TÓPICOS EM CIÊNCIA DE DADOS PARA O ESPORTE

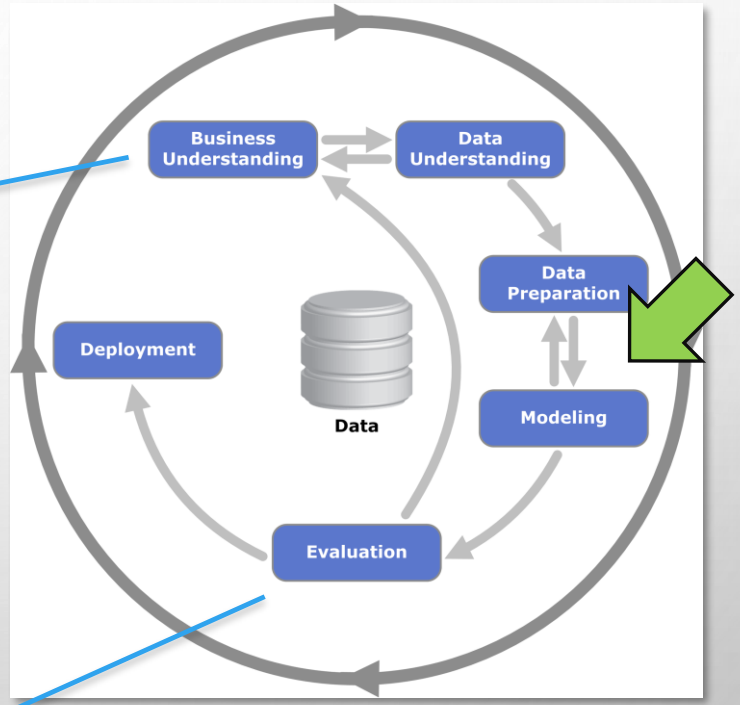
ESTATÍSTICAS PARA RANQUEAMENTO

DIEGO RODRIGUES DSC

INFNET

CRONOGRAMA

DIA	NÚMERO	ÁREA	AULA	TRABALHOS
30/1/2024	1	Intro	Introdução a Disciplina e Organização do Ambiente	
1/2/2024	2	Dados	Coleta de Dados e Sensoriamento	
6/2/2024	3	Estatística	Variáveis Aleatórias	Grupos
8/2/2024	4		Análise Exploratória	
15/2/2024	5		Estatísticas para Ranqueamento	
20/02/2024	6		Ranqueamento Estatístico : ELO	
22/02/2024	7		Ranqueamento Estatístico : Glicko	
27/2/2024	8		Ranqueamento Estatístico : TrueSkill	
29/2/2024	9		Ranqueamento Estatístico : XELO	Base de Dados
5/3/2024	10	ML	Modelos de Aprendizado de Máquina	
7/3/2024	11		Machine Learning: Classificação	
12/3/2024	12		Machine Learning: Regressão	
14/3/2024	13		Machine Learning: Agrupamento	Pesquisa
19/3/2024	14		Machine Learning: Visão Computacional	
21/3/2024	15	Esportes	Aplicações & Artigos: Esportes Independentes	Modelo
26/3/2024	16		Aplicações & Artigos: Esportes de Objeto	
28/3/2024	17		Aplicações & Artigos: Esportes de Combate	
2/4/2024	18		Aplicações & Artigos : Betting	
4/4/2024	19	Workshop	Workshop	
9/4/2024	20		Apresentações de Trabalhos I	Apresentação
11/4/2024	21		Apresentações de Trabalhos II	



CRISP-DM

AGENDA

- PARTE 1 : TEORIA
- ANÁLISE EXPLORATÓRIA MULTIVARIADA
  - TESTES DE HIPÓTESE
    - TESTE DE PROPORÇÃO
    - TESTE DE MÉDIAS
    - ANOVA E ETC
  - RELAÇÕES
    - CORRELAÇÃO
- VARIÁVEIS LATENTES
  - COMPARAÇÃO DE PARES
  - MODELO BRADLEY TERRY
- PARTE 2 : PRÁTICA
  - PANDAS + MATPLOTLIB → VARIÁVEIS ALEATÓRIAS NO ESPORTE

SETUP INICIAL DO AMBIENTE PYTHON

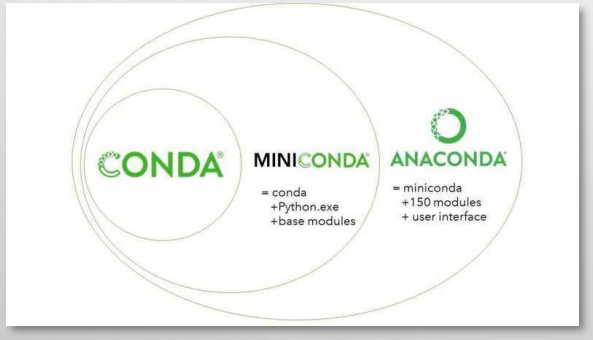
6. Estimação e Inferência

4. Variáveis Aleatórias

5. Visualização

statsmodels

1. Editor de Código



2. Gestor de Ambiente

3. Ambiente Python do Projeto

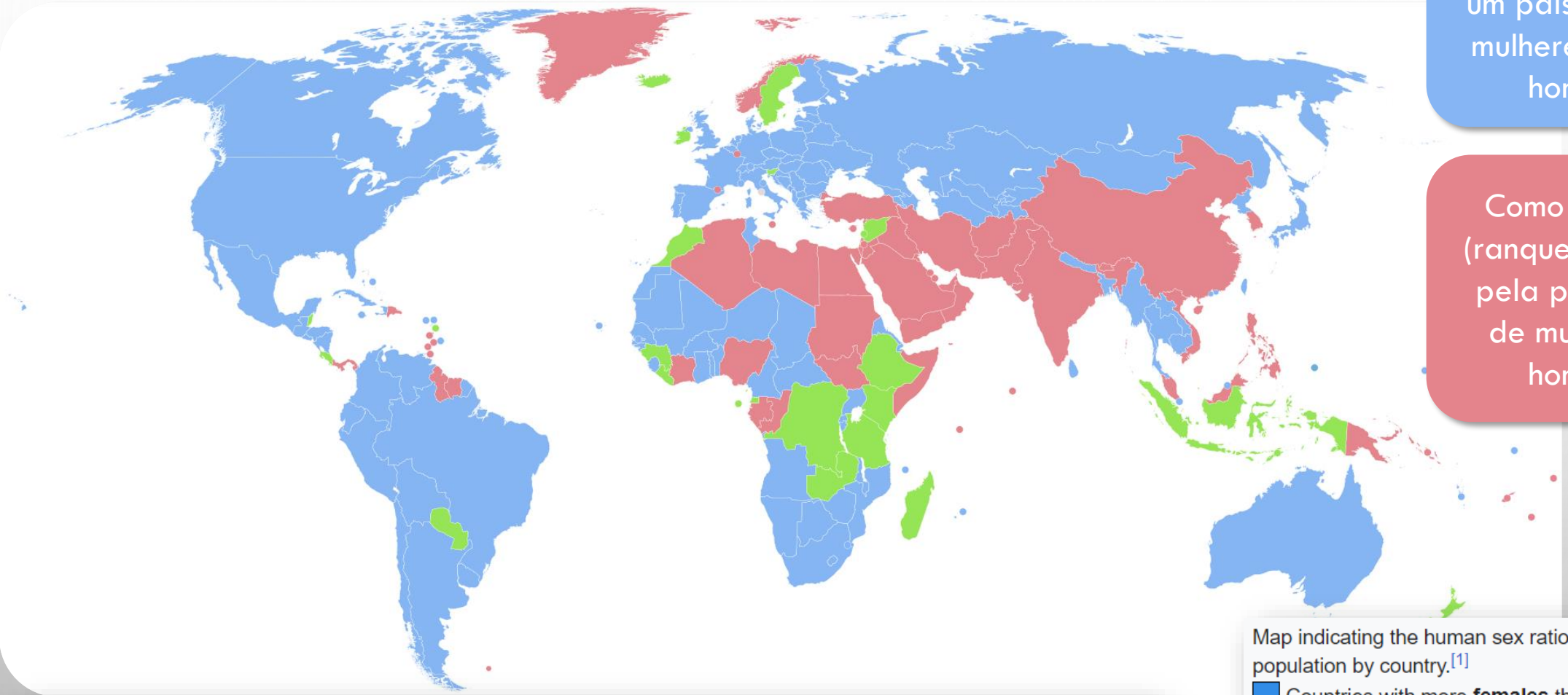
3. Notebook Dinâmico

ANÁLISE EXPLORATÓRIA MULTIVARIADA

TESTES DE HIPÓTESE

Como aferir que, estatisticamente, um país tem mais mulheres do que homens?

Como ordenar (ranquear) países pela proporção de mulheres e homens?



Map indicating the human sex ratio of total population by country.[1]

- Countries with more **females** than males
- Countries with more **males** than females
- Countries with very similar proportions of males and females (to 3 significant figures, i.e., 1.00 males to 1.00 females)
- No data

History [edit]

Early use [edit]

While hypothesis testing was popularized early in the 20th century, early forms were used in the 1700s. The first use is credited to John Arbuthnot (1710),[1] followed by Pierre-Simon Laplace (1770s), in analyzing the human sex ratio at birth; see § Human sex ratio.

TESTE DE HIPÓTESE - FORMULAÇÃO

CONTEXTO	*"QUAL TIME TEM A MAIOR PROBABILIDADE DE VENCER EM 2023, TIME A OU TIME B?"*
HIPÓTESE NULA H_0	*"TIME A E TIME B TEM A MESMA PROBABILIDADE DE VENCER EM 2023"*
HIPÓTESE ALTERNATIVA H_A	*"TIME A E TIME B TEM PROBABILIDADES DE VITÓRIAS DIFERENTES EM 2023"*
ESTATÍSTICA TESTE	*"TESTE DE DUAS PROPORÇÕES COM VARIÂNCIA DESCONHECIDA SOB HIPÓTESE NULA H0 → PA = PB"*
NÍVEL DE SIGNIFICÂNCIA	*"QUAL A PROBABILIDADE DE REJEITAR DA HIPÓTESE NULA, QUANDO ELA É VERDADEIRA"*
P VALOR	*"PROBABILIDADE DO RESULTADO OBTIDO SE A HIPÓTESE NULA FOR VERDADEIRA"*

TESTE DE HIPÓTESE – ESTATÍSTICAS PARA TESTES

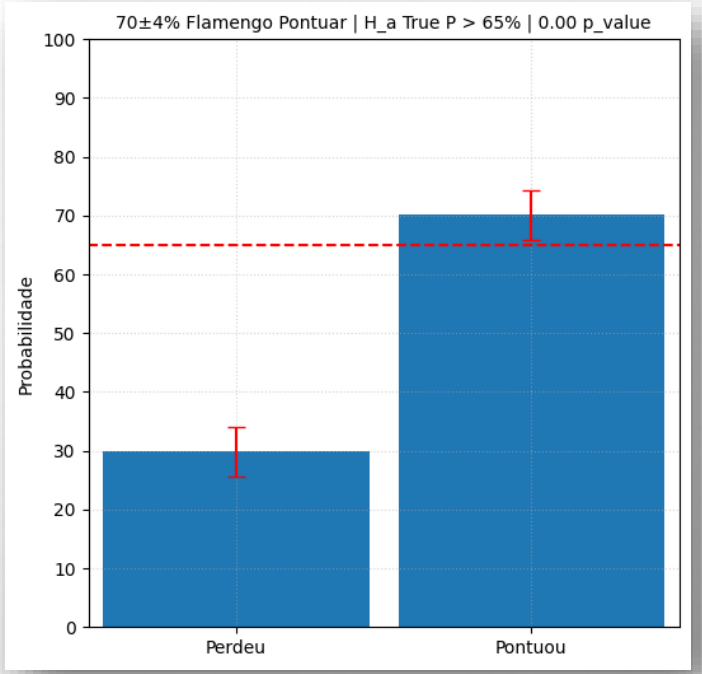
Name	Formula	Assumptions or notes
One-sample z -test	$z = \frac{\overline{x} - \mu_0}{(\sigma/\sqrt{n})}$	(Normal population **or** n large) **and** σ known. (z is the distance from the mean in relation to the standard deviation of the mean). For non-normal distributions it is possible to calculate a minimum proportion of a population that falls within k standard deviations for any k (see: *Chebyshev's inequality*).
Two-sample z-test	$z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Normal population **and** independent observations **and** σ_1 and σ_2 are known where d_0 is the value of $\mu_1 - \mu_2$ under the null hypothesis
One-sample t-test	$t = \frac{\overline{x} - \mu_0}{(s/\sqrt{n})}, \quad df = n - 1$	(Normal population **or** n large) **and** σ unknown
Paired t-test	$t = \frac{\overline{d} - d_0}{(s_d/\sqrt{n})}, \quad df = n - 1$	(Normal population of differences **or** n large) **and** σ unknown
Two-sample pooled t-test, equal variances	$t = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad df = n_1 + n_2 - 2$ [3]	(Normal populations **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 = \sigma_2$ unknown
Two-sample unpooled t-test, unequal variances (Welch's t-test)	$t = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ [3]	(Normal populations **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 \neq \sigma_2$ both unknown
One-proportion z-test	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n}$	$n \cdot p_0 > 10$ **and** $n\,(1 - p_0) > 10$ **and** it is a SRS (Simple Random Sample), see notes.
Two-proportion z-test, pooled for $H_0 : p_1 = p_2$	$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$	$n_1\,p_1 > 5$ **and** $n_1(1 - p_1) > 5$ **and** $n_2\,p_2 > 5$ **and** $n_2(1 - p_2) > 5$ **and** independent observations, see notes.

TESTE DE PROPORÇÕES – 1 PROPORÇÃO

H0 – PROBABILIDADE DO FLAMENGO PONTUAR É MENOR OU IGUAL A P

HA – PROBABILIDADE DO FLAMENGO PONTUAR É MAIOR QUE P

TESTE UNILATERAL E ALTERNATIVA “MAIOR QUE”



P_FLAMENGO = 70%

P_TESTE = 65%

SIGNIFICÂNCIA = 0.01

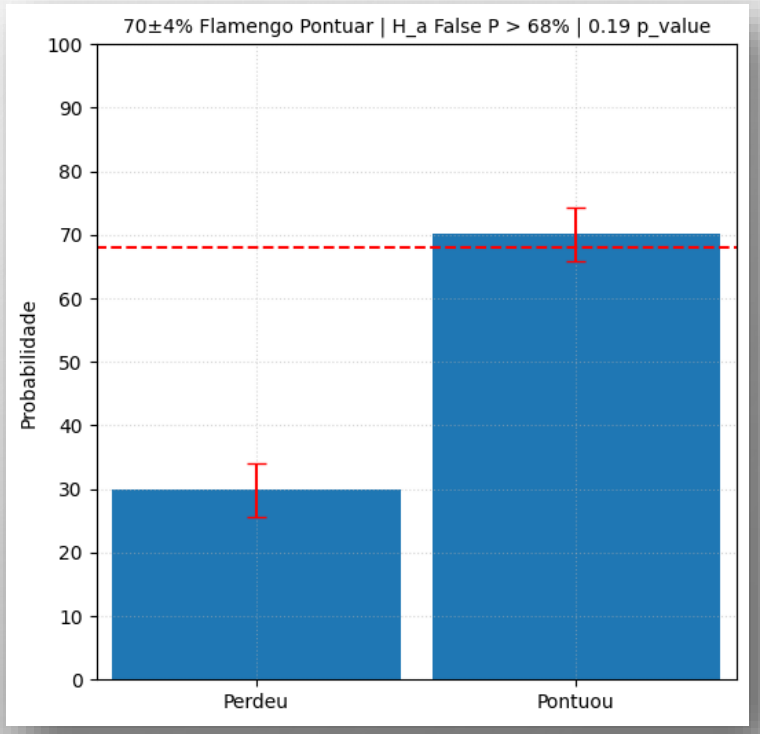
P VALOR = 0.00

P_FLAMENGO = 70%

P_TESTE = 68%

SIGNIFICÂNCIA = 0.01

P VALOR = 0.19

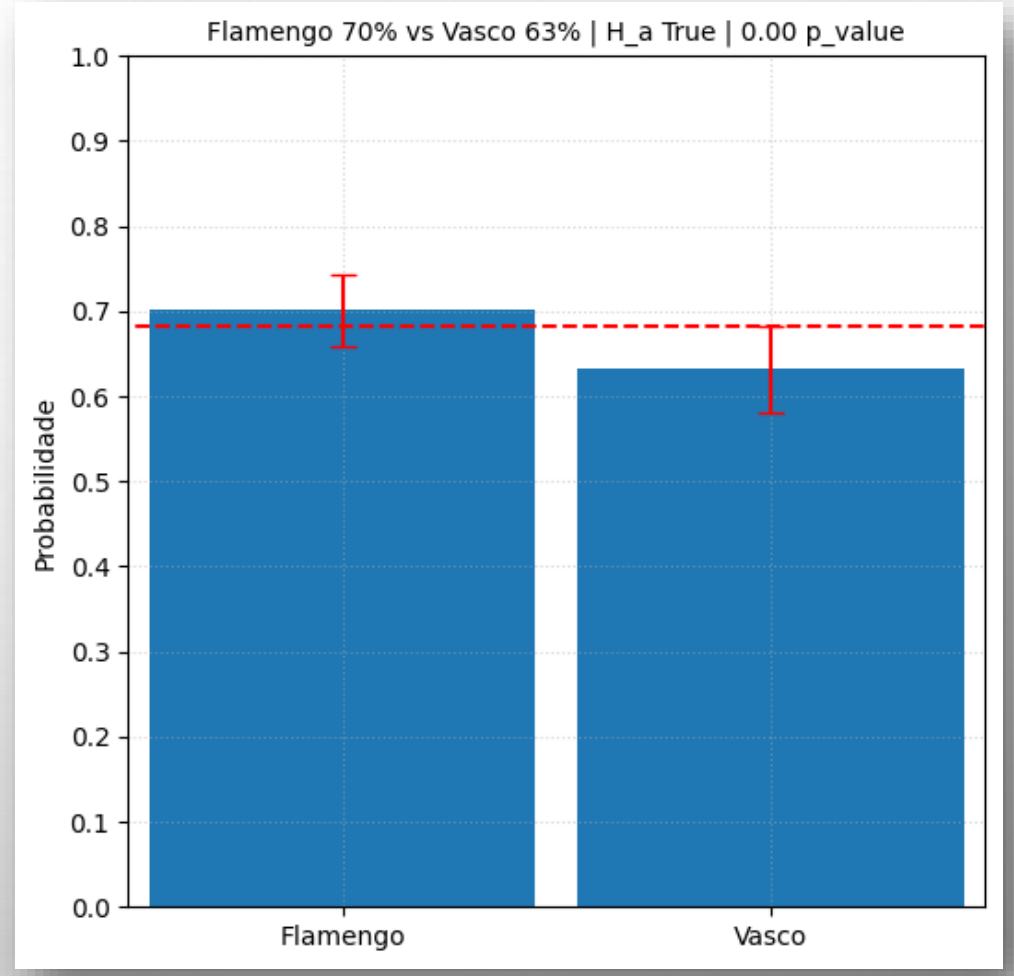


TESTE DE PROPORÇÕES – 2 PROPORÇÕES

H0 – PROBABILIDADE DO FLAMENGO PONTUAR É MENOR OU IGUAL AO DO VASCO

HA – PROBABILIDADE DO FLAMENGO PONTUAR É MAIOR QUE A DO VASCO

TESTE UNILATERAL 2 PROPORÇÕES E ALTERNATIVA “MAIOR QUE”



P_FLAMENGO = 70%

P_VASCO = 63%

N_FLAMENGO = 780

N_VASCO = 590

SIGNIFICÂNCIA = 0.01

P VALOR = 0.00

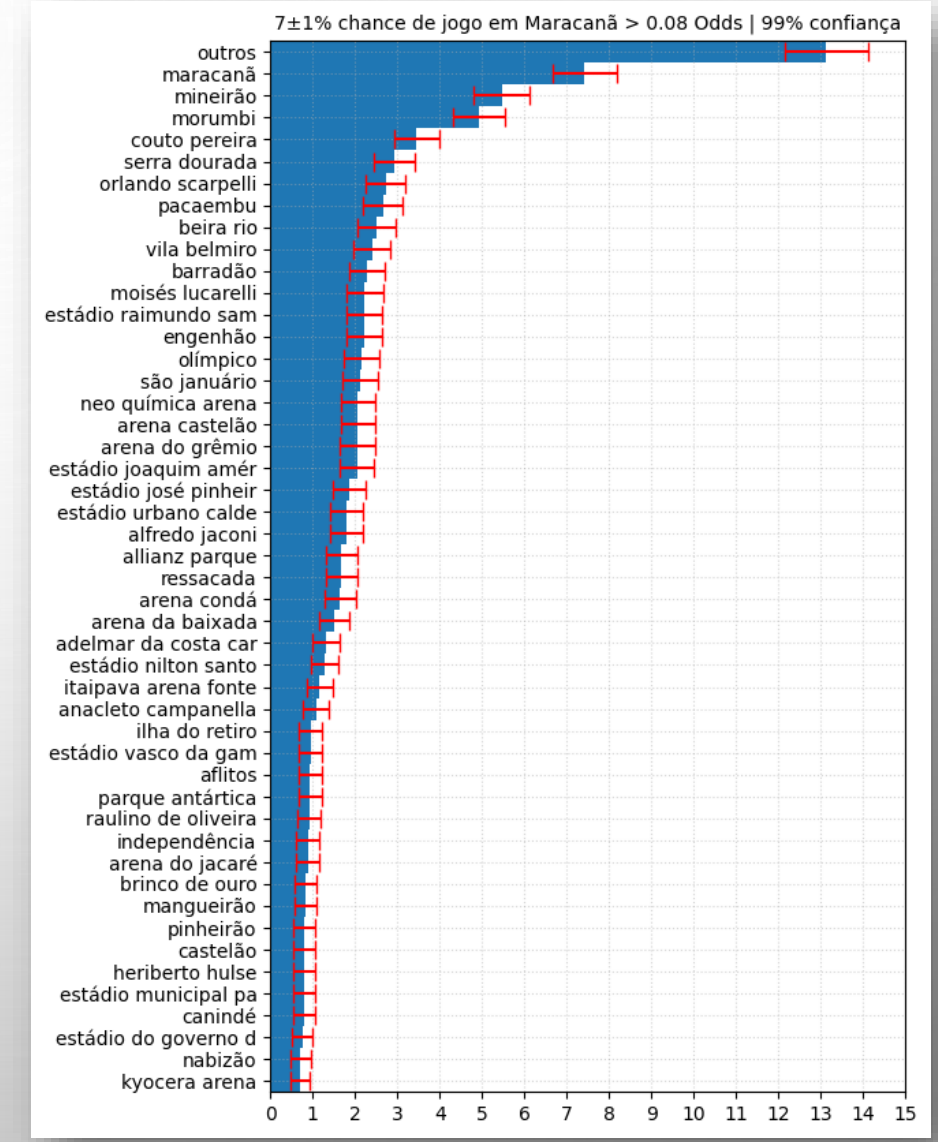
TESTE CHI QUADRADO

P VALOR = 0.00

H0 – TODOS OS ESTÁDIOS TEM A MESMA CHANCE DE TER JOGOS

HA – ESTÁDIOS TEM CHANCE DISTINTAS DE TER JOGOS

TESTE CHI QUADRADO DE K PROPORÇÕES

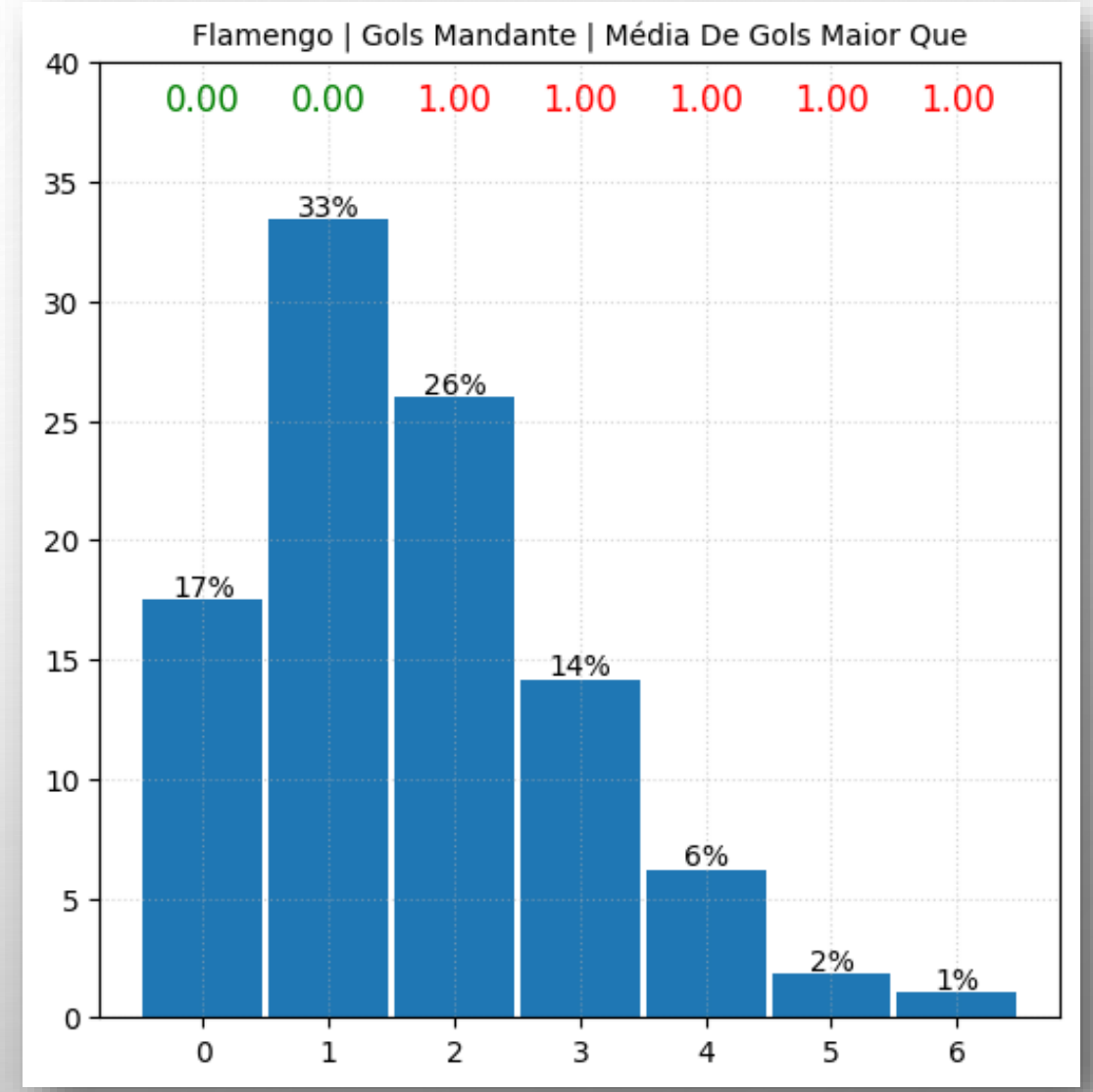


TESTE DE MÉDIA

H0 – MÉDIA DE GOLS DO FLAMENGO COMO MANDANTE É MENOR OU IGUAL A X

HA – MÉDIA DE GOLS DO FLAMENGO COMO MANDANTE É MAIOR QUE X

TESTE Z PARA MÉDIAS

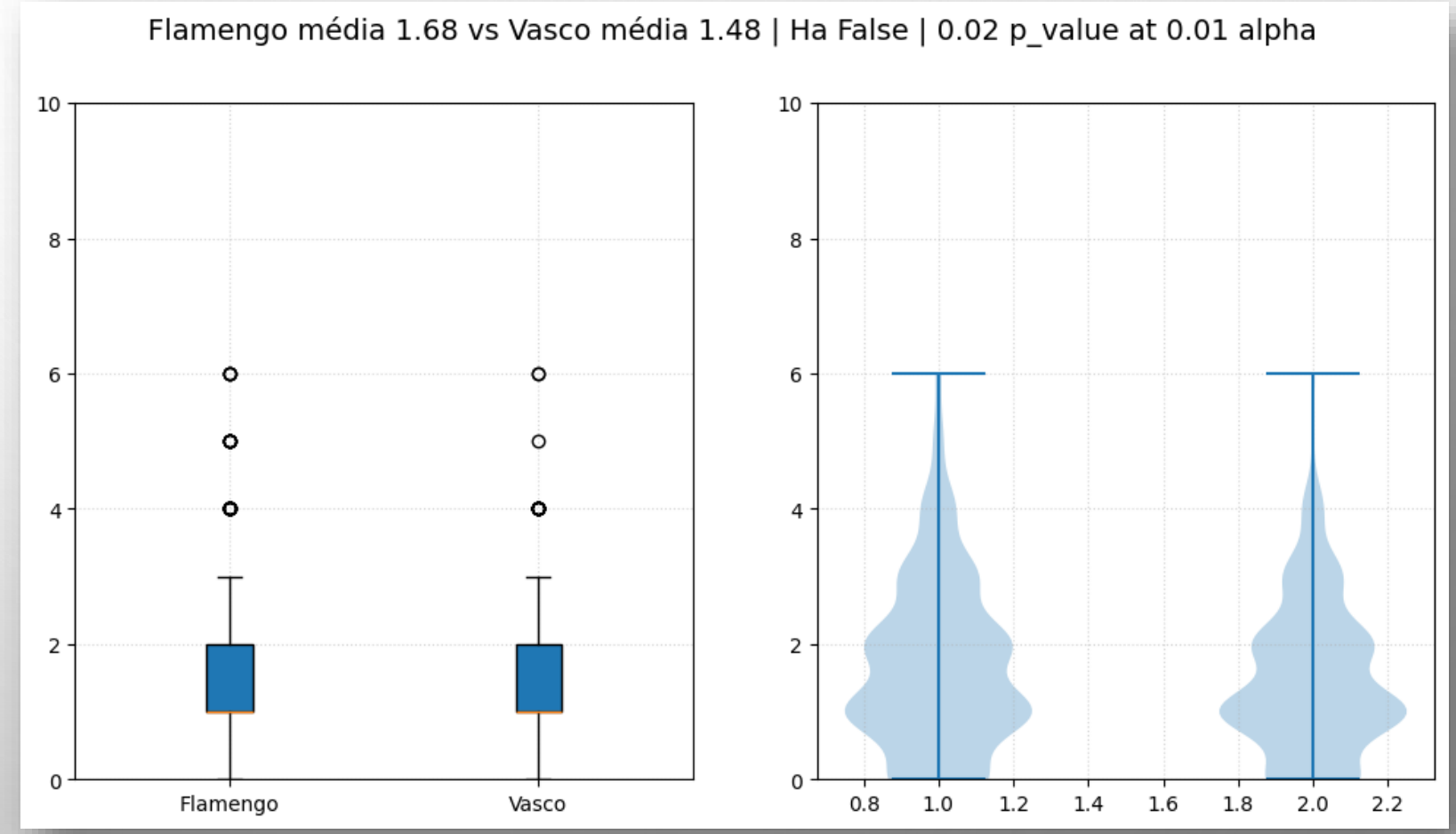


TESTE DE MÉDIA – DUAS POPULAÇÕES

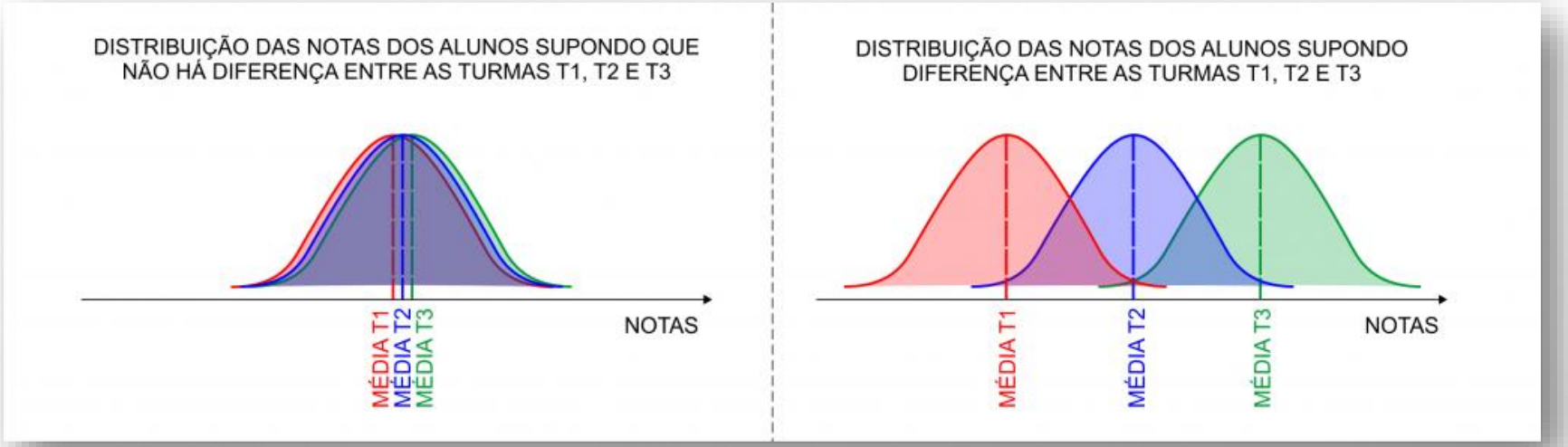
H0 – MÉDIA DE GOLS DO FLAMENGO COMO MANDANTE É MENOR OU IGUAL A DO VASCO COMO MANDANTE

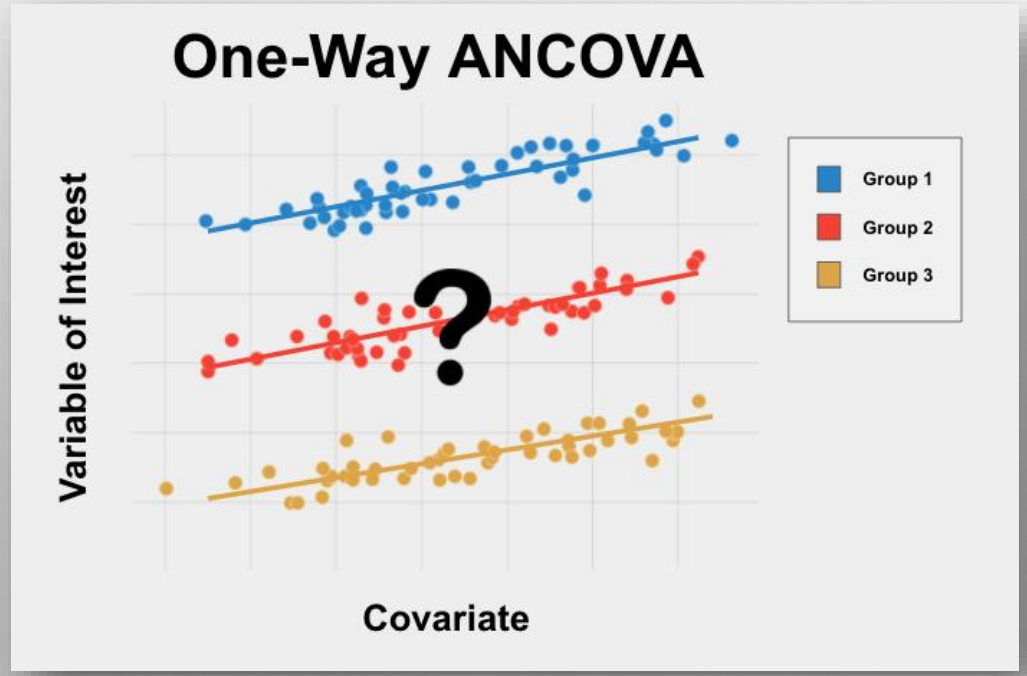
HA – MÉDIA DE GOLS DO FLAMENGO COMO MANDANTE É MAIOR QUE A DO VASCO COMO MANDANTE

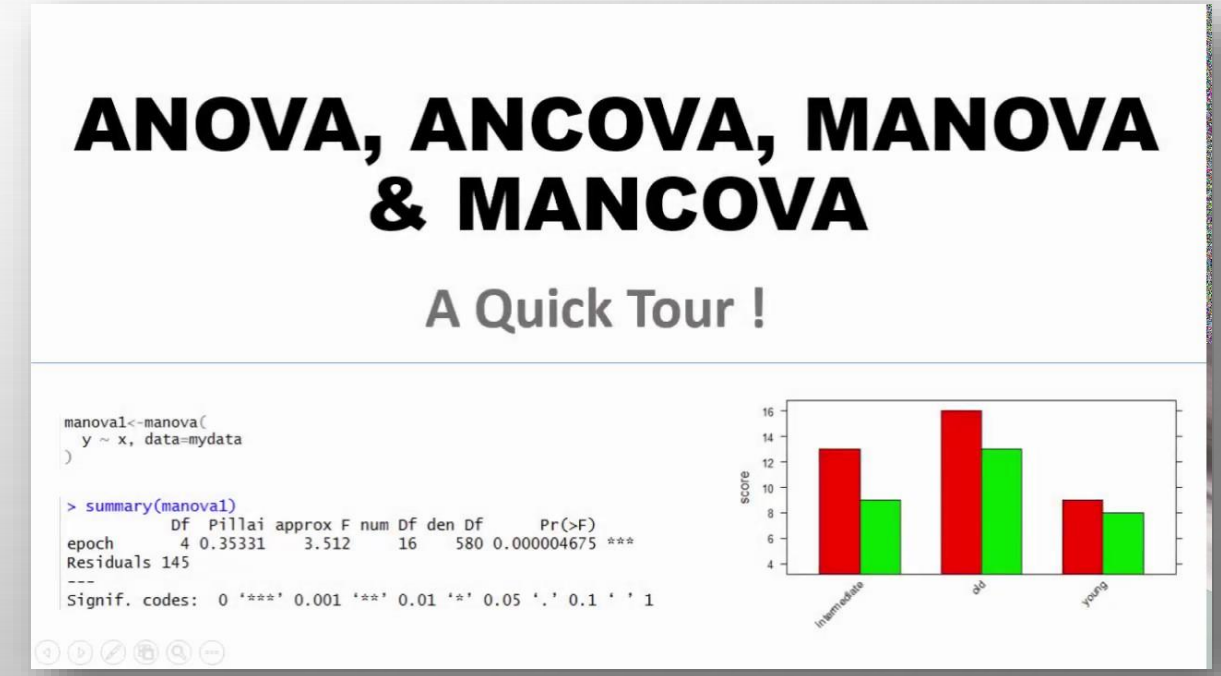
TESTE Z PARA MÉDIAS



ANOVA E ETC

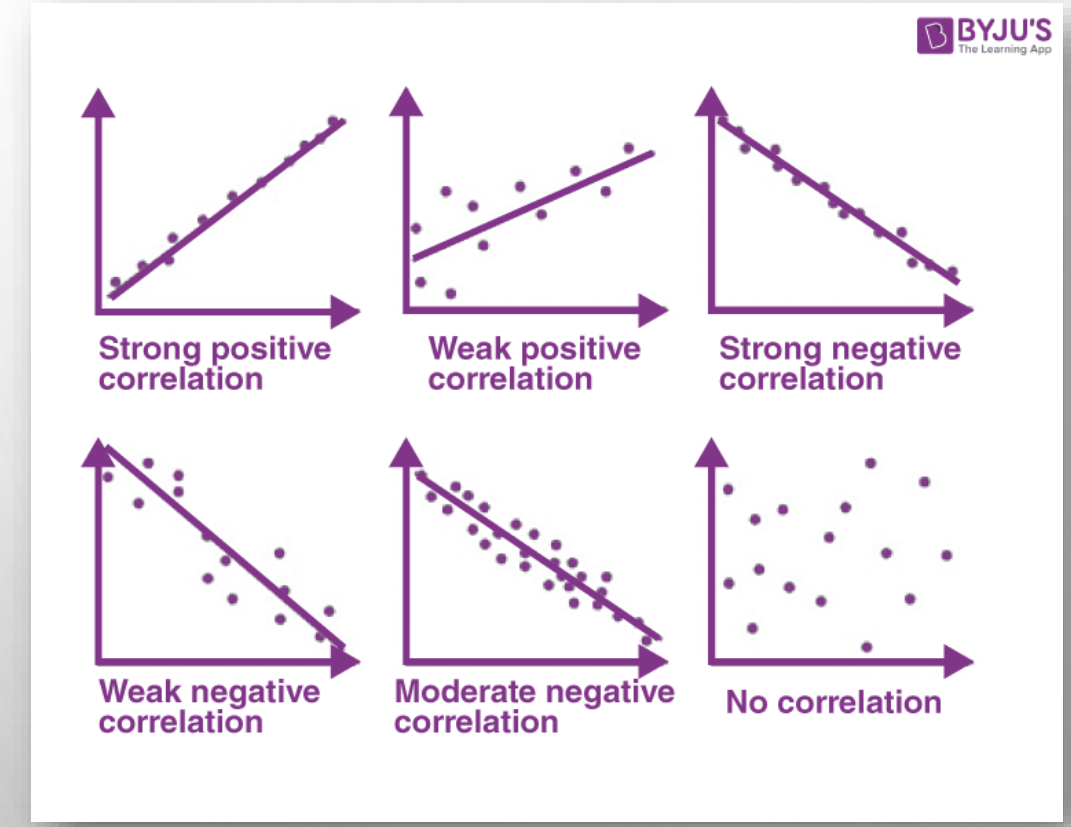

DISTRIBUIÇÃO DAS NOTAS DOS ALUNOS SUPONDO QUE NÃO HÁ DIFERENÇA ENTRE AS TURMAS T1, T2 E T3
MÉDIA T1
MÉDIA T2
MÉDIA T3
NOTAS
DISTRIBUIÇÃO DAS NOTAS DOS ALUNOS SUPONDO DIFERENÇA ENTRE AS TURMAS T1, T2 E T3
MÉDIA T1
MÉDIA T2
MÉDIA T3
NOTAS


One-Way ANCOVA
Variable of Interest
Covariate
Group 1
Group 2
Group 3
?


ANOVA, ANCOVA, MANOVA & MANCOVA
A Quick Tour !
manova1<-manova(
y ~ x, data=mydata
)
> summary(manova1)
Df Pillai approx F num Df den Df Pr(>F)
epoch 4 0.35331 3.512 16 580 0.000004675 ***
Residuals 145

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
score
intermediate
old
young

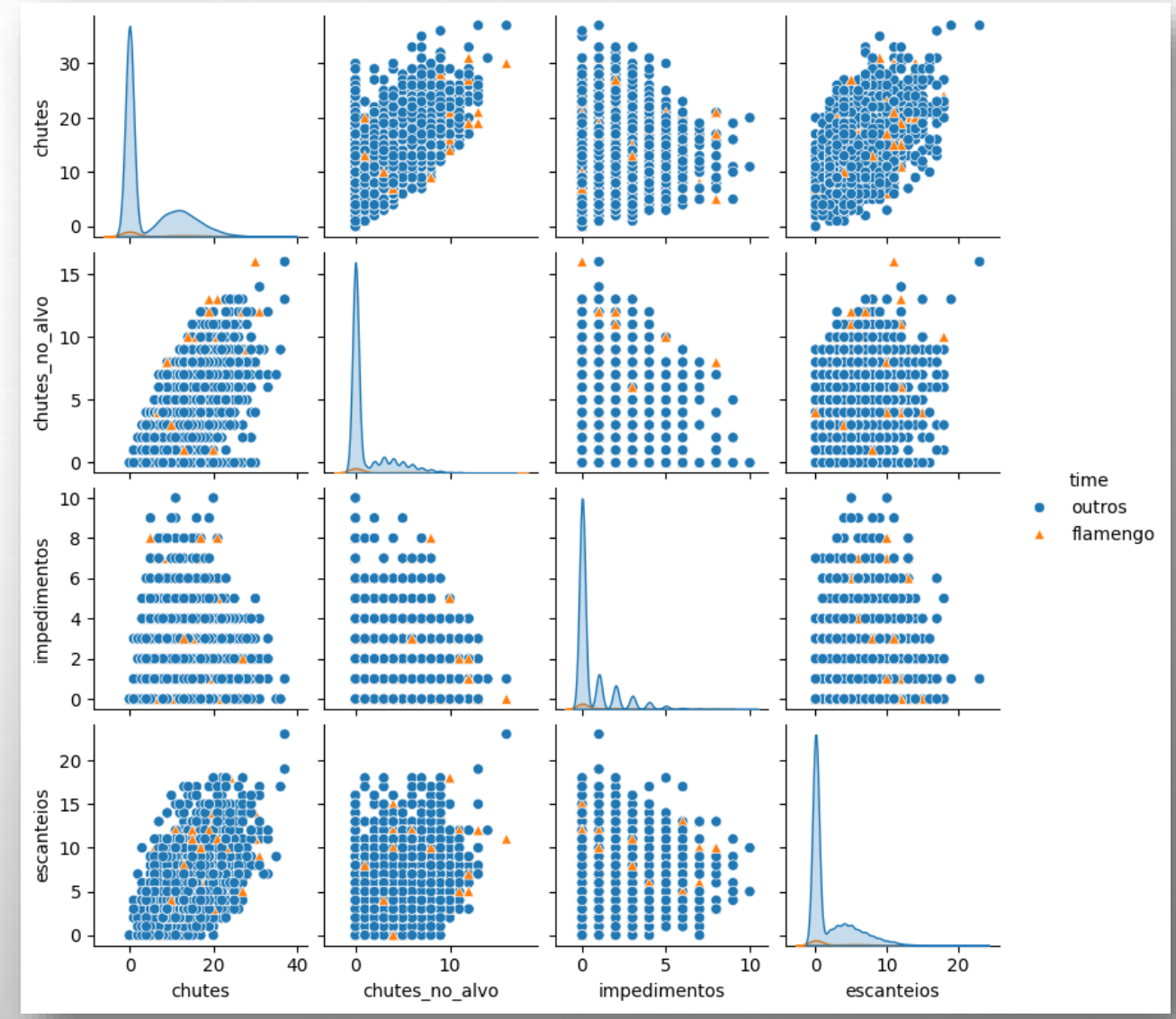
CORRELAÇÃO (DE PEARSON)



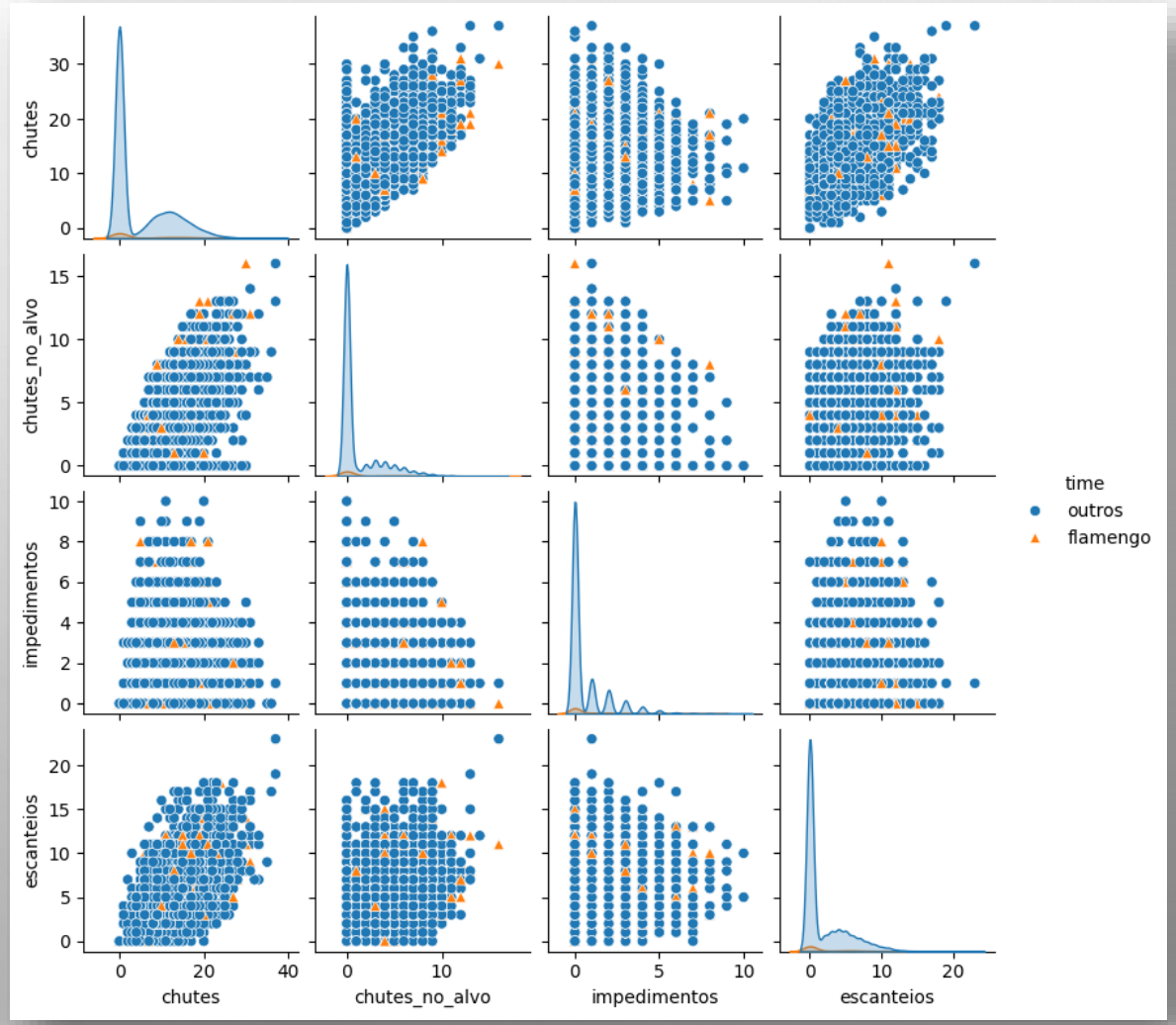
For a sample [edit]

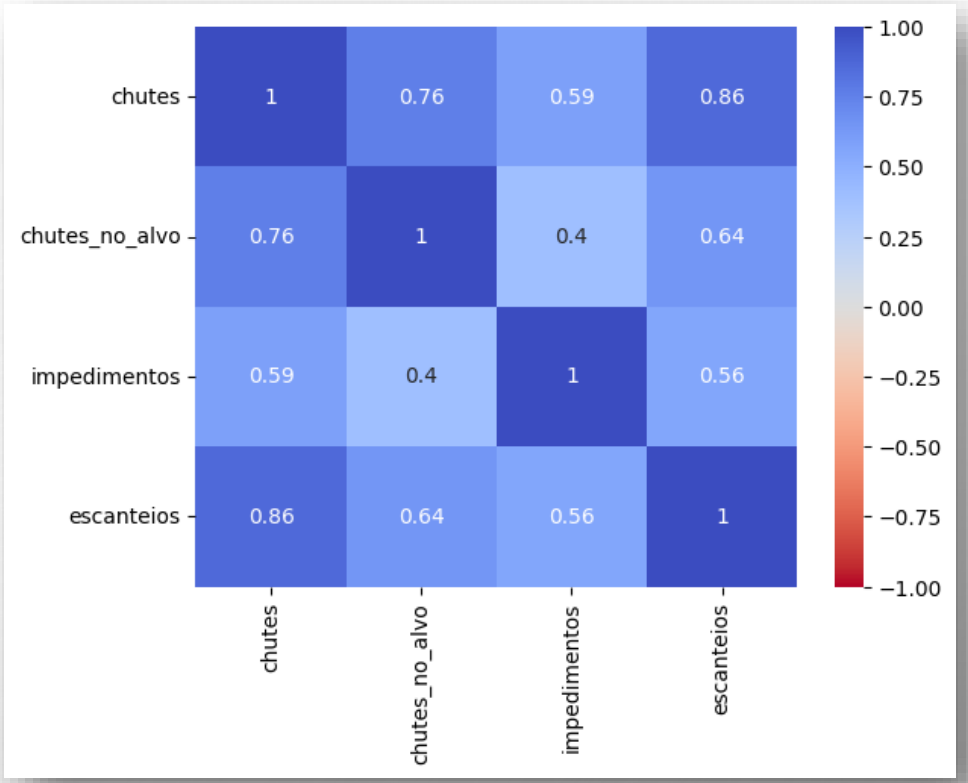
Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as

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



CORRELAÇÃO (DE PEARSON)

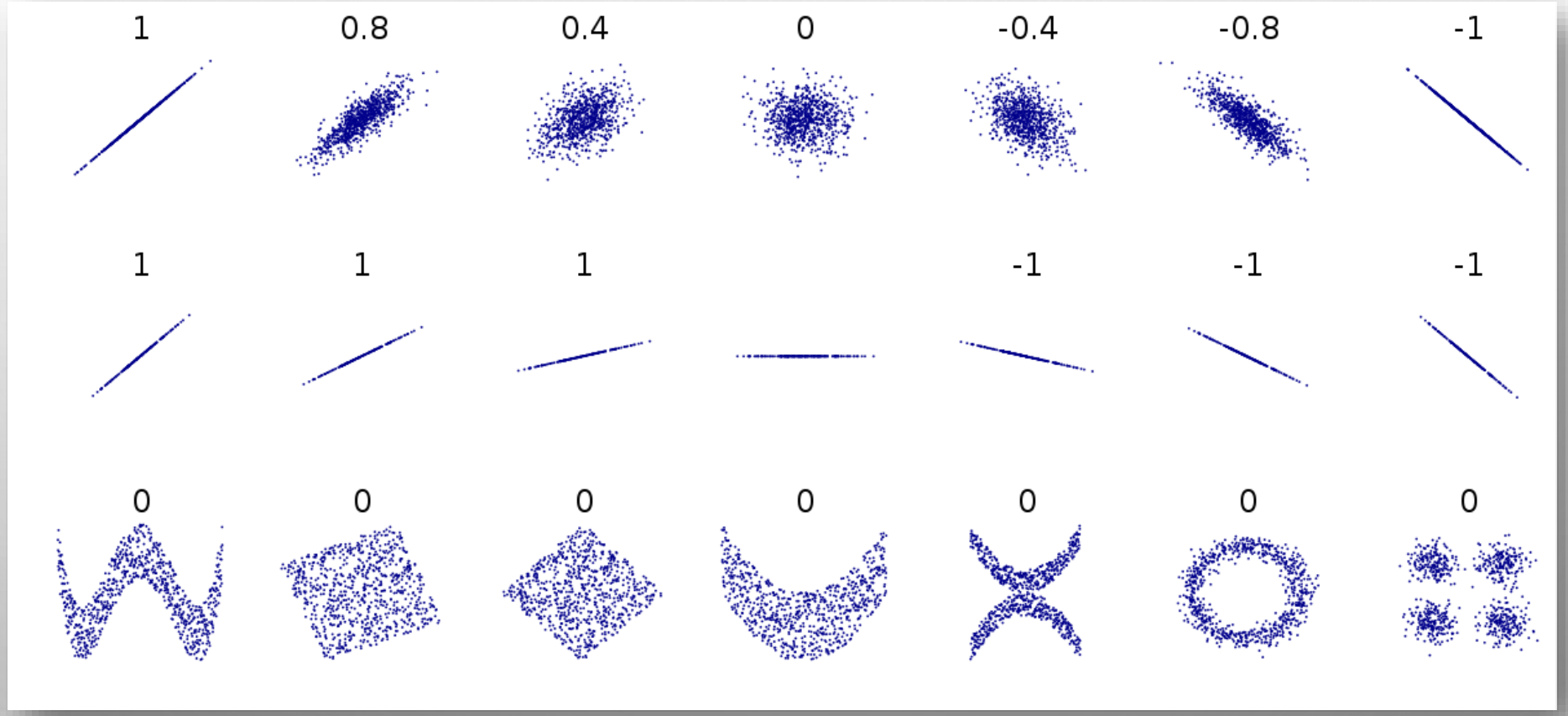




MAIOR CORRELAÇÃO : CHUTES COM ESCANTEIOS

MENOR CORRELAÇÃO : CHUTES NO ALVO COM IMPEDIMENTOS

CORRELAÇÃO (DE PEARSON)



INFORMAÇÃO MÚTUA

DIVERGÊNCIA DE KULLBACK LEIBER

COMPARAÇÃO DE PARES

Pairwise comparison

10 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

This article is about pairwise comparisons in psychology. For statistical analysis of paired comparisons, see paired difference test.

Pairwise comparison generally is any process of comparing entities in pairs to judge which of each entity is preferred, or has a greater amount of some quantitative property, or whether or not the two entities are identical. The method of pairwise comparison is used in the scientific study of preferences, attitudes, voting systems, social choice, public choice, requirements engineering and multiagent AI systems. In psychology literature, it is often referred to as **paired comparison**.

Prominent psychometrician L. L. Thurstone first introduced a scientific approach to using pairwise comparisons for measurement in 1927, which he referred to as the law of comparative judgment. Thurstone linked this approach to psychophysical theory developed by Ernst Heinrich Weber and Gustav Fechner. Thurstone demonstrated that the method can be used to order items along a dimension such as preference or importance using an interval-type scale.

Mathematician Ernst Zermelo (1929) first described a model for pairwise comparisons for chess ranking in incomplete tournaments, which serves as the basis (even though not credited for a while) for methods such as the Elo rating system and is equivalent to the Bradley–Terry model that was proposed in 1952.

Overview [edit]

If an individual or organization expresses a preference between two mutually distinct alternatives, this preference can be expressed as a pairwise comparison. If the two alternatives are *x* and *y*, the following are the possible pairwise comparisons:

The agent prefers *x* over *y*: "*x* > *y*" or "*xPy*"

The agent prefers *y* over *x*: "*y* > *x*" or "*yPx*"

The agent is indifferent between both alternatives: "*x* = *y*" or "*xIy*"

Probabilistic models [edit]

In terms of modern psychometric theory probabilistic models, which include Thurstone's approach (also called the law of comparative judgment), the Bradley–Terry–Luce (BTL) model, and general stochastic transitivity models,[1] are more aptly regarded as measurement models. The Bradley–Terry–Luce (BTL) model is often applied to pairwise comparison data to scale preferences. The BTL model is identical to Thurstone's model if the simple logistic function is used. Thurstone used the normal distribution in applications of the model. The simple logistic function varies by less than 0.01 from the cumulative normal ogive across the range, given an arbitrary scale factor.

In the BTL model, the probability that object *j* is judged to have more of an attribute than object *i* is:

$$\Pr\{X_{ji} = 1\} = \frac{e^{\delta_j - \delta_i}}{1 + e^{\delta_j - \delta_i}} = \sigma(\delta_j - \delta_i),$$

where δ_i is the scale location of object i; σ is the logistic function (the inverse of the logit). For example, the scale location might represent the perceived quality of a product, or the perceived weight of an object.

The BTL model, the Thurstonian model as well as the Rasch model for measurement are all closely related and belong to the same class of stochastic transitivity.

MODELO BRADLEY TERRY

Bradley–Terry model

1 language

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **Bradley–Terry model** is a probability model for the outcome of pairwise comparisons between individuals, teams, or objects. Given a pair of individuals i and j drawn from some population, it estimates the probability that the pairwise comparison $i > j$ turns out true, as

$$P(i > j) = \frac{p_i}{p_i + p_j} \tag{1}$$

where p_i is a positive real-valued score assigned to individual i. The comparison $i > j$ can be read as "i is preferred to j", "i ranks higher than j", or "i beats j", depending on the application.

For example, p_i might represent the skill of a team in a sports tournament and $P(i > j)$ the probability that i wins a game against j.[1][2] Or p_i might represent the quality or desirability of a commercial product and $P(i > j)$ the probability that a consumer will prefer product i over product j.

The Bradley–Terry model can be used in the forward direction to predict outcomes, as described, but is more commonly used in reverse to infer the scores p_i given an observed set of outcomes.[2] In this type of application p_i represents some measure of the strength or quality of i and the model lets us estimate the strengths from a series of pairwise comparisons. In a survey of wine preferences, for instance, it might be difficult for respondents to give a complete ranking of a large set of wines, but relatively easy for them to compare sample pairs of wines and say which they feel is better. Based on a set of such pairwise comparisons, the Bradley–Terry model can then be used to derive a full ranking of the wines.

Once the values of the scores p_i have been calculated, the model can then also be used in the forward direction, for instance to predict the likely outcome of comparisons that have not yet actually occurred. In the wine survey example, for instance, one could calculate the probability that someone will prefer wine i over wine j, even if no one in the survey directly compared that particular pair.

DESAFIO: RANKING DO BRASILEIRÃO POR ALGUM CRITÉRIO ESTATÍSTICO

PRÓXIMA AULA
LEITURA: ALGORITMO ELO