Estatística para Ciências de Dados

Aula 1: Introdução

Mariana Cúri
ICMC/USP
mcuri@icmc.usp.br

Programa

0. Introdução e motivação
1. Análise descritiva de dados
2. Probabilidade
3. Inferência estatística
4. Inferência bayesiana
5. Análise de regressão

Pergunta de interesse

3

Class

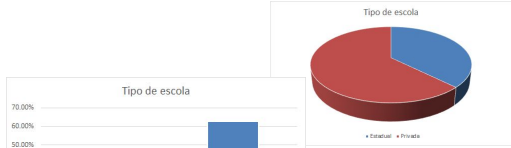
Bayesi

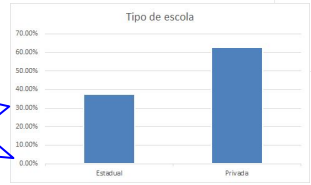
4

Análise Inferencial

(com base na Teoria de Probabi

2





População

O que avaliar:

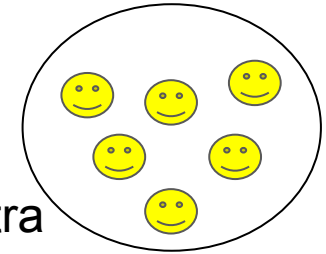
- quem: *UNIDADES AMOSTRAIS (samples/objetos)*
- o que: *VARIÁVEIS (atributos)*
- quando: *MOMENTOS / CONDIÇÕES* de avaliação

1

Análise Descritiva

Amostragem

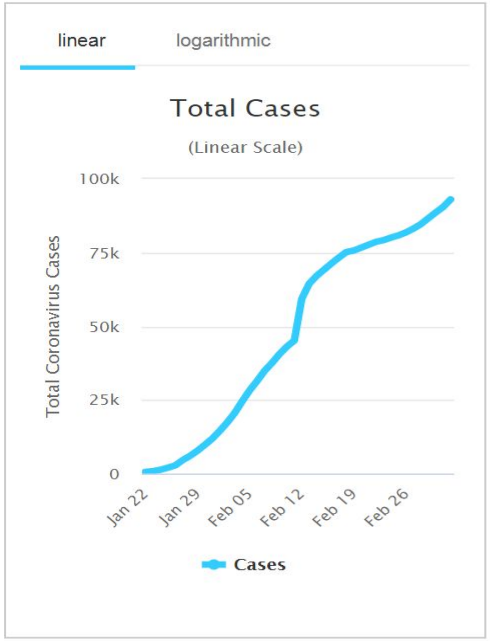
Amostra

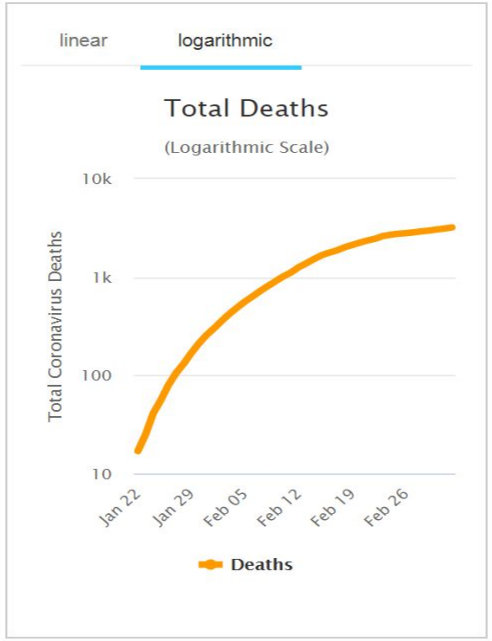


USP ICMC USP SÃO CARLOS CeMEAI CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

Motivação

A Estatística está em nosso dia-a-dia

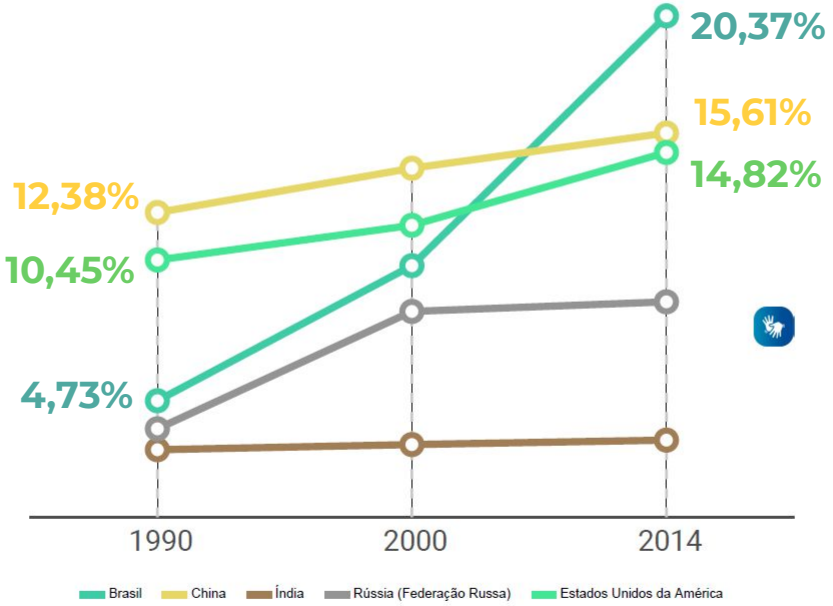




Fonte: https://www.worldometers.info/coronavirus/

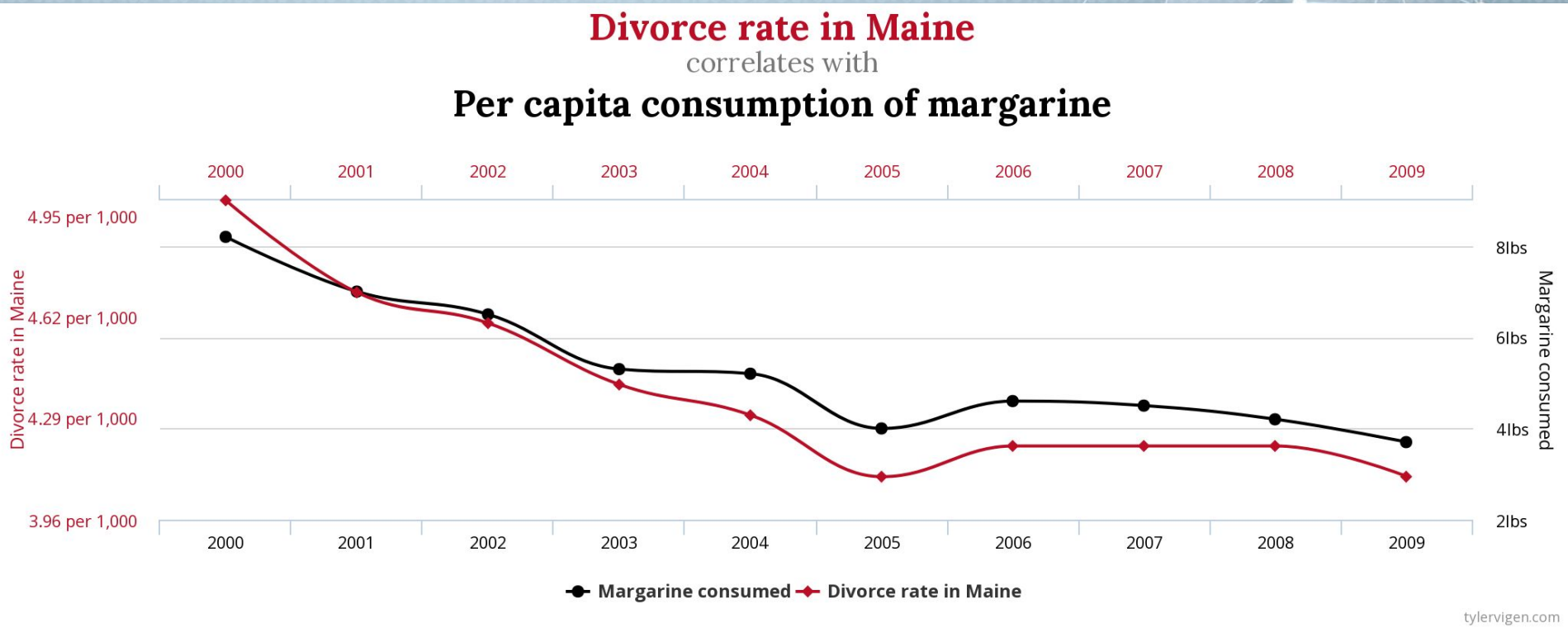
Áreas protegidas no total do território nacional

https://paises.ibge.gov.br/#/mapa/comparar/brasil?lang=pt



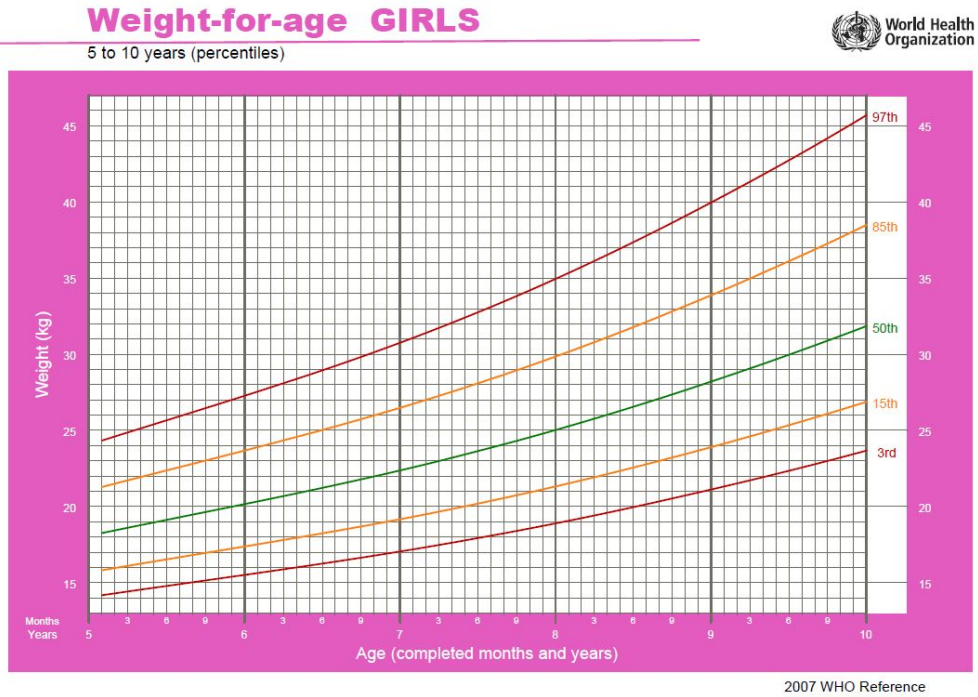
Fonte: Millennium Development Goals Indicators. United Nations Statistics Division, Department of Economic and Social Affairs. <http://mdgs.un.org/unsd/mdg/Data.aspx>. Acessado em: janeiro 2020 (http://mdgs.un.org/unsd/mdg/Data.aspx)

Motivação

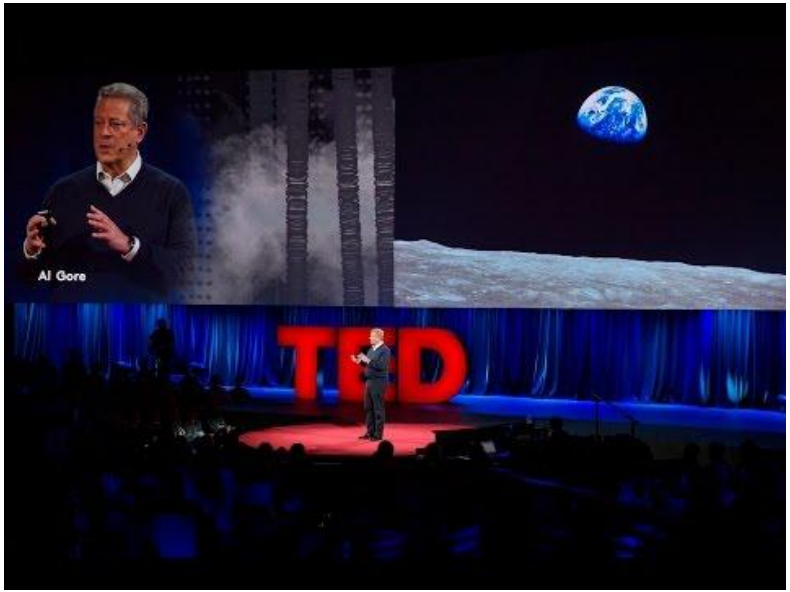


Fonte: https://www.tylervigen.com/spurious-correlations

Motivação



Fonte: https://www.who.int/growthref/cht_wfa_girls_perc_5_10years.pdf?ua=1

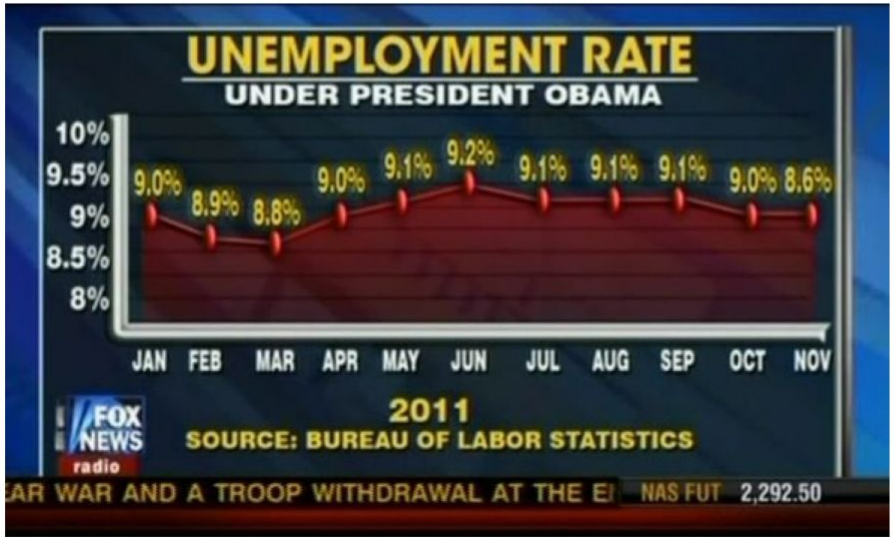


Intervalo: 3:15 a 4:00
Fonte:
https://www.youtube.com/watch?v=u7E1v24Dllk&t=1142s

Motivação

Fox News's unemployment chart: Better graphics?

By **Erik Wemple**
December 12, 2011



Screen grab of chart showing unemployment rate under President Obama. (Fox News)

Fonte: https://www.washingtonpost.com/blogs/erik-wemple/post/fox-newss-unemployment-chart-better-graphics/2011/12/12/gIQAUVgNqO_blog.html

Análise Descritiva

Tipos de variáveis

- Qualitativas: nominais ou ordinais
- Quantitativas: discretas ou contínuas

Variáveis qualitativas: Covid-19

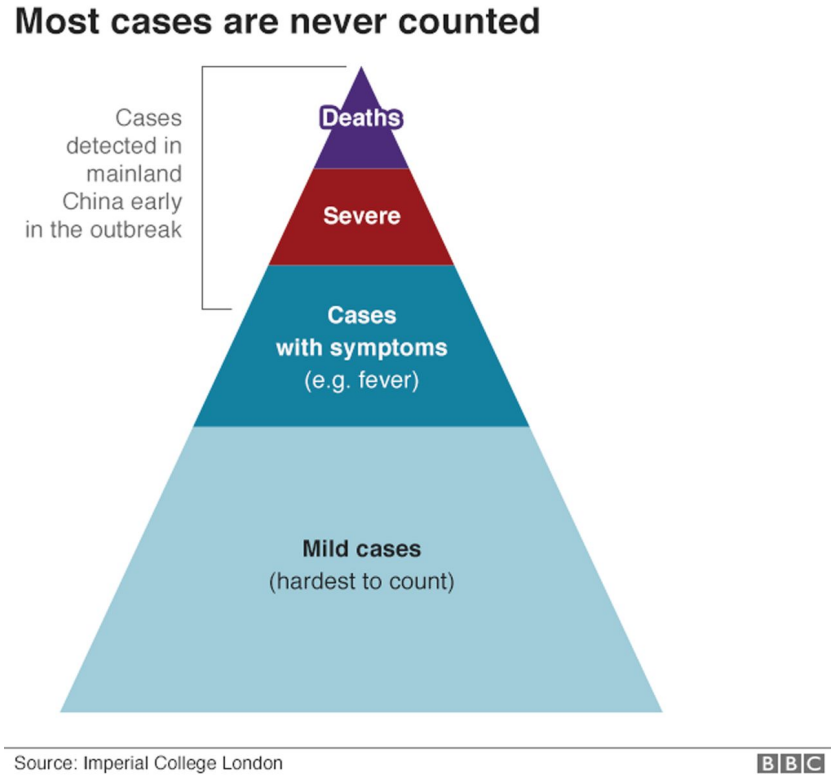
G1 CIÊNCIA E SAÚDE

Covid-19 comparada com Sars e Mers

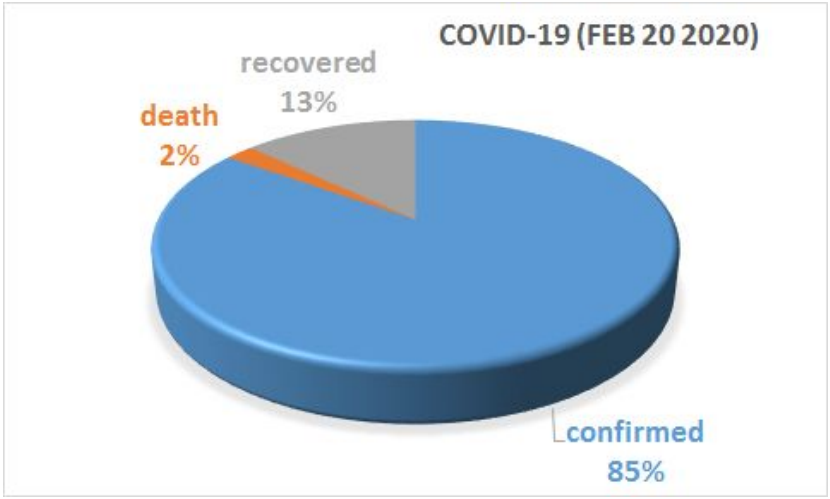
A pesquisa do CCDC afirma que cerca de 80,9% das novas infecções por coronavírus são classificadas como leves, 13,8% como graves e apenas 4,7% como críticas, o que inclui quadro de insuficiência respiratória, falência múltipla dos órgãos e sepse.

	# cases	%
confirmed	71226	85%
death	1770	2%
recovered	10865	13%
Total	**83861**	**100%**

05/03/2020 09h20 · Atualizado há 4 dias

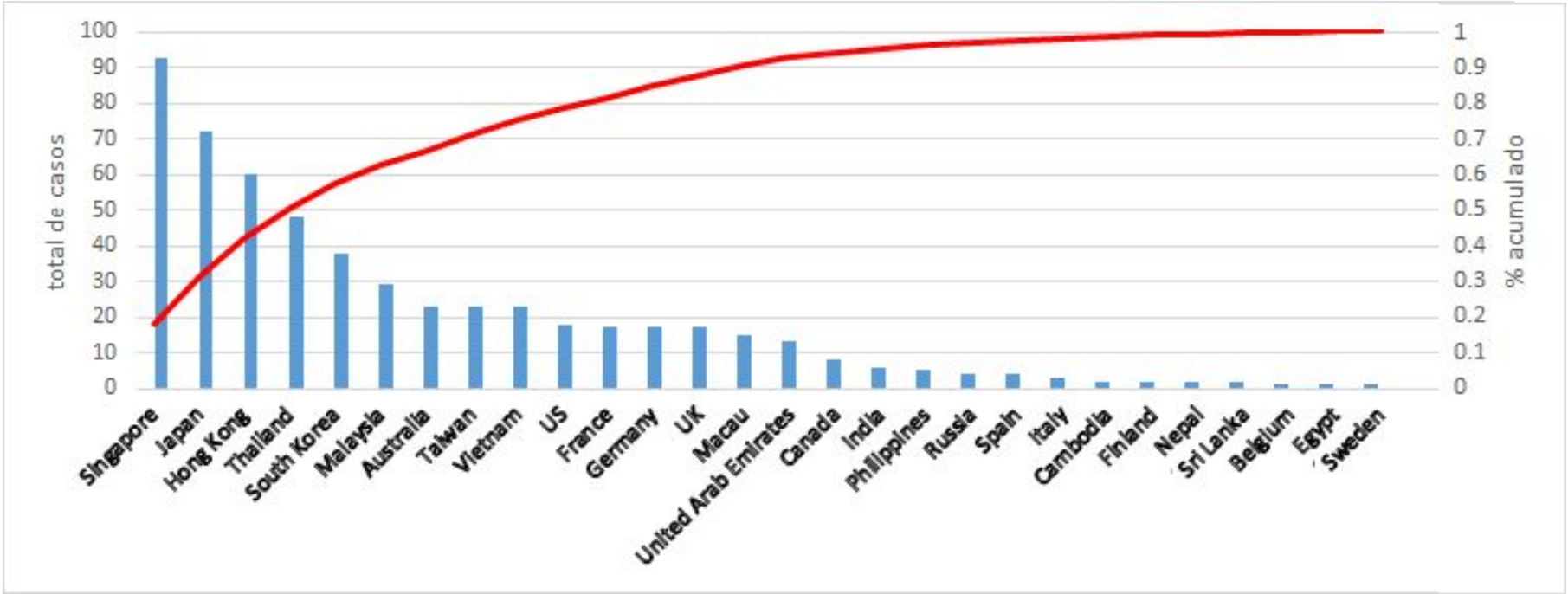


Variáveis qualitativas: Covid-19

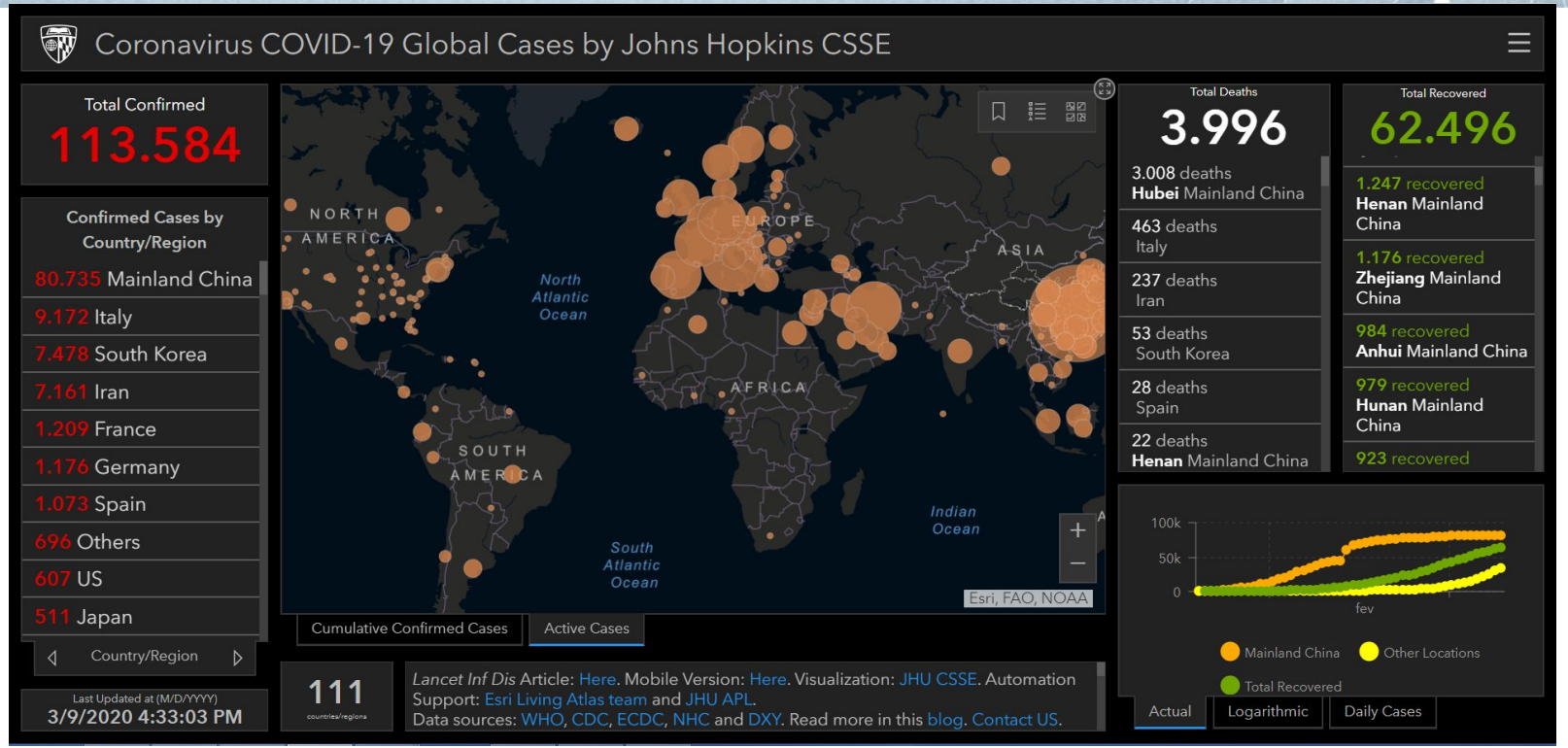


Variáveis qualitativas: Covid-19

Casos fora da China Continental: **Gráfico de Pareto**



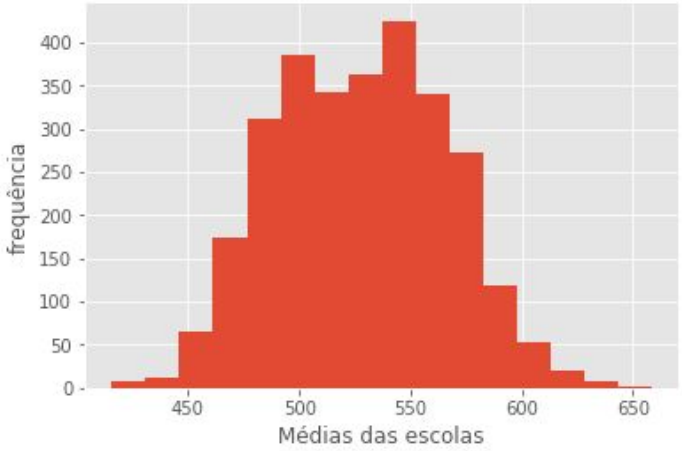
Variáveis qualitativas: Covid-19



Fonte: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Variáveis quantitativas:

- acesso à educação superior: Sisu, Prouni e convênios com instituições portuguesas
- 4 provas (180 questões) + 1 redação: notas “de 0 a 1000”, com média 500 e desvio padrão 100
- prova de matemática e suas tecnologias (2015): 45 questões, n=2901 escolas do estado de SP, 145.389 estudantes
- Perguntas: Existe diferença entre as escolas públicas e privadas? Tem alguma(s) escola(s) pública que se equipara às boas escolas privadas?



$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} = 527,6$$

$$S = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} = 38,1$$

$$X_{(1)} = 415,7$$

$$X_{(n)} = 658,3$$

$$Q_1 = 498,3$$

$$Q_2 = 528,4$$

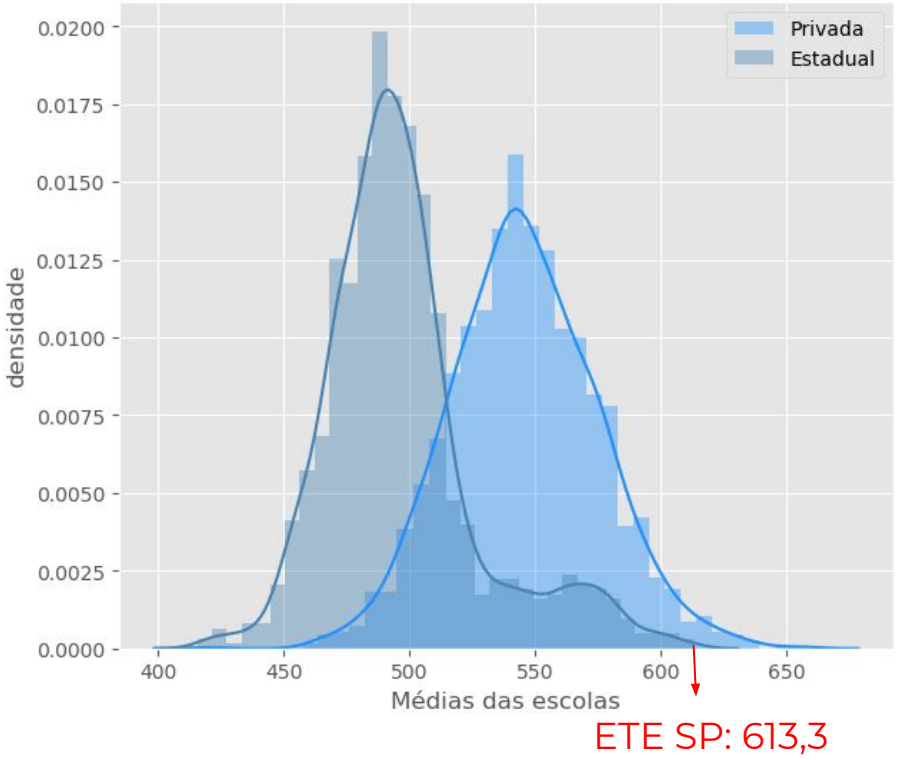
$$Q_3 = 555,7$$

$$CV = \frac{S}{\bar{X}} 100\% = 7,2\%$$

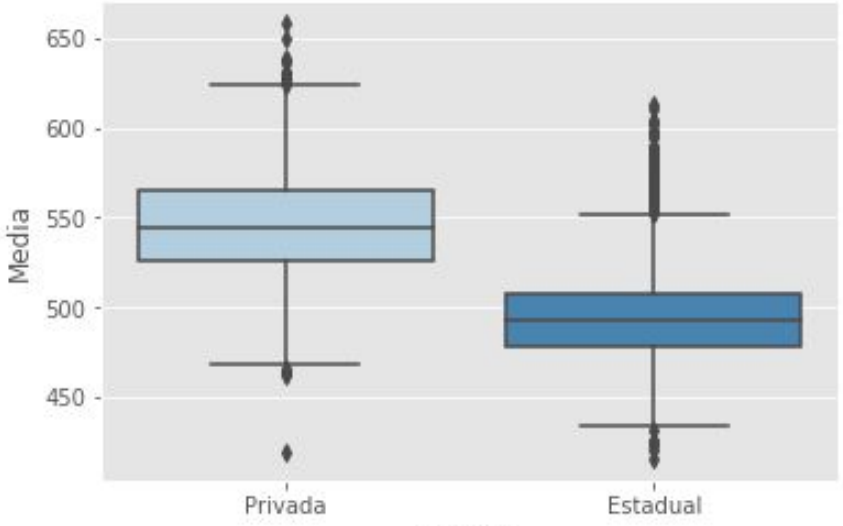
Art of Stat
http://www.artofstat.com/webapps.html

Variáveis quantitativas:

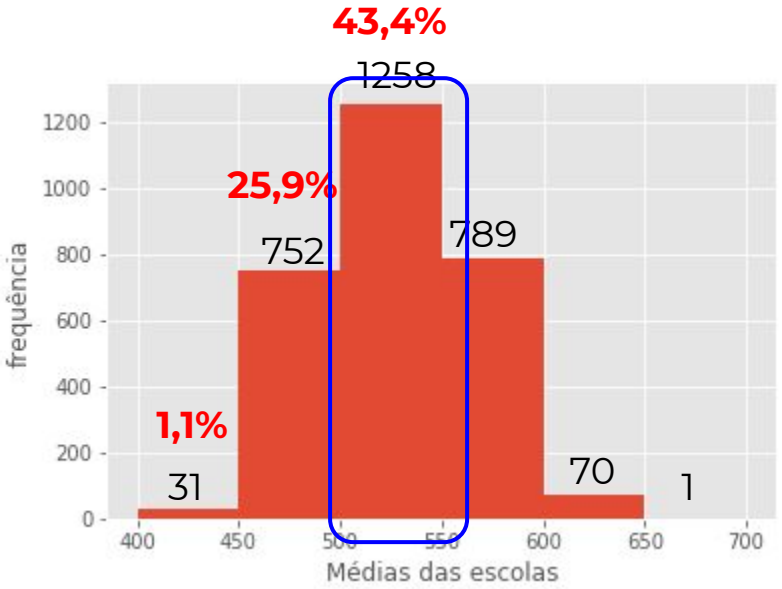
Tipo de escola	Média no ENEM	Desvio Padrão	nº de escolas
Estadual	497.1	30.7	1085
Privada	545.8	29.3	1816
Total Geral	**527.6**	**38.1**	**2901**



ETE SP: 613,3



Variáveis quantitativas: Histograma



$$\bar{X} = \frac{31}{2901}425 + \frac{752}{2901}475 + \frac{1258}{2901}525 + \frac{789}{2901}575 + \frac{70}{2901}625 + \frac{1}{2901}675 = 527,03$$

(550-500) —— **43,4%**

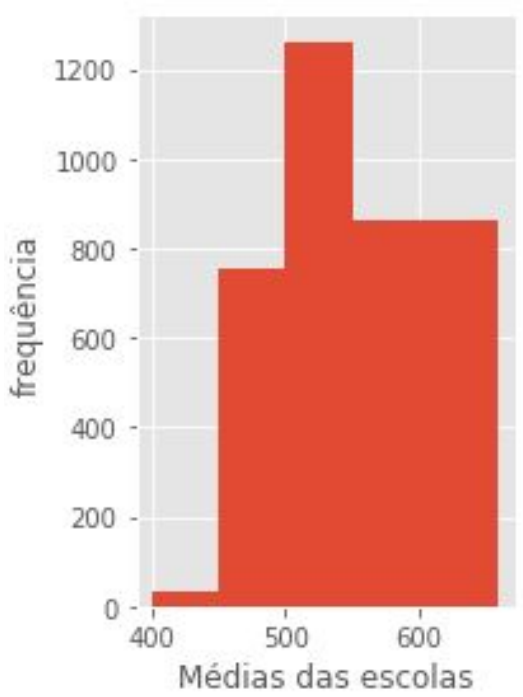
(Md-500) —— **23%**

Md = 526,5

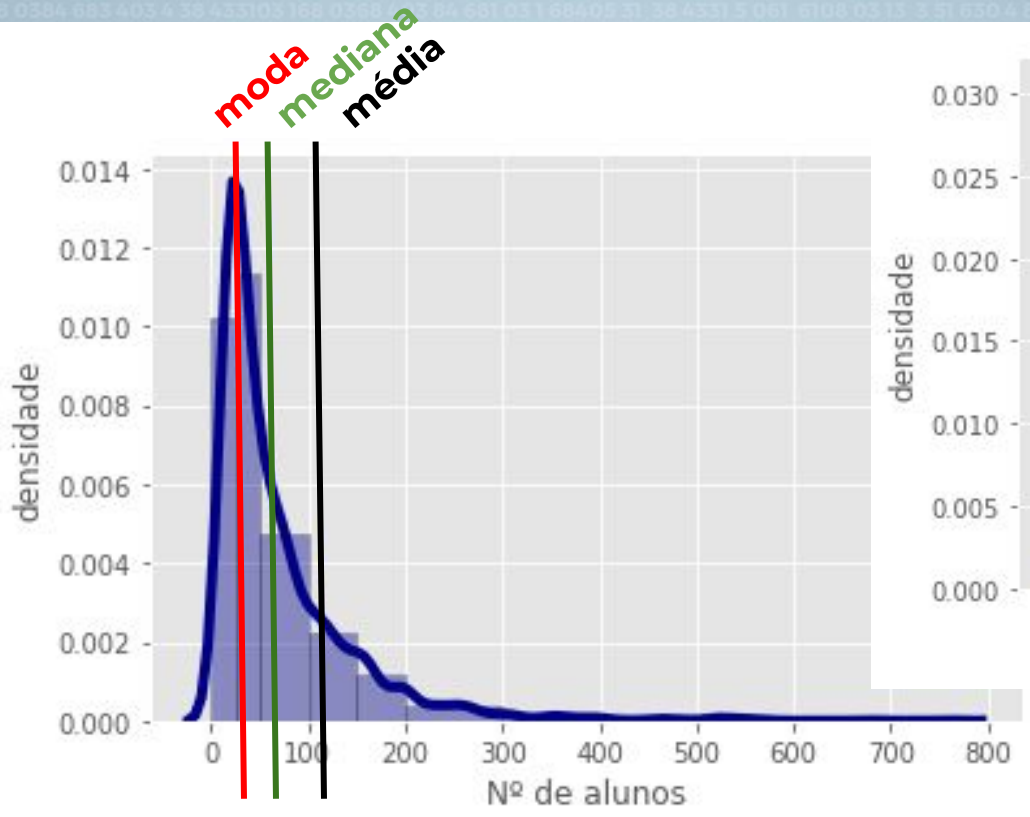
$$S = \sqrt{\frac{31}{2900}(425 - 527,03)^2 + \frac{752}{2900}(475 - 527,03)^2 + \ldots} \approx 41$$

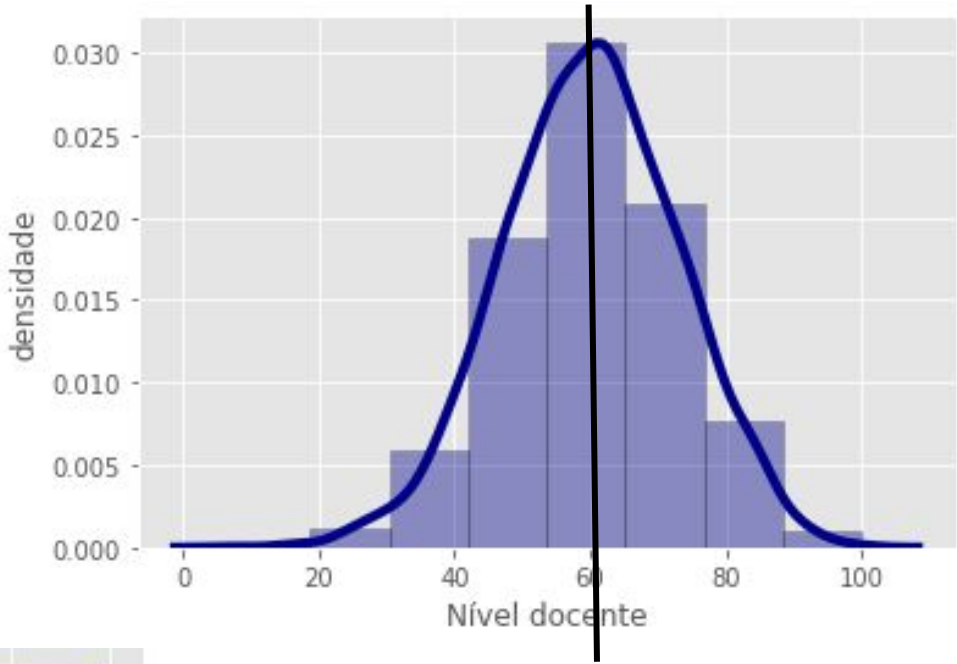
Classes de amplitudes diferentes

	count	%
(400, 450]	31	1.07
(450, 500]	752	25.92
(500, 550]	1258	43.36
(550, 600]	789	27.20
(600, 650]	70	2.41
(650, 700]	1	0.03

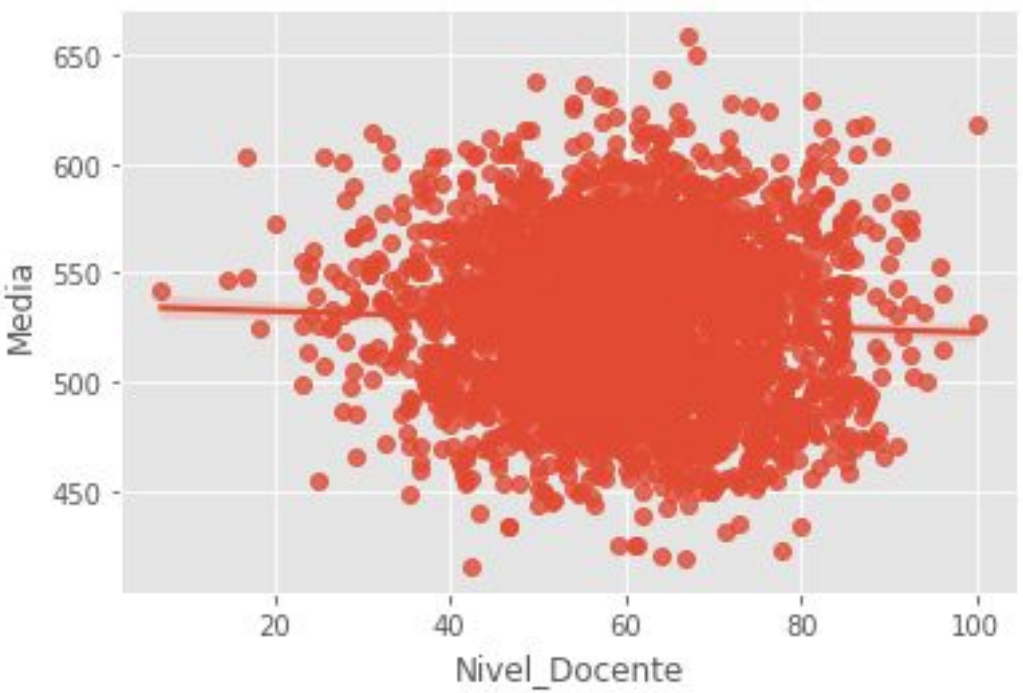


Variáveis quantitativas: (as)simétricas





Associação de variáveis



r=-0,04

Art of Stat

http://www.artofstat.com/webapps.html