

Estatística descritiva

Covariância e correlação

Prof. Dr. Tetsu Sakamoto

Instituto Metrópole Digital - UFRN

Sala A224, ramal 182

Email: tetsu@imd.ufrn.br





Slides e notebook em:

github.com/tetsufmbio/IMD0033/





Objetivos da aula

Obter noções básicas de:

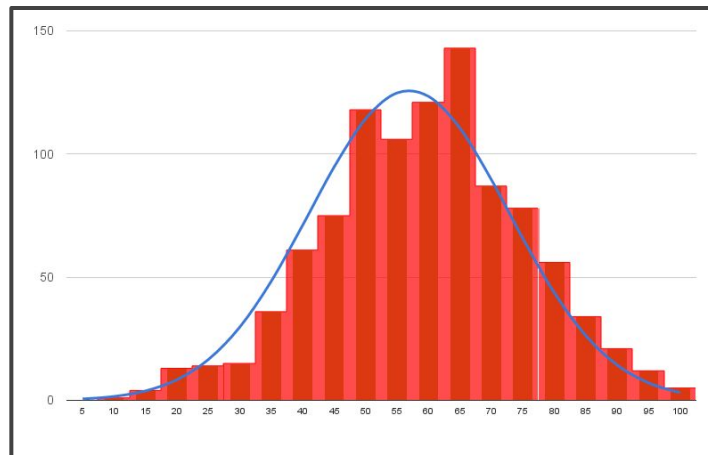
- Covariância;
- Correlação.



Até o momento...

Aprendemos a descrever estatisticamente dados que envolve uma variável (**análise univariada**);

aluno	nota
1	53,21
2	62,33
3	59,35
4	59,63
...	...

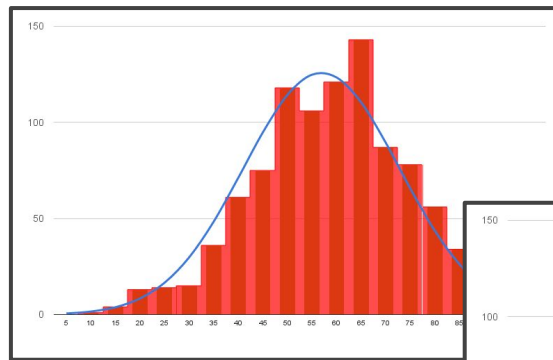




E se...

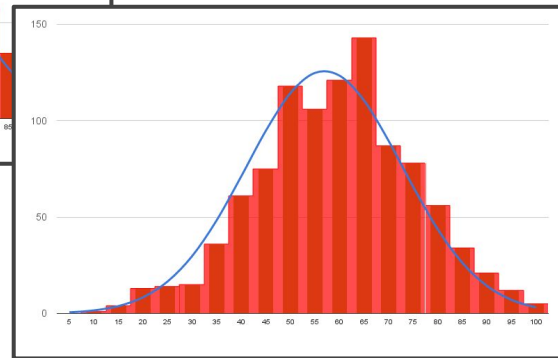
... além de termos os dados das notas, temos dados da altura de cada candidato que fez a prova. Que tipo de pergunta podemos fazer?

aluno	nota	altura
1	53,21	1,68
2	62,33	1,62
3	59,35	1,37
4	59,63	1,56
...



nota

altura



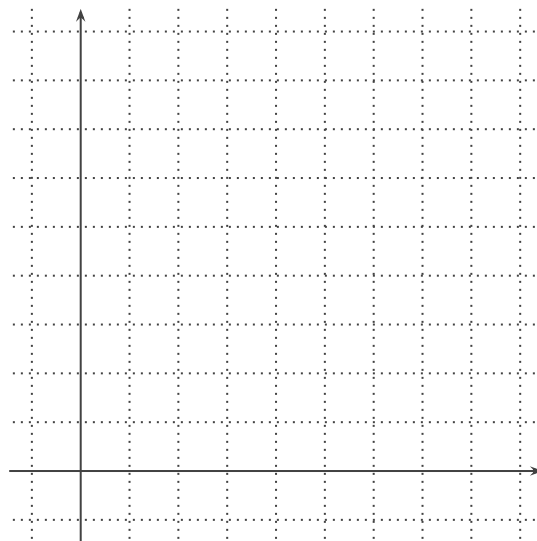
**Será que estas variáveis
se relacionam?**



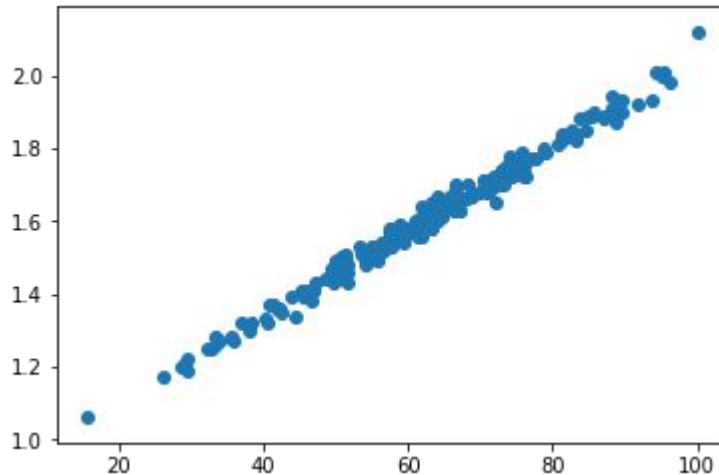


Será que estas variáveis se relacionam?

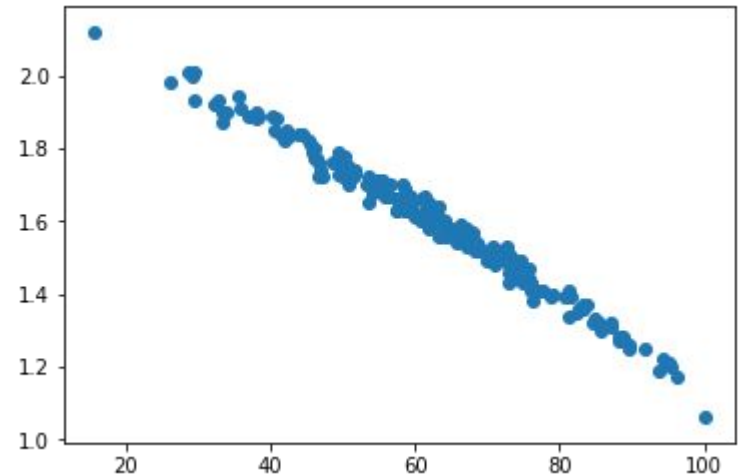
aluno	nota	altura
1	53,21	1,68
2	62,33	1,62
3	59,35	1,37
4	70,63	1,56
...



Como se comportariam duas variáveis que se relacionam?



Relação positiva



Relação negativa



Gráficos de dispersão (matplotlib.pyplot)

```
import matplotlib.pyplot as plt
```

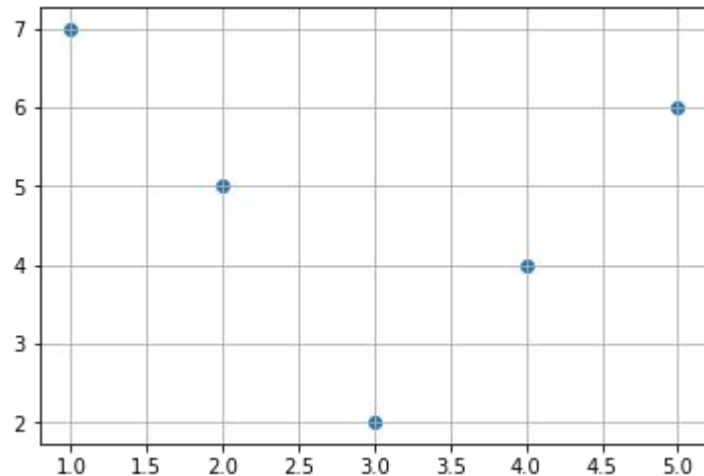
```
X = [1,2,3,4,5]
```

```
Y = [7,5,2,4,6]
```

```
plt.scatter(X, Y)
```

```
plt.grid()
```

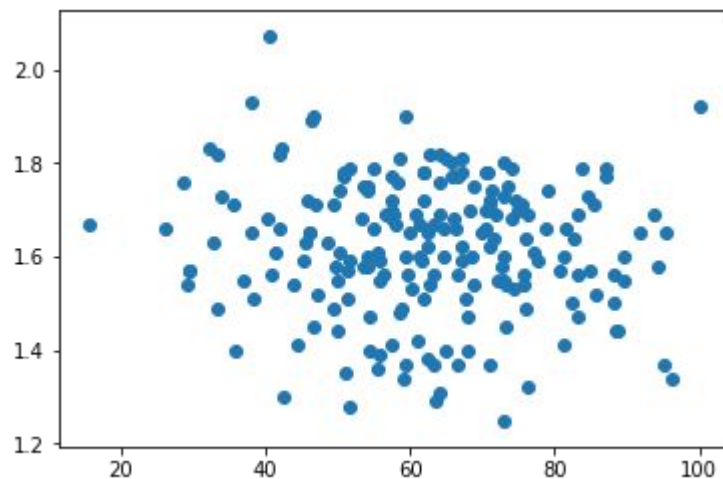
```
plt.show()
```





Quanto mais alto a pessoa, maiores/menores são as chances dele tirar notas boas?

Em outras palavras, as alturas e as notas dos candidatos relacionam?




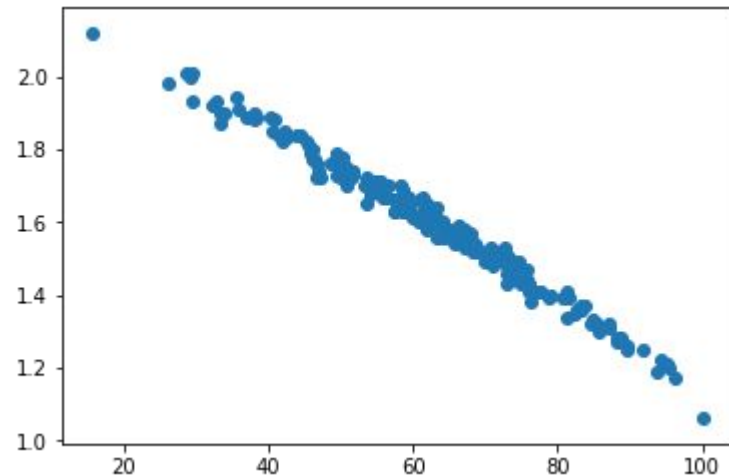
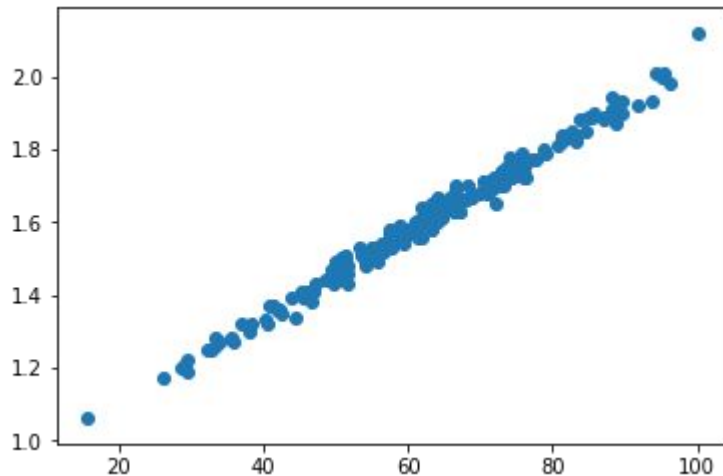



Como medir a relação entre as variáveis?

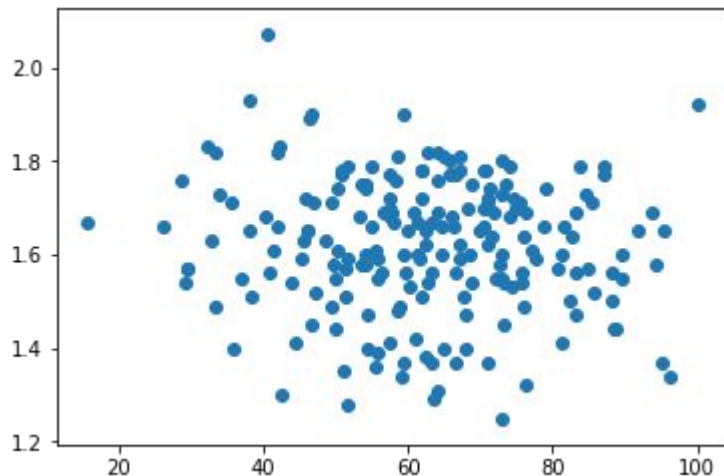
Covariância:

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$


$$\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$




$$\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$





Propriedades da Covariância

- $-\infty < \text{Cov}(X, Y) < +\infty$
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X, C) = 0$ se C é uma constante;

$$\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$



Sobre o valor numérico da covariância

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

A unidade da covariância seria: (unidade de X) * (unidade de Y);

Comparar covariância de diferentes pares de variável é difícil, pois se alterarmos a escala, a covariância muda também.

Como remover a escala dos dados? **Padronização (normalização)**



Covariância em dados padronizados

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$



Correlação (ρ) (Correlação de Pearson)

$$\textit{Correlation} = \frac{\textit{Cov} (x, y)}{\sigma x * \sigma y}$$



Propriedades da correlação

- ρ é a covariância dos dados padronizados de X e Y;
- Adimensional (lida com proporção);
- $-1 < \rho < 1$;

$$\textit{Correlation} = \frac{\textit{Cov}(x, y)}{\sigma x * \sigma y}$$



Exercícios do notebook em:

github.com/tetsufmbio/IMD0033/

