SME0803 Visualização e Exploração de Dados

Medidas de concentração e desigualdade

Prof. Cibele Russo

cibele@icmc.usp.br

Baseado em

Murteira, B. J. F., Análise Exploratória de Dados. McGraw-Hill, Lisboa, 1993.

Notas de aula de Análise Exploratória de Dados. Mário de Castro, ICMC-USP, 2010.

Como medir a desigualdade de renda?

Bairros de Paraisópolis e Morumbi, em São Paulo SP.

Fonte: `https://portal.fgv.br/noticias/desigualdade-renda-brasil-bate-recorde-aponta-levantar`

Medidas de concentração e desigualdade

A **concentração dos dados** está relacionada à **variabilidade** ou **dispersão** dos valores observados de uma determinada variável.

Medidas de concentração e desigualdade

A **concentração dos dados** está relacionada à **variabilidade** ou **dispersão** dos valores observados de uma determinada variável.

Sejam $x_1, ..., x_n$ os valores de uma variável na amostra. Estamos interessados em saber se $T = x_1 + ... + x_n$ tem maior contribuição de poucas observações ou se todas as observações contribuem de forma similar na soma T.

Medidas de concentração e desigualdade

Exemplo 1

Variável: renda de pessoas em uma amostra.

Valores: $x_1, ..., x_n$. Renda total: $T = x_1 + ... + x_n$.

Considere duas situações

a. A renda total pode estar igualmente repartida entre as n pessoas, cada uma com renda: $T/n\ (= \bar{x})$.

b. A renda total pode ser de uma única pessoa: $x_1 = T, x_2 = x_3 = ... = x_n = 0$.

Medidas de concentração e desigualdade

Exemplo 1

Variável: renda de pessoas em uma amostra.

Valores: $x_1, ..., x_n$. Renda total: $T = x_1 + ... + x_n$.

Considere duas situações

a. A renda total pode estar igualmente repartida entre as n pessoas, cada uma com renda: $T/n\ (= \bar{x})$.

b. A renda total pode ser de uma única pessoa: $x_1 = T, x_2 = x_3 = ... = x_n = 0$.

Estas duas situações são extremas.

Em (a), temos a mínima concentração de renda.

Em (b): temos a concentração máxima de renda.

Medidas de concentração e desigualdade

É mais comum encontrarmos situações intermediárias.

Exemplo 2. Variável: altura de pessoas.

Valores: $x_1, ..., x_n$.

Altura total: $T = x_1 + ... + x_n$.

Nesse caso,
$x_1 = T, x_2 = x_3 = ... = x_n = 0$ não faz sentido.

A curva de Lorenz

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

Total: $T = x_{(1)} + x_{(2)} + ... + x_{(n)}$.

Proporção acumulada de posições até a i-ésima posição ($p_0 = 0$) :
$p_i = i/n$:
$p_1 = 1/n$, $p_2 = 2/n$, ..., $p_{n-1} = (n-1)/n = 1 - 1/n$, $p_n = n/n = 1$.

Proporção acumulada de valores até a i-ésima posição ($q_0 = 0$) :
$q_i = (x_{(1)} + x_{(2)} + ... + x_{(i)})/T$.
$(q_n = T/T = 1)$.

A curva de Lorenz

Obs. Se $x_i \geq 0$, então $p_i \geq q_i$, $i = 1, ..., n$.

O gráfico formado pela união dos pontos
$(0,0), (p_1, q_1), (p_2, q_2), ..., (p_n, q_n)$ é chamado de **curva de Lorenz** ($p_n = q_n = 1$).

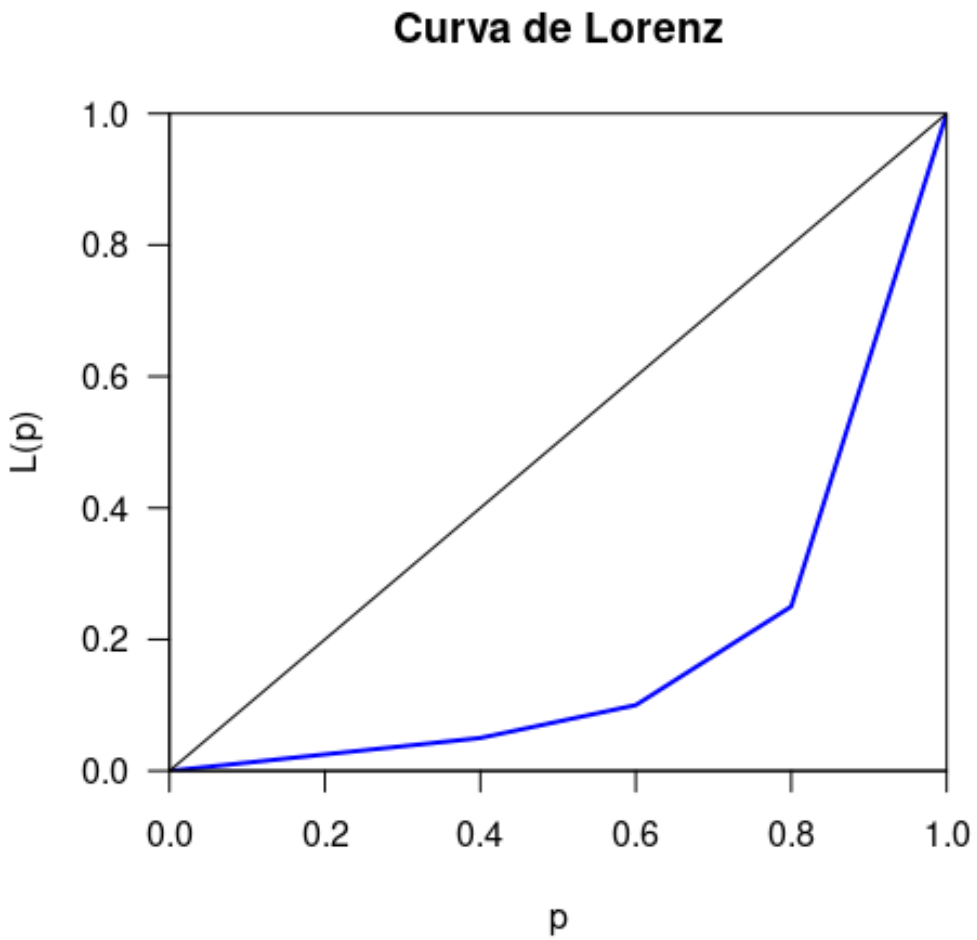
O segmento de reta unindo $(0,0)$ e $(1,1)$ também é incluído.

Exemplo

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

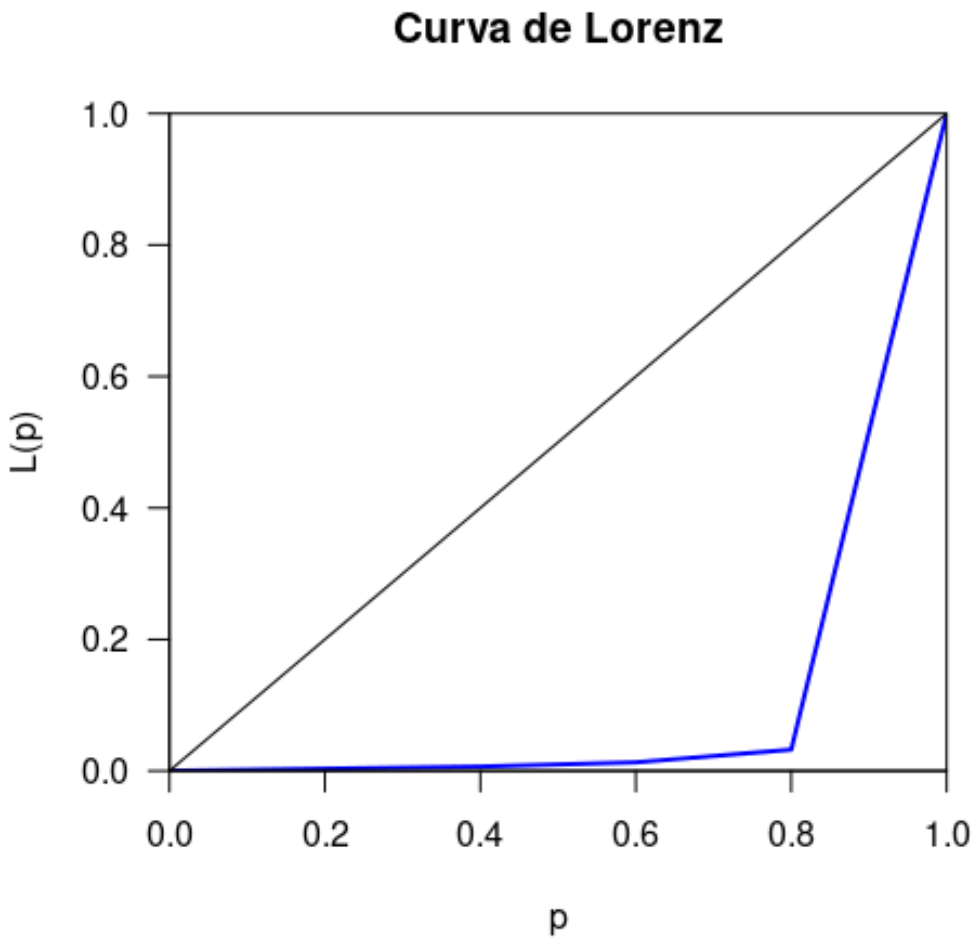
i	$X_{(i)}$	p_i	q_i
1	1	1/5 = 0,2	1/40 = 0,025
2	1	2/5 = 0,4	(1+1)/40 = 0,05
3	2	3/5 = 0,6	(1+1+2)/40 = 0,1
4	6	4/5 = 0,8	(1+1+2+6)/40 = 0,25
5	30	5/5 = 1	(1+1+2+6+30)/40 = 1

Exemplo - Curva de Lorenz



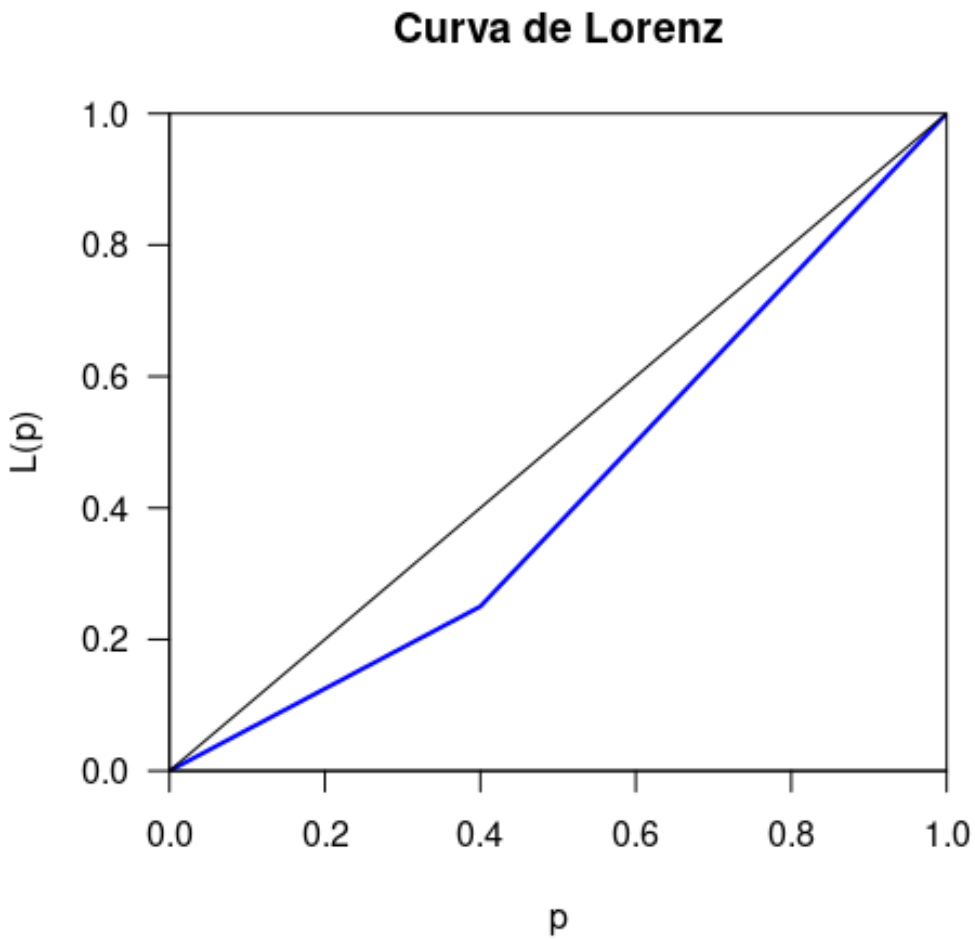
Curva de Lorenz para os dados 1, 1, 2, 6, 30

Exemplo - Curva de Lorenz



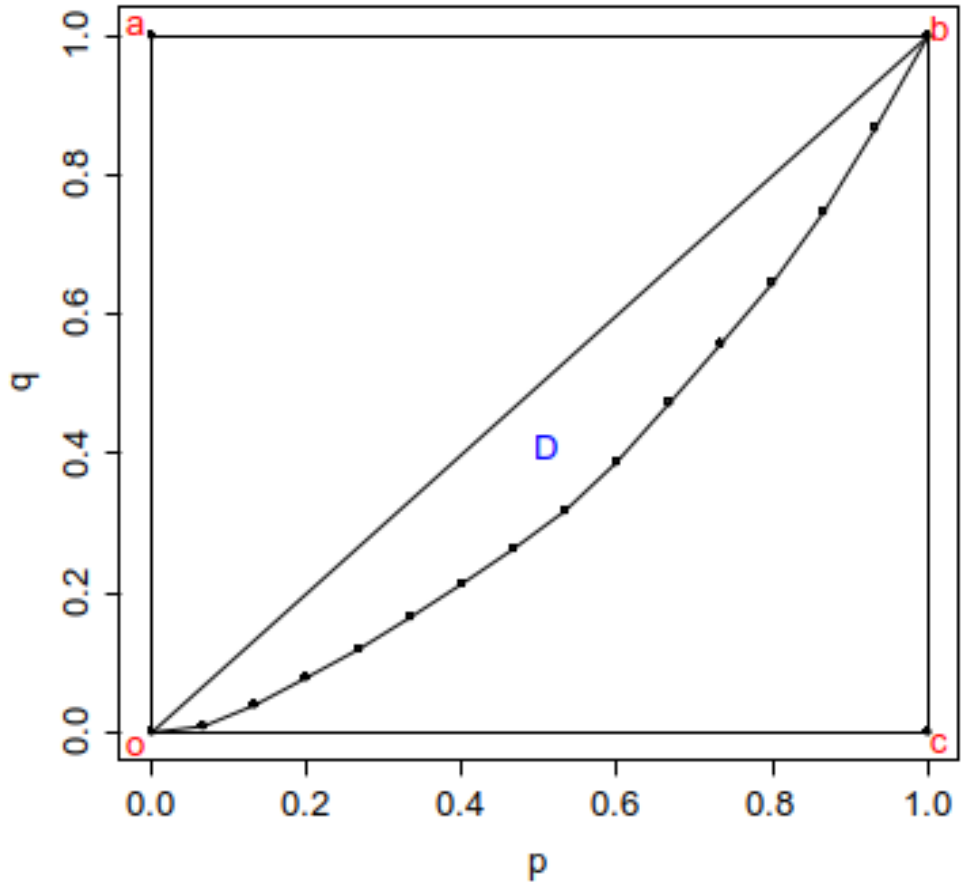
Curva de Lorenz para os dados 1, 1, 2, 6, 300

Exemplo - Curva de Lorenz



Curva de Lorenz para os dados 1, 1, 2, 2, 2

Área de desigualdade



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

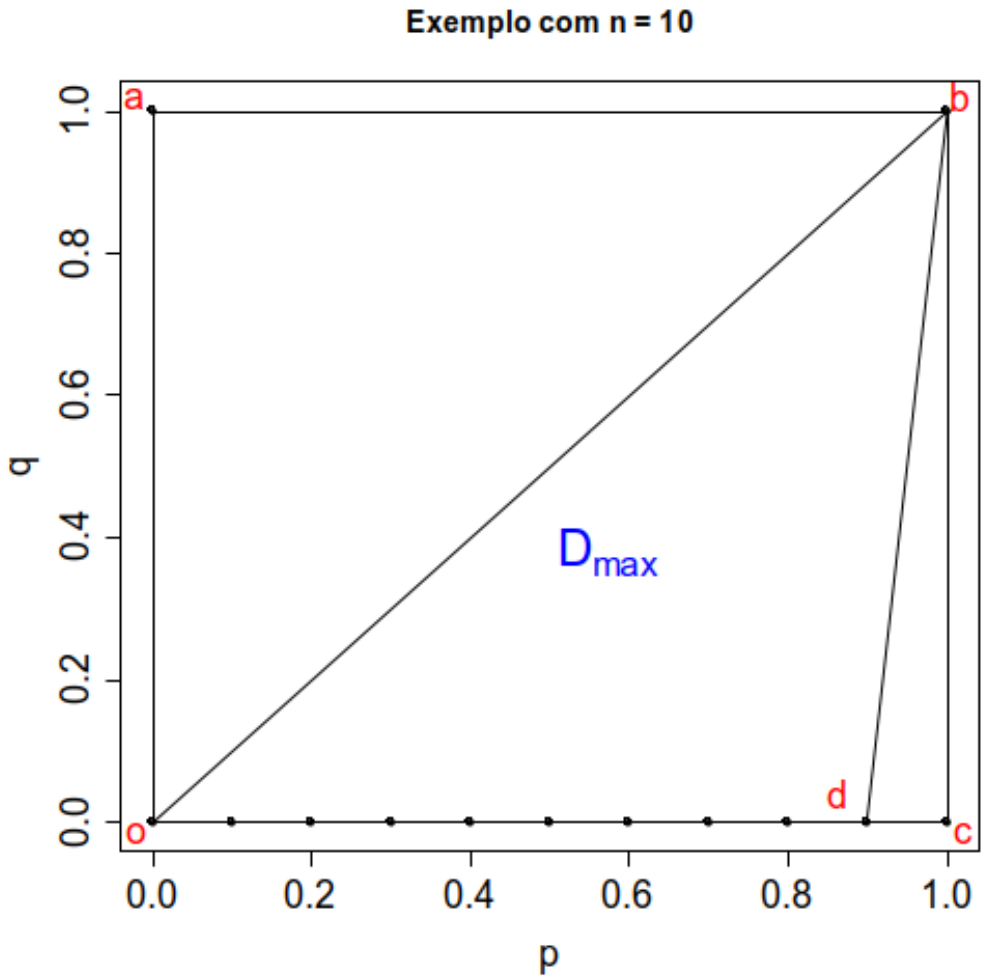
Área de desigualdade

Considere as situações:

- $x_{(1)} = x_{(2)} = ... = x_{(n)} = T/n$: proporções de posições = proporções acumuladas de valores ($q_i = p_i, i = 1, ..., n$). $\Longrightarrow$ curva de Lorenz = segmento ob (**linha da igualdade perfeita**).
- $x_{(1)} = x_{(2)} = ... = x_{(n-1)} = 0$ e $x(n) = T$: $\Longrightarrow$ curva de Lorenz é formada pelos pontos $(0, 0), (1 - 1/n, 0)$ e $(1, 1)$: **curva da desigualdade perfeita.** Quando $n \to \infty$: curva da desigualdade perfeita coincide com ocb.

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

Índice de Gini



Índice de Gini

Índice de Gini

Curva da desigualdade perfeita: odb.

Como a área do triângulo $ocb = 1/2$, temos que $0 \leq D < 1/2$.

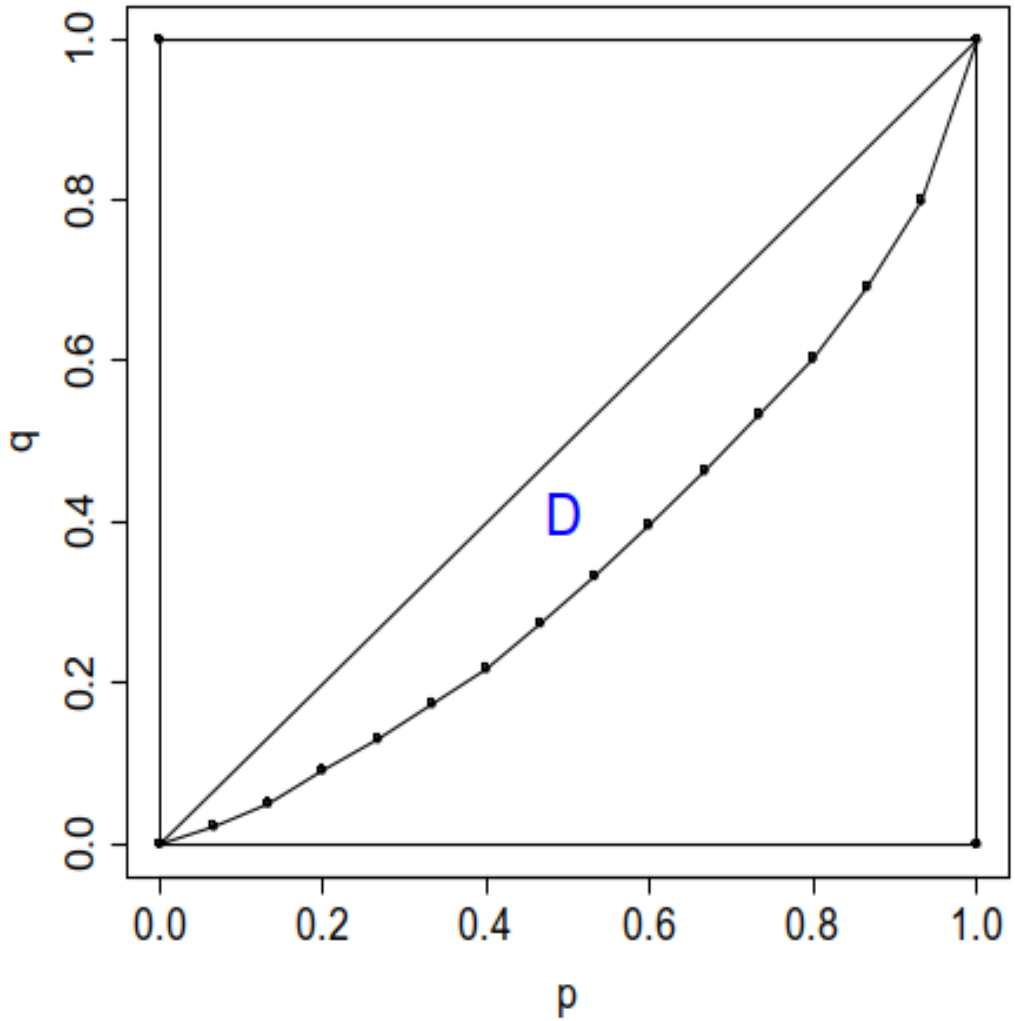
Valor máximo de D (desigualdade perfeita):

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

$D_{max} \to 1/2$ quando $n \to \infty (d \to c)$.

$max\ D_{max} = 1/2$.

Índice de Gini



Índice de Gini

Proposto por C. Gini em 1914.

Índice de Gini

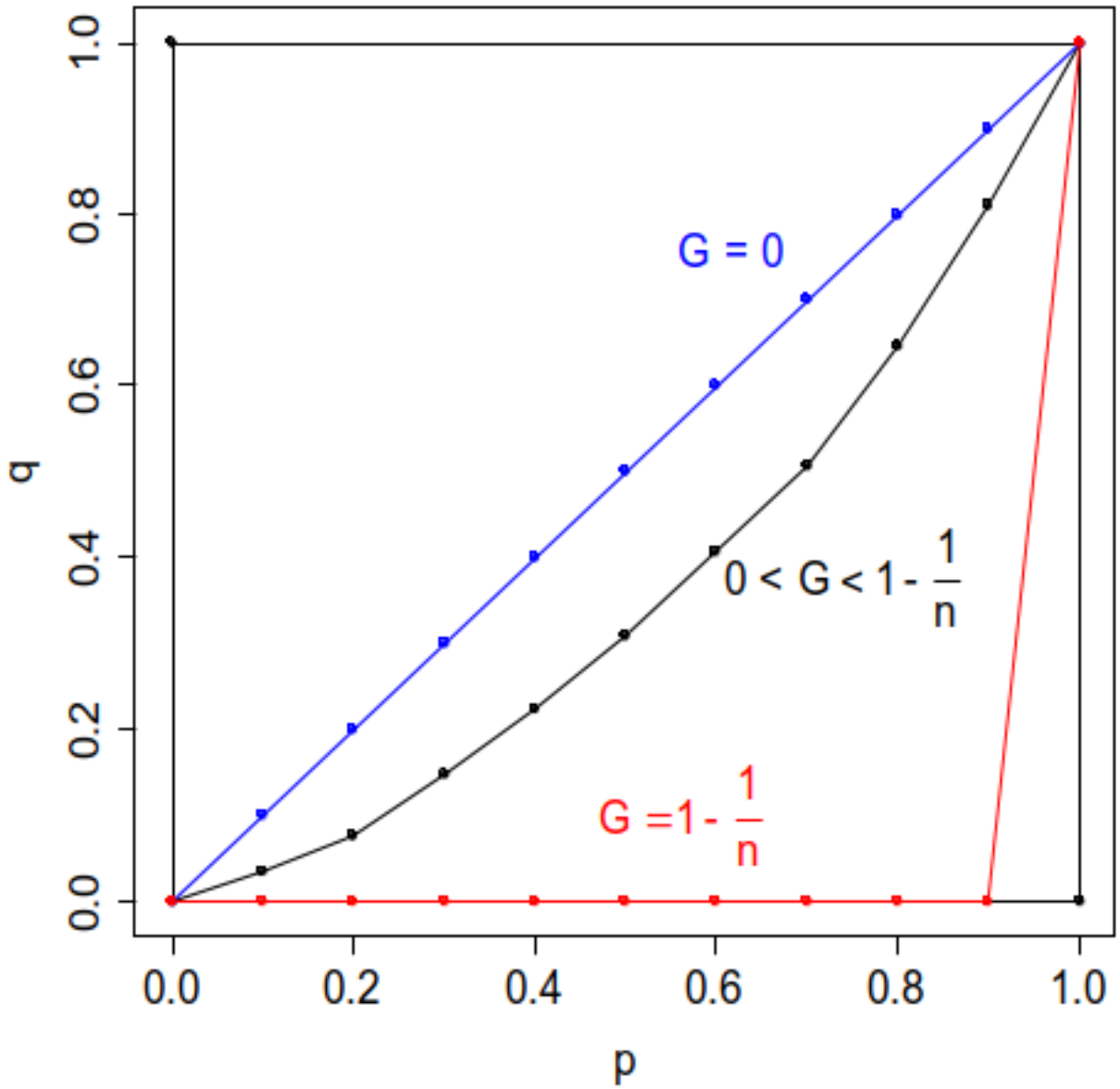
$$G = D/maxD_{max} = D/(1/2) = 2D.$$

Propriedades

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

.

Índice de Gini



Índice de Gini

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

Como calcular G?

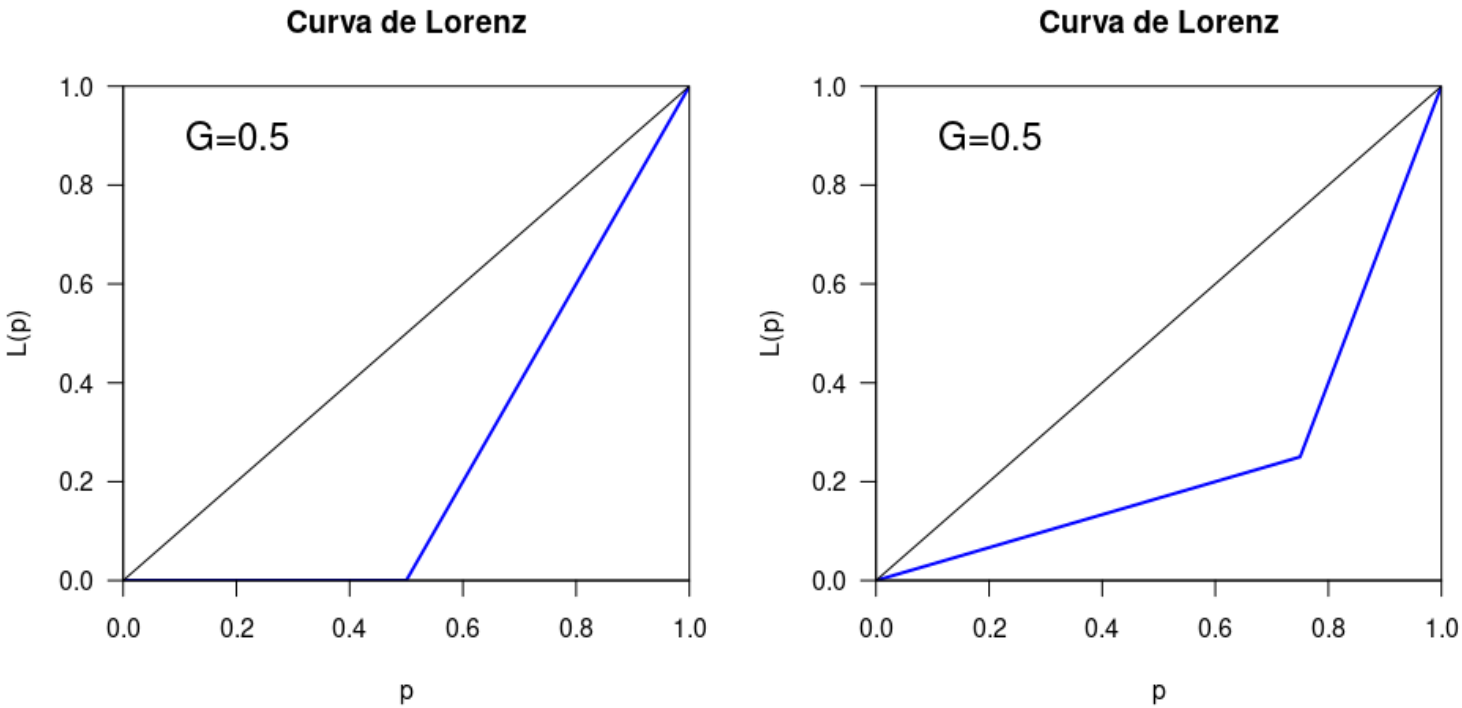
$$G = 1 - \frac{1}{n} \sum_{i=1}^{n} (q_i + q_{i-1})$$

sendo que $q_0 = 0$ e

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

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.

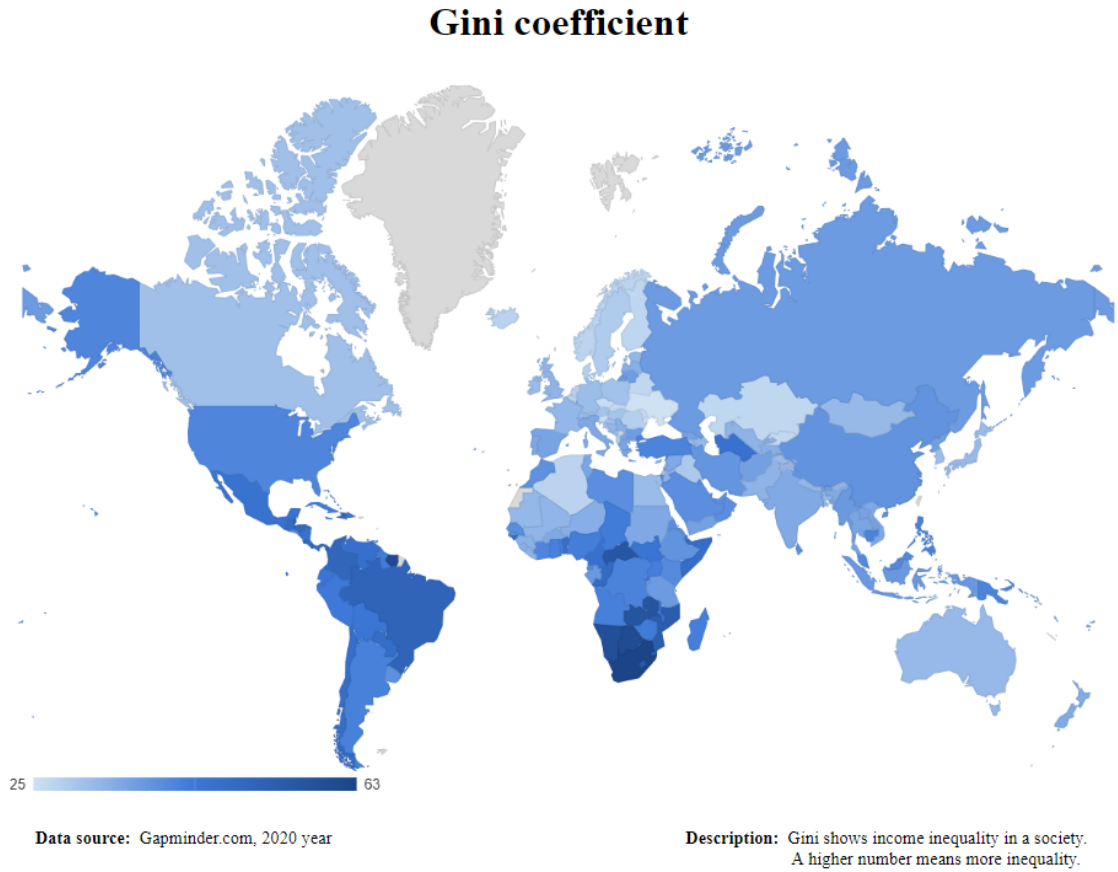
Índice de Gini

O índice de Gini é indicador de desigualdade de renda que varia de 0 a 1, sendo 0 em uma situação na qual toda a população possuísse uma renda equivalente, e 1 se apenas uma pessoa detivesse toda a riqueza do país.

Fonte: https://www.br.undp.org/content/dam/brazil/docs/RelatoriosDesenvolvimento/PressReleases/undp-br-rdh_desig-2006.pdf.

https://www.br.undp.org/content/brazil/pt/home/presscenter/articles/2018/brasil-mantem-tendencia-de-avanco-no-desenvolvimento-l
html

Índice de Gini



Cartograma do índice de Gini em países do mundo. Fonte:

https://www.reddit.com/r/dataisbeautiful/comments/f8f938/oc_gini_coefficient_by_country_2020_year/

Índice de Gini

Gráfico de linhas do índice de Gini de 2012 a 2019.

Fonte: `https://portal.fgv.br/noticias/desigualdade-renda-brasil-bate-recorde-aponta-levantar`

Índice de Gini

Diferença média. Medida de dispersão dada por

$$\bar{d} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

Diferenças $(n = 5)$:

	x_1	x_2	x_3	x_4	x_5
x_1	$x_1 - x_1 = 0$	$x_1 - x_2$	$x_1 - x_3$	$x_1 - x_4$	$x_1 - x_5$
x_2	$x_2 - x_1$	$x_2 - x_2 = 0$	$x_2 - x_3$	$x_2 - x_4$	$x_2 - x_5$
x_3	$x_3 - x_1$	$x_3 - x_2$	$x_3 - x_3 = 0$	$x_3 - x_4$	$x_3 - x_5$
x_4	$x_4 - x_1$	$x_4 - x_2$	$x_4 - x_3$	$x_4 - x_4 = 0$	$x_4 - x_5$
x_5	$x_5 - x_1$	$x_5 - x_2$	$x_5 - x_3$	$x_5 - x_4$	$x_5 - x_5 = 0$

Pode ser provado que $G = \dfrac{\bar{d}}{2\bar{x}}$

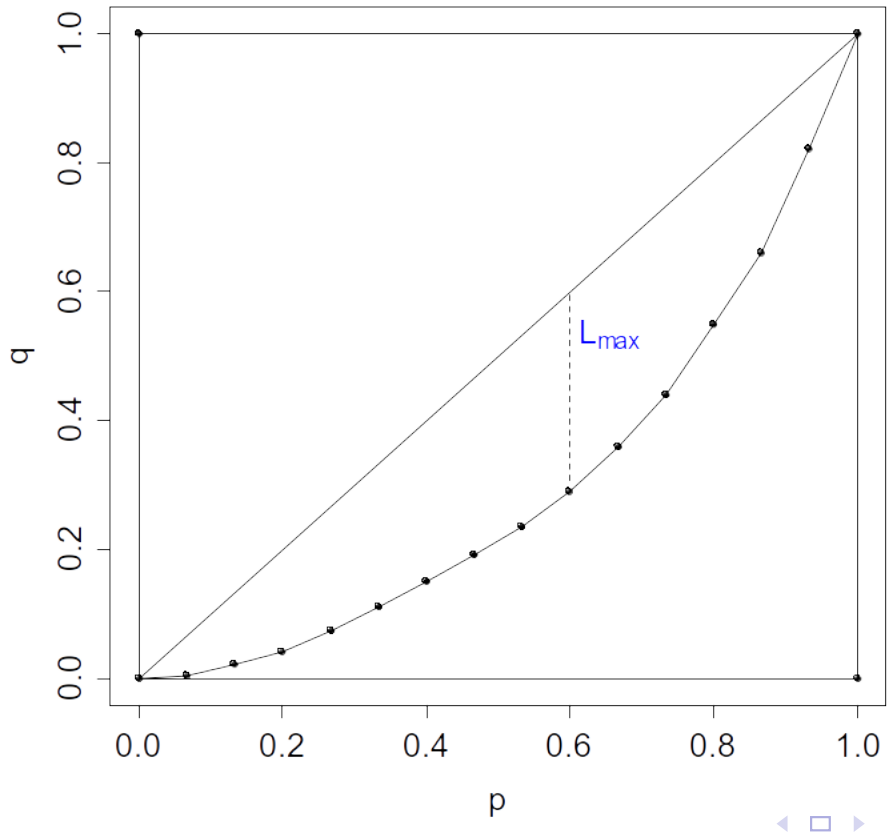
G é uma medida de dispersão relativa.

Discrepância máxima

Medida associada à curva de Lorenz.

Valor máximo da diferença entre a proporção acumulada de posições e a proporção acumulada de valores:

$L_{max} = max(p_i - q_i), i = 1, ..., n.$



Discrepância máxima

Declividade da curva:

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

$$x_{(i)} \leq \bar{x} \implies B_i \leq 1$$

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

Encontrar j tal que

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(j)} \leq \bar{x} \leq x_{(j+1)} \leq \ldots x_{(n)}$$

$L_{max} = p_i - q_i$.

Pode ser provado que $L_{max} = \frac{d_m}{2\bar{x}}$, $d_m = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$: desvio médio

L_{max} é uma medida de dispersão relativa.