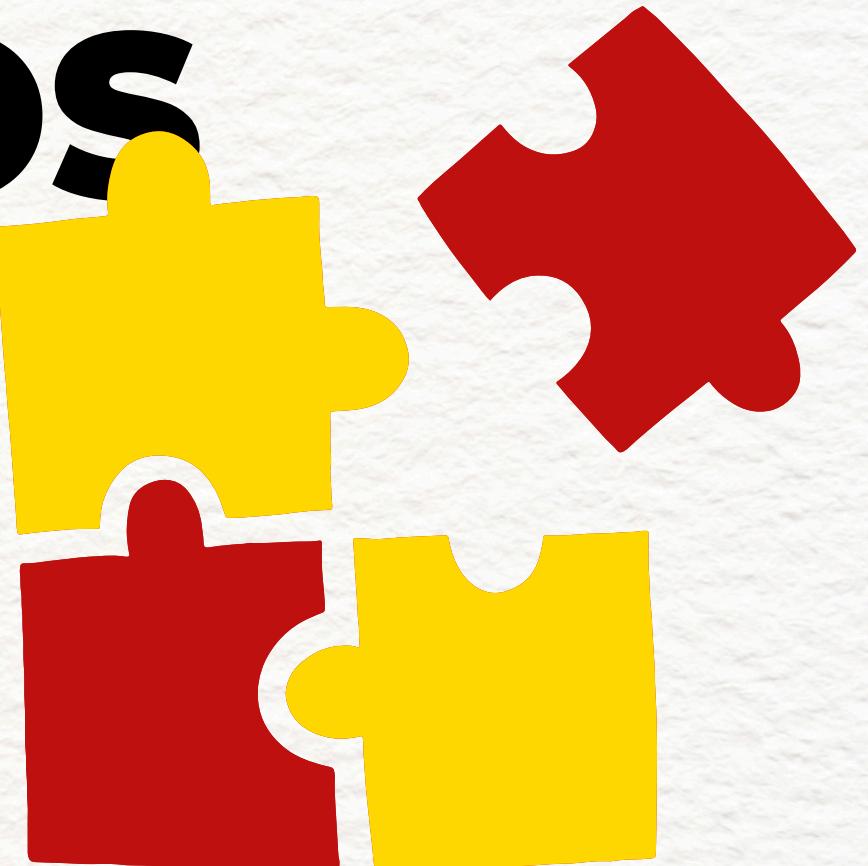


INTRODUÇÃO A ANÁLISE DE DADOS EM AMBIENTE "R"



Francisco J. Castelhano

francisco.castelhano@ufrn.br - Sala 503 CCHLA

Sobre o R



A linguagem R surgiu em 1995, derivada da linguagem S, e é orientada a objetos.

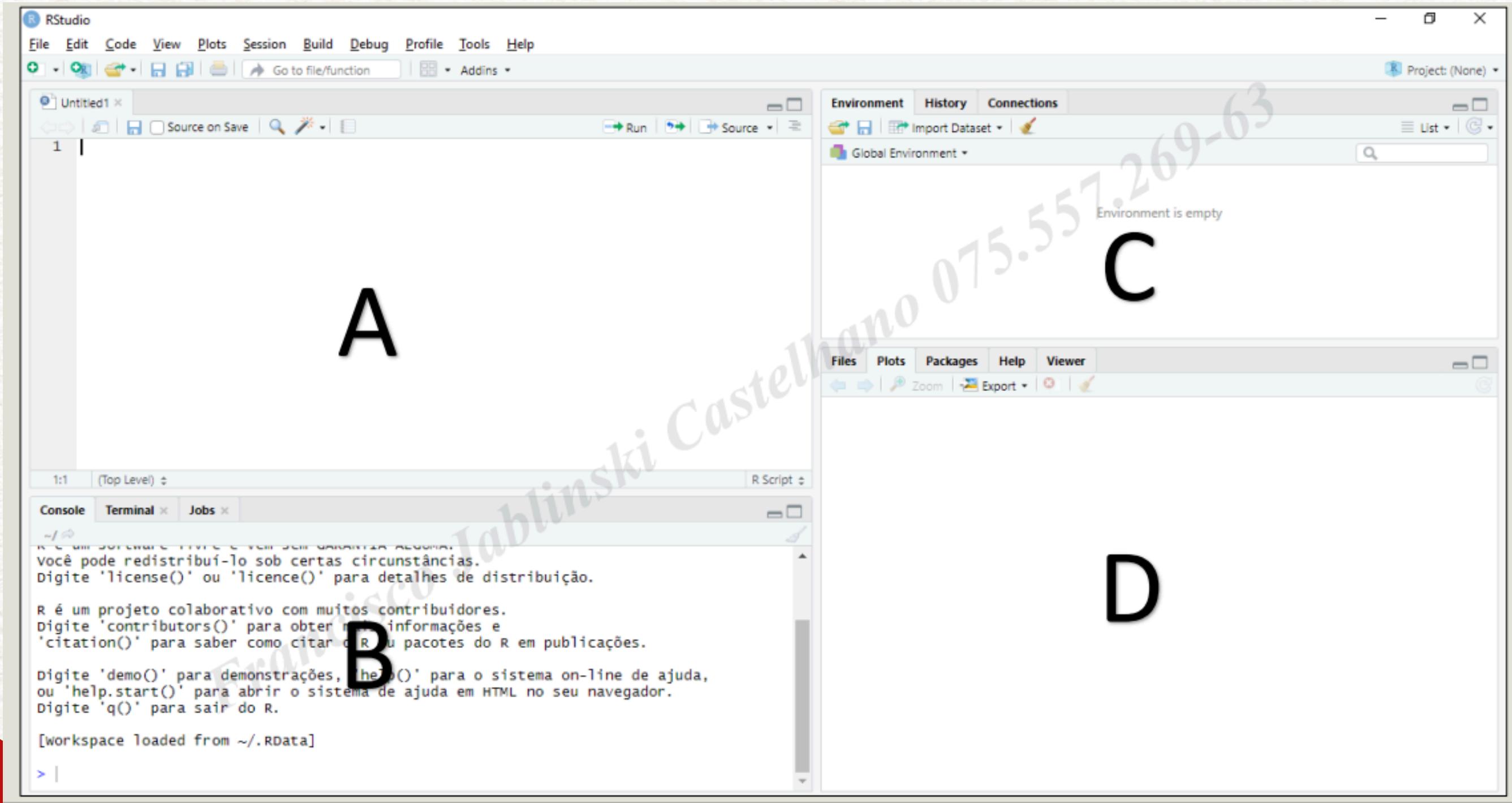
Livre e OpenSource

Possui inúmeros pacotes (mais de 17 mil), com vantagem para a aplicação da Estatística Avançada e uma vasta comunidade de suporte, além de fortes capacidades voltadas ao Data Science.

Compartilhado

Comprehensive R Archive Network (CRAN) é o repositório da linguagem R em que cada usuário pode contribuir com novos pacotes (coleções de funções em R com código compilado). Esses pacotes podem ser facilmente instalados com uma linha de código

Sobre o R



Sobre o R

Objetos

A maneiras simples de se acessar algo que foi salvo na memória da máquina.

Pode ser um valor, uma palavra, uma ou mais variáveis, uma URL, uma base de dados amostral ou populacional, uma lista de coisas distintas contendo informações e tamanhos distintos, um gráfico, um mapa, uma imagem, um novo comando, etc.

No R TUDO que criamos é um objeto! Cada um desses objetos possui uma classe!

Funções

Correspondem a ações, a ordens direcionadas à máquina e implementadas sobre um objeto

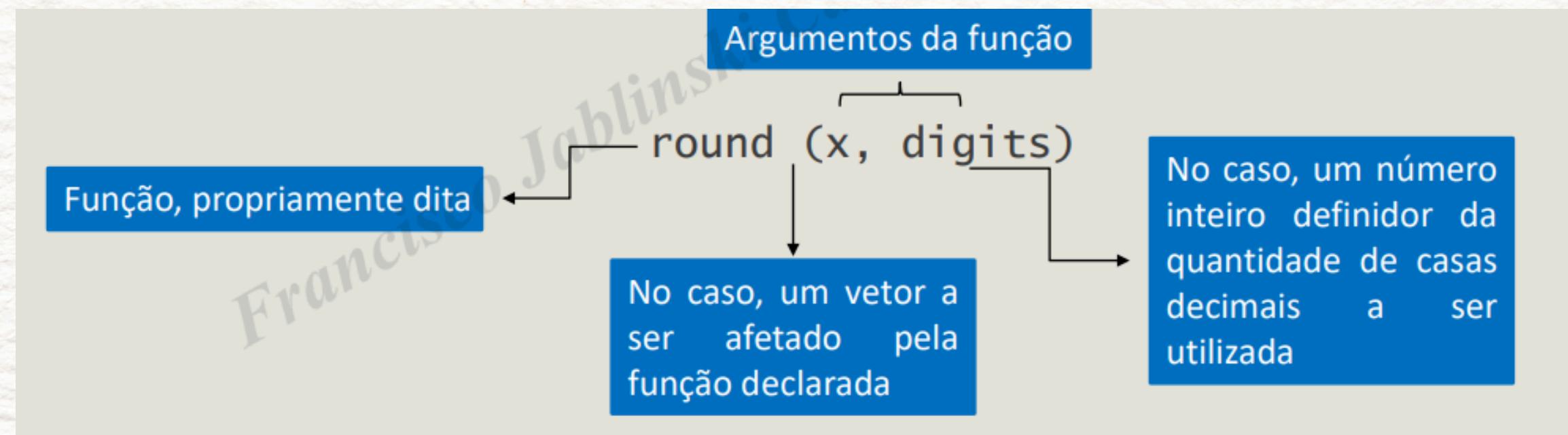
Argumento

Correspondem a um refinamento ou um melhor direcionamento das ações ou ordens propostas

Sobre o R

Utilizando funções no R

Para utilizar uma função no R, devemos conhecer sua forma funcional, isto é, devemos, em regra, declarar os argumentos inerentes a ela. Exemplo de utilização da função round():



Sobre o R

Utilizando funções no R

Mãos a Obra -> Itens 1 e 2 do Script

Sobre o R

Importando Dados

Conforme mencionamos, é possível abrir um banco de dados pré-existente no R. Algumas funções de leitura são utilizadas para tanto.
ex.: `read.csv` ; `read.rds` dentre outras

- > Atente-se a organização dos dados (separados por vírgula, ponto e vírgula, ponto ...)
- > Cuidado com a formatação (casa decimal separada por ponto ou por vírgula?)
- > A edição dos dados em si não é fácil de ser feita no R, nestes casos, opte por editá-los em ambiente Excel

Sobre o R

Importando Dados

Existem CINCO grande tipos de dado que o R comprehende

- 1.Numeric (1.2, 5, 7, 3.14159)
- 2.Integer (1, 2, 3, 4, 5)
- 3.Complex (i + 4)
- 4.Logical (TRUE / FALSE)
- 5.Character ("a", "apple")

Sobre o R

Utilizando funções no R

Mãos a Obra -> Item 3 do Script

Medidas de Posição

1

Média -> função "mean"

2

Moda

3

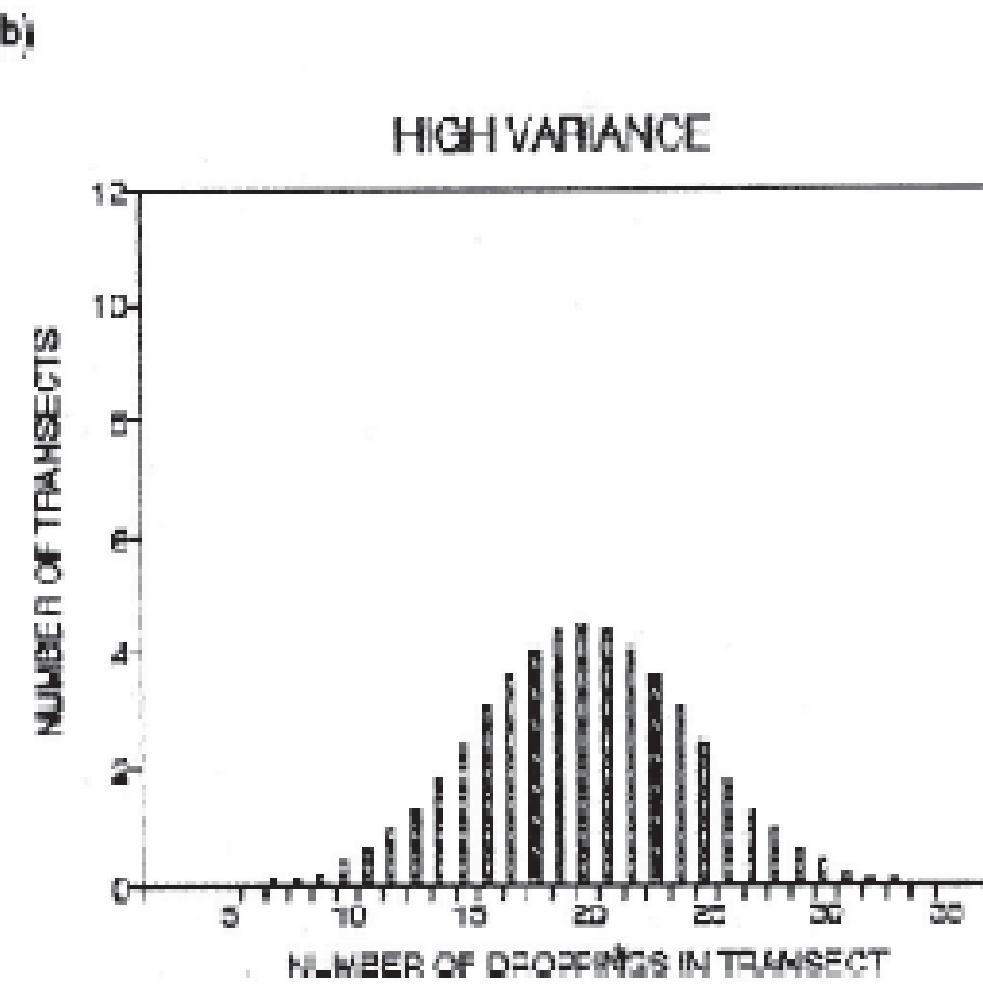
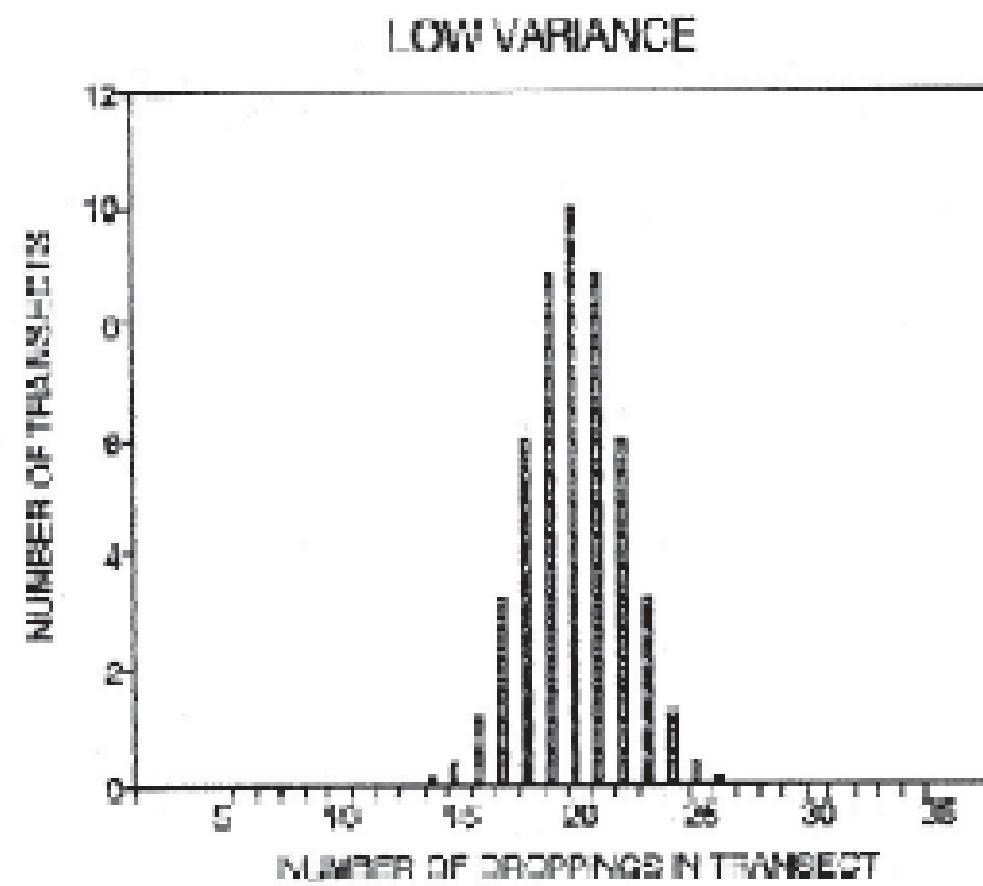
Mediana -> função "median"

4

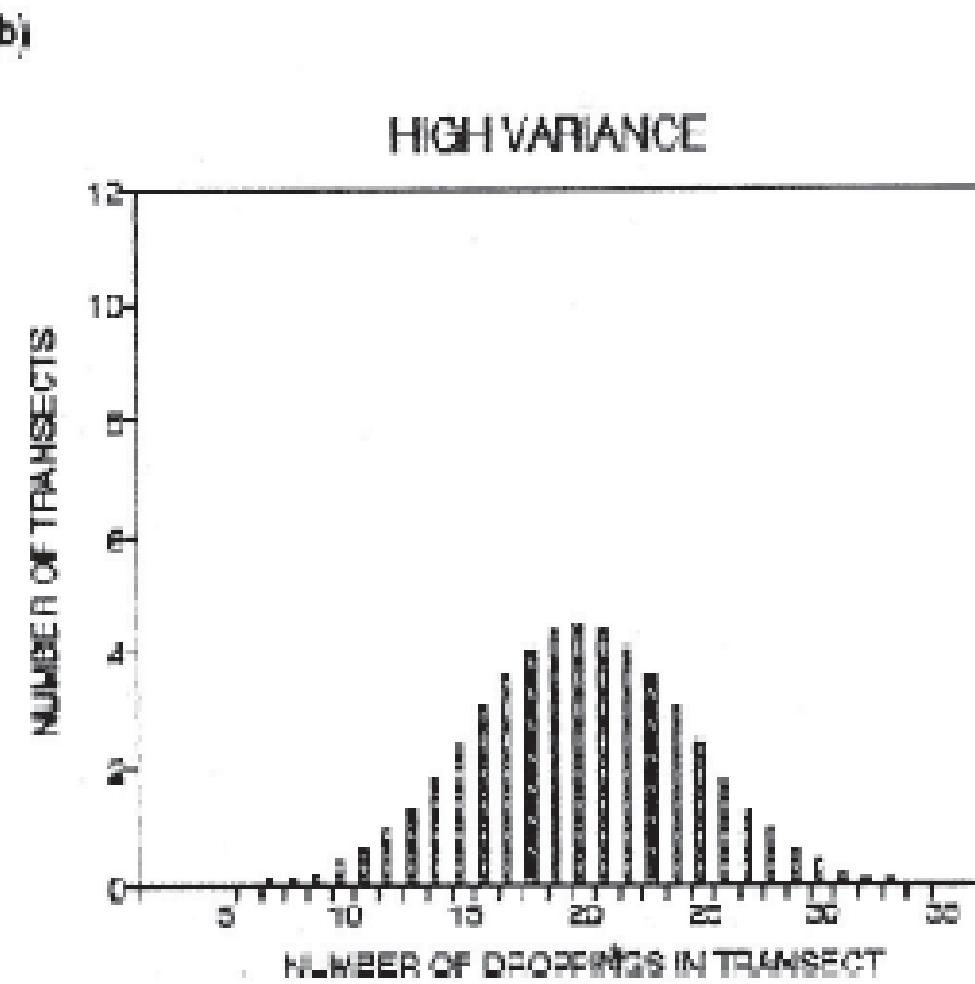
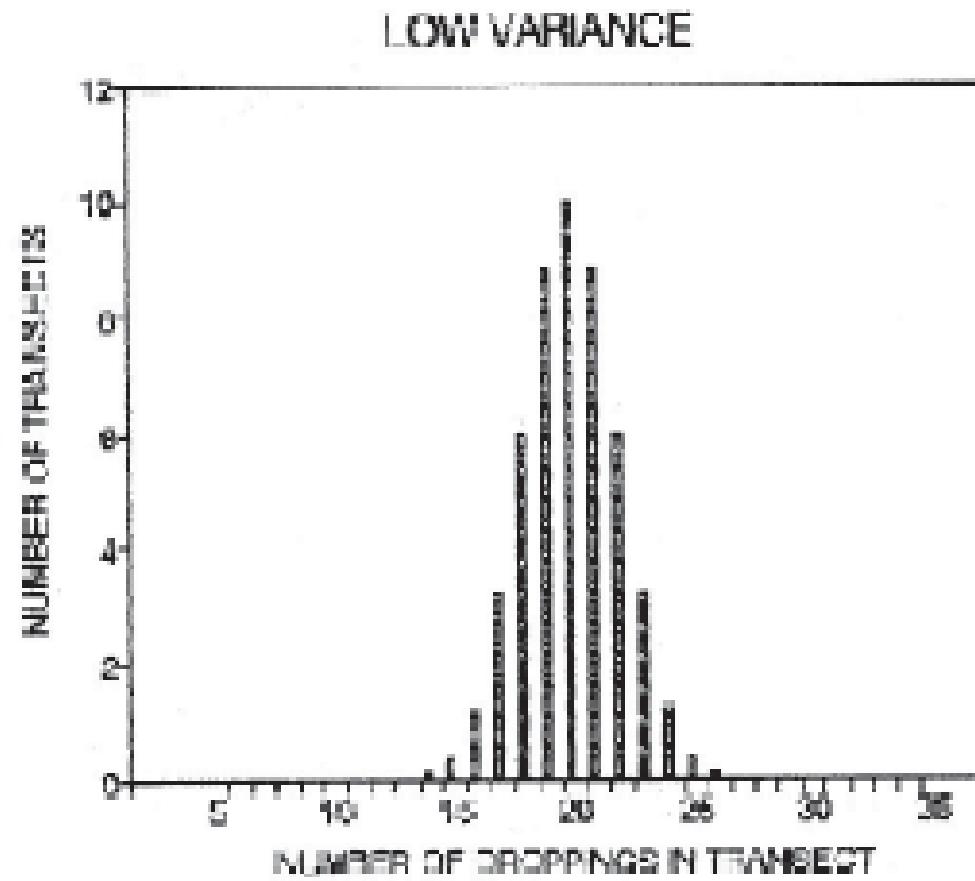
Percentis e Quartis -> função "quantile"

CORRELAÇÃO

- - Duas variáveis estão relacionadas se a mudança de uma provoca mudança na outra
- - Por exemplo:
- - Velocidade x Consumo de combustível
- - Energia x Calor
- - Inflação x Poder aquisitivo
- - Enfermidades x Variável geográfica
- - Parte-se do princípio de que uma variável (independente) se relaciona com outra variável (dependente)

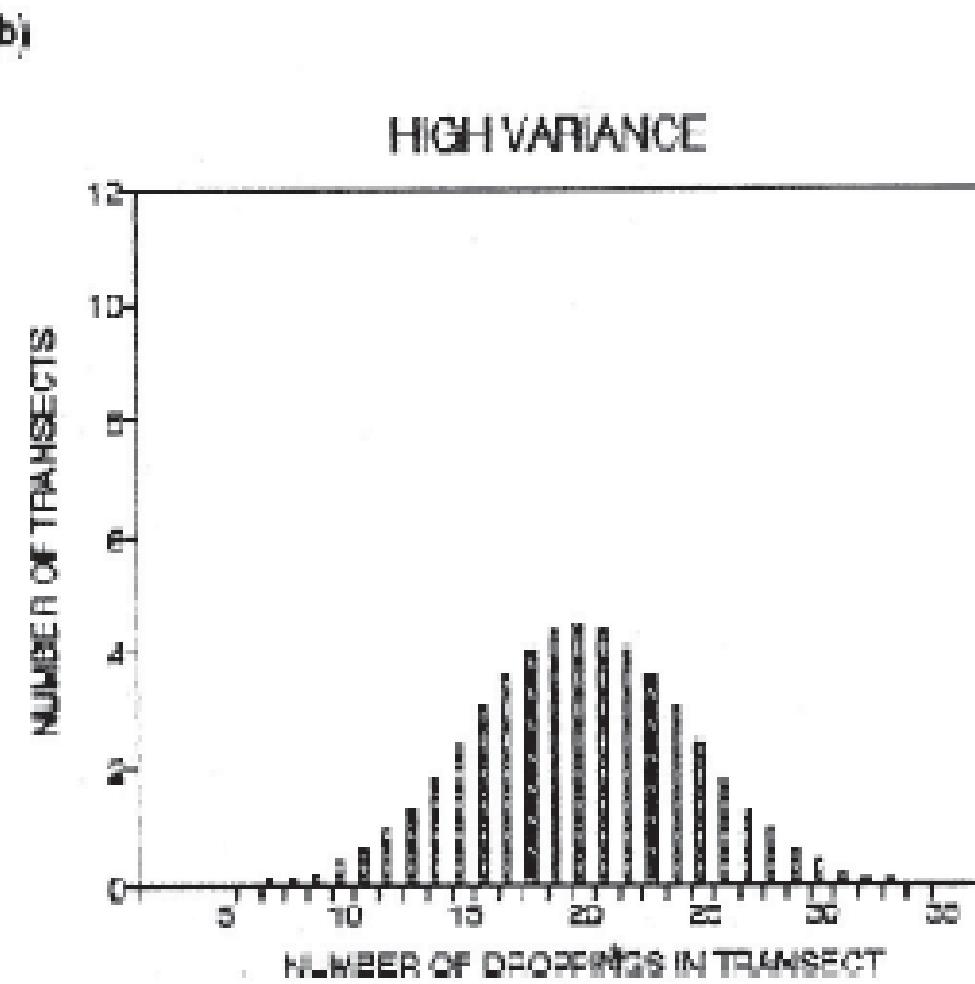
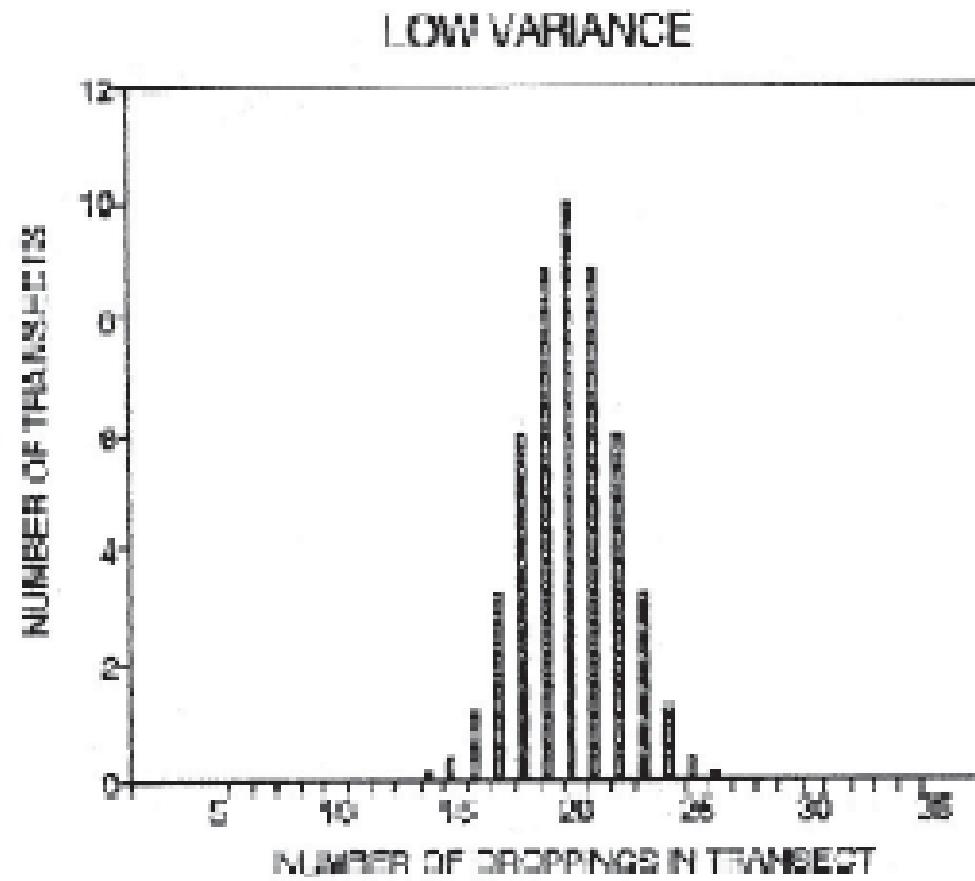


CORRELAÇÃO



- Representação gráfica: Diagramas de dispersão
- Demonstra a relação entre duas variáveis quantitativas, medidas sobre os mesmos indivíduos.
- Os valores de uma variável aparecem no eixo horizontal, e os da outra, no eixo vertical.
 - Comumente, coloca-se no eixo x a variável independente
- Cada indivíduo aparece como o ponto do gráfico definido pelos valores de ambas as variáveis para aquele indivíduo

CORRELAÇÃO

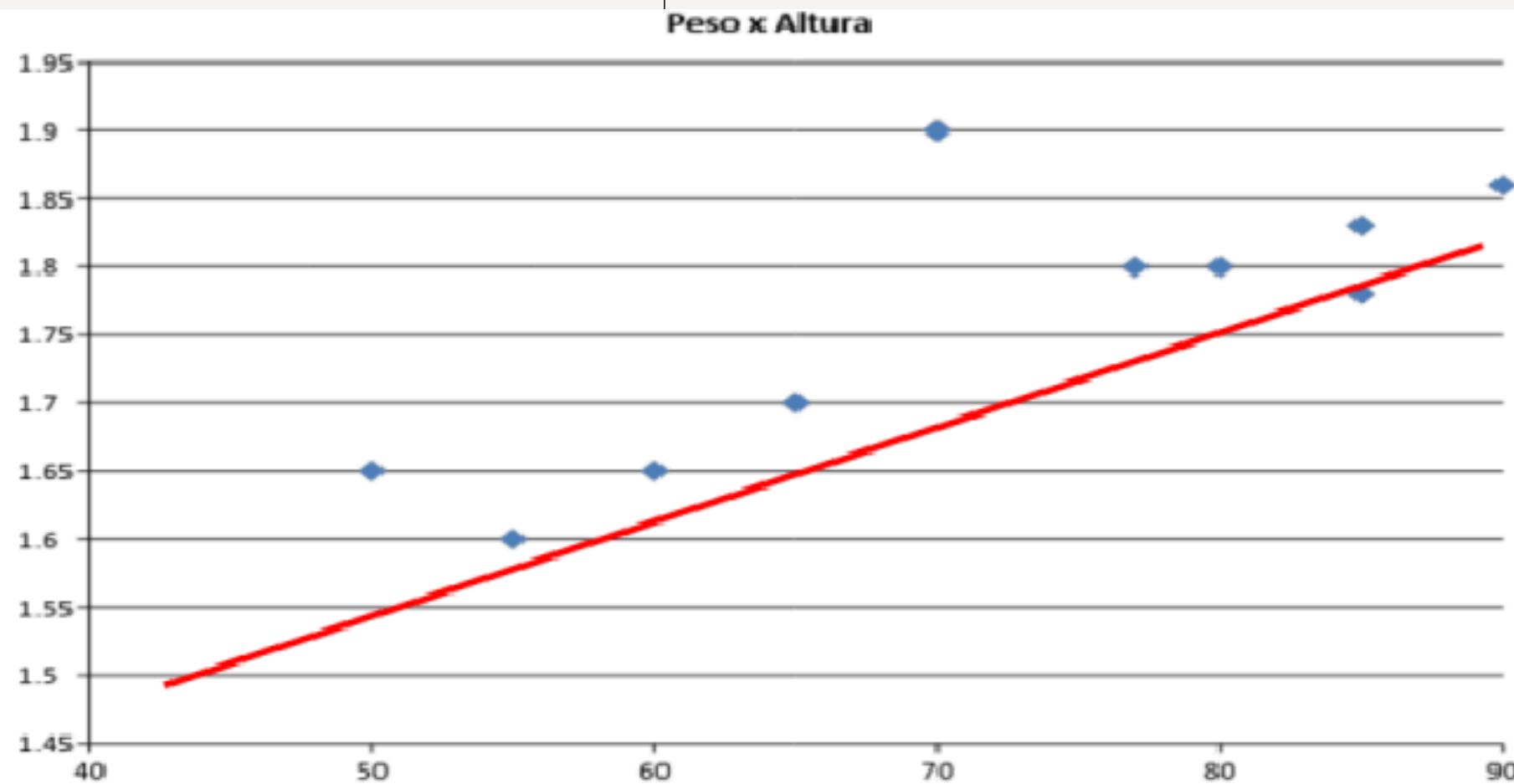


- Representação gráfica: Diagramas de dispersão
- Demonstra a relação entre duas variáveis quantitativas, medidas sobre os mesmos indivíduos.
- Os valores de uma variável aparecem no eixo horizontal, e os da outra, no eixo vertical.
 - Comumente, coloca-se no eixo x a variável independente
- Cada indivíduo aparece como o ponto do gráfico definido pelos valores de ambas as variáveis para aquele indivíduo

CORRELAÇÃO

Gráfico de dispersão

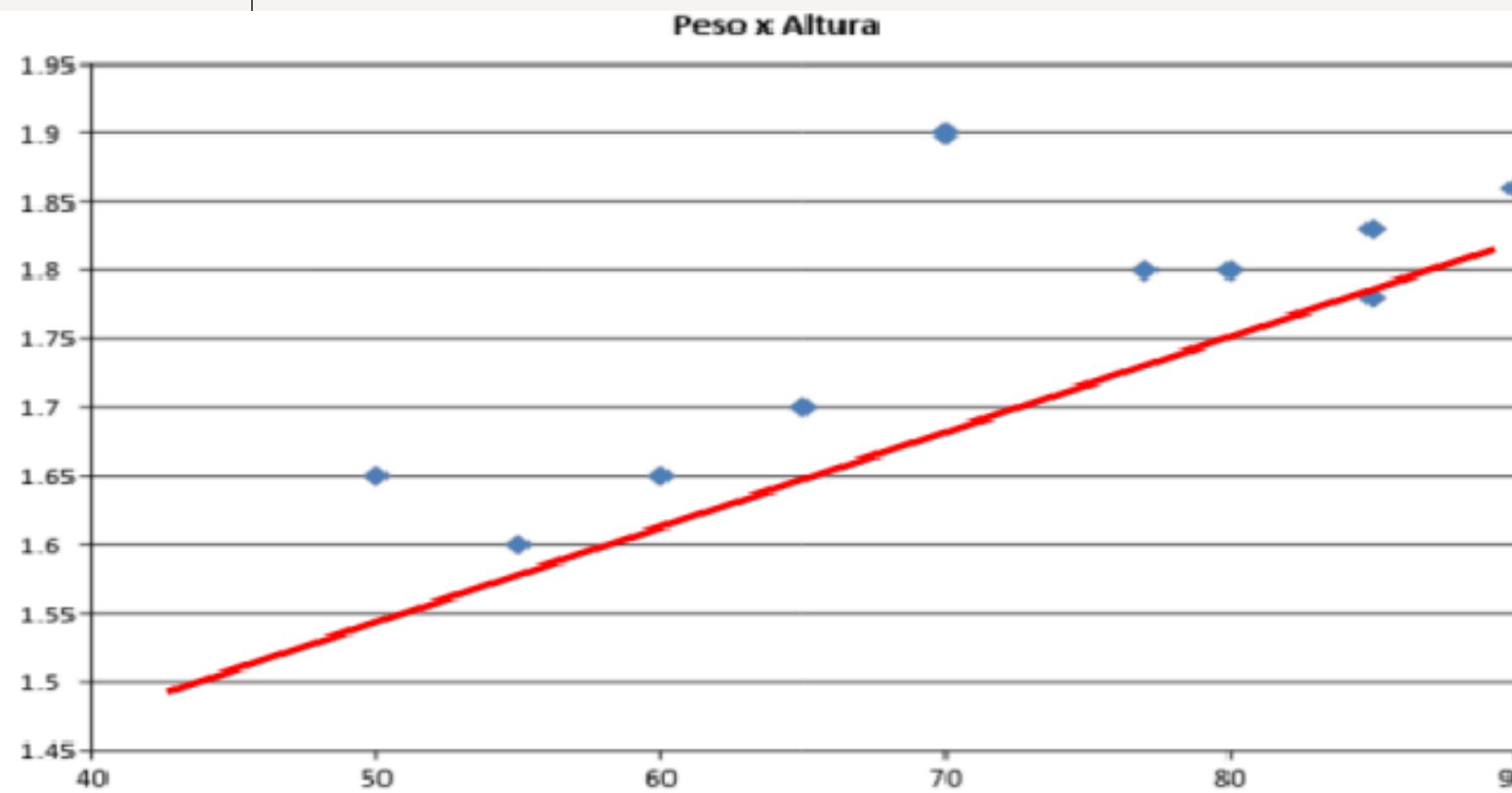
Peso	Altura
80	1,8
85	1,83
50	1,65
70	1,9
55	1,6
77	1,8
85	1,78
93	1,86
65	1,7
60	1,65



CORRELAÇÃO

- - Aspectos relevantes na análise dos Diagramas
- - DIREÇÃO (crescente, decrescente)
- - FORMA (linear, não-linear, aglomerados)
- PONTOS DISCREPANTES

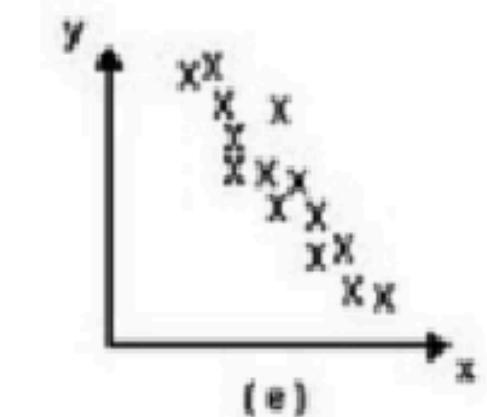
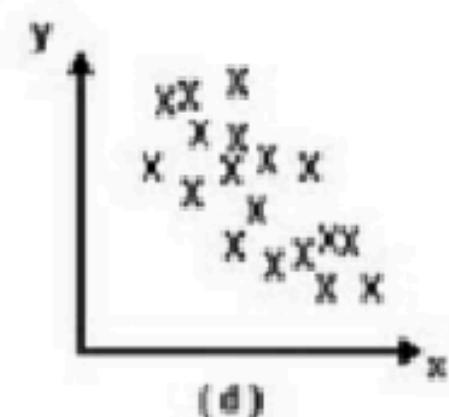
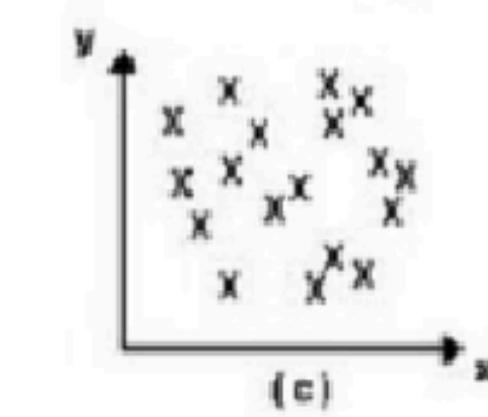
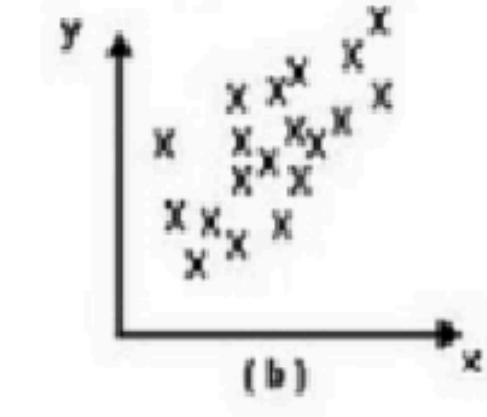
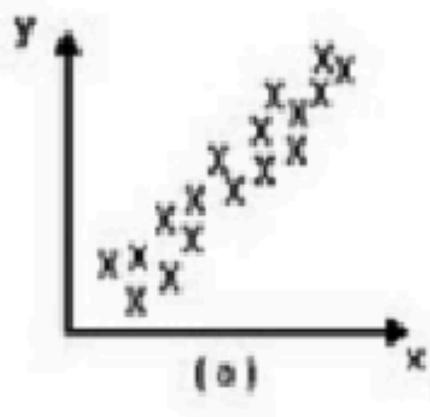
Peso	Altura
80	1,8
85	1,83
50	1,65
70	1,9
55	1,6
77	1,8
85	1,78
93	1,86
65	1,7
60	1,65



CORRELAÇÃO

- Padrões de Diagramas de Dispersão

Possíveis Padrões para Diagramas de Dispersão.



Legenda

- (a) - Elevada correlação positiva
- (b) - Moderada correlação positiva
- (c) - Ausência de correlação
- (d) - Moderada correlação negativa
- (e) - Elevada correlação negativa

CORRELAÇÃO

- Padrões de Diagramas de Dispersão
- Correlação Positiva -> Variáveis se atraem
- Correlação Negativa -> Variáveis se repelem
- Força da Correlação não implica na sua direção

CORRELAÇÃO

- - Antes do calculo da Correlação, importante analisar os tipos de variaveis que temos:
- -O tipo de variável altera o cálculo a ser feito

COEFICIENTE	SÍMBOLO	INTERVALO DE VARIAÇÃO	VARIÁVEIS	
			X	Y
Pearson	ρ	$-1 \leq \rho \leq 1$	Continua	Continua
Ponto Bisserial	ρ_{pb}	$-1 \leq \rho_{pb} \leq 1$	Continua	Dicotômica
Bisserial	ρ_b	$-1 \leq \rho_b \leq 1$	Continua	Continua, mas dicotomizada
Tetracórico	ρ_t	$-1 \leq \rho_t \leq 1$	Continua, mas dicotomizada	Continua, mas dicotomizada
Phi	ϕ	$-1 \leq \phi \leq 1$	Dicotômica	Dicotômica
Spearman	ρ_s	$-1 \leq \rho_s \leq 1$	Dados em <i>ranks</i> ou passíveis de serem transformados	Dados em <i>ranks</i> ou passíveis de serem transformados
Kendall	τ	$-1 \leq \tau \leq 1$	Dados em <i>ranks</i>	Dados em <i>ranks</i>
Contingência	C	$0 \leq C < 1$	Nominal	Nominal
Eta	η	$0 \leq \eta \leq 1$	Continua	Continua ou discreta; pode assumir valores nominais ou outros tipos de valores

CORRELAÇÃO

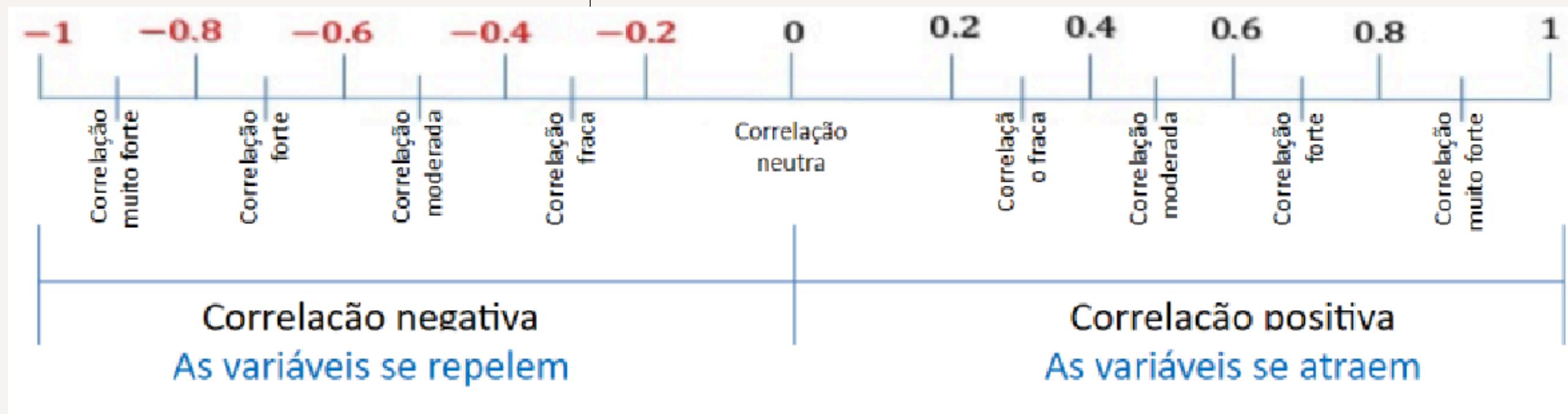
- Correlação de Pearson
- as variáveis X e Y co-variam, pois a variação de
- uma em relação à sua média está associada à variação da outra em relação à sua média

CORRELAÇÃO

- Coeficiente de correlação linear r
- Mede o grau de influência que a variável independente tem sobre a variável dependente
- Mede a intensidade (valor) e a direção (sinal) da correlação
- Quanto maior o valor absoluto de r, maior é a influência
- Sinal positivo = correlação positiva = as variáveis se atraem
- O aumento de x -> aumento em y
- Sinal negativo = correlação negativa = as variáveis se repelem
- O aumento de x ->a diminuição de y

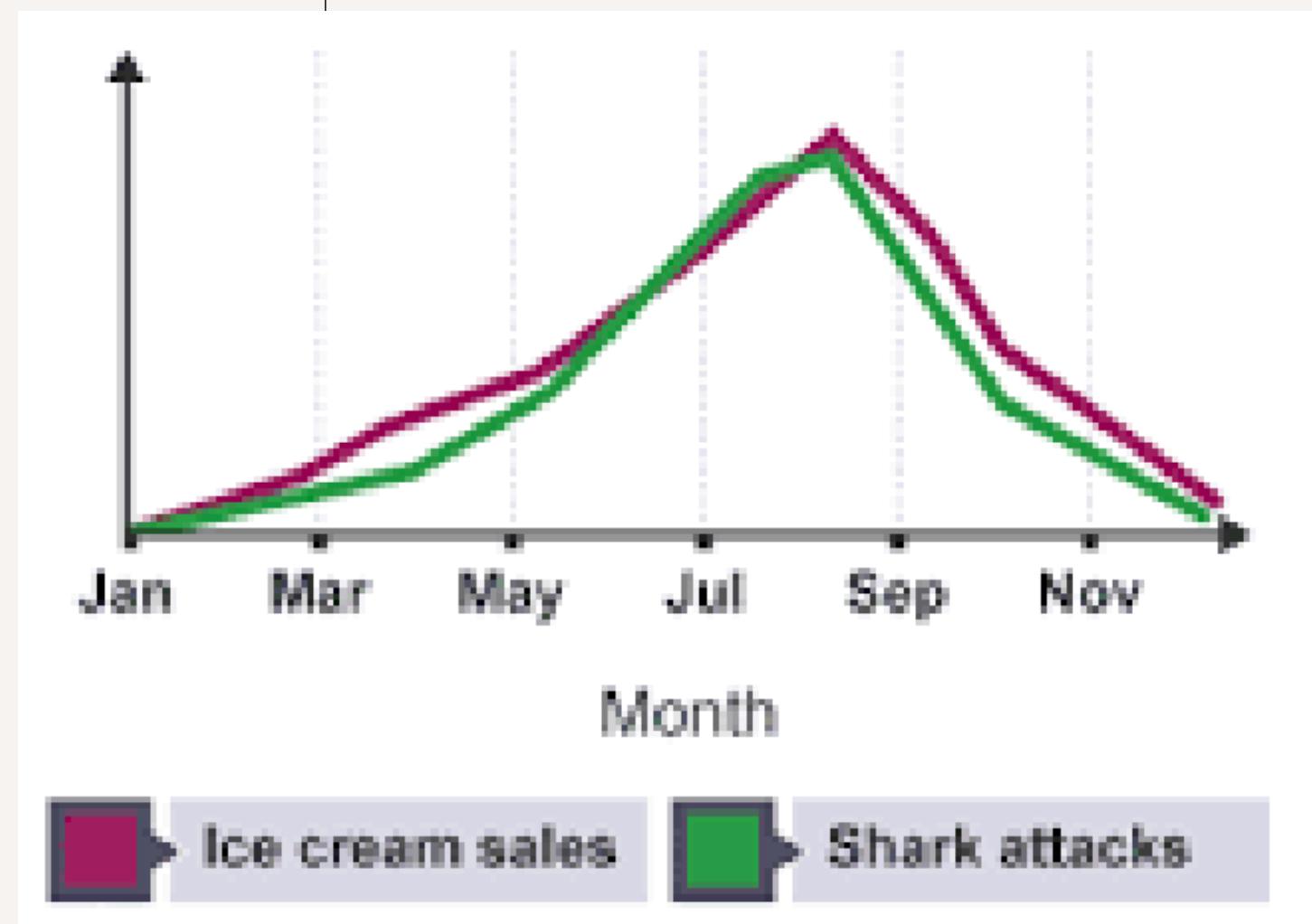
CORRELAÇÃO

- O intervalo de r (coeficiente de correlação) vai de -1 (correlação negativa perfeita) até 1 (correlação positiva perfeita)

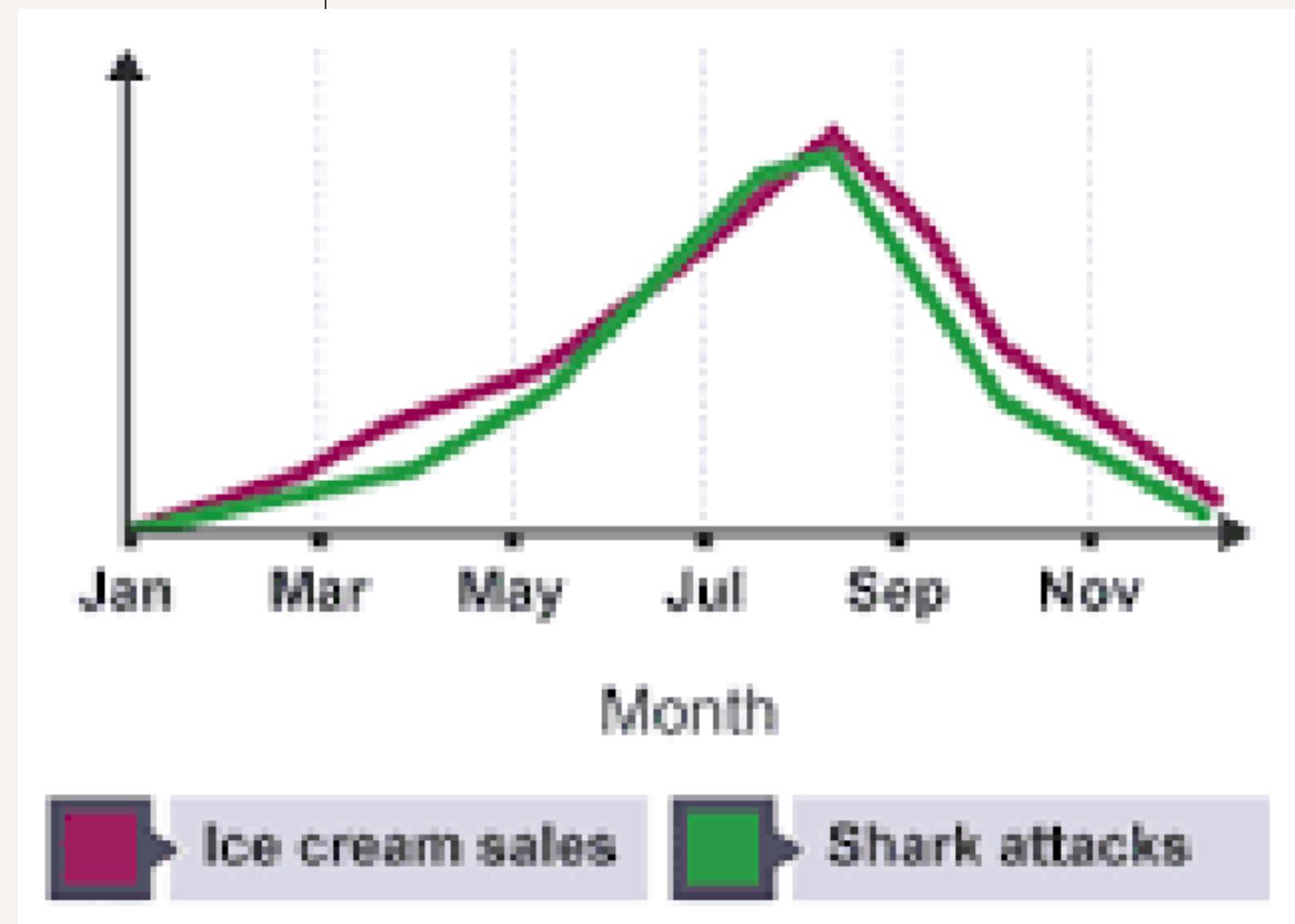


CORRELAÇÃO

- ATENÇÃO
- - Correlação \neq Causalidade
- - A Altura seria a causa do aumento de peso?
- - NÃO, se relacionam, mas não são causa e efeito.



- - Existe relação entre as variáveis?
- - Sim
- - Existe causalidade entre as variáveis?
- - Não
- - Existe um terceiro fator responsável por desencadear ambos



CORRELAÇÃO

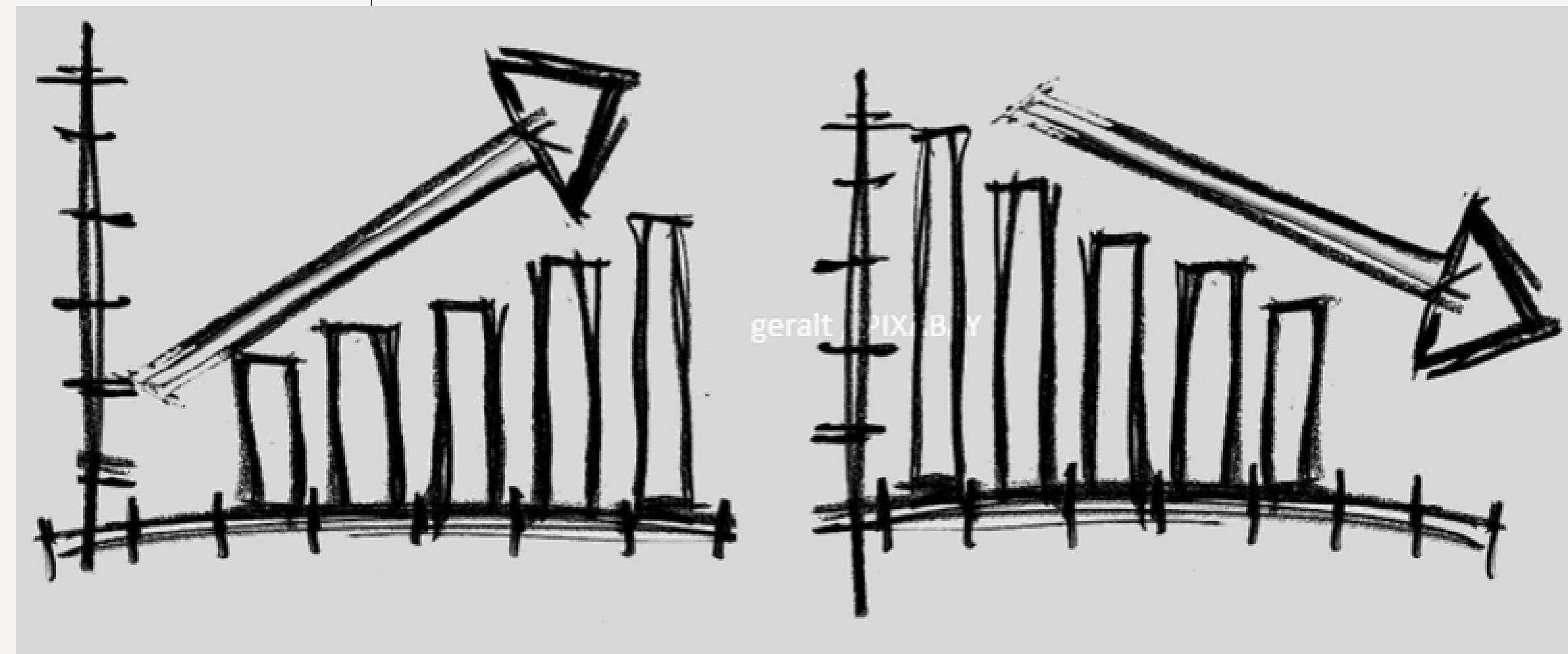
- <https://tylervigen.com/spurious-correlations>

CORRELAÇÃO

- ATENÇÃO II
 - Cálculo para uma Correlação LINEAR simples, isto é, se as variáveis tem uma relação entre si DE FORMA LINEAR

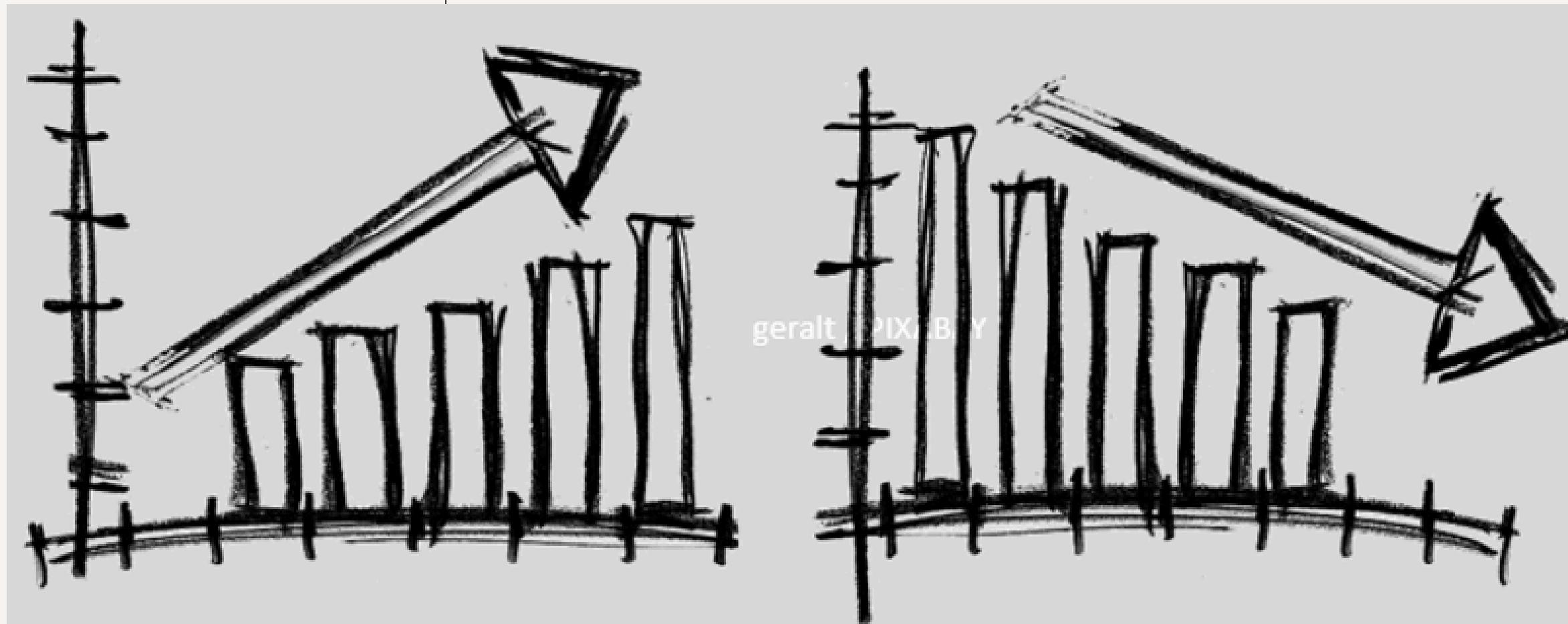
TESTES DE TENDÊNCIA

- Como constatar se uma série de dados temporal está crescendo ou não?
- Será que o crescimento é significativo?
- Será que não apresenta tendência?



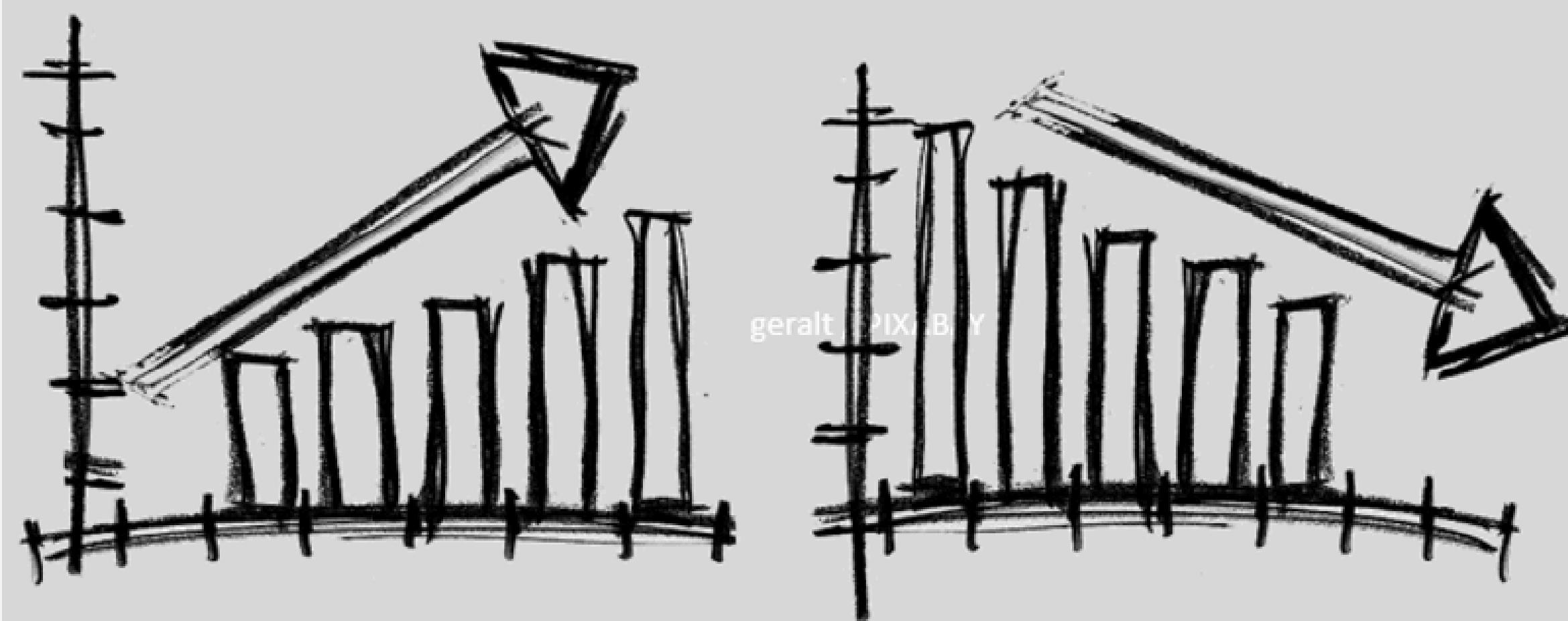
TESTES DE TENDÊNCIA

- Testes de tendência nos fornecem informações referentes
 - a:
 - Direção da tendência (positiva ou negativa)
 - Intensidade
 - Significância



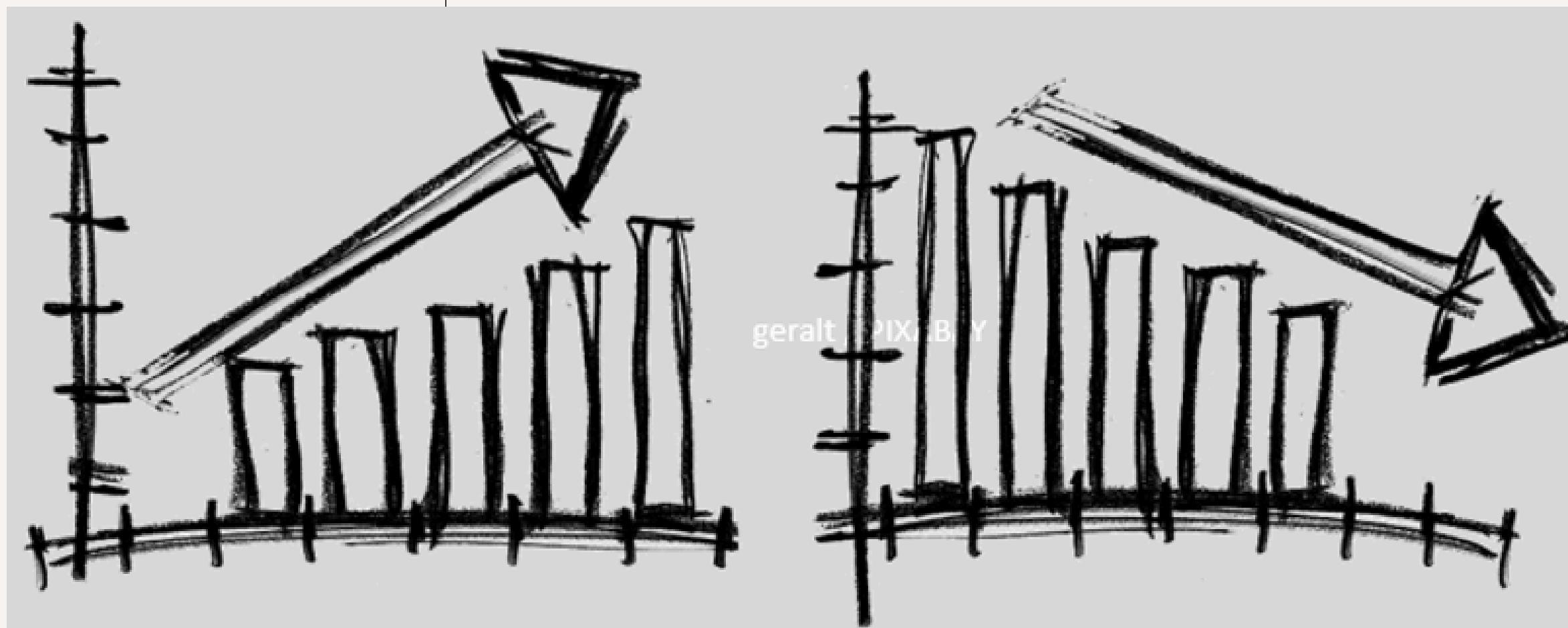
TESTES DE TENDÊNCIA

- Significância?
- Probabilidade da hipótese ser coincidência ou não
- Me responde se meu resultado é representativo ou não
- Me dá um nível de confiança em relação a minha inferência



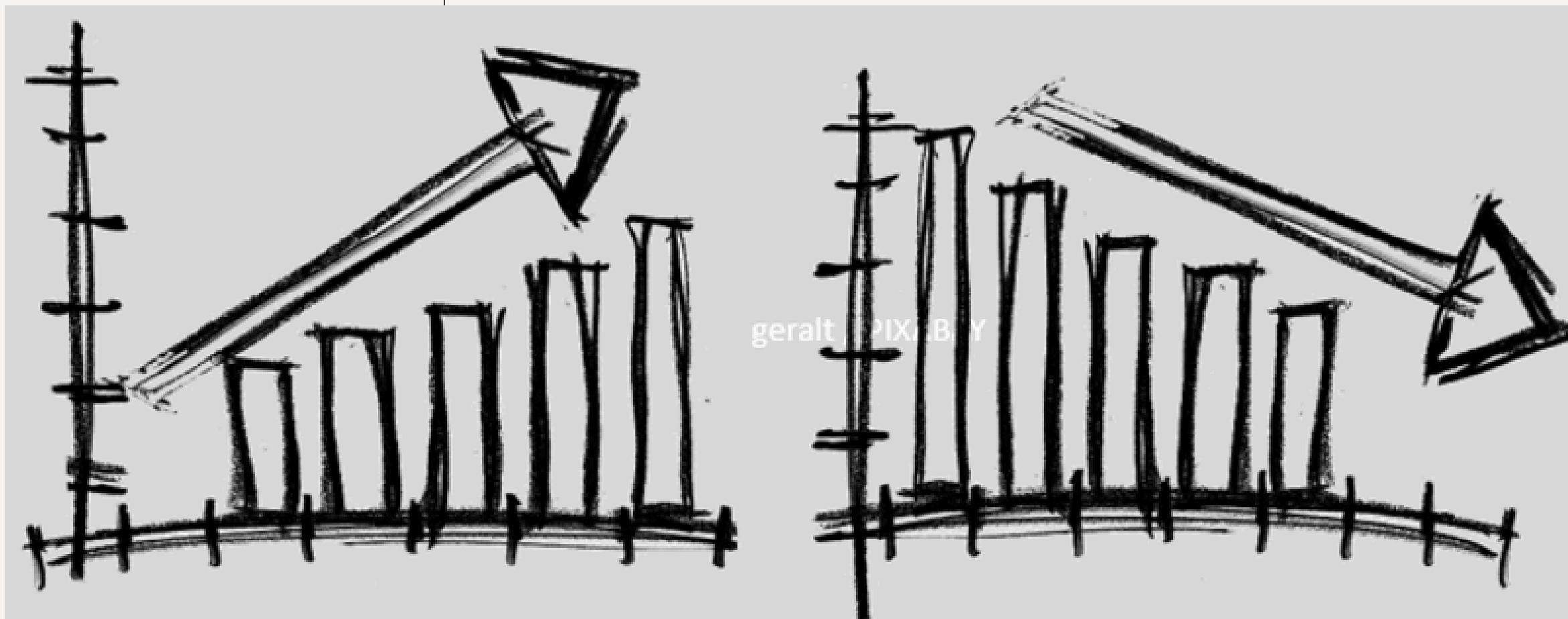
TESTES DE TENDÊNCIA

- Significância?
- Hipotetizamos a existência de relação entre duas variáveis a partir do coeficiente de correlação
- O teste de significância me dirá se a chance de minha
- Uma hipótese nula é normalmente o pressuposto padrão e é definida como a previsão de que não existe nenhuma interação entre as variáveis ou a relação encontrada é coincidência



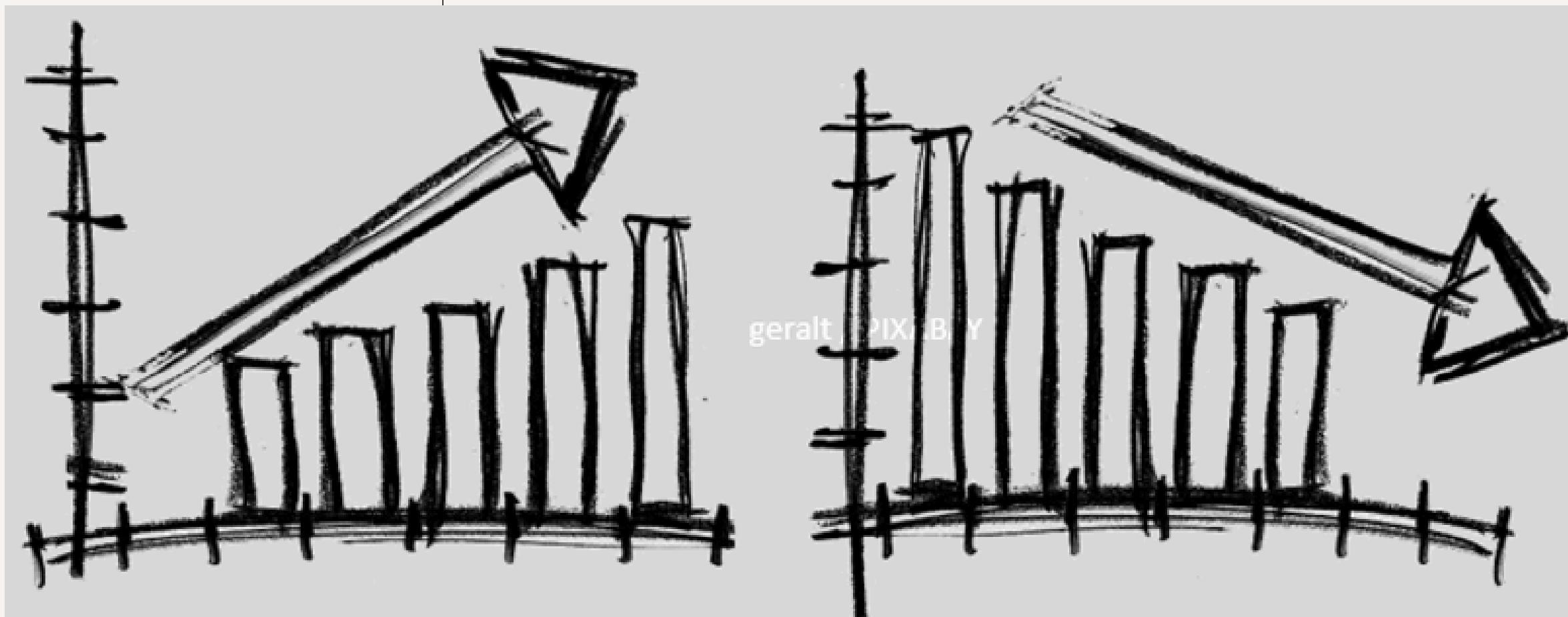
TESTES DE TENDÊNCIA

- Significância?
- Na estatística, considera-se 95% de chance da hipótese nula ser rejeitada como aceitável
- "Temos 95% de probabilidade de que nossa hipótese está correta"
- Teste t de student é um teste de significância



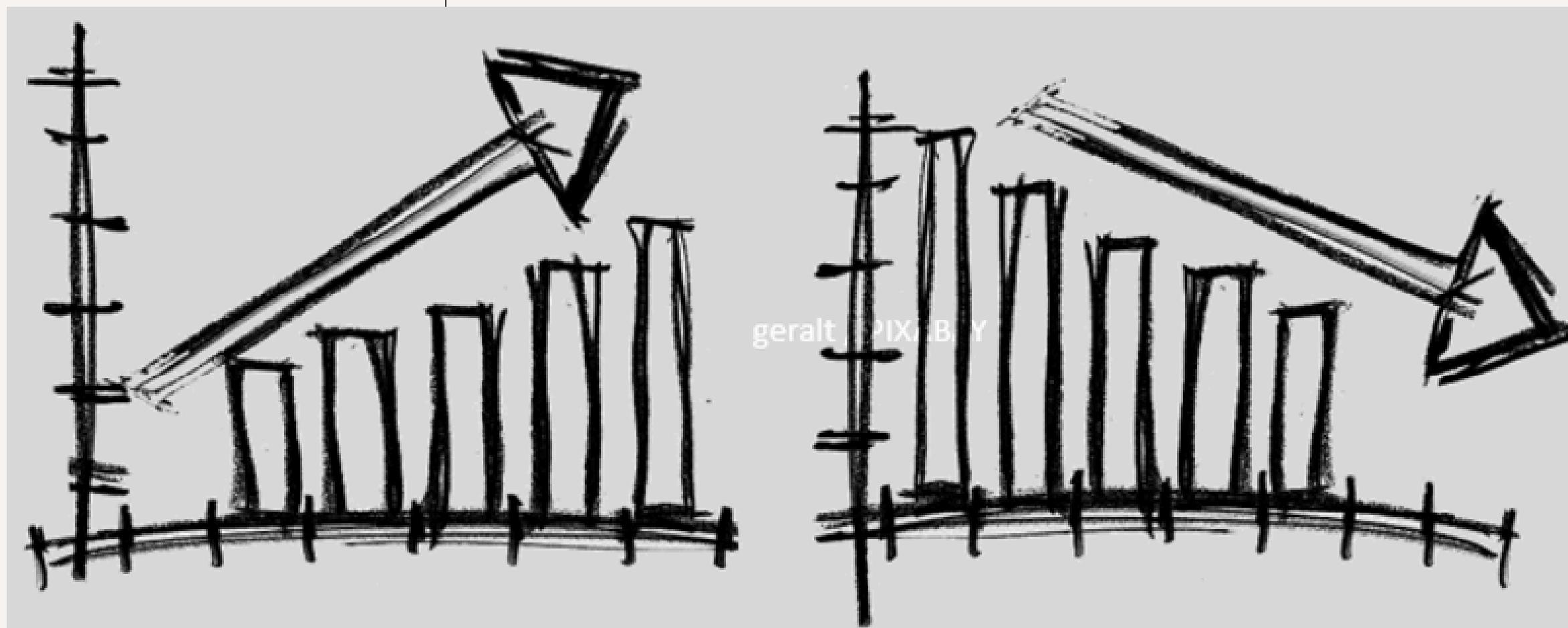
TESTES DE TENDÊNCIA

- Significância?
- Teste de significância terá como resultado o chamado p-valor ou valor de probabilidade
- Probabilidade de que? de sua hipótese ser nula.
- Exemplo -> p-valor = 0.05 -> 5% de chance da hipótese ser nula e 95% dela ser validada



TESTES DE TENDÊNCIA

- Significância?
- P-valor = 0.5 → 50% de chance da hipótese ser nula, rejeito a minha hipótese original
- As chances do meu resultado ser coincidência são muito elevadas.
- consideramos baixa significância a partir de 10% de chances da hipótese ser nula, 5% é um valor ideal, mas quanto menor o p-valor maior a significância da minha hipótese



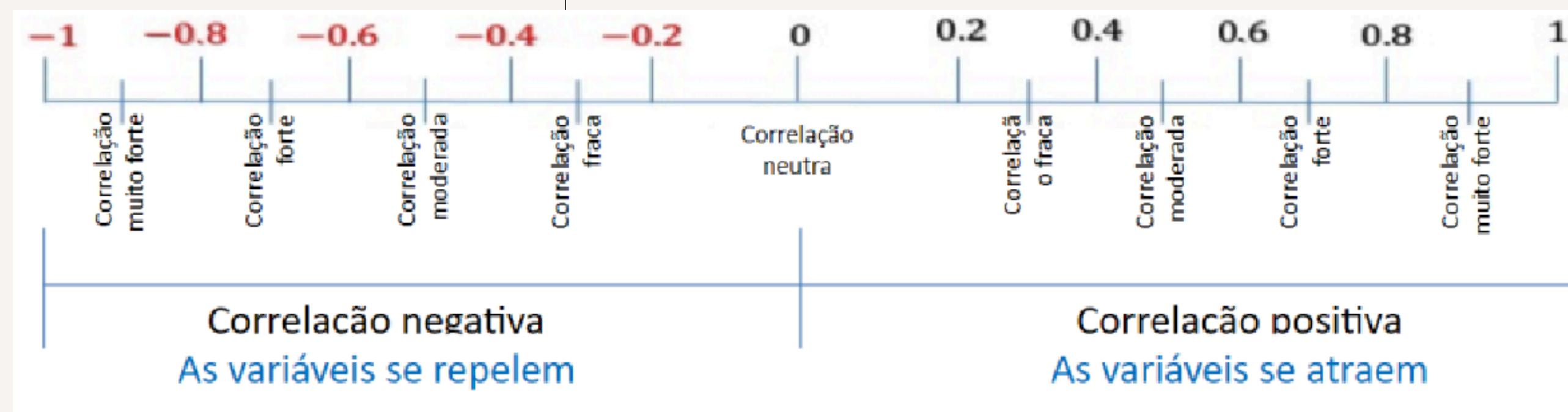
TESTES DE TENDÊNCIA

- Teste de Mann Kendall
- Retorna dois valores
- Tau = Coeficiente de correlação entre a variável e o tempo cronológico
- Tau -> Varia de -1 a 1 (igual o coeficiente de Pearson)
- p-value -> Significância
- Consideramos a tendência como significativa quando o p-valor é inferior a 0.05

Prática 3 -> RStudio

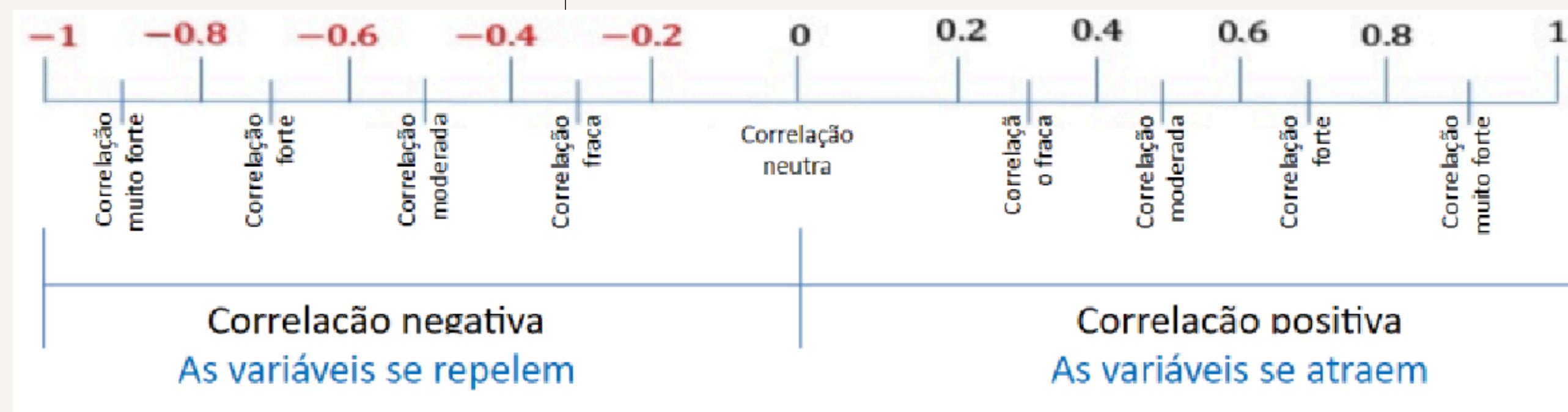
TESTES DE TENDÊNCIA

- Teste de Mann Kendall
- Retorna dois valores
- Tau = Coeficiente de correlação entre a variável e o tempo cronológico
- Tau -> Varia de -1 a 1 (igual o coeficiente de Pearson)
- p-value -> Significância
- Consideramos a tendência como significativa quando o p-valor é inferior a 0.05



TESTES DE TENDÊNCIA

- Teste de Mann Kendall
- Retorna dois valores
- Tau = Coeficiente de correlação entre a variável e o tempo cronológico
- Tau -> Varia de -1 a 1 (igual o coeficiente de Pearson)
- p-value -> Significância
- Consideramos a tendência como significativa quando o p-valor é inferior a 0.05



CORRELAÇÃO

- FINALIZAR ITENS 4 E 5

ÍNDICES ATRAVÉS DE PACOTES

- Rainfall Anomaly Index

	RAI range	Classification
Rainfall Anomaly Index (RAI)	Above 4	Extremely humid
	2 to 4	Very humid
	0 to 2	Humid
	-2 to 0	Dry
	-4 to -2	Very dry
	Below -4	Extremely dry

Para anomalias positivas

$$IAC = 3 \left[\frac{(N-\bar{N})}{(\bar{M}-\bar{N})} \right] \quad (5)$$

Para anomalias negativas

$$IAC = -3 \left[\frac{(N-\bar{N})}{(\bar{X}-\bar{N})} \right] \quad (6)$$

Onde:

N = precipitação (mm) atual do mês ou ano que será calculado o IAC;

\bar{N} = precipitação média mensal ou anual da série histórica (mm);

\bar{M} = média das dez maiores precipitações mensais ou anuais da série histórica;

\bar{X} = média das dez menores precipitações mensais ou anuais da série histórica.

ÍNDICES ATRAVÉS DE PACOTES

- O Índice Padronizado de Precipitação (SPI-n)
- indicador estatístico que compara a precipitação total recebida em um determinado local durante um período de n meses com a distribuição de chuvas de longo prazo para o mesmo período de tempo naquele local.

SPI	Cumulative Probability	Interpretation
-3.0	0.0014	extremely dry
-2.5	0.0062	extremely dry
-2.0	0.0228	extremely dry ($SPI < -2.0$)
-1.5	0.0668	severely dry ($-2.0 < SPI < -1.5$)
-1.0	0.1587	moderately dry ($-1.5 < SPI < -1.0$)
-0.5	0.3085	near normal
0.0	0.5000	near normal
0.5	0.6915	near normal
1.0	0.8413	moderately wet ($1.0 < SPI < 1.5$)
1.5	0.9332	very wet ($1.5 < SPI < 2.0$)
2.0	0.9772	extremely wet ($2.0 < SPI$)
2.5	0.9938	extremely wet
3.0	0.9986	extremely wet

ÍNDICES ATRAVÉS DE PACOTES

- O Índice Padronizado de Precipitação (SPI-n)
- indicador estatístico que compara a precipitação total recebida em um determinado local durante um período de n meses com a distribuição de chuvas de longo prazo para o mesmo período de tempo naquele local.

	Value of index
Uniform	≤ 10
Moderate	$>10 \leq 15$
Irregular	$>15 \leq 20$
Strongly irregular	>20

ÍNDICES ATRAVÉS DE PACOTES

- Índice de Desconforto de Thom

$$DI = (T - 0.55) \times [(1 - 0.01) \times RH] \times (T - 14.5)$$

Interpretation of category	DI
Uncomfortable	$DI \leq 14.9$
Comfortable	$15.0 \leq DI \leq 19.9$
Partially comfortable	$20.0 \leq DI \leq 26.4$
Uncomfortable	$DI \geq 26.5$

INDICES ATRAVÉS DE PACOTES

- Indice de Desconforto HUMIDEX

$$Pas = 6.112 \times \left(10 \times \frac{7.5 \times T}{237.7 \times T} \right) \times \frac{RH}{100}$$

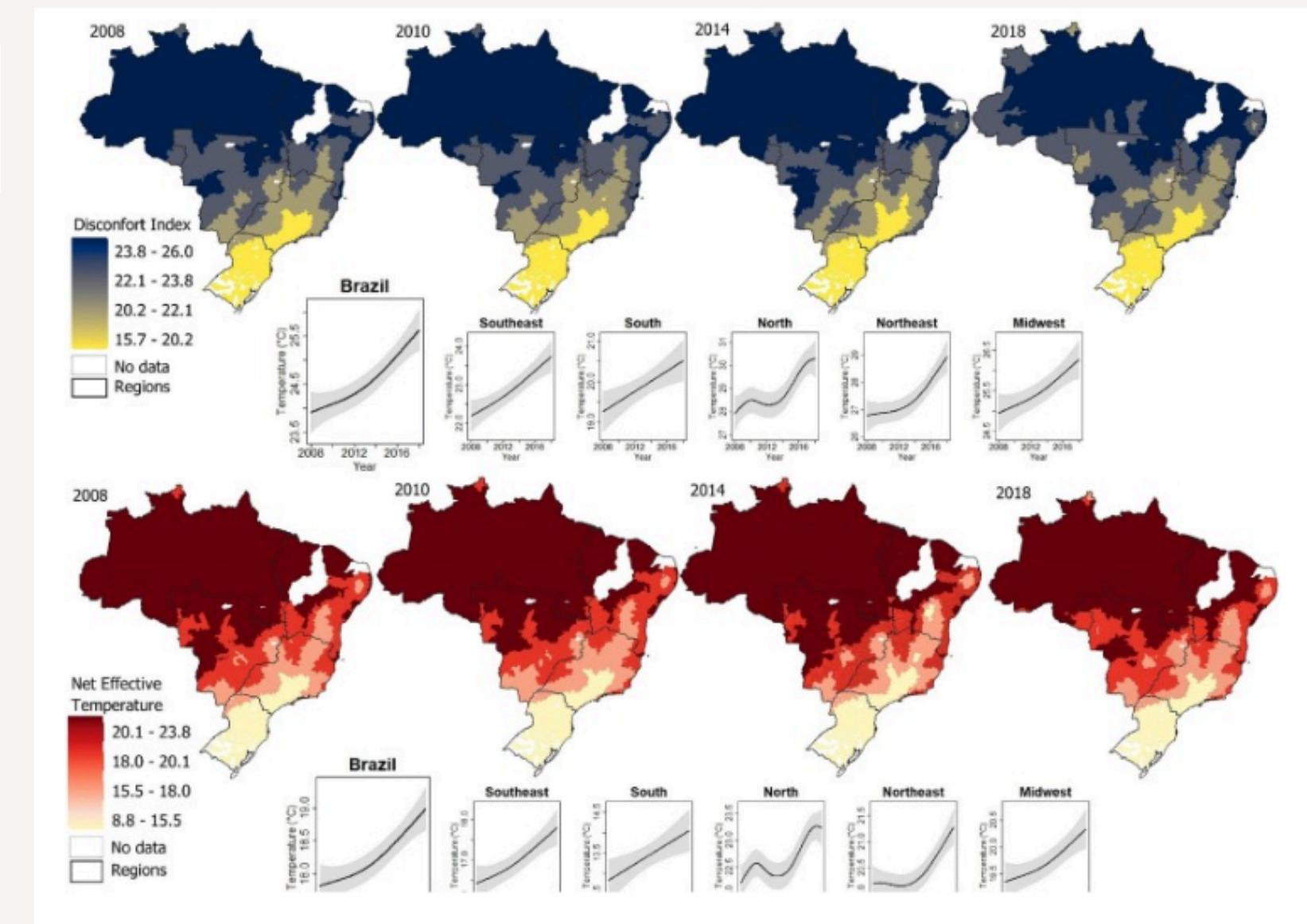
$$H = T + \frac{5}{9} \times (Pas - 10)$$

		Relative Humidity (%)															
		100	95	90	85	80	75	70	65	60	55	50	45	40	35	30	25
T E M P °C	43	56	54	51	49												
	42	56	54	52	50												
	41	56	54	52	50												
	40	57	54	52	51	49	47	45	43	42	41	39	37	35	33	31	30
	39	56	54	53	51	49	47	46	45	43	42	40	38	37	36	34	33
	38	57	56	54	52	51	49	48	47	45	43	42	41	39	38	37	36
	37	55	53	51	50	48	47	46	45	43	42	41	39	38	37	36	34
	36	58	57	56	55	54	53	51	50	48	47	45	43	42	40	38	37
	35	58	57	56	55	54	52	51	49	48	47	45	43	42	41	38	37
	34	58	57	56	55	53	52	51	49	48	47	45	43	42	41	39	37
	33	55	54	52	51	50	48	47	46	44	43	42	40	38	37	36	34
	32	52	51	50	49	47	46	45	43	42	41	39	38	37	36	34	33
	31	50	49	48	46	45	44	43	41	40	39	38	36	35	34	33	31
	30	48	47	46	44	43	42	41	40	38	37	36	35	34	33	31	30
	29	46	45	44	43	42	41	39	38	37	36	34	33	32	31	30	29
	28	43	42	41	41	39	38	37	36	35	34	33	32	31	30	29	28
	27	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26
	26	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25	24
	25	37	36	35	34	33	33	32	31	30	29	28	27	26	25	24	23
	24	35	34	33	33	32	31	30	29	28	28	27	26	26	25	24	23
	23	33	32	32	31	30	29	28	27	27	26	25	25	24	23	22	21
	22	31	29	29	28	28	27	26	26	24	24	23	23	22	22	21	
	21	29	29	28	27	27	26	26	24	24	23	23	22	22	21		
	20	27	27	26	25	25	25	24	23	23	22	22	21				

INDICES ATRAVÉS DE PACOTES

- Indice de Desconforto Temperatura Efetiva

$$NET = 37 - \frac{37-T}{[(0.68-0.0014) \times RH] + \frac{1}{1.76+1.4V^{0.75}}} - (0.29 \times T) \times [1 - (0.01 \times RH)]$$



ÍNDICES ATRAVÉS DE PACOTES

- Finalizar Itens 6 a 9