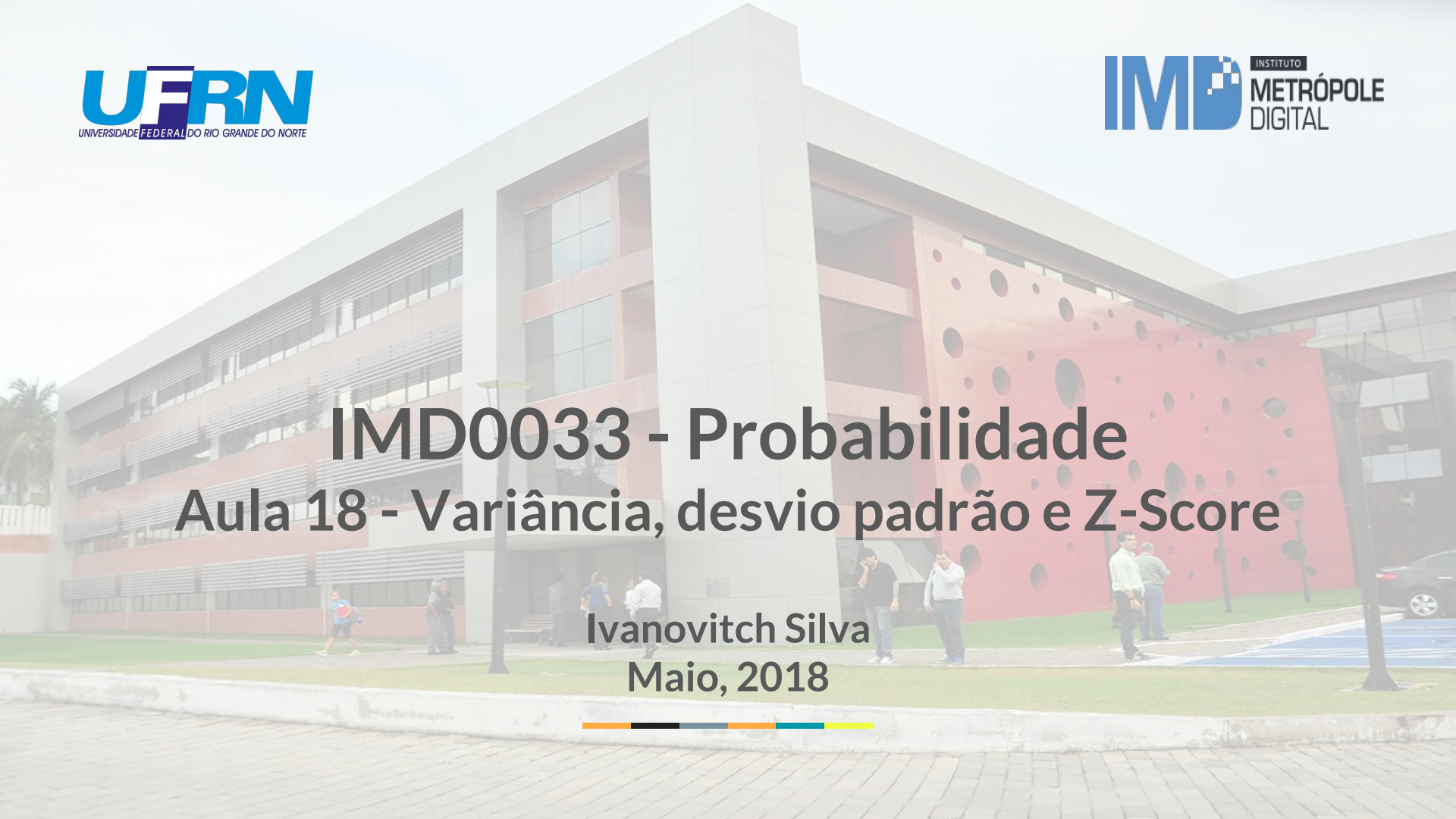


IMD0033 - Probabilidade

Aula 18 - Variância, desvio padrão e Z-Score

Ivanovitch Silva

Maio, 2018

Agenda

- Quartil e limitações
- Variância
- Desvio Padrão
- Z-Score

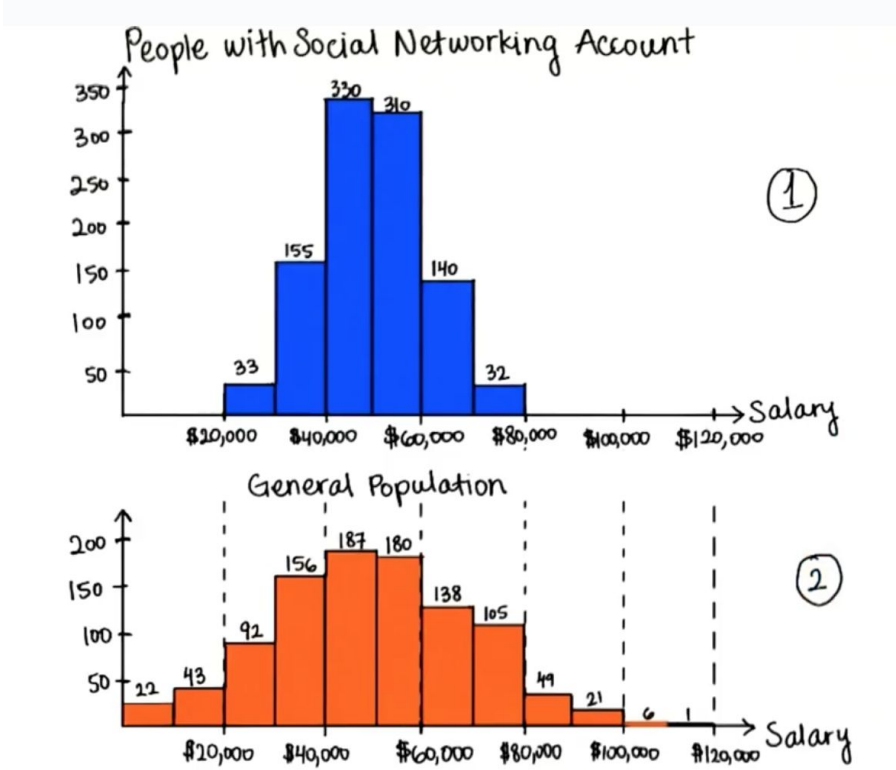
Atualizar o repositório

```
git clone https://github.com/ivanovitchm/imd0033_2018_1.git
```

Ou ....

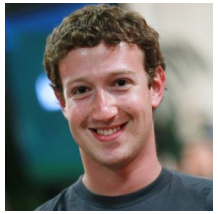
```
git pull
```

Salários de usuários com/sem redes sociais



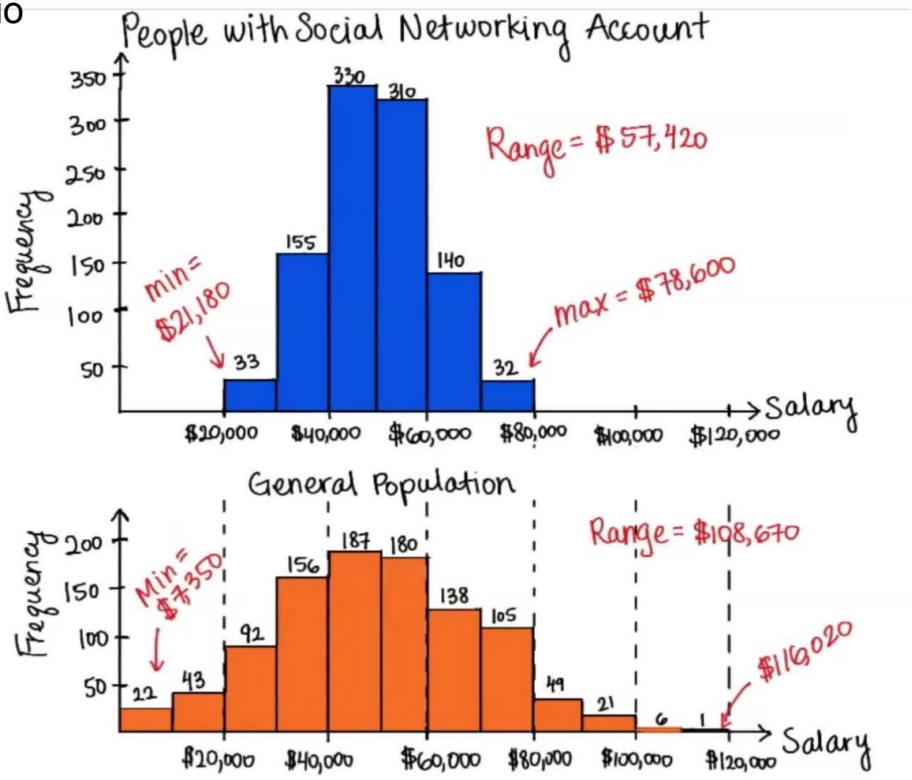
É interessante ter uma conta em redes sociais?

Qual a diferença entre as distribuições?



Quantificar o espalhamento (range)

Salário
$10M

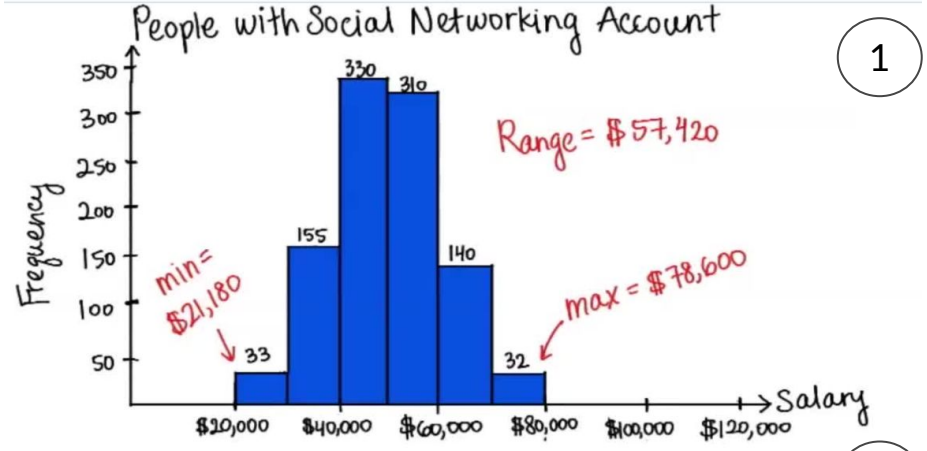


Os limites da distribuição são alterados quando novos dados são inseridos?

1) Sempre
2) Algumas vezes
3) Nunca

Quartil (Q1 - Q3)

Chop of the tails



Amostra 1		Amostra 2	
38,946		33,219	
43,420		36,254	
49,191	Q1	38,801	Q1
50,430		46,335	
50,557	Q2 Mediana	46,840	Q2 Mediana
52,580		47,596	
53,595		55,130	
54,135	Q3	56,863	Q3
60,181		78,070	
10,000,000		88,830	

Q3 - Q1 = Interquartile range (IQR)

O que é um ponto fora da curva?

Amostra 1

38,946
43,420
49,191 Q1
50,430
50,557 Q2 Mediana
52,580
53,595
54,135 Q3
60,181
10,000,000

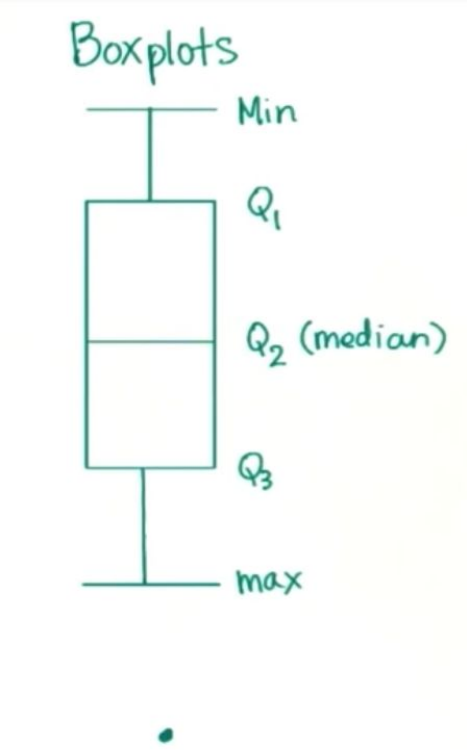
Q3 - Q1 = IQR
4944

Quais valores são considerados pontos fora da curva?

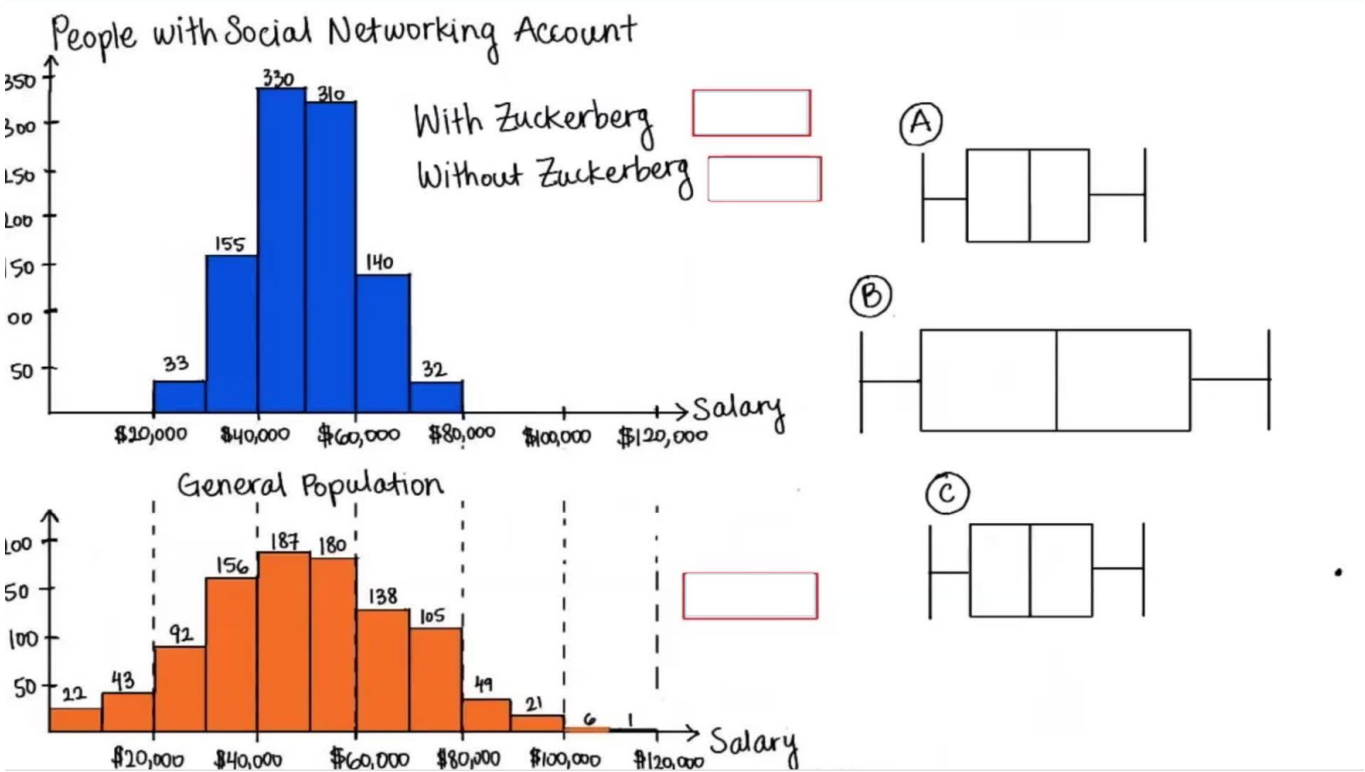
- $60,000
- $80,000
- $100,000
- $200,000

Definição:
Outlier < Q1 - 1.5 x IQR → 41,775
Outlier > Q3 + 1.5 x IQR → 61,551

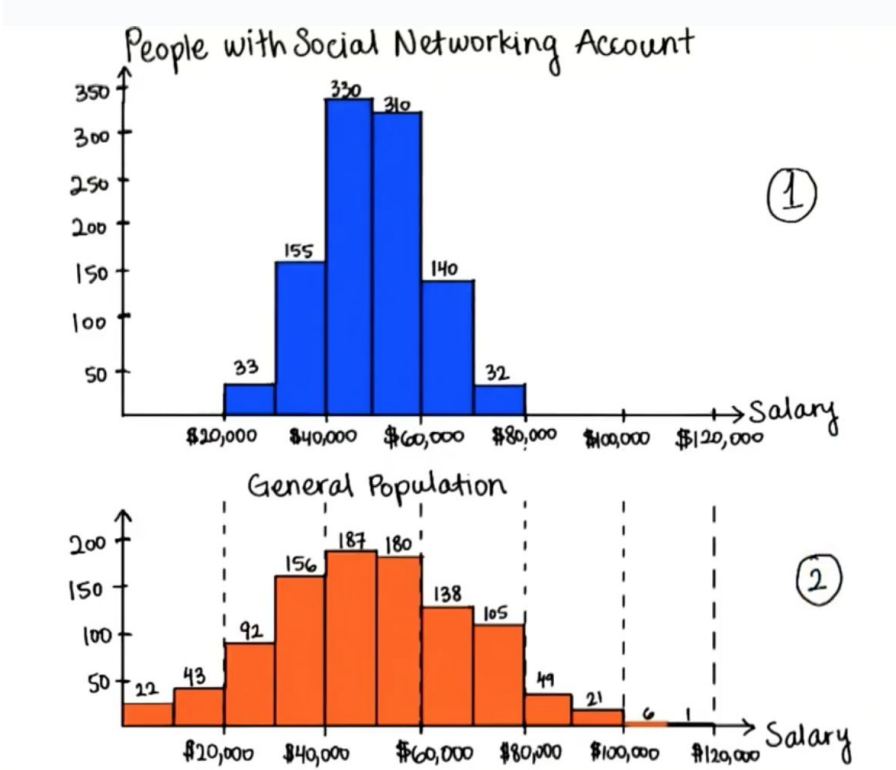
Gráficos de Caixa

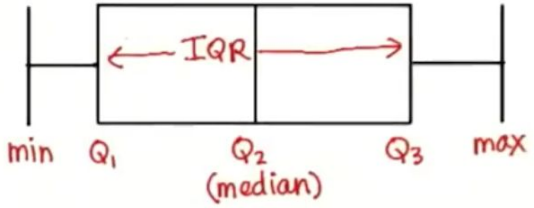


Quiz: gráficos de caixa



Problemas com o IQR





A média sempre será entre Q1 and Q3?

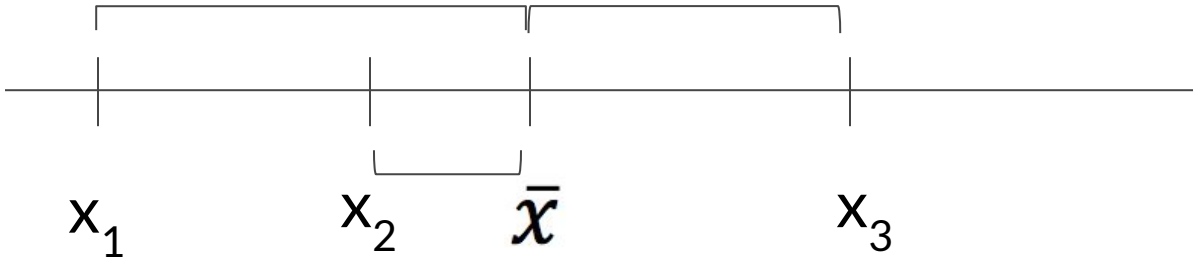
- Sim
- Não

Medidas de variabilidade

Nós precisamos de uma métrica que avalie o espalhamento da amostra estatística levando em consideração todos os dados.

~~Range~~
~~IQR~~

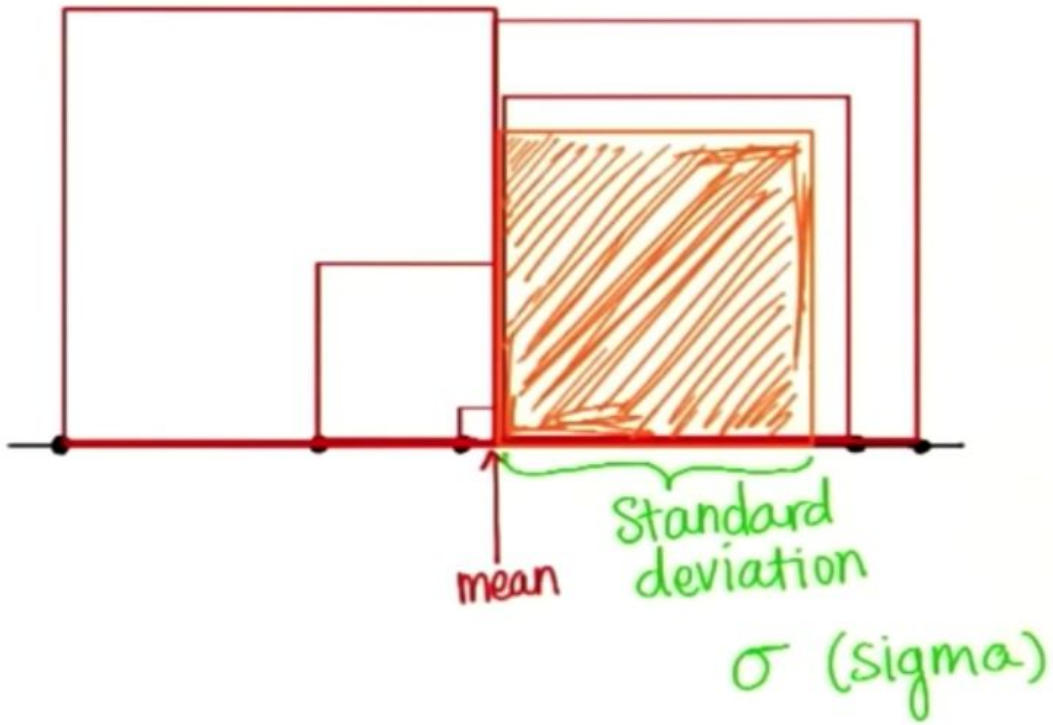
Medida de variabilidade (ideia)

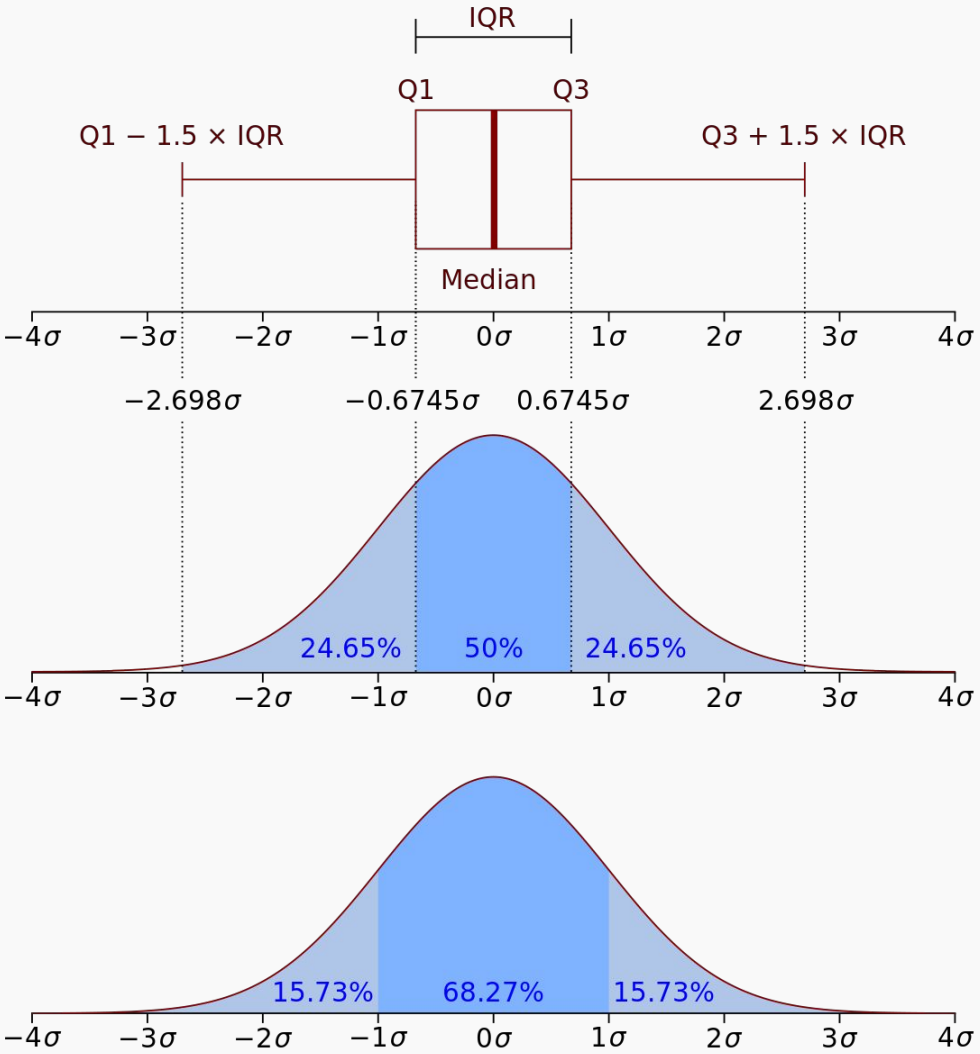


Variância $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Desvio Padrão σ $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Measure Variability (idea)




IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

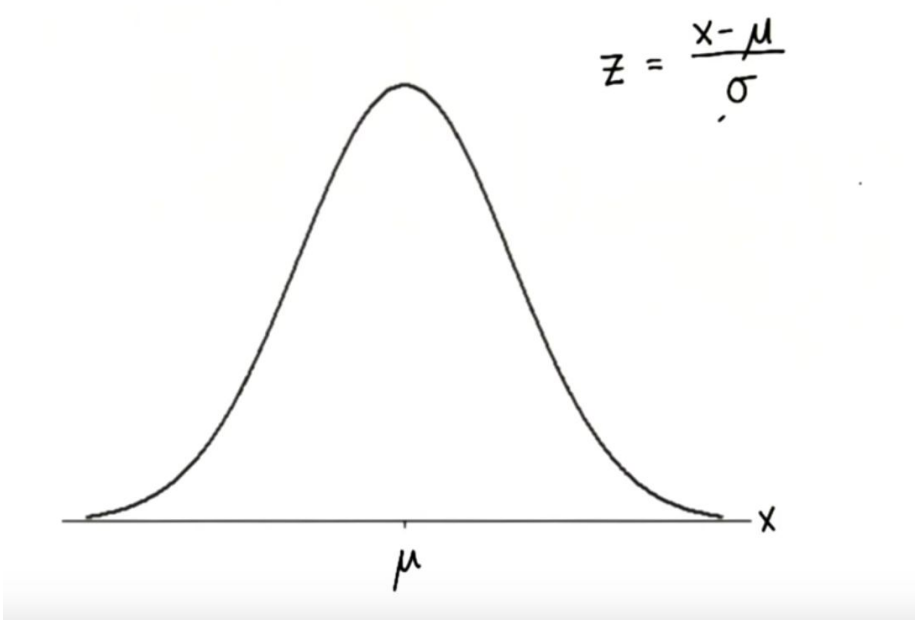
Desafio

- Criar uma função *percentagem*(*a*,*b*)
  - *a* - um vetor com as distâncias em desvios padrões a partir da média para cada ponto do vetor.
    - E.g: média = 2 e desvio padrão = 5
    - Pontos da distribuição: 1,2,3,4,5
    - *a* = [(1-2)/5, (2-2)/5, (3-2)/5, (4-2)/5, (5-2)/5]
  - *b* - a faixa de desvio padrão que se deseja (1, 2 ou 3)
    - 1 - corresponderia a todos os dados entre -1 e 1 desvios padrões de distância
  - Retorno: a percentagem de significância da amostra para o valor de *b*
  - Se *b* = 1, o retorno deverá ser 0.6827.

Dica

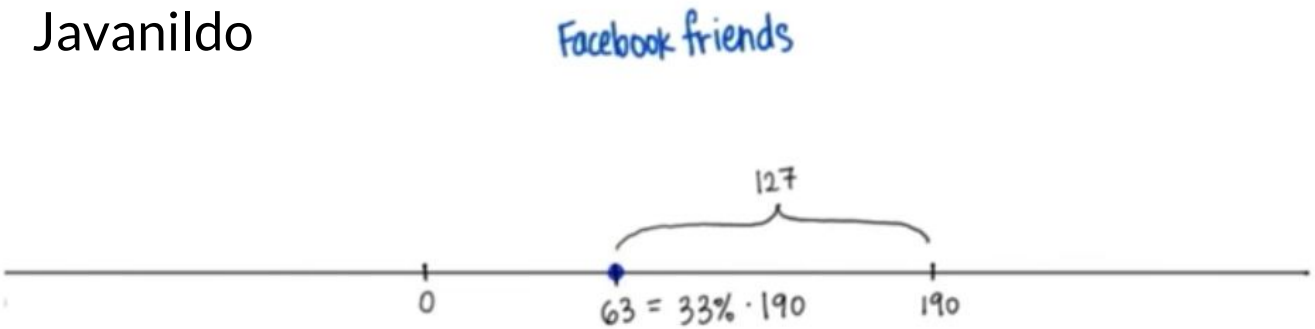
```
standard_deviations = [(i - mean) / standard_deviation for i in
wing_lengths]
def within_percentage(deviations, count):
    within = [i for i in deviations if i <= count and i >= -count]
    count = len(within)
    return count / len(deviations)
```

Z-Score

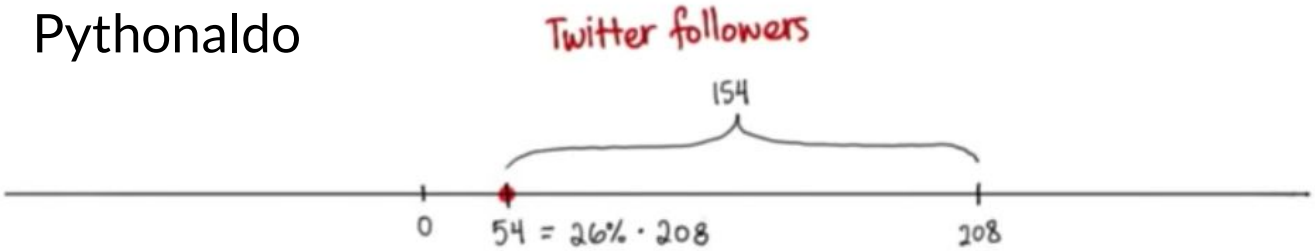


Quiz: Quem é mais popular?

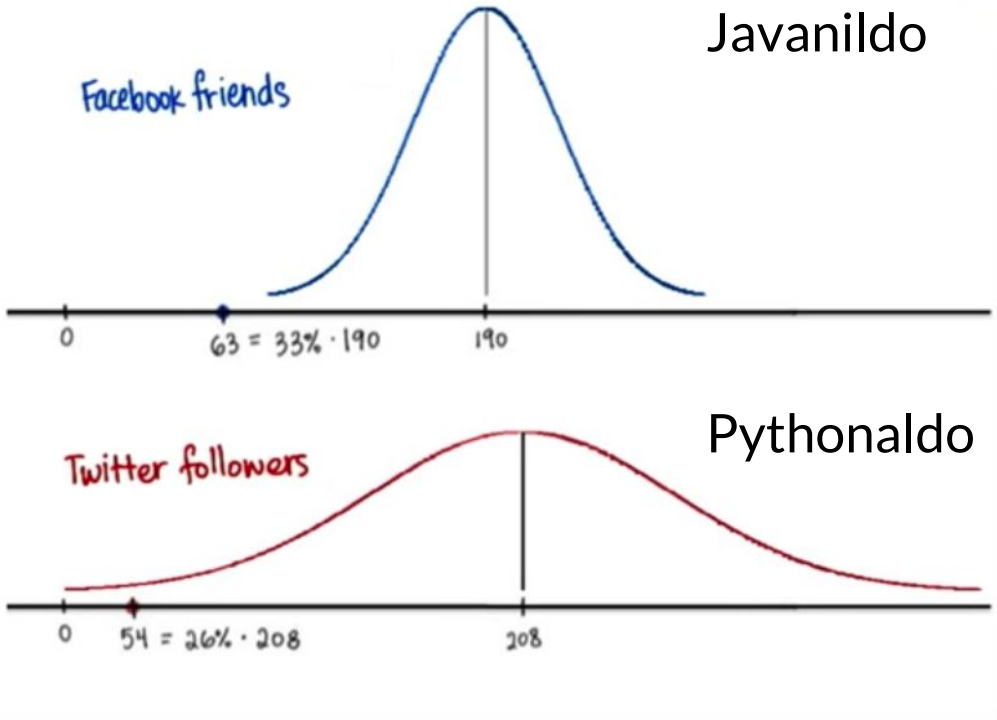
Javanildo



Pythonaldo



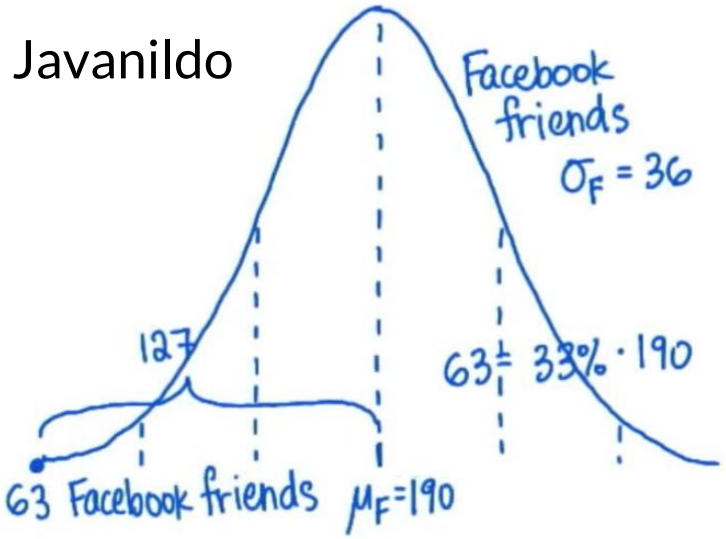
Quiz: Quem é mais popular?



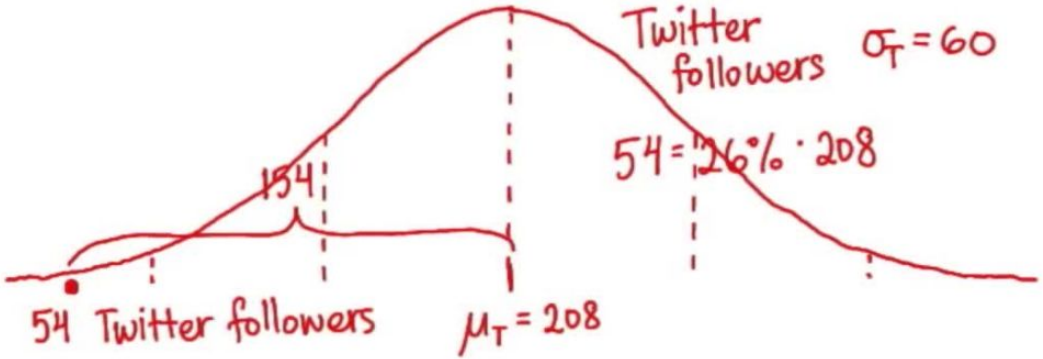
Quiz: Quem é mais popular?

Qual a distância em desvios padrões do número de amigos em relação a média?

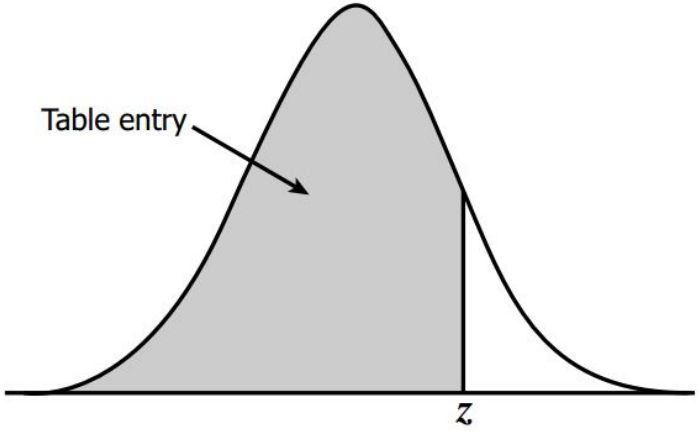
Javanildo



Pythonaldo



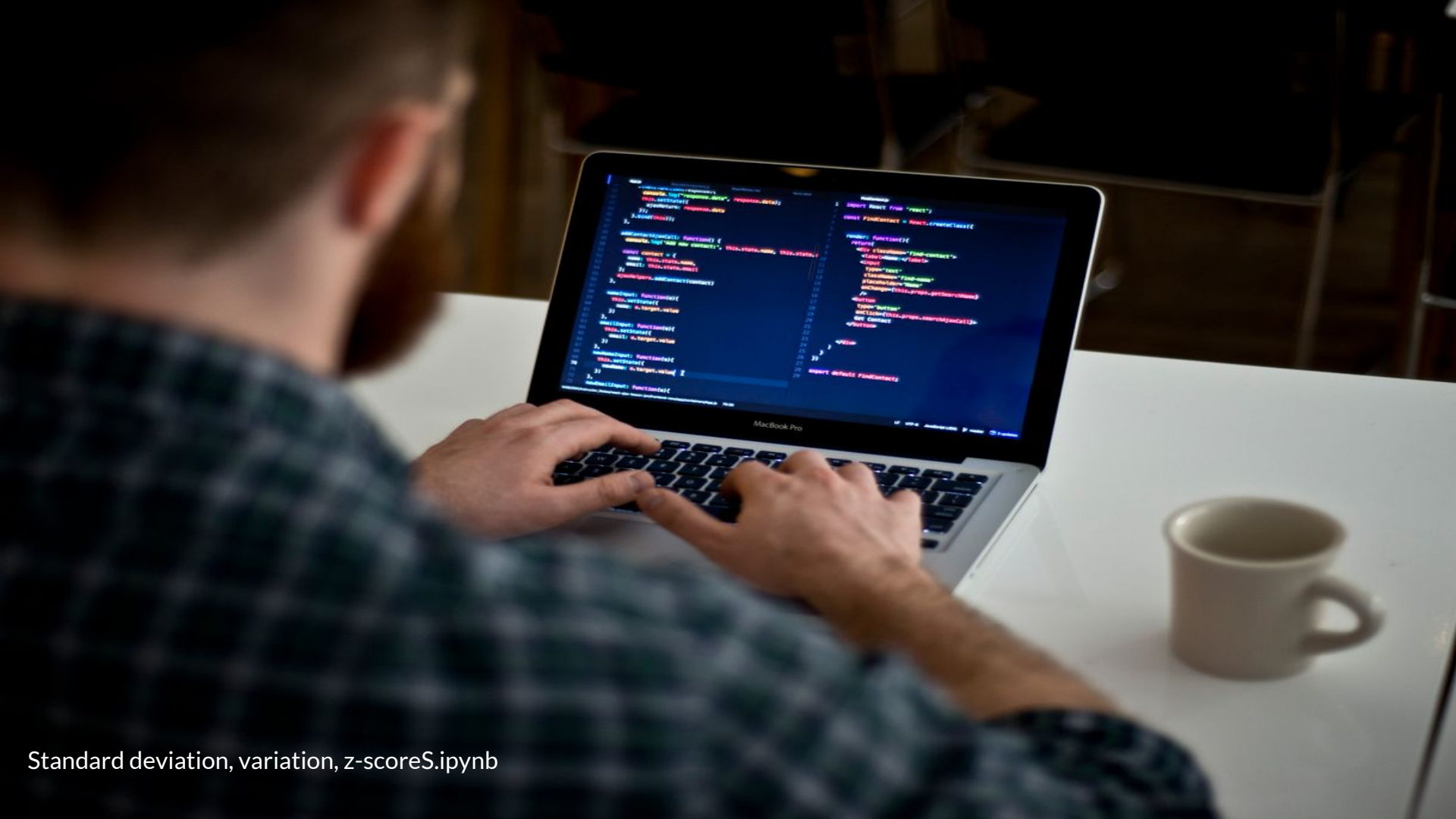
Z-Table



http://www.z-table.com/

Facebook example:
μ = 190
σ = 36
Xi = 240

Qual a percentagem de pessoas que possuem menos de 240 amigos no facebook?



Standard deviation, variation, z-scoreS.ipynb