


IMD0033 - Probabilidade

Aula 15 - Variância, desvio padrão e Z-Score

Ivanovitch Silva
Outubro, 2017



Agenda

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

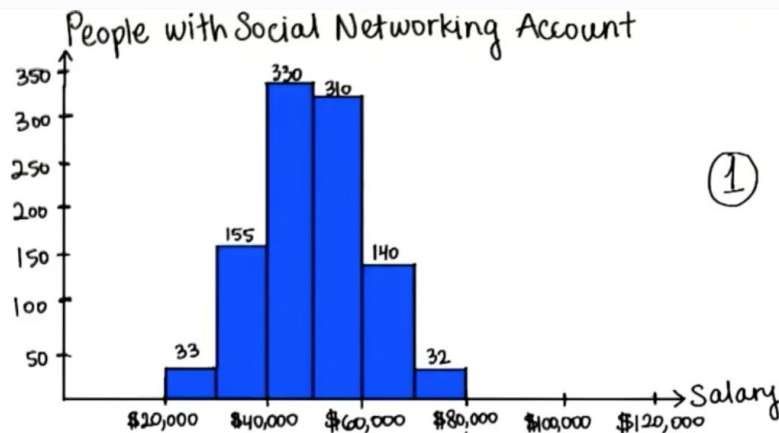
Atualizar o repositório

```
git clone https://github.com/ivanovitchm/IMD0033_Probabilidade.git
```

Ou

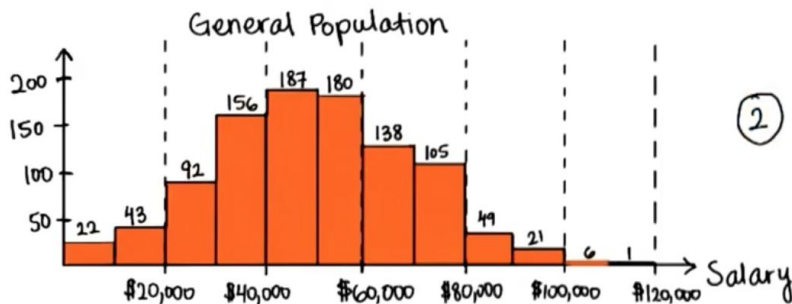
```
git pull
```

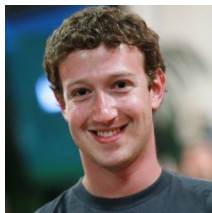
Salários de usuários com/sem redes sociais



É interessante ter uma conta em redes sociais?

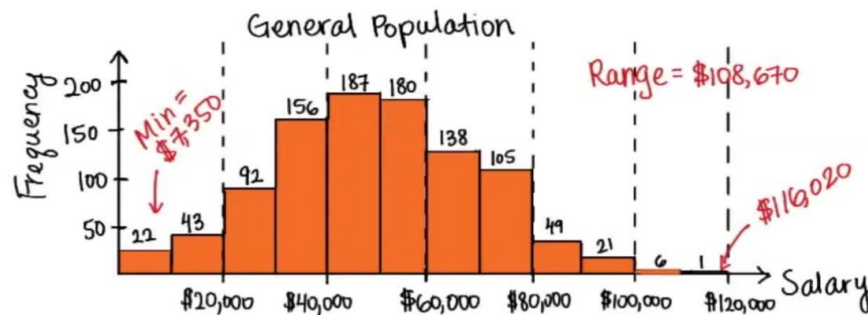
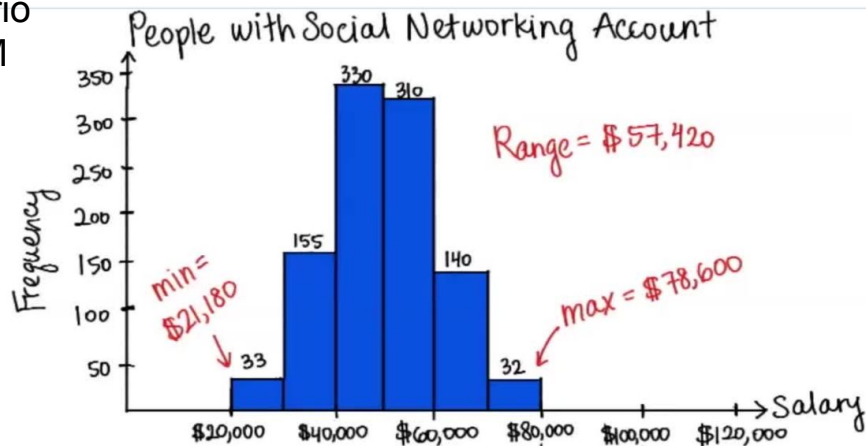
Qual a diferença entre as distribuições?





Quantificar o espalhamento (range)

Salário
\$10M

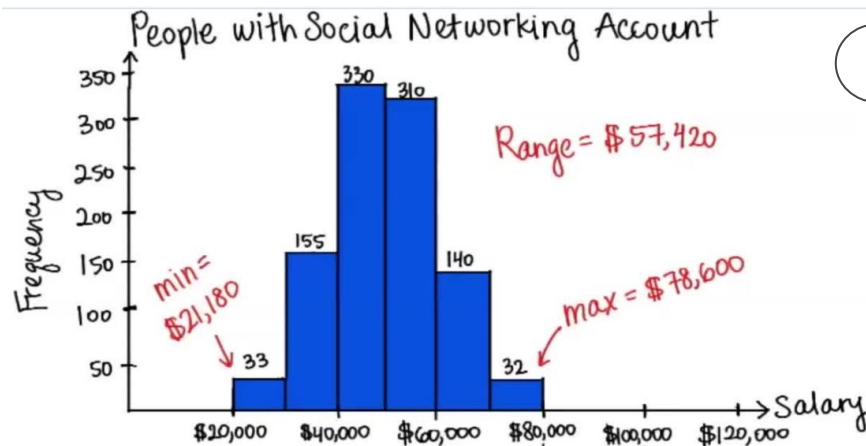


Os limites da distribuição são alterados quando novos dados são inseridos?

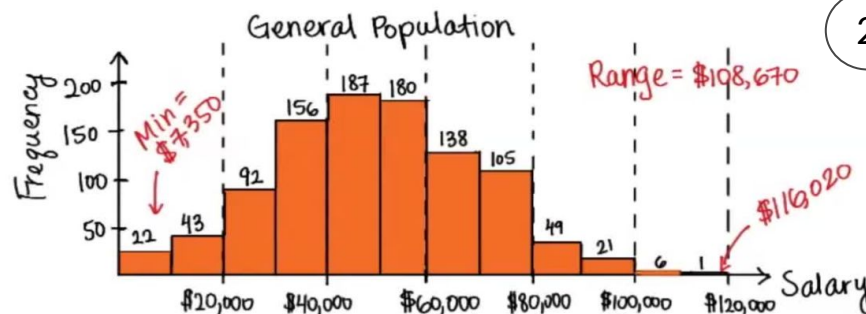
- 1) Sempre
- 2) Algumas vezes
- 3) Nunca

Quartil (Q1 - Q3)

Chop of the tails



1



2

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

Amostra 2

33,219
 36,254
38,801 Q1
 46,335
46,840 Q2 Mediana
 47,596
 55,130
56,863 Q3
 78,070
 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

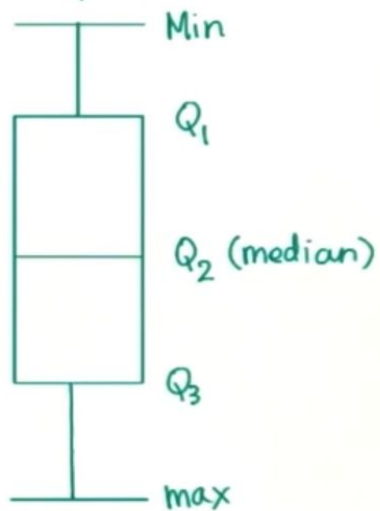
Outlier > Q3 + 1.5 x IQR

41,775

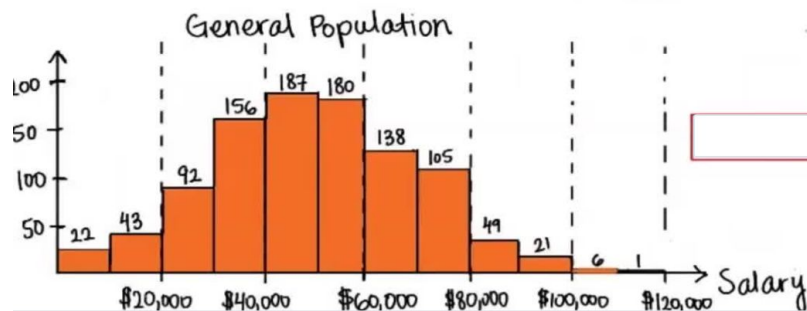
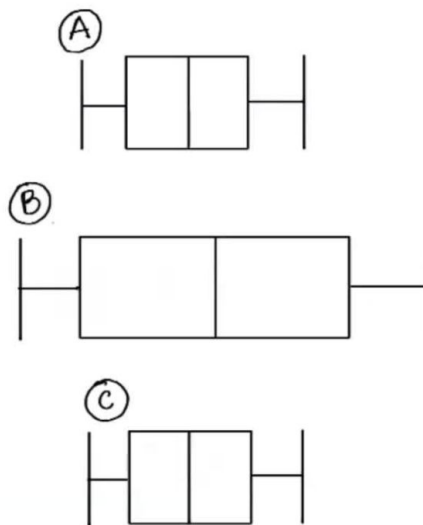
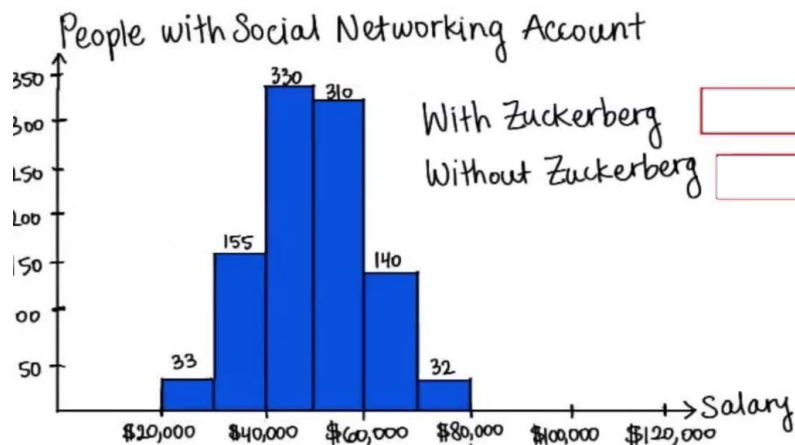
61,551

Gráficos de Caixa

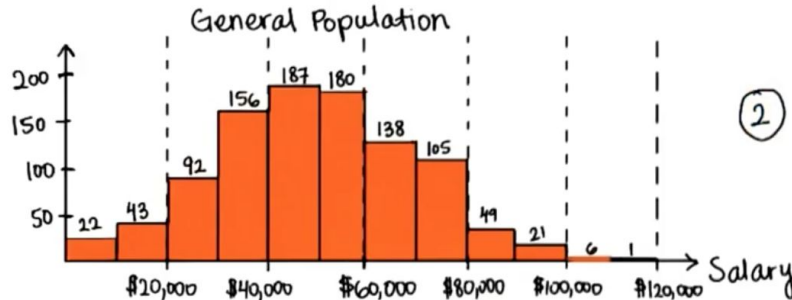
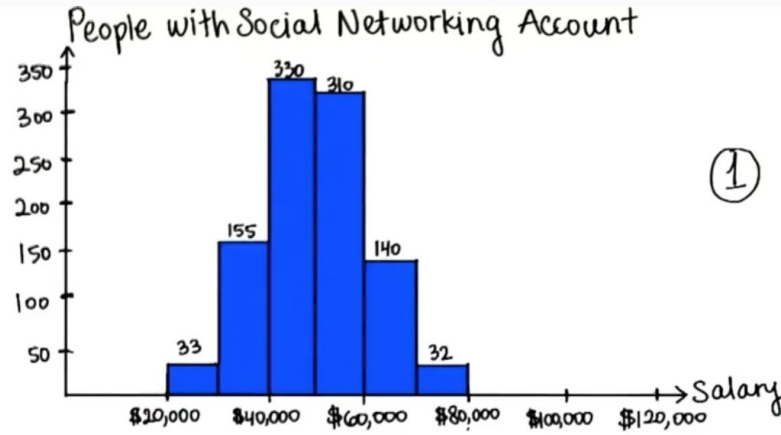
Boxplots



Quiz: gráficos de caixa



Problemas com o IQR



A média sempre será entre Q_1 and Q_3 ?

- Sim
- Não

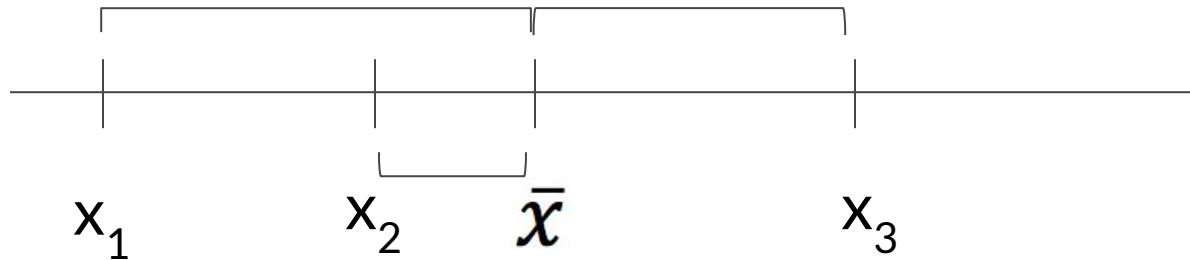
Medidas de variabilidade

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

~~Range~~

~~IQR~~

Medida de variabilidade (ideia)



Variancia

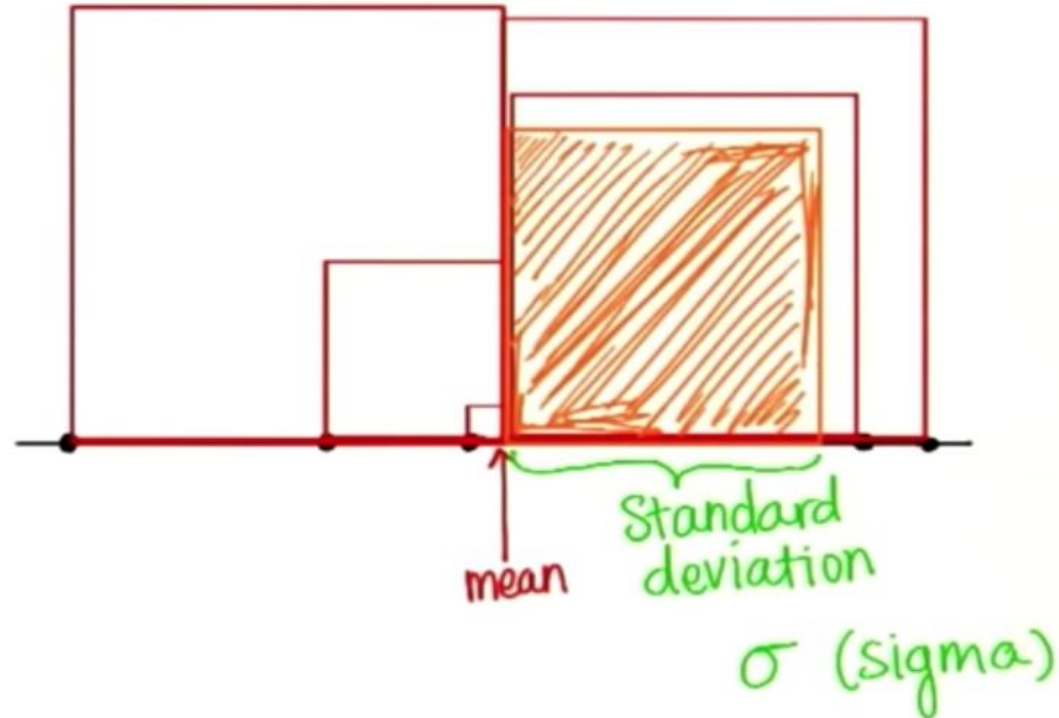
$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

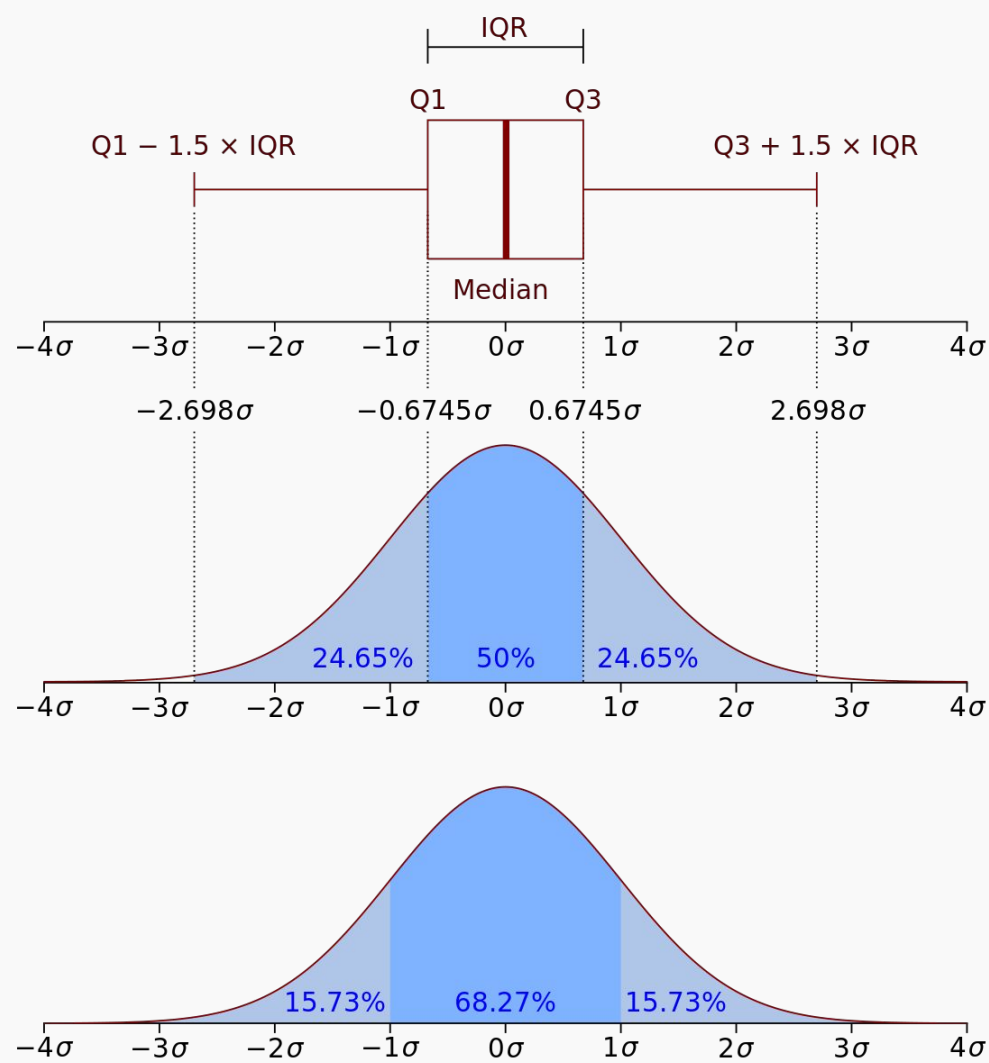
Desvio
Padrão

σ

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

Measure Variability (idea)

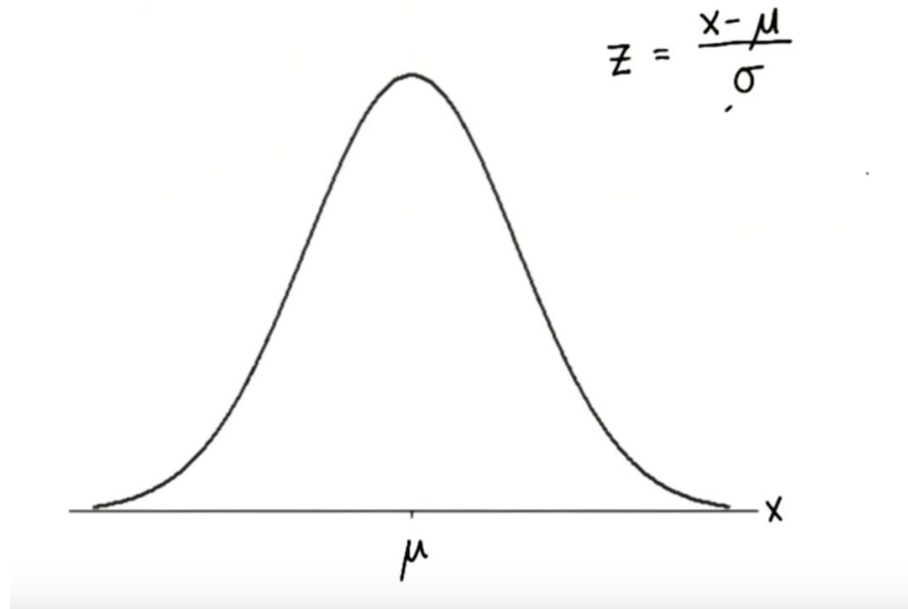




Desafio

- 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.

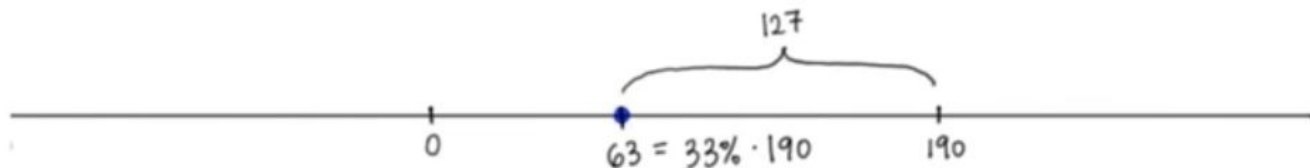
Z-Score



Quiz: Quem é mais popular?

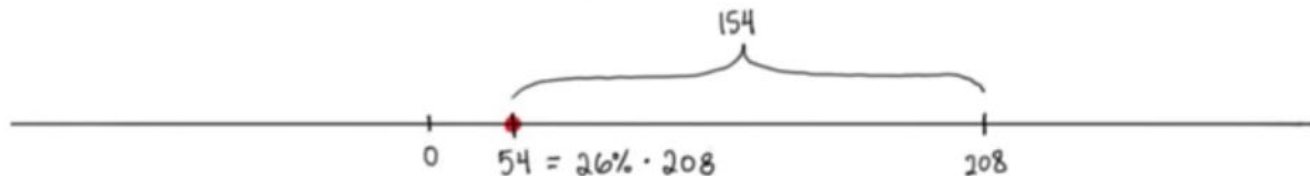
Javanildo

Facebook friends

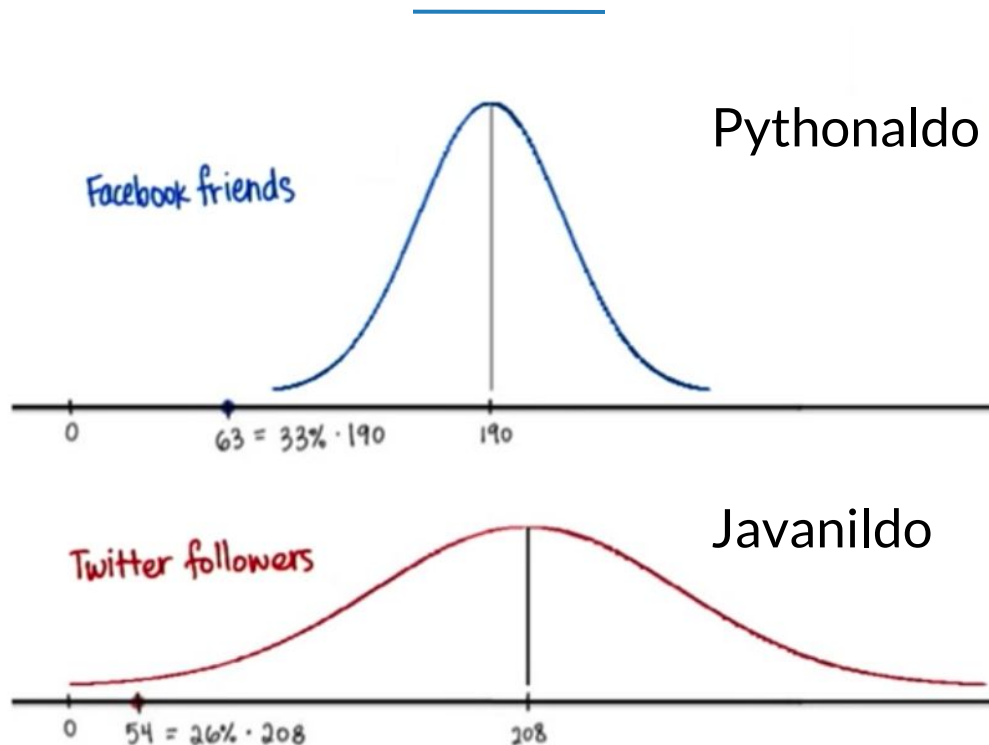


Pythonaldo

Twitter followers



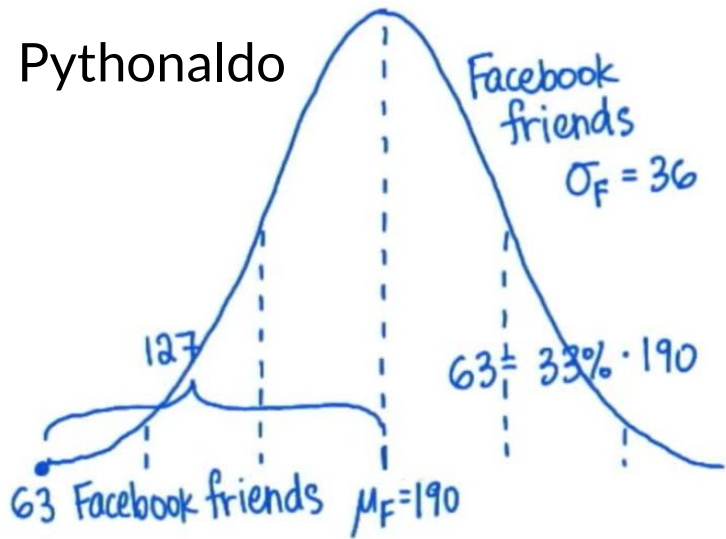
Quiz: Quem é mais popular?



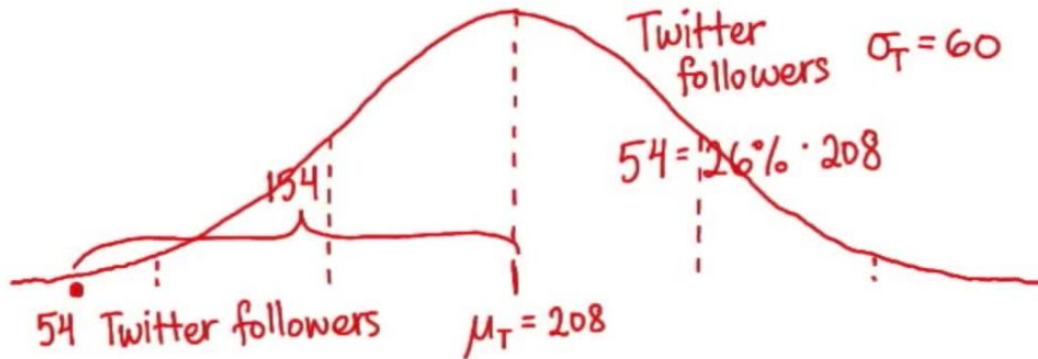
Quiz: Quem é mais popular?

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

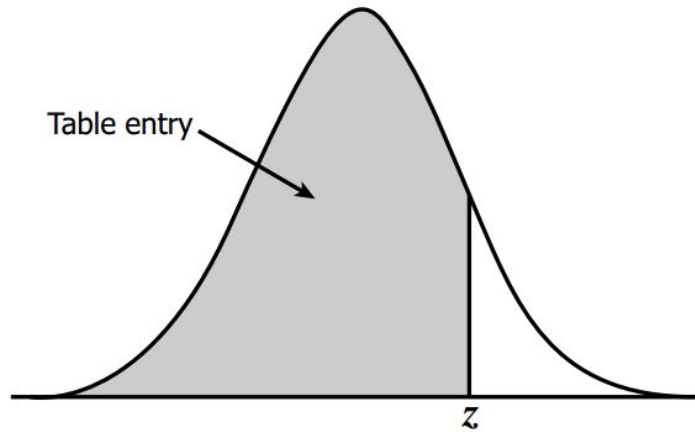
Pythonaldo



Javanildo



Z-Table



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

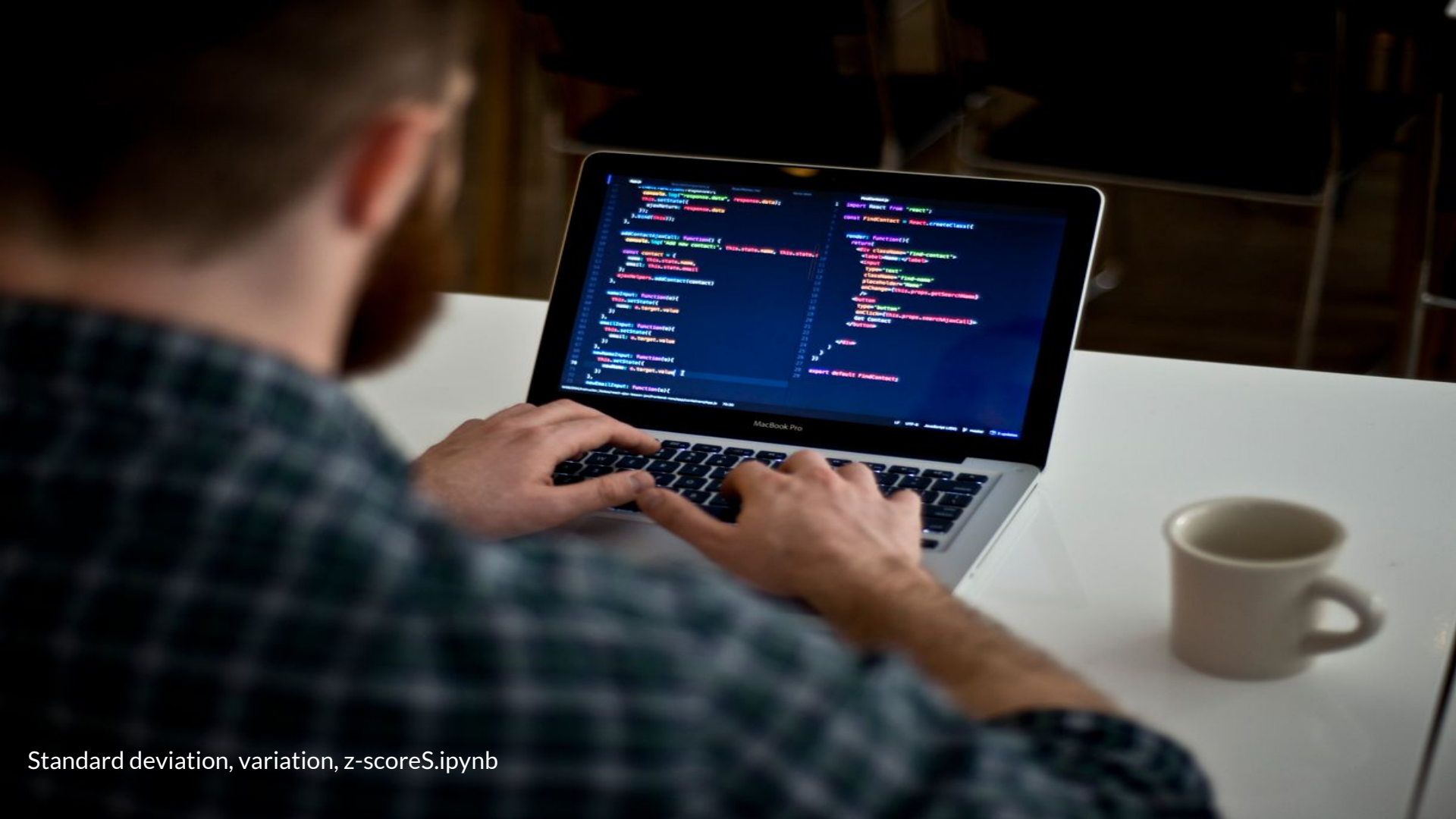
Facebook example:

$$\mu = 190$$

$$\sigma = 36$$

$$X_i = 240$$

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



Standard deviation, variation, z-scoreS.ipynb