

TÓPICOS EM CIÊNCIA DE DADOS PARA O ESPORTE

ESTATÍSTICAS PARA RANQUEAMENTO

DIEGO RODRIGUES DSC

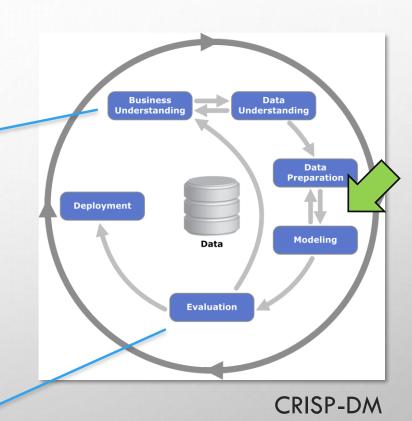
INFNET

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

CRONOGRAMA

DIA	NÚMERO	ÁREA	AULA	TRABALHOS
10/10/2023	1	Intro	Introdução a Disciplina e Organização do Ambiente	
17/10/2023	2	Dados	Coleta de Dados e Sensoriamento	
19/10/2023	3		Variáveis Aleatórias	Grupos
24/10/2023	4		Análise Exploratória	
26/10/2023	5		Estatísticas para Ranqueamento	
31/10/2023	6	Estatística	Ranqueamento Estatístico : ELO	Base de Dados
07/11/2023	7		Ranqueamento Estatístico : Glicko	
09/11/2023	8		Ranqueamento Estatístico : TrueSkill	
14/11/2023	9		Ranqueamento Estatístico : XELO	
16/11/2023	10		Modelos de Aprendizado de Máquina	Pesquisa
21/11/2023	11		Machine Learning: Classificação	
23/11/2023	12	ML	Machine Learning: Regressão	
28/11/2023	13		Machine Learning: Agrupamento	
30/11/2023	14		Machine Learning: Visão Computacional	Modelo
5/12/2023	15		Aplicações & Artigos: Esportes Independentes	
7/12/2023	16	Fanantas	Aplicações & Artigos: Esportes de Combate	
12/12/2023	17	Esportes	Aplicações & Artigos: Esportes de Objeto	
14/12/2023	18		Aplicações & Artigos : Betting	
19/12/2023	19	Workshop	Workshop	
21/12/2023	20	www.siiop	Apresentações de Trabalhos	Apresentação



SETUP INICIAL DO AMBIENTE PYTHON









4. Variáveis Aleatórias

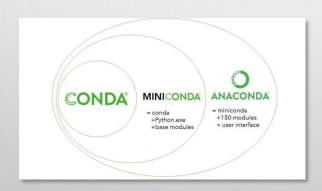


Inferência





1. Editor de Código



2. Gestor de Ambiente



3. Ambiente
Python do Projeto



3. Notebook Dinâmico



ANÁLISE EXPLORATÓRIA MULTIVARIADA

TESTES DE HIPÓTESE



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

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.

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

TESTE DE HIPÓTESE - FORMULAÇÃO

CONTEXTO

HIPÓTESE NULA Ho

HIPÓTESE ALTERNATIVA HA

ESTATÍSTICA TESTE

NÍVEL DE SIGNIFICÂNCIA

P VALOR

"QUAL TIME TEM A MAIOR PROBABILIDADE DE VENCER EM 2023, TIME A OU TIME B?"

"TIME A E TIME B TEM A MESMA
PROBABILIDADE DE VENCER EM 2023"

"TIME A E TIME B TEM PROBABILIDADES DE VITÓRIAS DIFERENTES EM 2023"

"TESTE DE DUAS PROPORÇÕES COM VARIÂNCIA DESCONHECIDA SOB HIPÓTESE NULA HO → PA = PB"

> "QUAL A PROBABILIDADE DE REJEITAR DA HIPÓTESE NULA, QUANDO ELA É VERDADEIRA"

"PROBABILIDADE DO RESULTADO OBTIDO SE A HIPÓTESE NULA FOR VERDADEIRA"

TESTE DE HIPÓTESE – ESTATÍSTICAS PARA TESTES

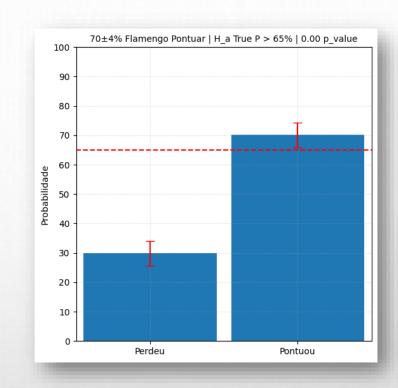
Name	Formula	Assumptions or notes	
One-sample z -test	$z=rac{\overline{x}-\mu_0}{(\sigma/\sqrt{n})}$	(Normal population or n large) and σ known. (z is the distance from the mean in relation to the standard deviation of the mean). For non-normal distributions it is possible to calculate a minimum proportion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

HO – 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

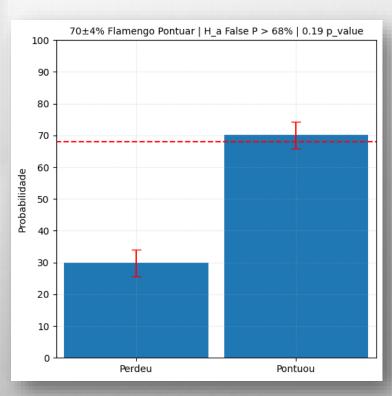
P VALOR = 0.00

 $P_FLAMENGO = 70\%$

 $P_{\text{TESTE}} = 68\%$

SIGNIFICÂNCIA = 0.01

P VALOR = 0.19

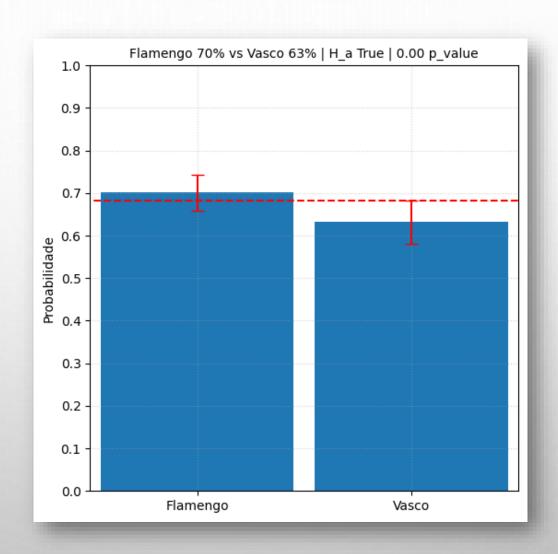


TESTE DE PROPORÇÕES – 2 PROPORÇÕES

HO – 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

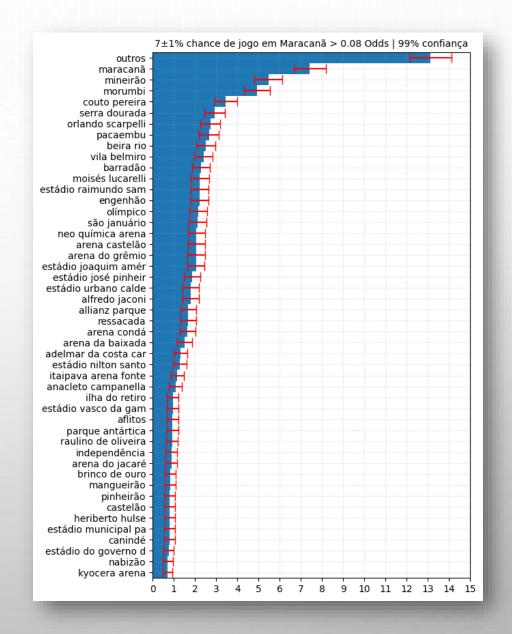
SIGNIFICÂNCIA = 0.01

P VALOR = 0.00

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

HA – ESTÁDIOS TEM CHANCE DISTINTAS DE TER JOGOS

TESTE CHI QUADRADO DE K PROPORÇÕES

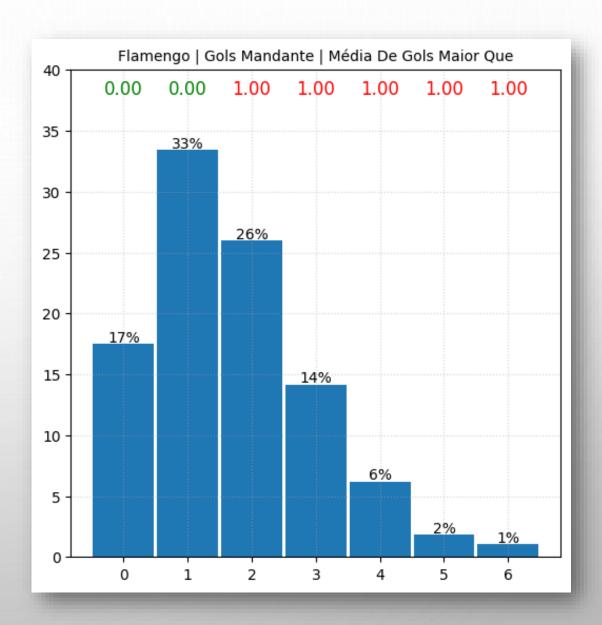


TESTE DE MÉDIA

HO – 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

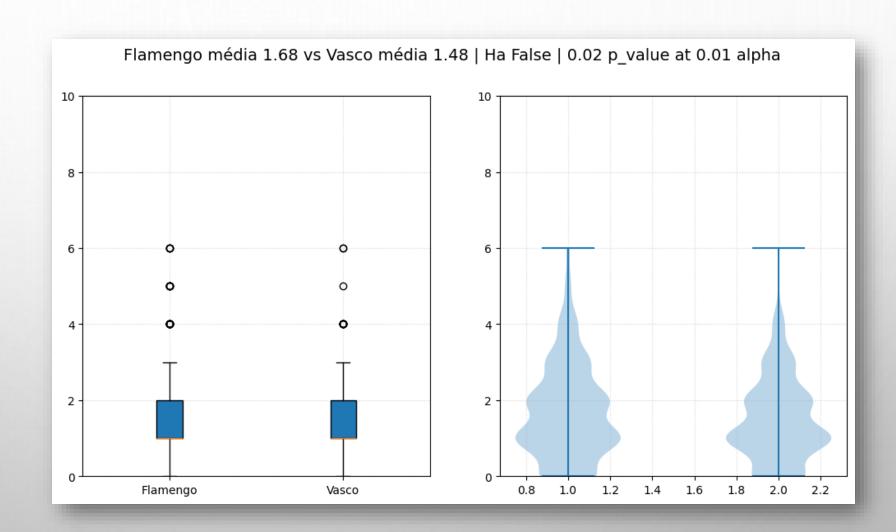


TESTE DE MÉDIA – DUAS POPULAÇÕES

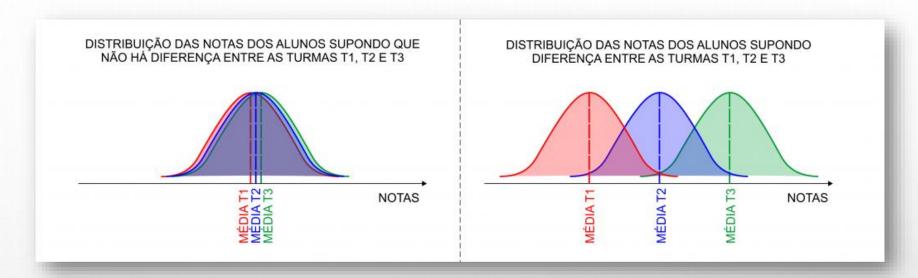
HO – MÉDIA DE GOLS DO
FLAMENGO COMO MANDANTE É
MENOR OU IGUAL A DO VASCO
COMO MANDANTE

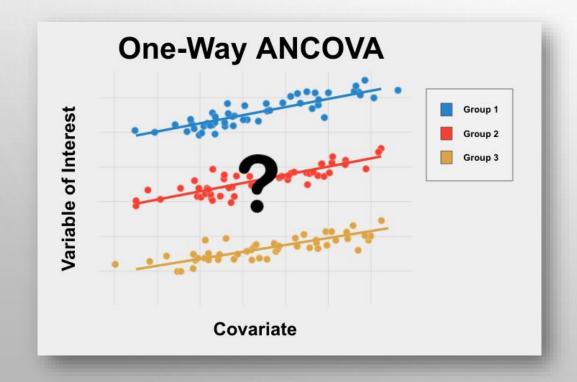
HA – MÉDIA DE GOLS DO
FLAMENGO COMO
MANDANTE É MAIOR QUE A
DO VASCO COMO
MANDANTE

TESTE Z PARA MÉDIAS



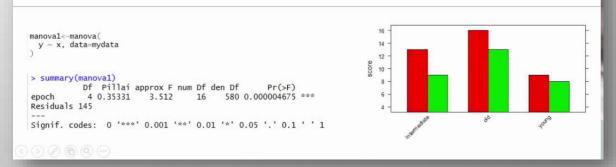
ANOVA E ETC



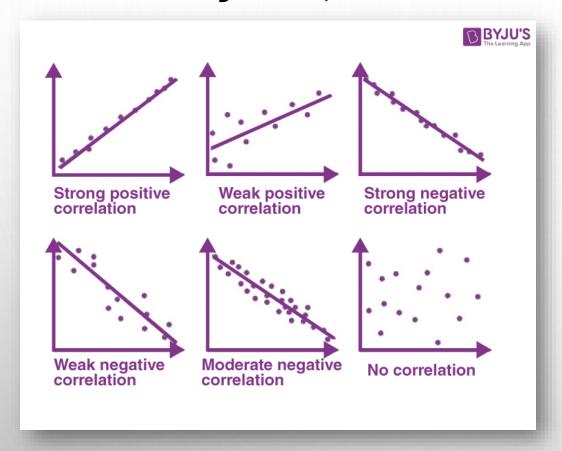


ANOVA, ANCOVA, MANOVA & MANCOVA

A Quick Tour!



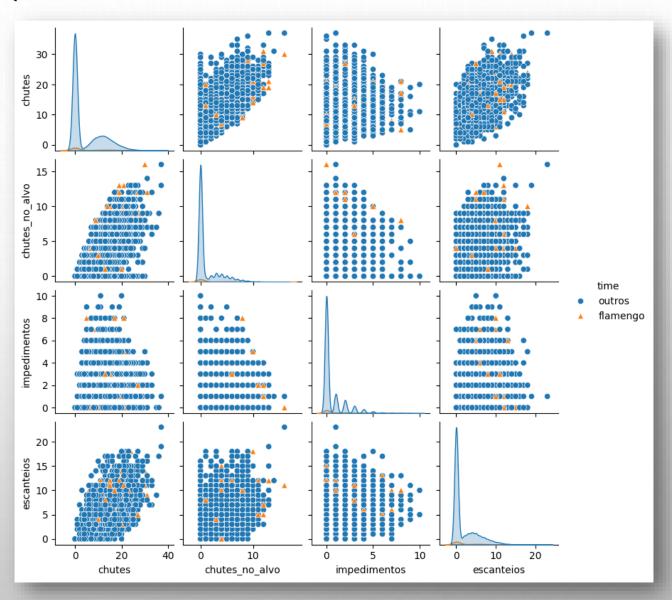
CORRELAÇÃO (DE PEARSON)



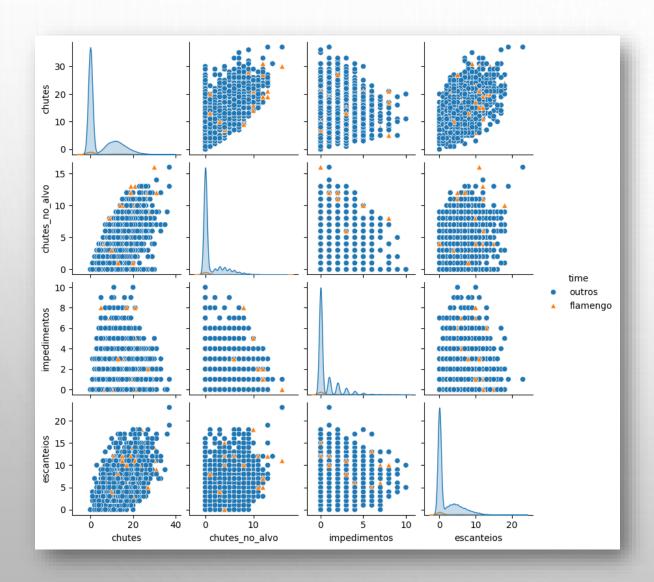
For a sample [edit]

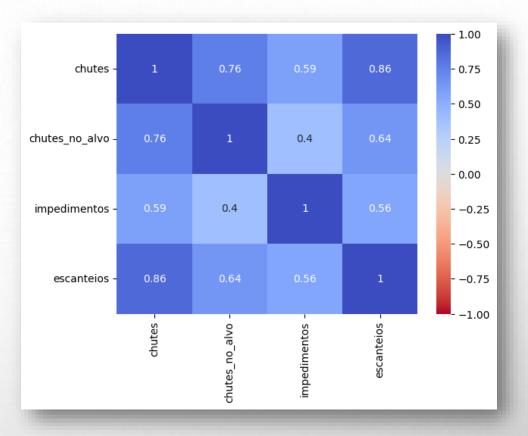
Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based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

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



CORRELAÇÃO (DE PEARSON)

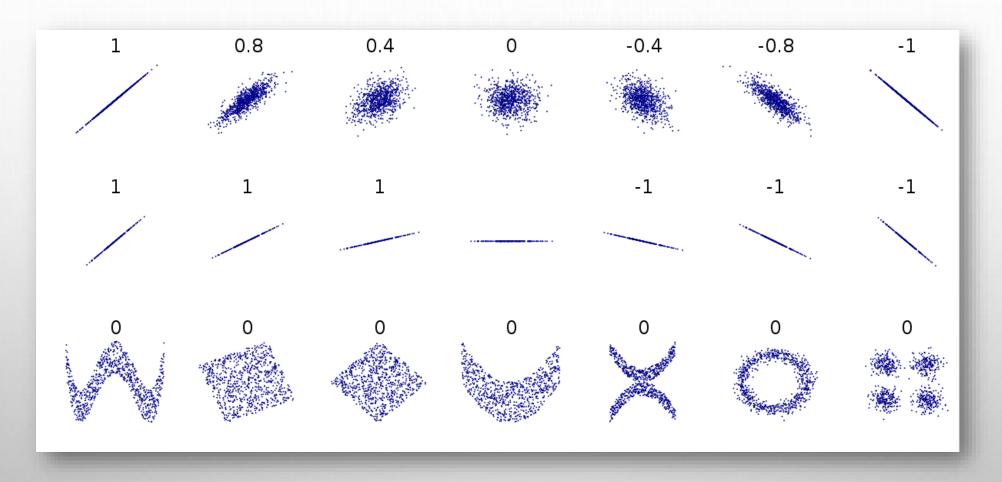




MAIOR CORRELAÇÃO: CHUTES
COM ESCANTEIOS

MENOR CORRELAÇÃO: CHUTES
NO ALVO COM IMPEDIMENTOS

CORRELAÇÃO (DE PEARSON)



INFORMAÇÃO MÚTUA

DIVERGÊNCIA DE KULLBACK LEIBER

COMPARAÇÃO DE PARES

≡ Pairwise comparison

文 10 languages ~

Article Talk Read Edit Viewhistory 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 "x/y"

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:

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

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

文 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

$$P(i>j)=rac{p_i}{p_i+p_j}$$
 (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. 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