DCA- 0131: Ciência de Dados

Fundamentos de Estatística

Luiz Affonso Guedes - affonso@dca.ufrn.br

Conteúdo

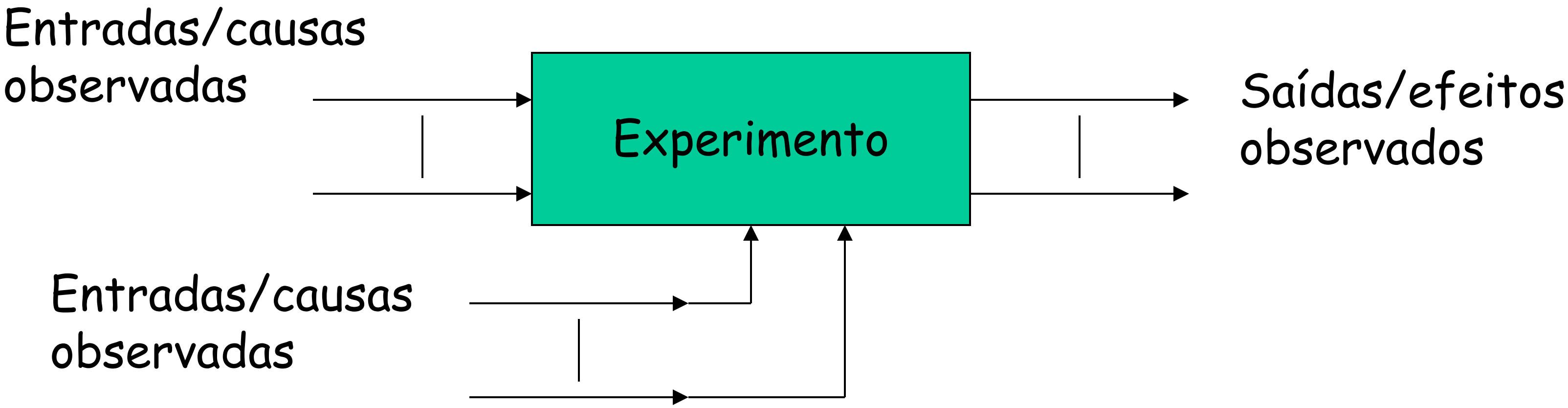
- Conceitos de estatística
- Métricas estatísticas

Introdução

- Fenômenos Determinísticos
  - Conhecidos com certeza
  - Não sujeitos às leis do acaso
    - Ex.: o ano atual, idade de uma pessoa jovem
- Fenômenos Probabilísticos
  - Não conhecidos com certeza
  - Sujeitos às leis do acaso
    - Ex.: face de um dado, se vai chover amanhã, se o time de futebol vai ser campeão

Introdução

- Experimentos que ao serem repetidos nas mesmas condições não produzem o mesmo resultado são denominados de experimentos aleatórios.
- Mas por quê isto ocorre?



Introdução

- Sistemas Reais
  - Composição de parte determinística com parte probabilística (randômica)

Introdução

- Sistemas Determinísticos:
  - Modelagem baseada em leis físicas e matemática clássica (eqs. diferenciais, ...)
  - Como obter modelos e seus parâmetros do modelo?

Introdução

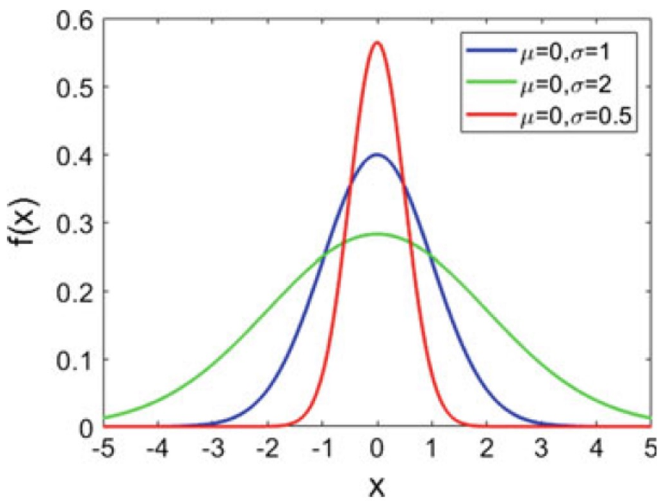
- Sistemas Randômicos
  - Modelagem baseada em dados.
  - Como obter os dados?
  - Quantos dados são necessários para se ter um bom modelo?

$$P(x_i) = \lim_{K \to \infty} \frac{K_i}{K}$$

  - Como obter modelos e parâmetros?

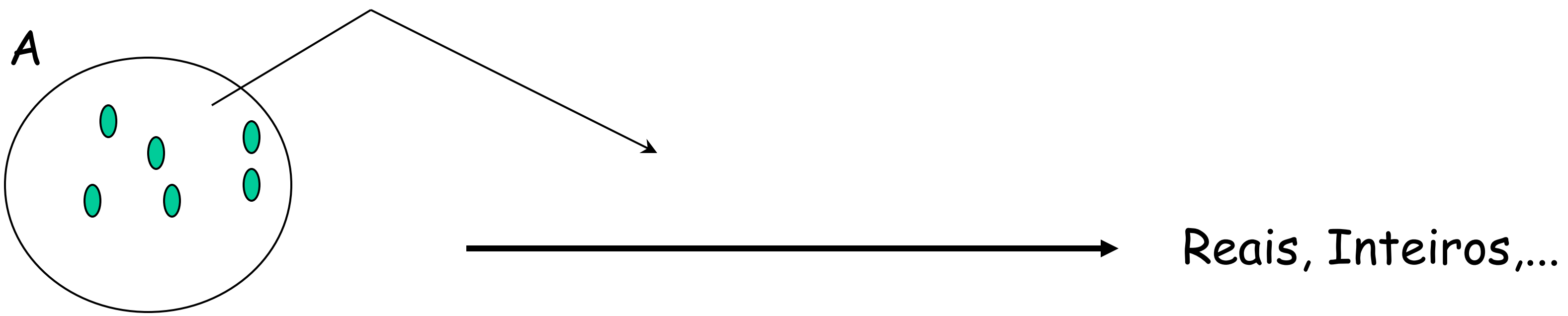
$$f(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(\boldsymbol{x};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$



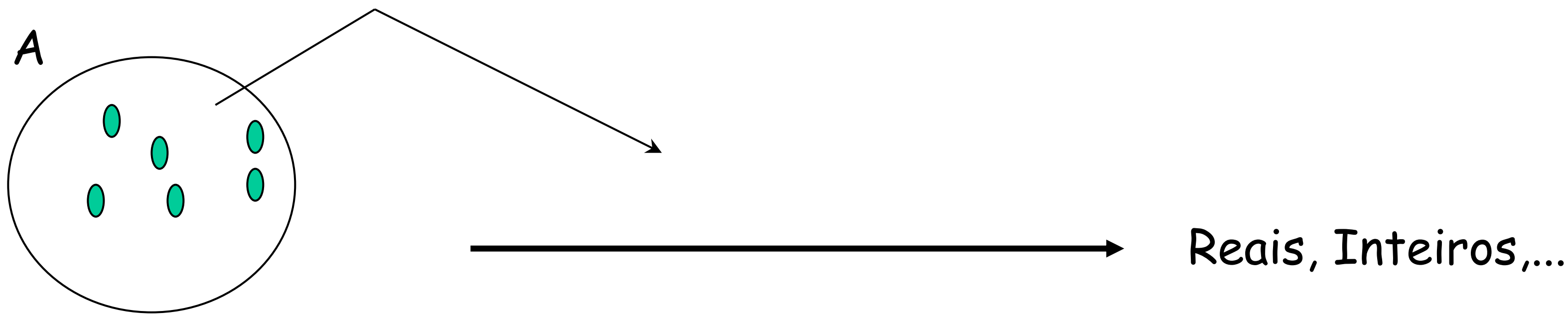
Variáveis Aleatórias

- Como devemos descrever um experimento aleatório?
  - Uma boa forma seria associar cada experimento a valores numéricos.



Variáveis Aleatórias

- **Definição:**
  - Dado um espaço de probabilidade descrito por (S,P), uma variável aleatória (v.a.) sobre esse espaço é uma função sobre S.
  - $X(A) = f(A)$
  - Variável aleatória é uma função dos eventos, não uma variável.



Variáveis Aleatórias

- Exemplo:

  Experimento Probabilístico de se lançar dois dados.

  - Então, seu espaço amostral é:
    - S = {(1,1), (1,2), (1,3), ..., (2,1), (2,2), ..., (6,6)}
    - X(e) – V.A. correspondendo a soma dos valores de cada face.
    - Y(e) – V.A. par ou ímpar, caso a soma dê par ou ímpar, respectivamente.

Variáveis Aleatórias

e	X(e)	Y(e)	P(e)
(1,1)	2	par	1/36
(1,2)	3	ímpar	1/36
...			
(2,1)	3	ímpar	1/36
...			
(6,6)	12	par	1/36

Espaço de X(e) é {2,3,4,5,6,7,8,9,10,11 ,12)

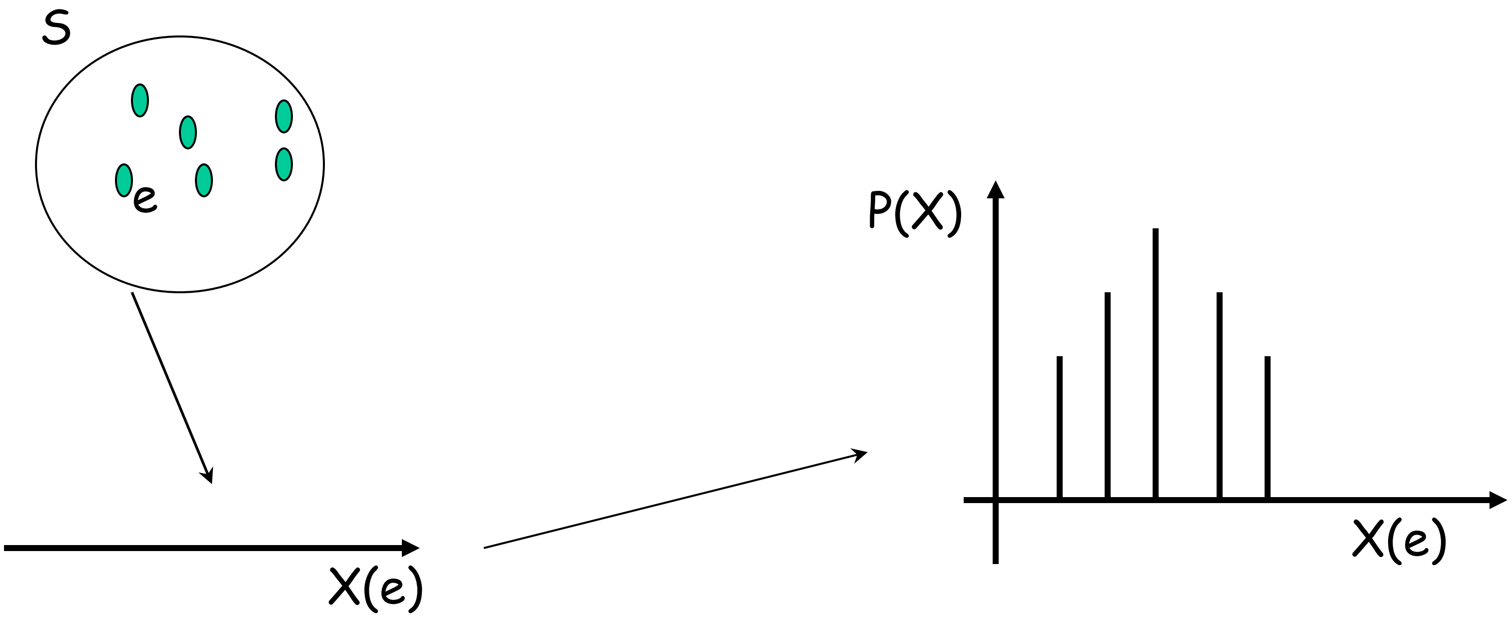
Espaço de Y(e) é {par, ímpar)

Y(e)	P(Y)
par	18/36
ímpar	18/36

P(X=7) = 1/6

X(e)	P(X)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Variáveis Aleatórias



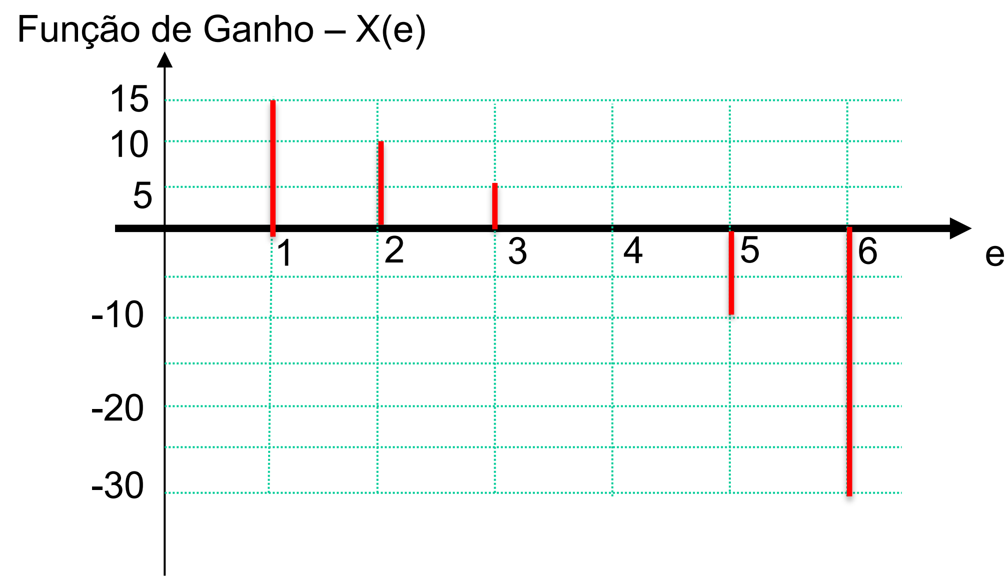
Variáveis Aleatórias

- V.A. Discreta – quando os valores da variável formam um conjunto enumerável.
  - V.A. Discreta Ordinal - quando os valores têm conceito de ordem.
    - Exemplos: idade, qualidade (ruim, razoável, bom excelente)
  - V.A Discretas Categórica – quando os valores pertencem a categorias sem conceito de ordem.
    - Exemplos: sexo, cor da pele, profissão.
- V.A. Contínua – quando os possíveis resultados do experimento são representados por infinitos valores em um intervalo contínuo.
  - Exemplos: temperatura, tempo, valor de tensão.

Esperança Matemática de V.A.

- Colocação de um problema:
  - Seja um jogo de dado expresso pela V.A. descrita na tabela abaixo:
    - Quais são as suas chances nesse jogo?

Face	Ganho
1	+15
2	+10
3	+5
4	0
5	-10
6	-30



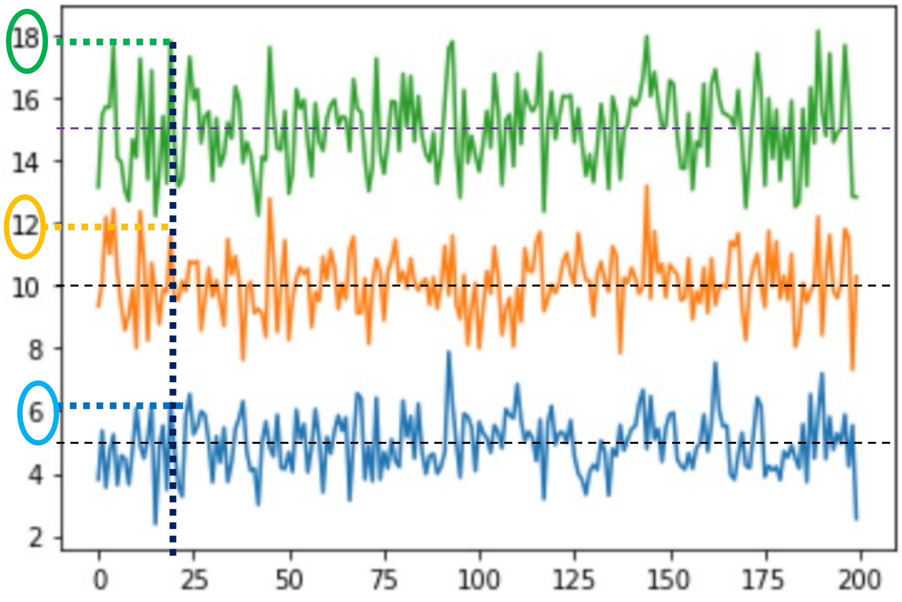
Esperança Matemática de V.A.

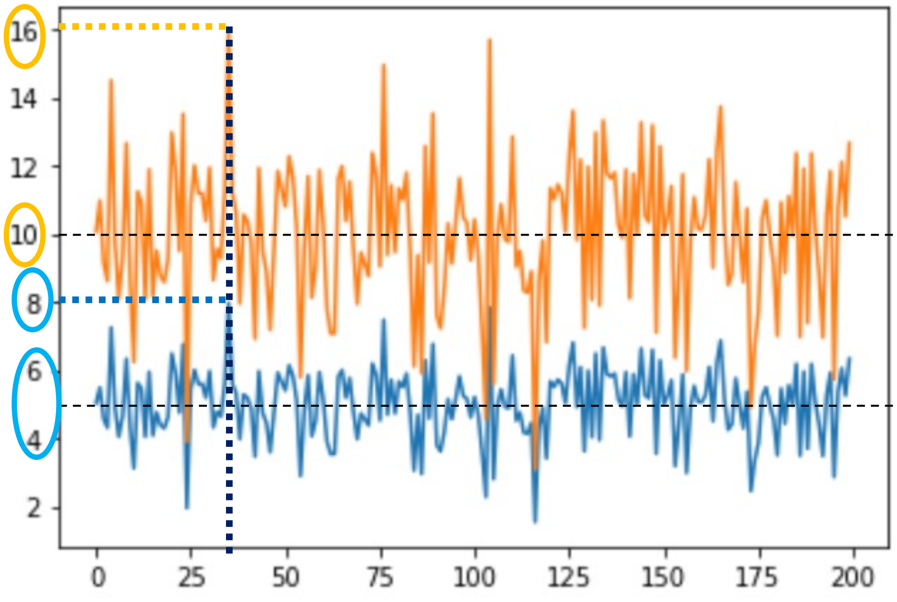
- Esperança para V.A. Discretas:
  - $E[X(e)] = \sum_{e \in S} X(e)P(e)$
  - ou em termos de um conjunto enumerável,
  - $E[X] = \sum_{i=1\ldots\infty} \{x_i P[X=x_i]\}$
  - $E[X] = x_1P(x_1) + x_2P(x_2) + \ldots + x_nP(x_n)$
  - Desde que $\sum_{i=1\ldots\infty} P(x_i) = 1$,

    Então, $E[X]$ é basicamente uma média ponderada

Esperança Matemática de V.A.

- Propriedade da linearidade:
  - $E[X+Y] = E[X] + E[Y]$
  - $E[cX] = c.E[X]$





- $E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$

Esperança Matemática de V.A.

- Exercício:
  - Uma loteria vende 100 bilhetes a R$1,50 cada. Sendo que o prêmio é de R$100,00, qual é a sua esperança de ganho se você jogou 1, 2,10 ou 100 bilhetes?

Esperança Matemática de V.A.

- Exercício: Uma loteria dá 200 prêmios de R$5,00, 20 de R$25,00 e 5 de R$100,00. Admitindo-se que sejam emitidos e vendidos 10.000 bilhetes, qual seria o preço justo para um bilhete?

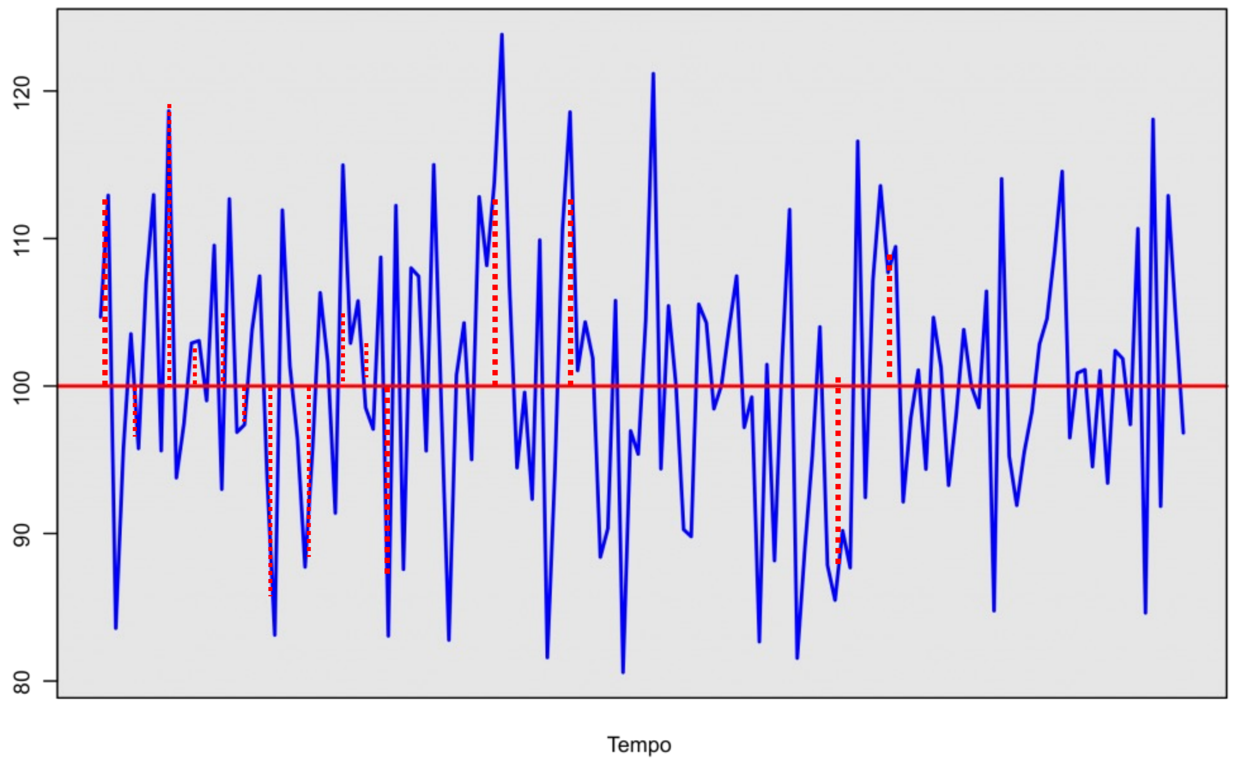
- Exerício: Uma moeda viciada, de modo que P(cara)=3/4 e P(coroa)=1/4, é lançada 3 vezes. Seja X a v.a. que representa o número de caras ocorrido. Ache a distribuição de probabilidade e a média de X.

Esperança Matemática de V.A.

- Esperança como valor médio

$$E[X] : \overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\overline{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

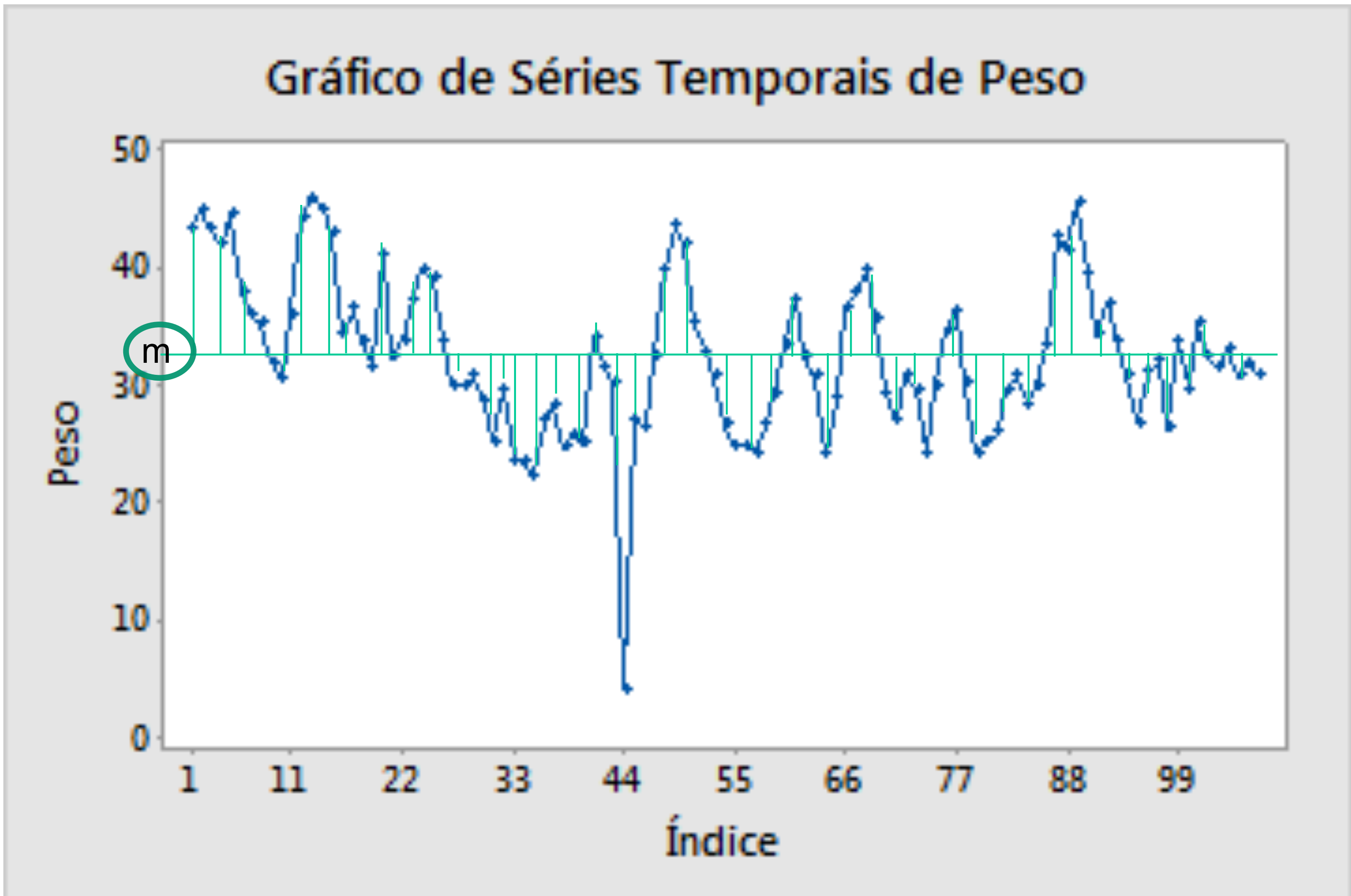


Funções de uma Variável Aleatória

- Variância:
  - Denotada por $\sigma^2(X)$, a variância de uma V.A. X é definida matematicamente por:
    - $\sigma^2(X) = E[(X - E[X])^2]$,
    - Onde, $E[X] = m_x$, valor médio de X
    - Interpretação para variância.
    - Mostrar que:
      - $\sigma^2(X) = E[X^2] - (E[X])^2 = E[X^2] - m_x^2$
    - O que significa: $E[X^2]$?
- Desvio-Padrão: é raiz quadrada da variância
  - $\sigma(X)$

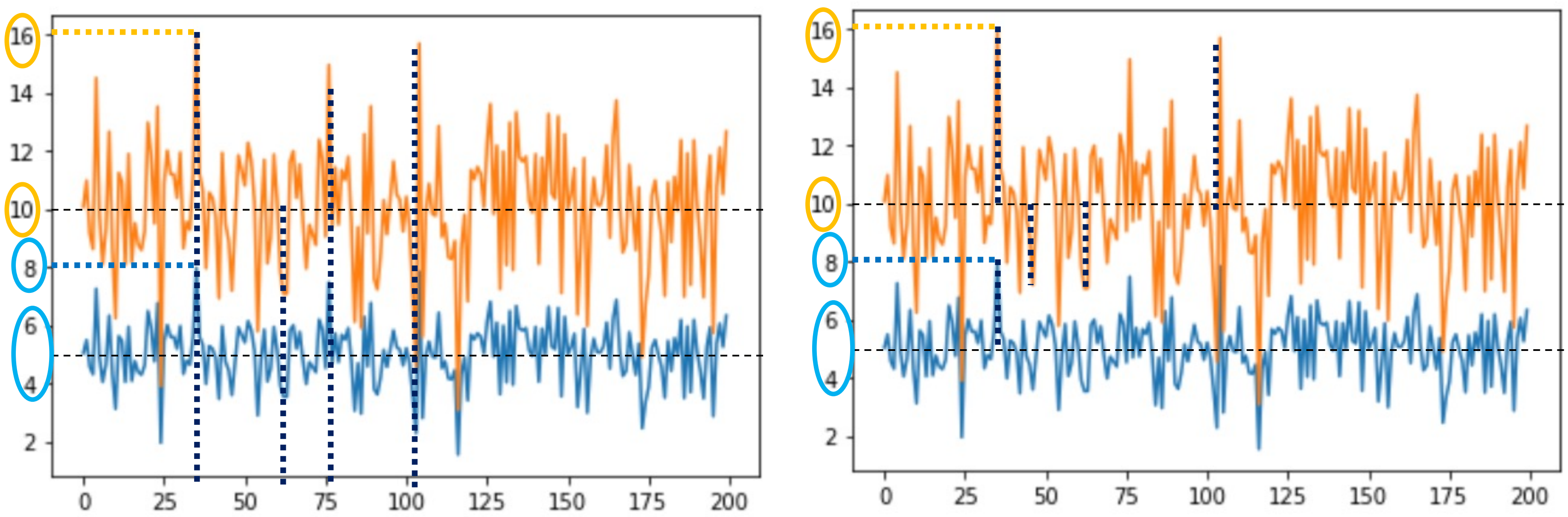
Funções de uma Variável Aleatória

- Variância:
  - $\sigma^2(X) = E[(X - m)^2] = E[X^2] - m^2 = \frac{\sum_{i=1}^{n}(xi - m)^2}{n}$
  - $E[X^2]$?



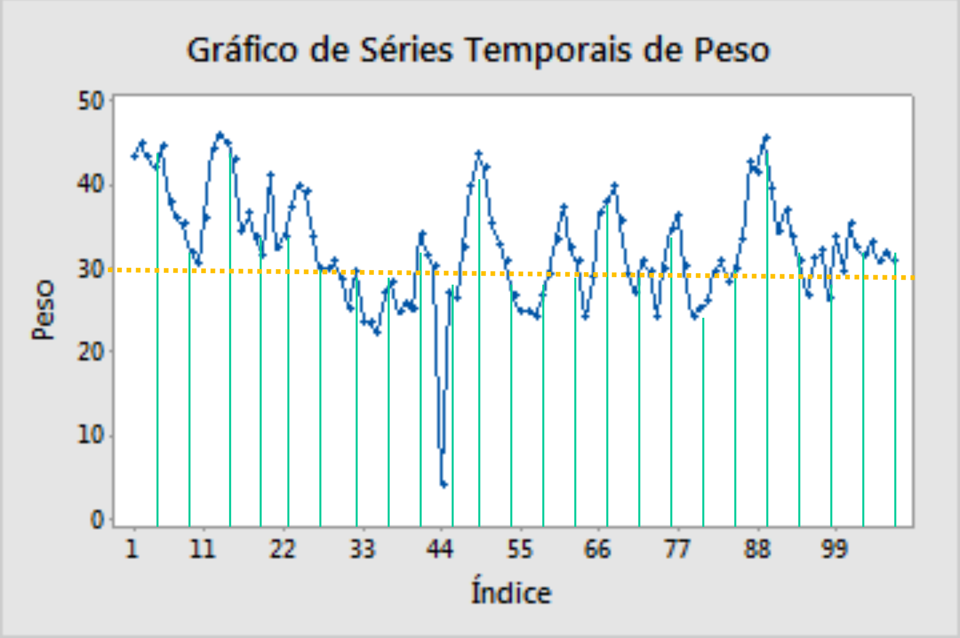
Funções de uma Variável Aleatória

- Se Y= a X, qual é o valor de σ²(Y)?
- $\sigma^2(X) = \frac{\sum_{i=1}^{n}(xi-m)^2}{n}$
- $E[x^2] = \frac{\sum_{i=1}^{n}(xi)^2}{n}$



Funções de uma Variável Aleatória

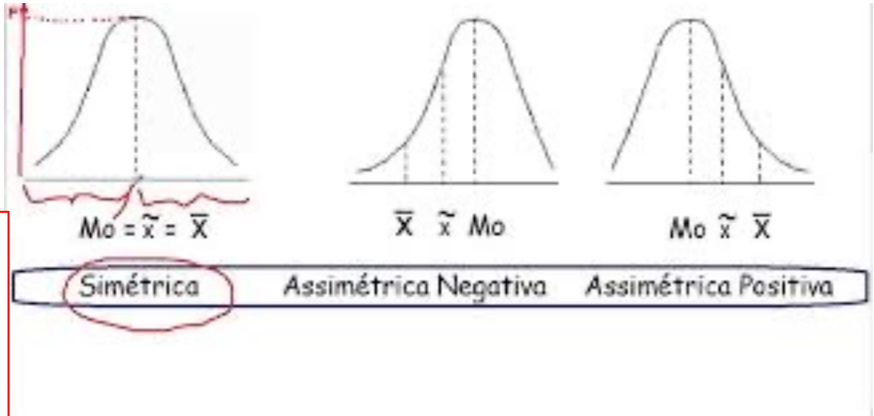
- Momentos de V.A.
  - E[X] - primeiro momento de X → $M1 = (X_1 + X_2 + ... + X_n)/n$
  - E[X²] - segundo momento de X → $M2 = (X_1^2 + X_2^2 + ... + X_n^2)/n$
  - E[X³] - terceiro momento de X → $M3 = (X_1^3 + X_2^3 + ... + X_n^3)/n$
  - E[X⁴] - quarto momento de X → $M4 = (X_1^4 + X_2^4 + ... + X_n^4)/n$
  - $\sigma^2(X) = E[(X - E[X])^2]$, segundo momento centralizado → $\sigma^2(X) = [(X_1 - M1)^2 + (X_2 - M1)^2 + ... + (X_n - M1)^2]/n$



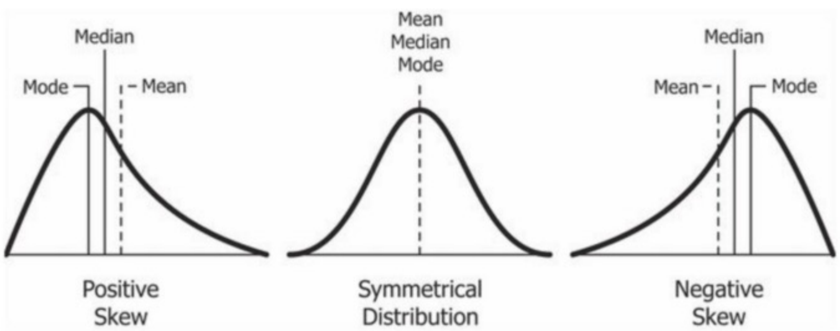
  - $\sigma^2(X) = M2 - M1^2$
  - $\sigma^2(X) = E[X^2] - E[X]^2$

Funções de uma Variável Aleatória

- Medias Estatísticas - Momentos de V.A.
  - Média → Primeiro Momento
  - Variância → Dispersão → Segundo momento central
  - Assimetria
    - Skewness



$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}.$$



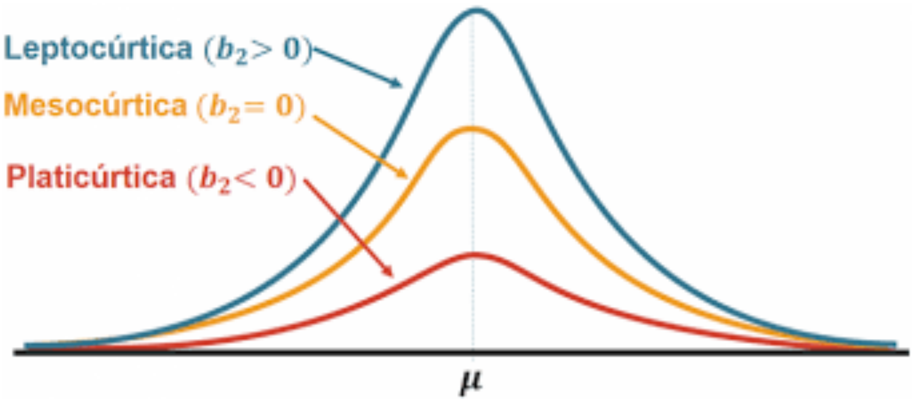
- Medida de Forma

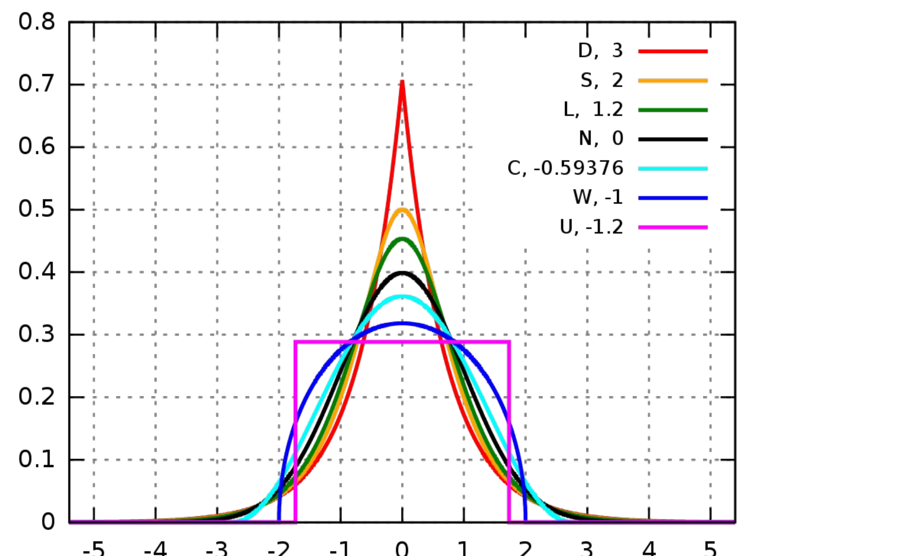
Funções de uma Variável Aleatória

- Medias Estatísticas - Momentos de V.A.
  - Média → Primeiro Momento
  - Variância → Dispersão → Segundo momento central
  - Assimetria
  - Medida de Forma
    - Curtose

$$\boxed{C_m = \frac{m_4}{(m_2)^2} = \frac{m_4}{S^4}}$$

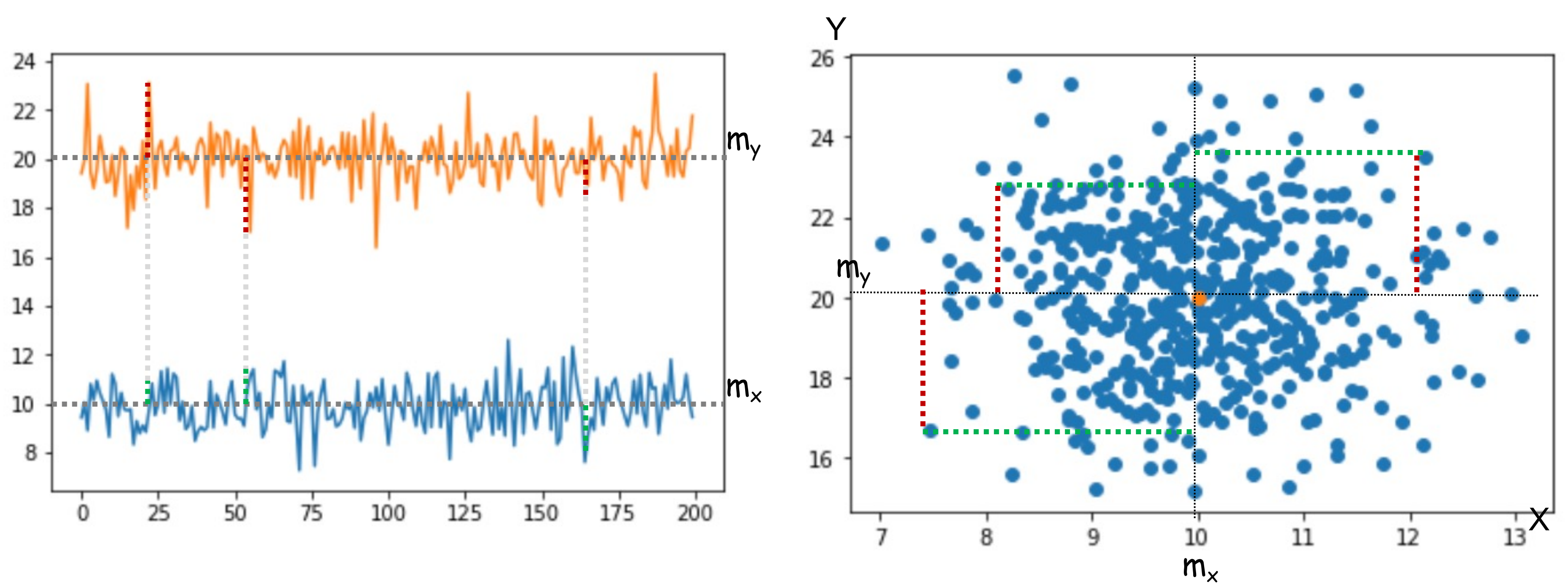
$$\boxed{\frac{m_4(\mu)}{\sigma^4} + (-3)}$$





Covariância

- Sejam duas v.a. X e Y.
  - A covariância fornece uma medida de dispersão em relação às suas médias.
  - Cov(X,Y) = E{ (X-E[X]) (Y-E[Y]) }
  - Cov(X,Y) = E{ (X-m_x) (Y-m_y)}
  - $$C_{x,y} = \frac{\sum_{i=1}^{n}(xi - mx)(yi - my)}{n}$$



Covariância

The **covariance** between X and Y is defined as

$$\mathrm{Cov}(X, Y) = E\Big[(X - EX)(Y - EY)\Big] = E[XY] - (EX)(EY).$$

Covariância

- Covariância, Correlação e Variância
  - Cov(X,Y) = E{ (X-E[X]) (Y-E[Y]) }
  - Cov(X,Y) = E[XY] – E[X] E[Y]
  - $C_{x,y} = R_{x,y} - m_x.m_y$; $C_{x,y} = \frac{\sum_{i=1}^{n}(xi-mx)(yi-my)}{n}$
  - $R_{x,y} = \frac{\sum_{i=1}^{n}(xi)(yi)}{n} =$ E[XY] ←Correlação
  - Se m_x=0 ou m_y=0 → Covariância é igual a Correlação
  - Se X=Y → Cov(X,Y) = Cov(X,X) = Var(X)

Covariância

- Propriedades

Lemma 5.3
The covariance has the following properties:

1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$;
2. if X and Y are independent then $\mathrm{Cov}(X, Y) = 0$;
3. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$;
4. $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$;
5. $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$;
6. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$;
7. more generally,

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

$$\begin{aligned}\mathrm{Cov}(X + Y, Z) &= E[(X + Y)Z] - E(X + Y)EZ \\ &= E[XZ + YZ] - (EX + EY)EZ \\ &= EXZ - EXEZ + EYZ - EYEZ \\ &= \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z).\end{aligned}$$

$$\mathrm{Cov}(X, Y) = E[XY] - EXEY = 0.$$

Covariância

- Variância da Soma de 02 VAs
  - Z = X+Y

$$
\begin{aligned}
\mathrm{Var}(Z) &= \mathrm{Cov}(Z, Z) \\
&= \mathrm{Cov}(X + Y, X + Y) \\
&= \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, Y) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).
\end{aligned}
$$

  - Para o caso geral:

$$
\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y) \qquad (5.21)
$$

Covariância

- Dado:
  - $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
    - Se X e Y são independente → $Cov(X,Y) = 0$
    - Se X e Y tendem a variar no mesmo sentido, a $Cov(X,Y)$ será positiva.
    - Se X e Y tendem a variar em sentidos opostos, a $Cov(X,Y)$ será negativa.

Covariância

- Covariância e Coeficiente de Correlação
  - Sejam duas VAs X e Y. Se X'= aX e Y'= bY:
    - Cov(X',Y') = ab Cov(X,Y)
      - Cov(X,Y) = E[XY] – E[X] E[Y]
    - Então, a covariância depende das escalas das VAs.
- Seria interessante se trabalhar com uma medida de dispersão independente de escala!

Covariância

- Coeficiente de Correlação
  - $\rho(X,Y) = Cov(X,Y) / (\sigma(X)\, \sigma(Y))$
  - $|\rho(X,Y)| \leq 1 \leftarrow$ normalização da covariância

  - Se $X' = aX$ e $Y' = bY$:
    - $Cov(X',Y') = ab\, Cov(X,Y)$
    - $\rho(X',Y') = ab\, Cov(X,Y) / (a\, \sigma(X)\, b\sigma(Y))$
    - $= \rho(X,Y)$
  - Se $Y = aX + b \rightarrow \rho(X,Y) = signal(a) \cdot 1$

Covariância

- Coeficiente de Correlação
  - Propriedades

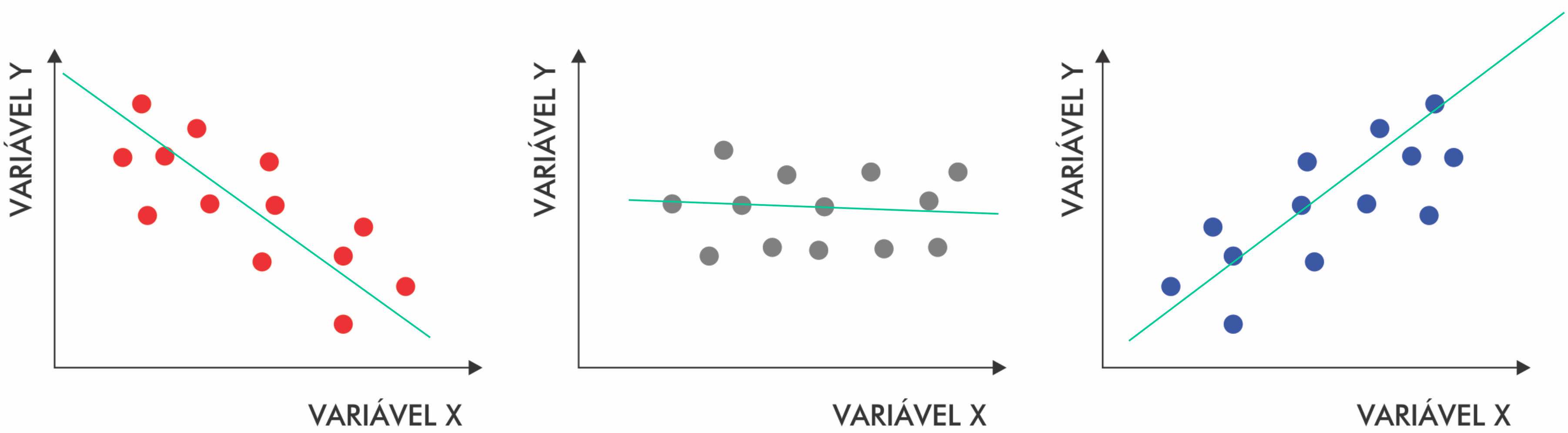
Properties of the correlation coefficient:

1. $-1 \leq \rho(X, Y) \leq 1$;
2. if $\rho(X, Y) = 1$, then $Y = aX + b$, where $a > 0$;
3. if $\rho(X, Y) = -1$, then $Y = aX + b$, where $a < 0$;
4. $\rho(aX + b, cY + d) = \rho(X, Y)$ for $a, c > 0$.

Covariância

- Coeficiente de Correlação

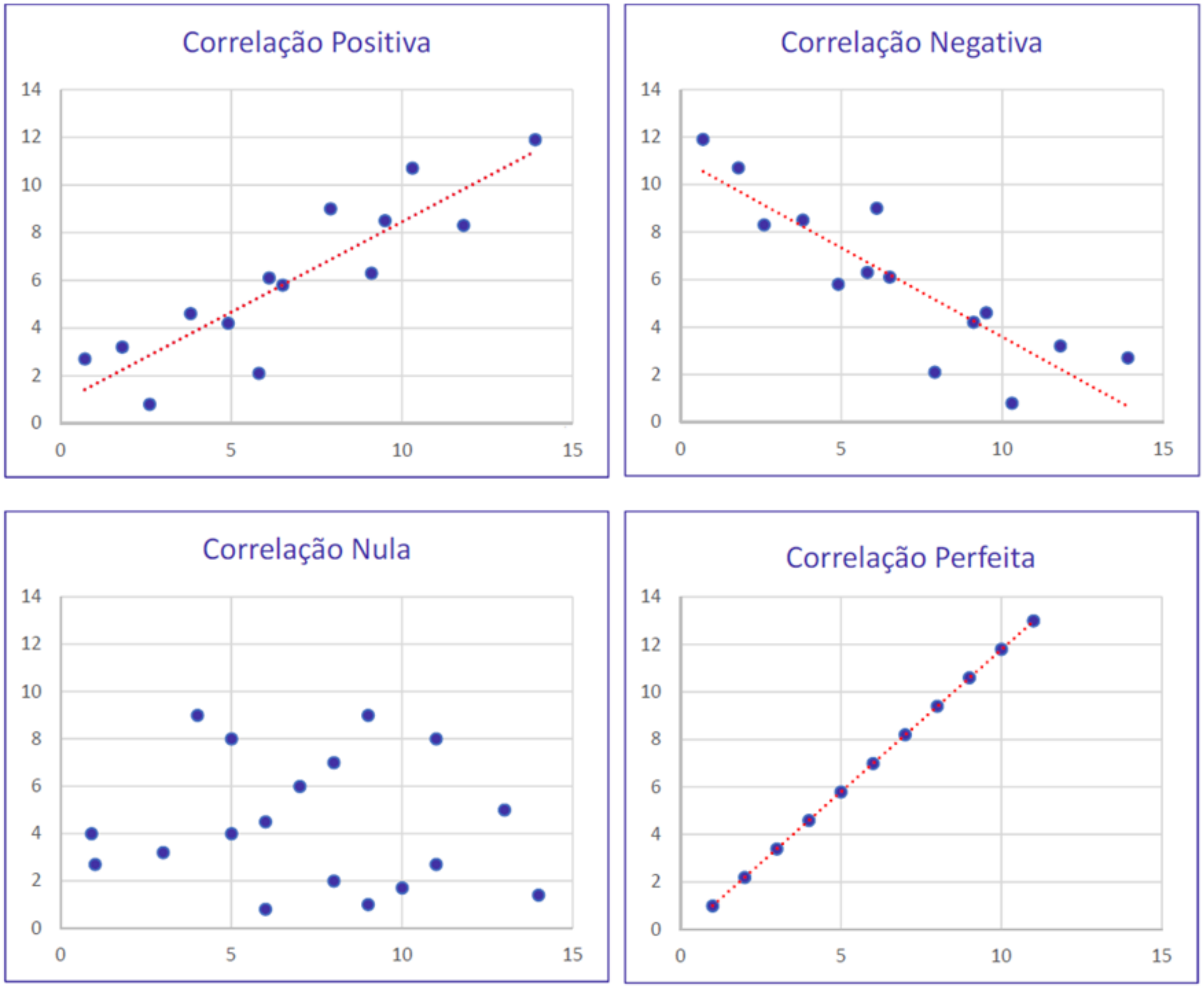
$$r_{xy} = \frac{Cov(X,Y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Covariância

- Coeficiente de Correlação

$$r_{xy} = \frac{Cov(X,Y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

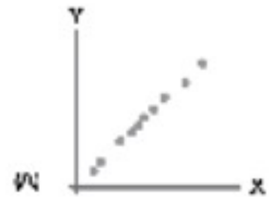


Covariância

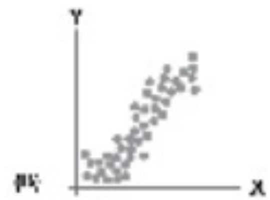
- Exercício

$$r_{xy} = \frac{Cov(X,Y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

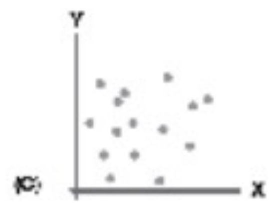
(I) correlação positiva entre x e y



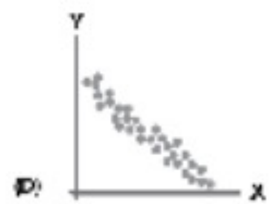
(II) correlação positiva entre x e y



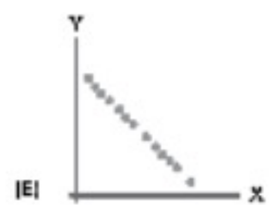
(III) correlação positiva entre x e y



(III) Forte correlação positiva entre x e y



(V) Nenhuma correlação entre x e y



Descorrelação

- **Descorrelação**
  - $Cov(X,Y) = C_{x,y} = 0 \rightarrow \rho(X,Y) = 0$;
    - $Cov(X,Y) = E[XY] - E[X]\,E[Y]$
  - $E[XY] = E[X]\,E[Y] \leftarrow$ independência implica em descorrelação, mas o inverso não é verdadeiro.
  - $Var(X+Y) = Var(X) + Var(Y) + 2.Cov(X,Y)$
    - Basta ser descorrelacionado, não necessariamente independente.

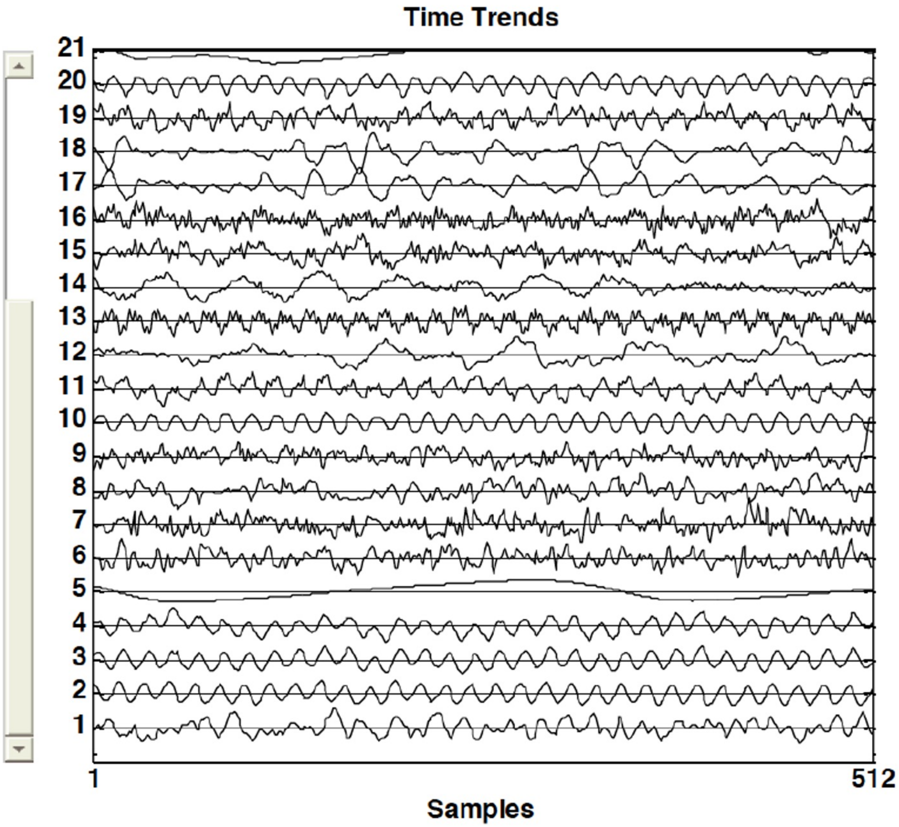
Matriz de Covariância

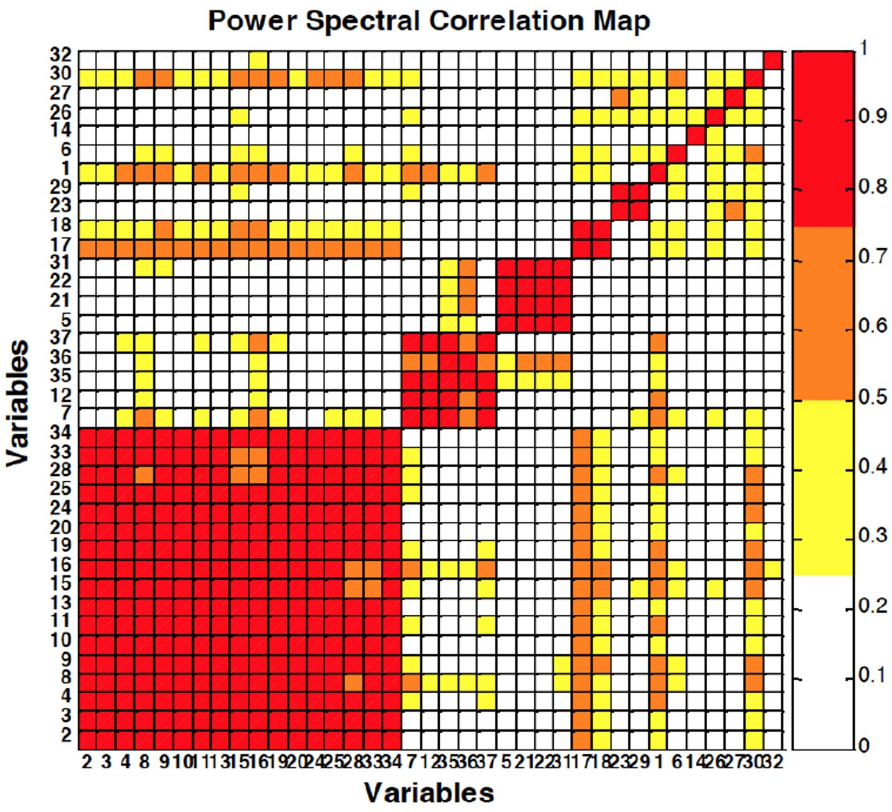
- Dada 03 VAs: X, Y, e Z.
- Temos:
  - $Cov(X,X) = C_{x,x} = Var(X)$
  - $\boxed{Cov(X,Y) = C_{x,y} = C_{y,x} = \frac{\sum_{i=1}^{n}(xi-mx)(yi-my)}{n}}$
  - $\boxed{Cov(X,Z) = C_{x,z} = C_{z,x} = \frac{\sum_{i=1}^{n}(xi-mx)(zi-my)}{n}}$
  - $\boxed{Cov(Y,Z) = C_{y,z} = C_{z,y} = \frac{\sum_{i=1}^{n}(yi-mx)(zi-my)}{n}}$
- A Matriz de Covariância P é dada por:

$$P = \begin{bmatrix} C_{x,x} & C_{x,y} & C_{x,z} \\ C_{y,x} & C_{y,y} & C_{y,z} \\ C_{z,x} & C_{z,y} & C_{z,z} \end{bmatrix} = \begin{bmatrix} Var(X) & \boxed{C_{x,y}} & \boxed{C_{x,z}} \\ \boxed{C_{x,y}} & Var(Y) & \boxed{C_{y,z}} \\ \boxed{C_{x,z}} & \boxed{C_{y,z}} & Var(Z) \end{bmatrix}$$

Covariância

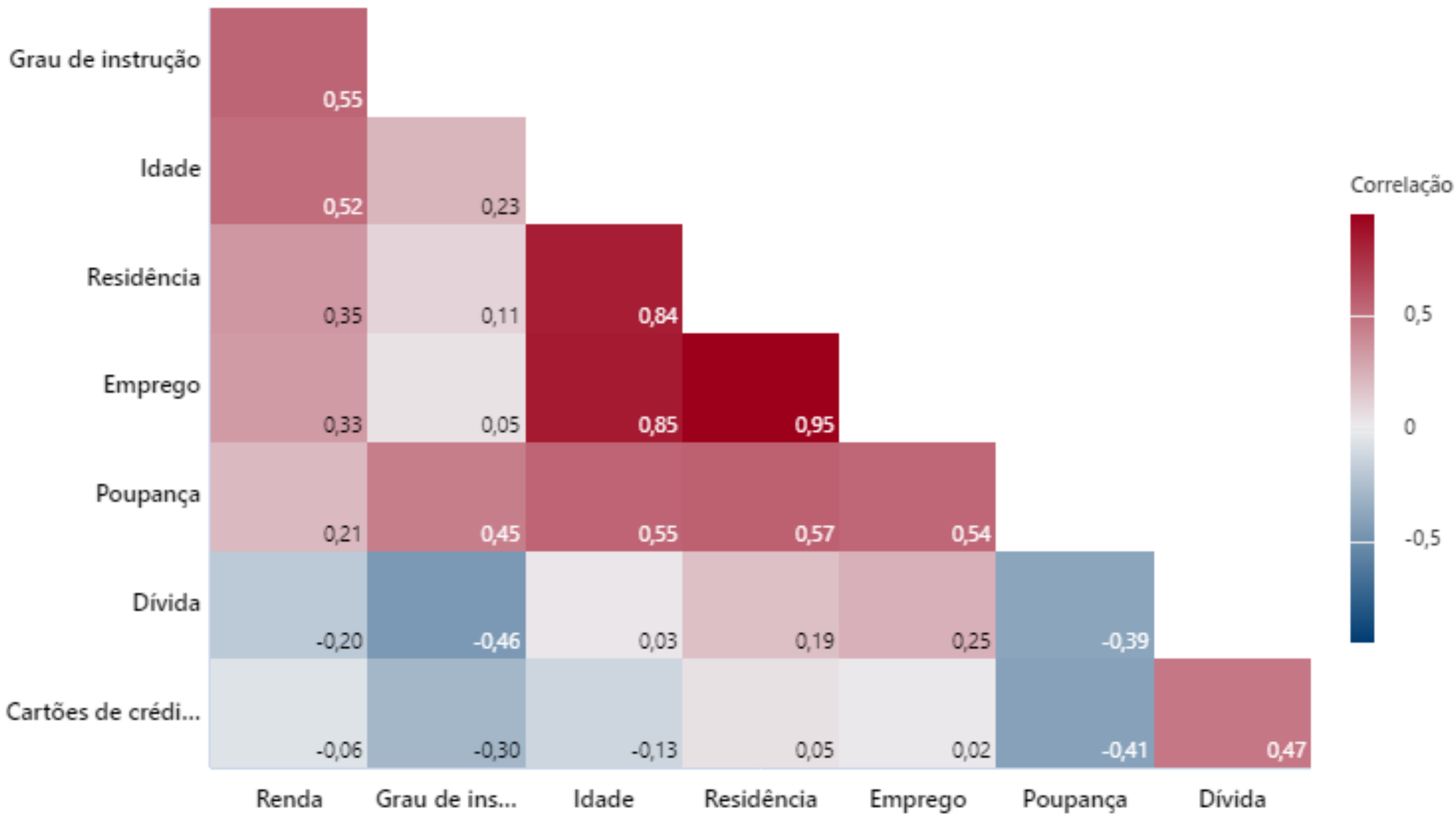
– Exemplo: Aplicação em automação





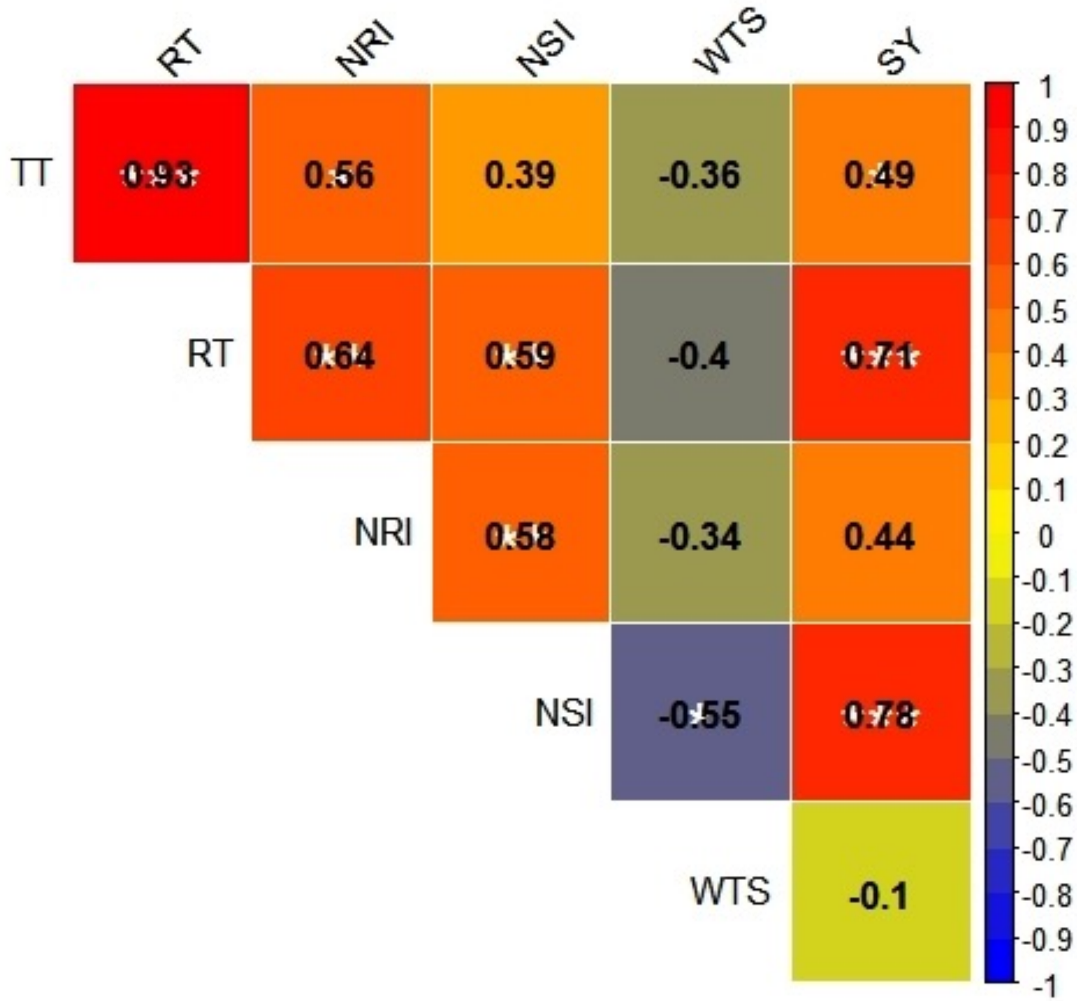
Covariância

- Exemplo: Mapa de correlação



Covariância

- Exemplo: Mapa de coeficiente de correlação




"That's all Folks!"