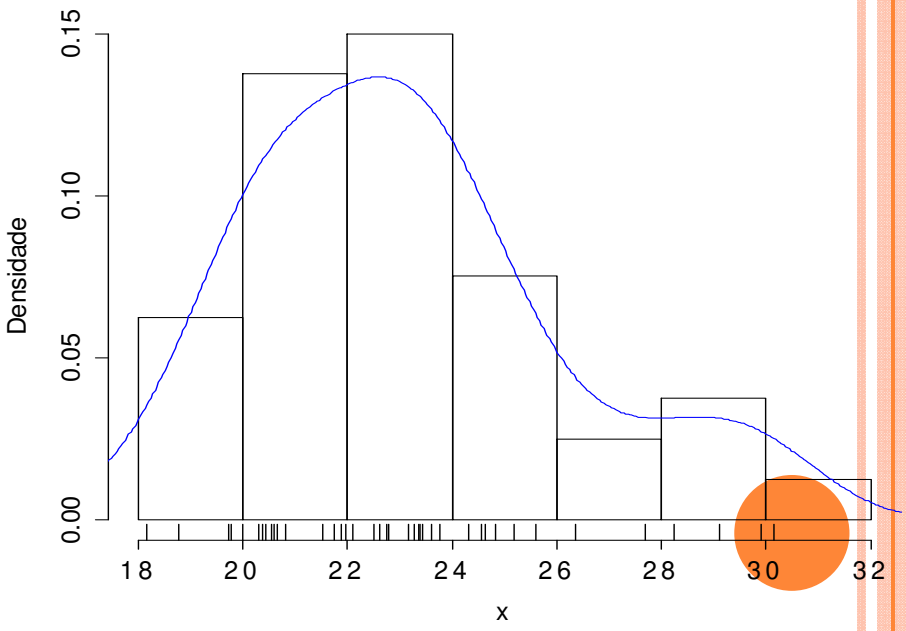
Curtose

```
> kurtosis(x)
[1] 3.03822
```

Qual o resultado de `mom4[5] / mom4[3]^2`?

```
> kurtosis(x) - 3
[1] 0.03822
```

```
> hist(x, main = "", xlab = "x", freq = FALSE, ylab = "Densidade")
> rug(x)
> lines(density(x), col = "blue")
```

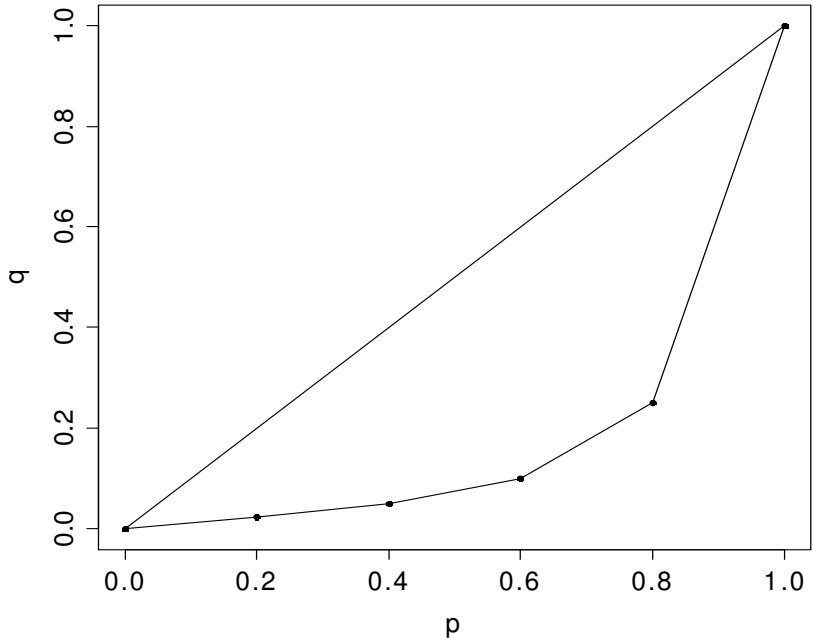


3. Concentração e Desigualdade

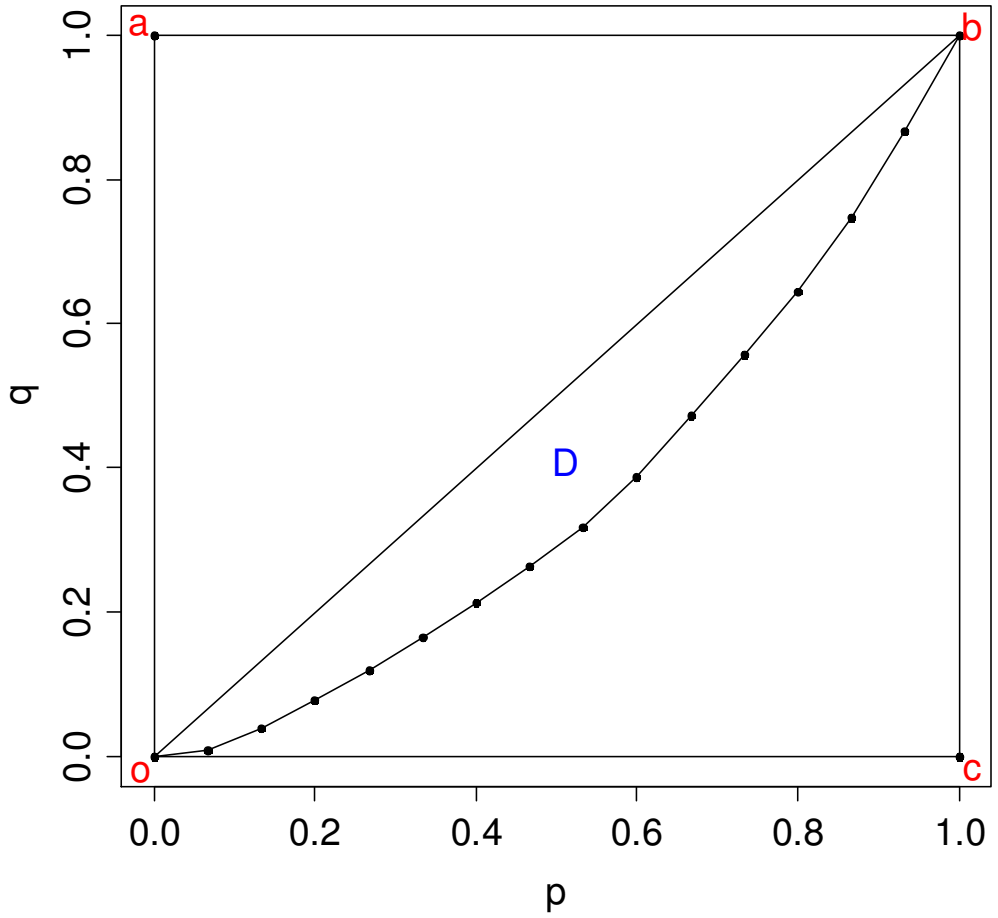
Exemplo

Dados ordenados: 1, 1, 2, 6, 30 ($n = 5$, $T = 40$ e média = T / n = 8).

i	$X_{(i)}$	p_i	q_i
1	1	1 / 5 = 0,2	1 / 40 = 0,025
2	1	2 / 5 = 0,4	(1 + 1) / 40 = 0,05
3	2	3 / 5 = 0,6	(1 + 1 + 2) / 40 = 0,1
4	6	4 / 5 = 0,8	(1 + 1 + 2 + 6) / 40 = 0,25
5	30	5 / 5 = 1	(1 + 1 + 2 + 6 + 30) / 40 = 1



Área de desigualdade



Área compreendida entre ob e a curva de Lorenz: área de desigualdade (D).

(a) $x_{(1)} = x_{(2)} = ... = x_{(n)} = T / n$: proporções de posições = proporções acumuladas de valores ($q_i = p_i$, i = 1,...,n).

$\Rightarrow$ curva de Lorenz = segmento ob (linha da igualdade perfeita).

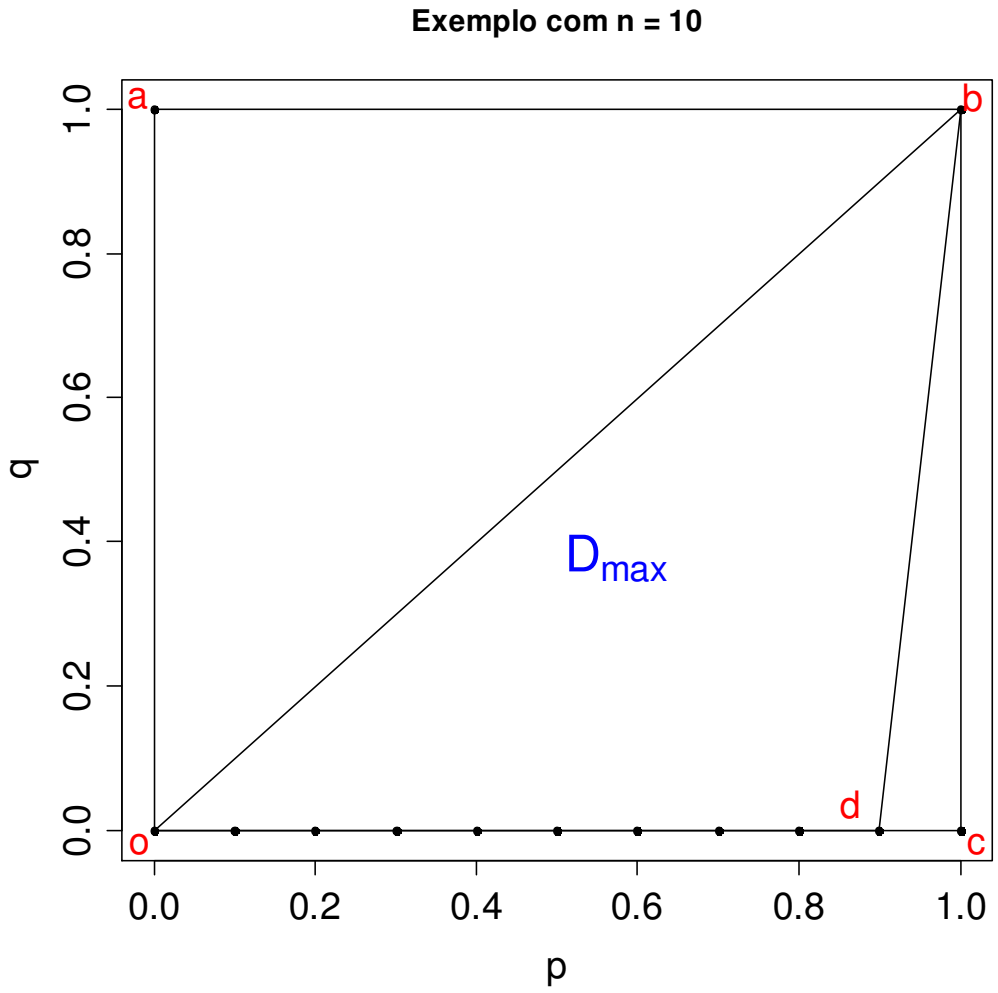
(b) $x_{(1)} = x_{(2)} = ... = x_{(n-1)} = 0$ e $x_{(n)} = T$:

curva de Lorenz é formada pelos pontos (0, 0), (1 – 1 / n, 0) e (1, 1): curva da desigualdade perfeita.

Quando $n \rightarrow \infty$: curva da desigualdade perfeita coincide com ocb.

Quanto mais a curva de Lorenz estiver afastada de ob, maior o grau de desigualdade.

7.2. Índice de Gini



Curva da desigualdade perfeita: odb.

Como a área do triângulo obc = ½, temos que $0 \leq D < ½$.

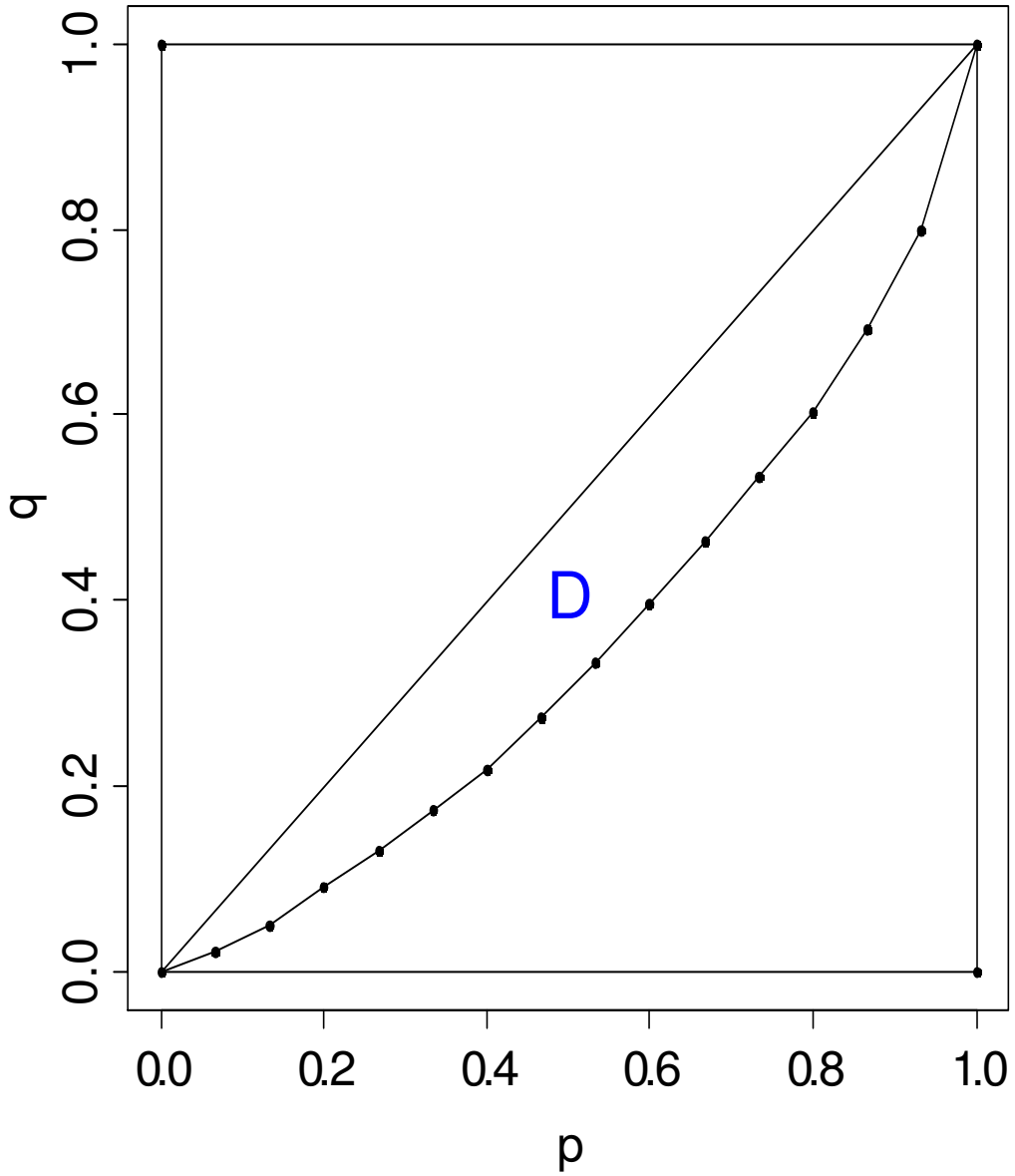
Valor máximo de D (desigualdade perfeita):

$$D_{\max} = \frac{1}{2}\left(1 - \frac{1}{n}\right).$$

$D_{max} \to ½$ quando $n \to \infty$ ($d \to c$).

$\max D_{max} = ½$.

7.2. Índice de Gini



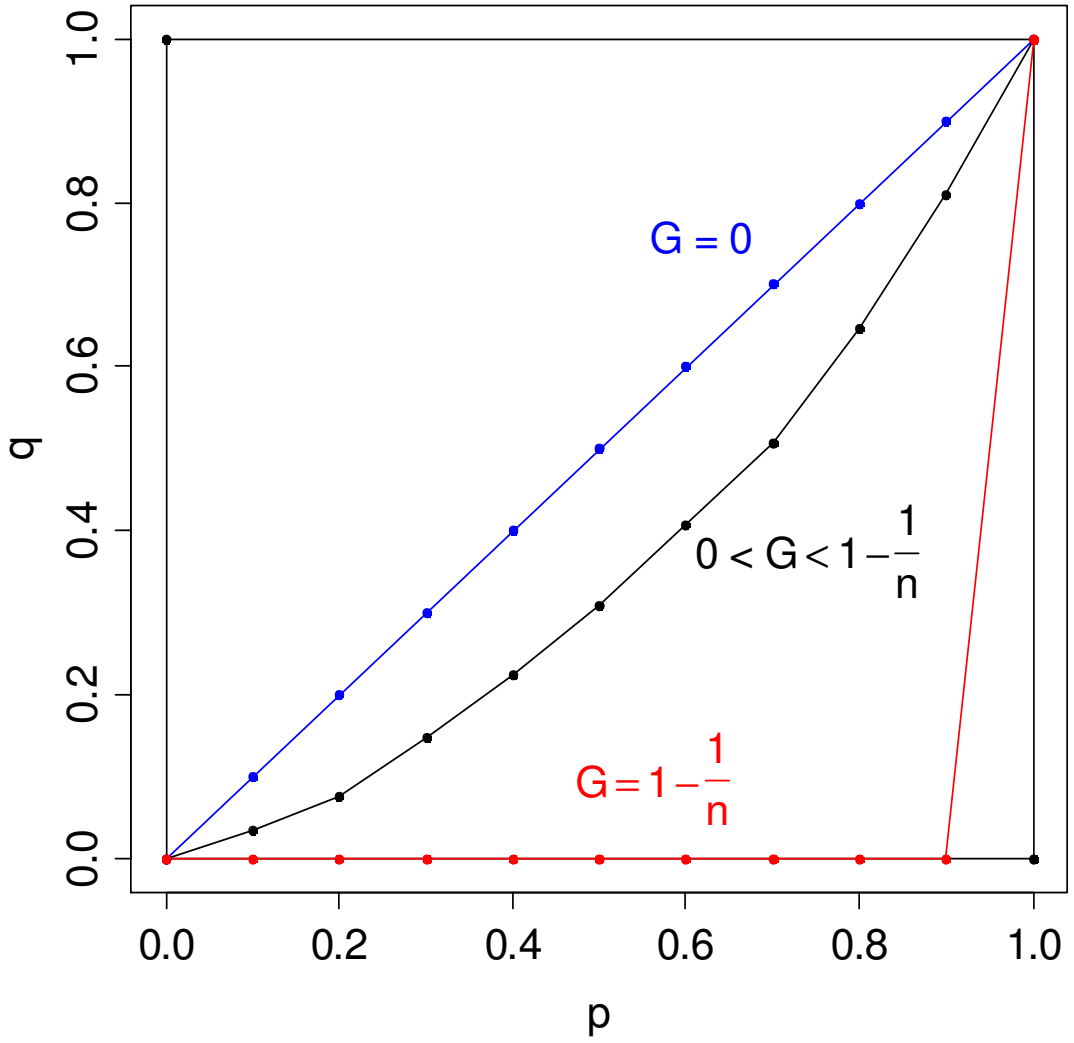
Proposto por C. Gini em 1914.

$G = D / \max D_{max} = D / ½ = 2\,D$.

Propriedades. (a) $0 \leq G < 1$ e (b) $0 \leq G \leq 1 - 1/n$.

Igualdade perfeita: $G = 0$.
Desigualdade perfeita:
$G = 1 - 1 / n$
($\to 1$ quando $n \to \infty$).

7.2. Índice de Gini



Valores ordenados:
$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Como calcular G?

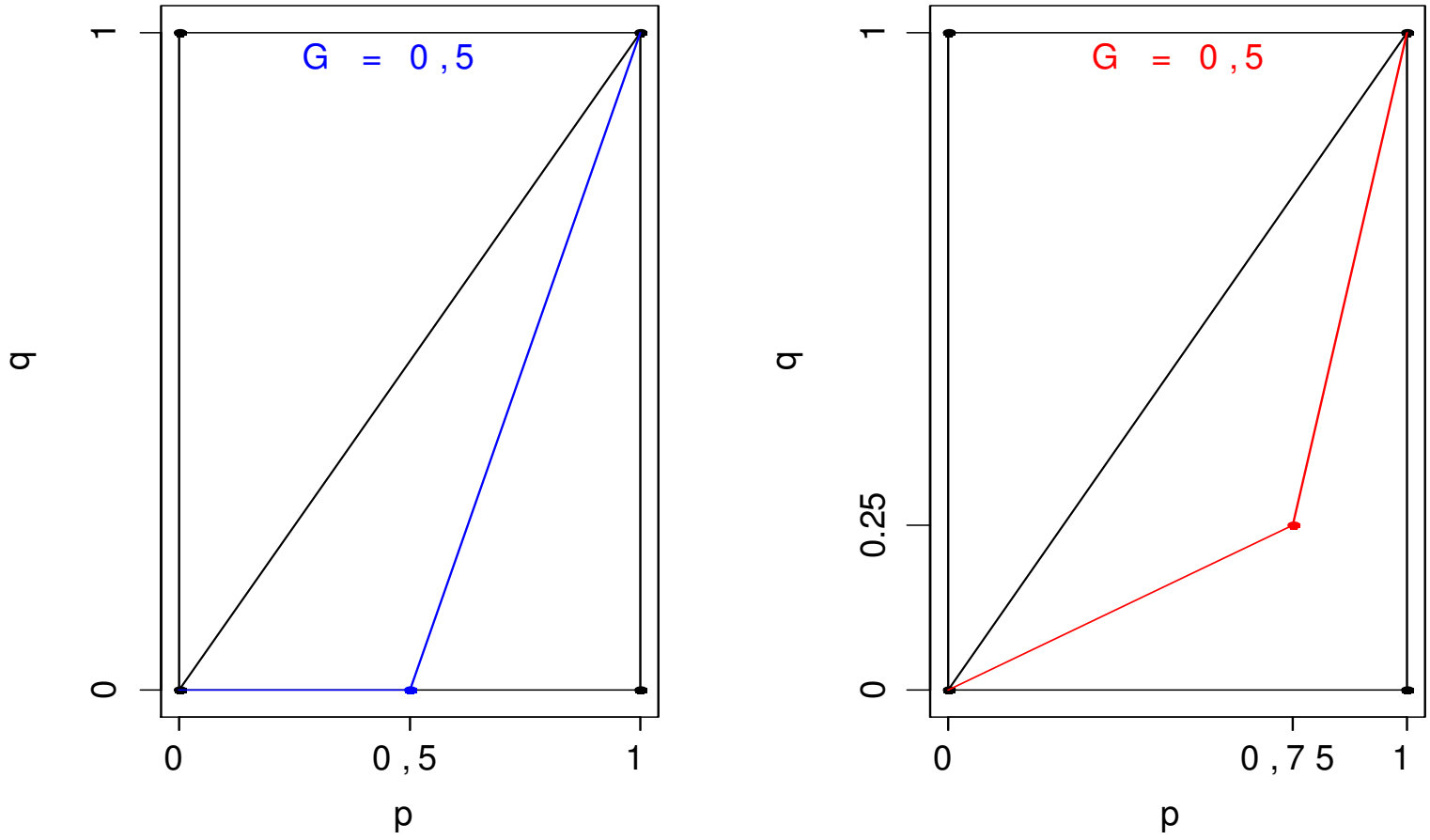
$$G = 1 - \frac{1}{n} \sum_{i=1}^{n} (q_i + q_{i-1}),$$

sendo que $q_0 = 0$ e

$$q_i = \frac{1}{T} \sum_{j=1}^{i} x_{(j)}.$$

7.2. Índice de Gini

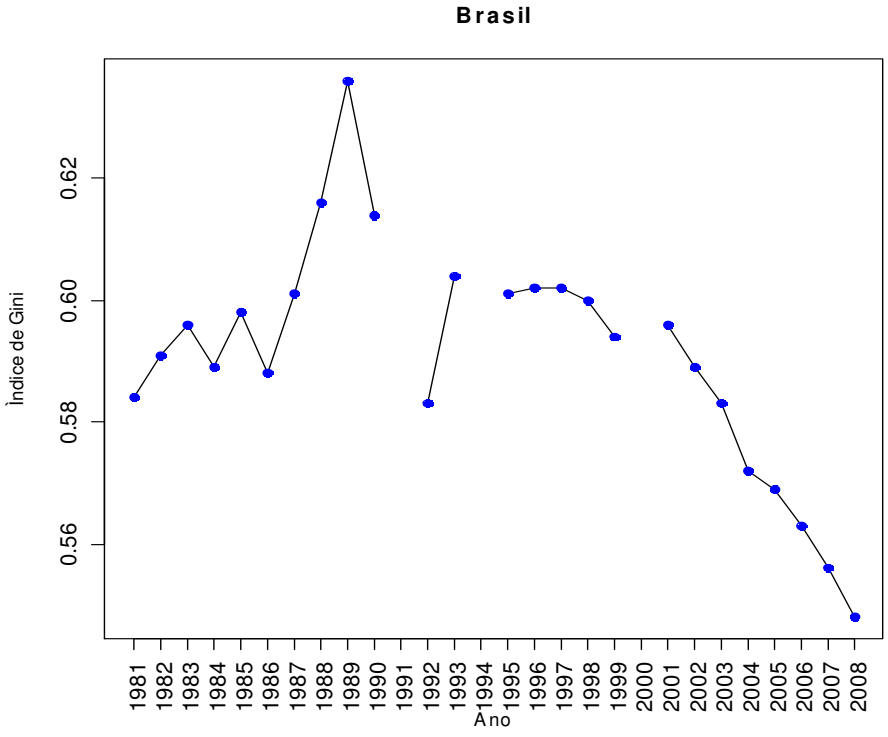
Obs. (a) Diferentes curvas de Lorenz podem gerar o mesmo valor de G.



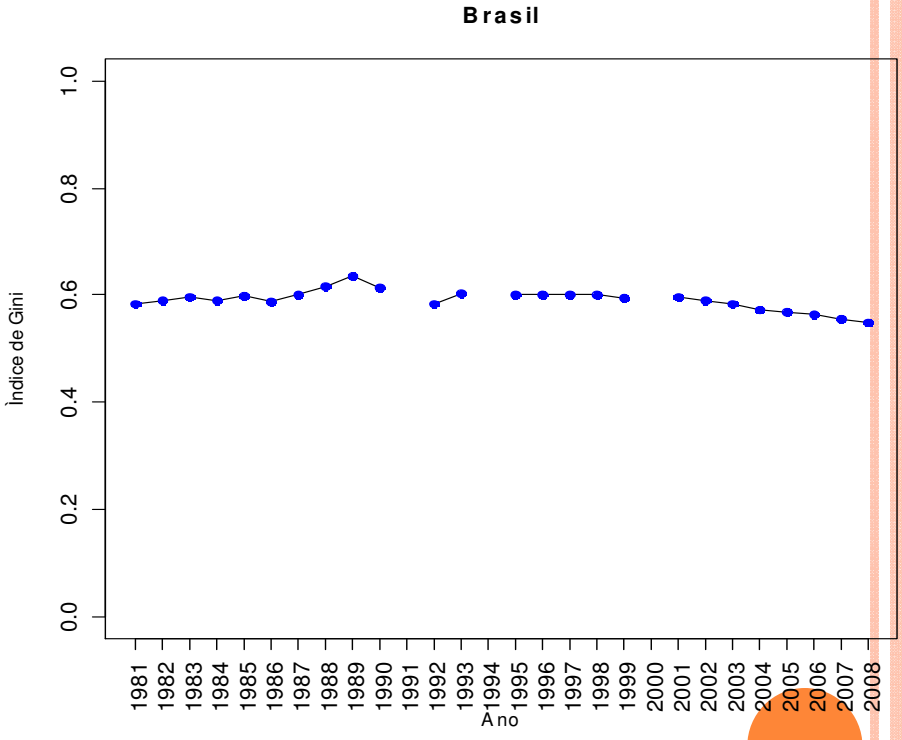
(b) G mede apenas desigualdade. Por exemplo, diferentes países podem ter valores de G semelhantes e diferentes níveis de riqueza.

7.2. Índice de Gini

Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar *per capita*. Seu valor varia de 0, quando não há desigualdade (a renda de todos os indivíduos tem o mesmo valor), a 1, quando a desigualdade é máxima (apenas um um indivíduo detém toda a renda da sociedade e a renda de todos os outros indivíduos é nula). Fonte: http://www.pnud.org.br/popup/pop.php?id_pop=97.

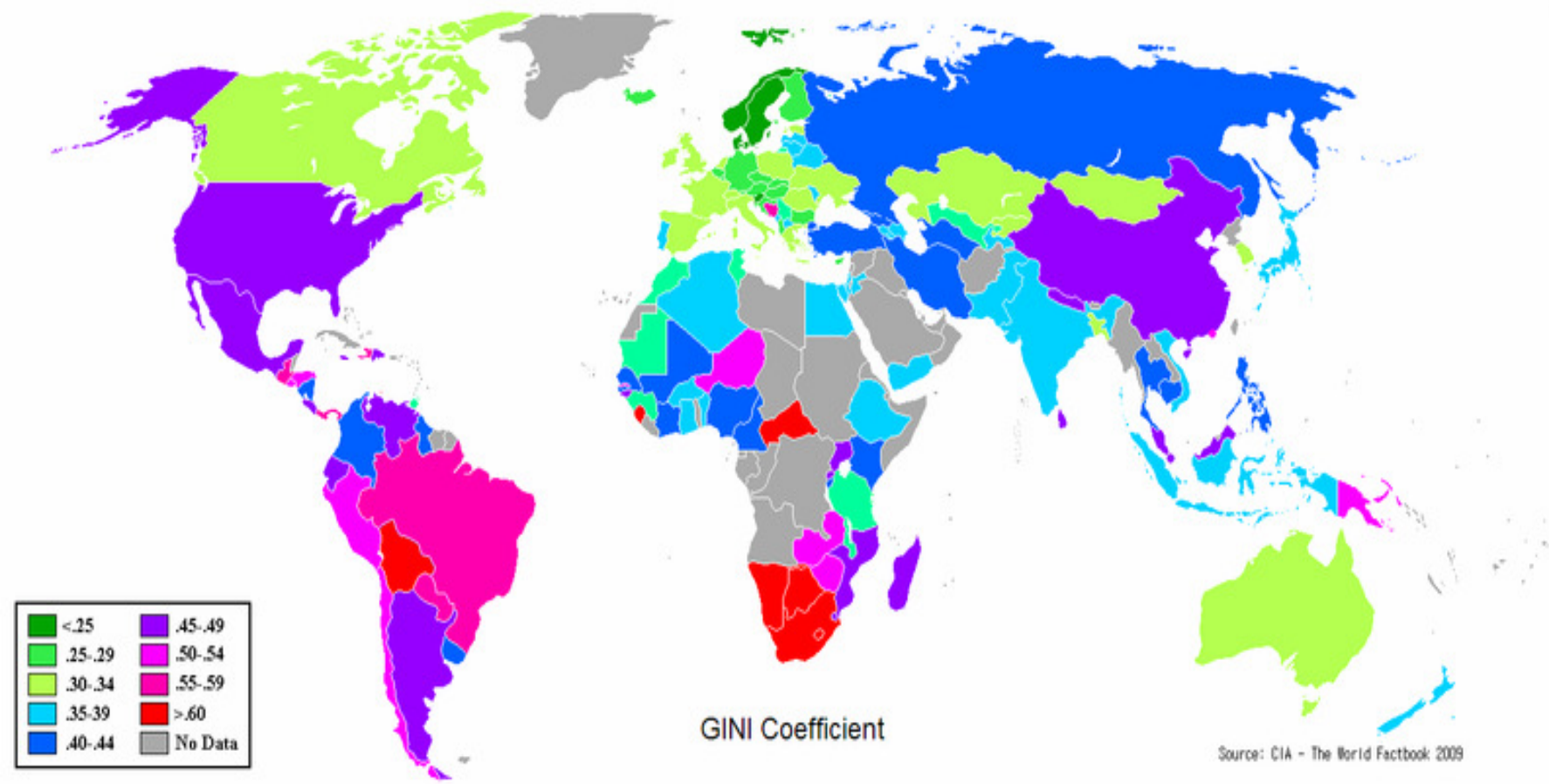


Redução importante nos últimos anos.



Pouca variação.

7.2. Índice de Gini

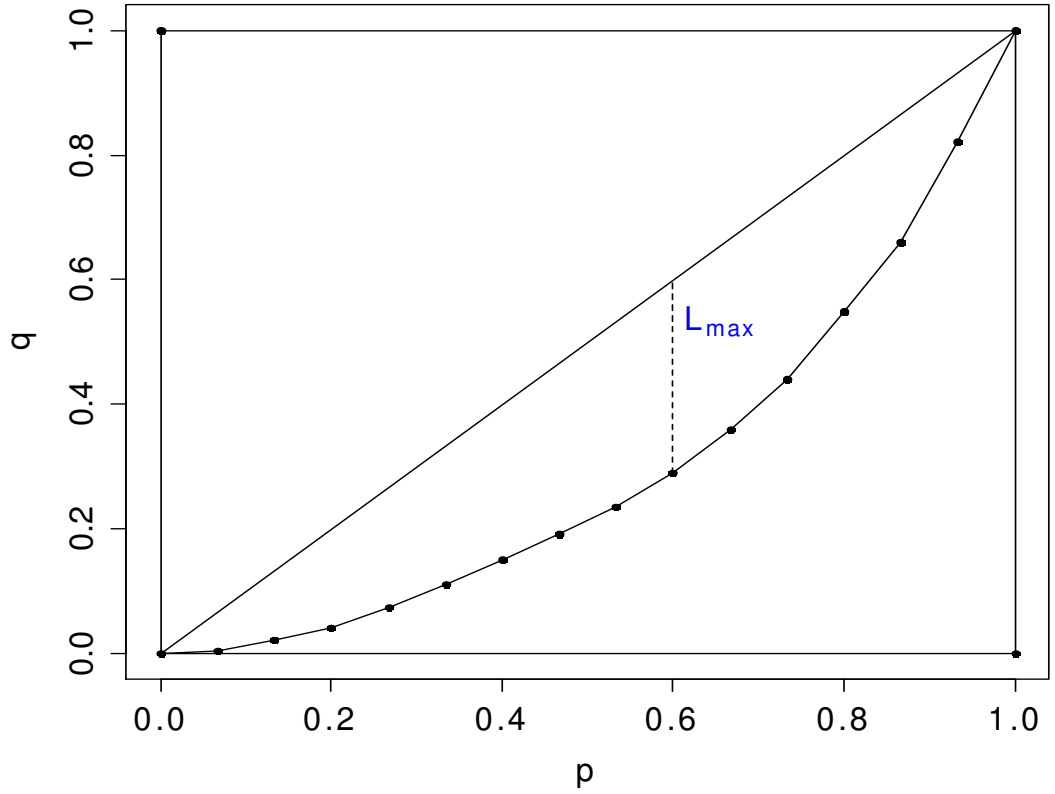


Gini Coefficient World CIA Report 2009.

Obs. Exemplo de um cartograma.

7.3. Discrepância máxima

Medida associada à curva de Lorenz. Valor máximo da diferença entre a proporção acumulada de posições e a proporção acumulada de valores: L_{max} = max $(p_i - q_i)$, i = 1,...,n.



Declividade da curva:

$$B_i = \frac{q_i - q_{i-1}}{p_i - p_{i-1}} = \frac{x_{(i)}}{\bar{x}}, \quad i = 1,...,n.$$

$$x_{(i)} \le \bar{x} \Rightarrow B_i \le 1.$$

$$x_{(i)} > \bar{x} \Rightarrow B_i > 1.$$

Encontrar j tal que $x_{(1)} \le x_{(2)} \le ... \le x_{(j)} \le \bar{x} < x_{(j+1)} \le ... \le x_{(n)}$. $L_{max} = p_j - q_j$.

Pode ser provado que $L_{max} = \frac{dm}{2\bar{x}}$. L_{max} é uma medida de dispersão relativa.

Medidas de desigualdade em R

Pacote `ineq`

```
> library(ineq)
```

15 observações

```
> x = c(2.8, 13.7, 6.8, 12.1, 1.1, 5.9, 4.5, 9.6, 2.3, 28.9, 6.7, 0.4, 5.6, 8.0, 10.3)
```

```
> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.400   3.650   6.700   7.913   9.950  28.900
```

Curva de Lorenz: função `Lc`.

```
> clorenz = Lc(x)
```

Índice de Gini:

```
> Gini(x)
```

```
[1] 0.4213423
```

```
> names(clorenz)
```

```
[1] "p"         "L"         "L.general"
```

= p = q

```
> (jmax = which.max(clorenz$p - clorenz$L))
```

```
[1] 10
```

```
> (Lmax = clorenz$p[jmax] - clorenz$L[jmax])
```

```
[1] 0.2958719
```

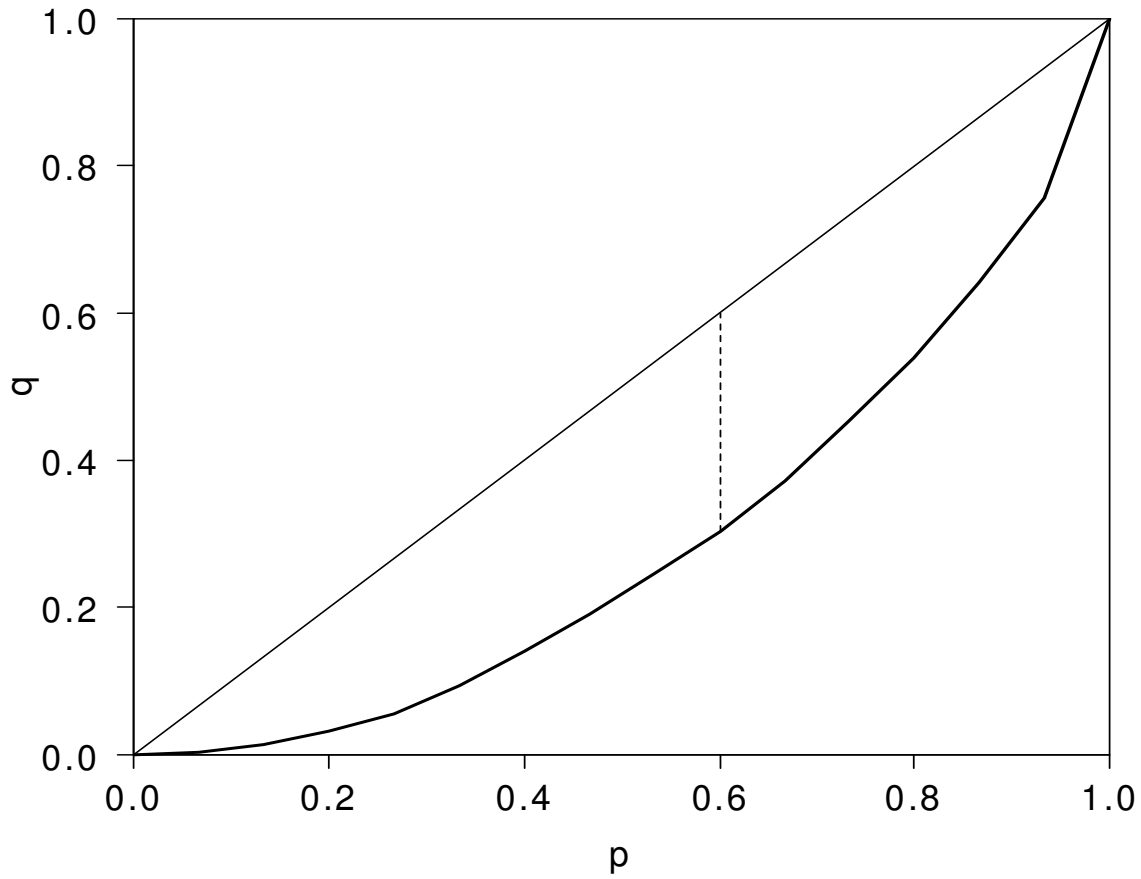
```
> c(clorenz$p[jmax], clorenz$L[jmax])
```

```
[1] 0.6000000 0.3041281
```

Medidas de desigualdade em R

Curva de Lorenz e discrepância máxima (L_{max}):

```
> plot(clorenz, main = "", ylab = "q")
> segments(clorenz$p[jmax], clorenz$L[jmax],  clorenz$p[jmax],
clorenz$p[jmax],lty = 2)
```



Exercício

Analise os seguintes conjuntos de dados considerando medidas de posição, dispersão, assimetria, curtose, concentração e desigualdade.

Temperatura em cidades portuguesas

```
> dados<- read.table("http://wiki.icmc.usp.br/images/b/bd/Temperatura.txt",
   header=TRUE)
> attach(dados)
```

Distâncias ortodônticas em crianças

```
> dados <- read.table("http://wiki.icmc.usp.br/images/1/1a/Ortodontia.txt", header=TRUE)
> attach(dados)
```

Salários na CompanhiaMB

```
> dados <- read.table(
   "http://wiki.icmc.usp.br/images/f/f4/CompanhiaMB.txt",header=TRUE)
> attach(dados)
```