

FIAP

NBA

MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

STATISTICS WITH R



Dra. Regina Tomie Ivata Bernal

Cientista de Dados na área da Saúde

Formação Acadêmica:

Estatístico - UFSCar

Mestre em Saúde Pública – FSP/USP

Doutor em Ciências – Epidemiologia - FSP/USP

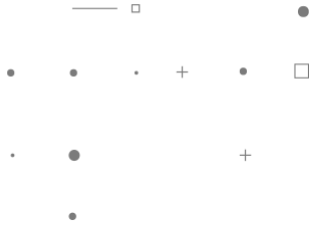
Atividades Profissionais:

Professora de pós-graduação na FIAP

Consultora externa da SVS/MS

Cientista de Dados em Saúde

profregina.bernal@fiap.com.br
reginabernal@terra.com.br



NOÇÕES DE PROBABILIDADE



PROBABILIDADE

Fenômenos aleatórios: situação ou acontecimentos cujos resultados não podem ser previstos com certeza.

Exemplos:

- Condições climáticas no próximo domingo.
- Faturamento da empresa no próximo mês.
- Quantidade de clientes cancelados nos próximos seis meses.
- Taxa de inflação no próximo mês.



Modelos podem ser estabelecidos
para quantificar as INCERTEZAS.



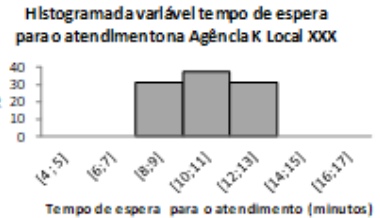
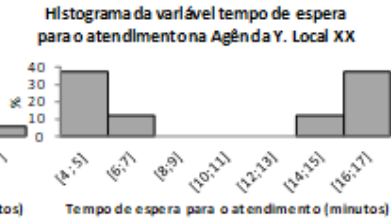
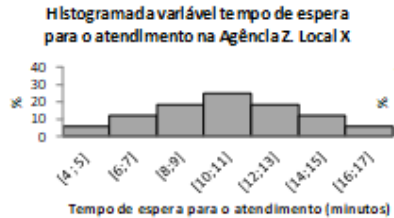
MODELOS
PROBABILÍSTICOS

PROBABILIDADE

Inferência clássica

Frequência é uma estimativa da probabilidade de ocorrência de certos eventos de interesse.

- Exemplo: Qual a probabilidade de um cliente ser atendido entre 4 e 5 minutos na agência Z? E na agência Y? E na agência K?



PROBABILIDADE

Propriedades:

1. Para cada experiência define-se um espaço amostral
2. Probabilidade de um evento E: $0 \leq P(E) \leq 1$
3. $P(S) = \text{Soma das probabilidades dos eventos simples} = 1$

PROBABILIDADE

Definição 1:

“A probabilidade simplesmente determina qual é a chance de algo acontecer.”

“Toda vez que não temos certeza sobre o resultado de algum evento, estamos tratando da probabilidade de certos resultados acontecerem—ou quais as chances de eles acontecerem.”

Fonte: <https://pt.khanacademy.org/math/probability/probability-geometry/probability-basics/a/probability-the-basics>

PROBABILIDADE

Definição 2:

“As frequências relativas são estimativas de probabilidade de ocorrência de certos eventos de interesse. Com suposições adequadas, e sem observarmos diretamente o fenômeno aleatório de interesse, podemos criar um modelo teórico que reproduza de maneira razoável a distribuição das frequências, quando o fenômeno é observado diretamente. Tais modelos são chamados de modelos probabilísticos.”

(Bussab, WO e Morettin, PA, Estatística Básica. 5 ed. São Paulo: Saraiva, 2002, página 103).

Noções de Probabilidade

Exemplo 1:

De um grupo de duas mulheres (M) e três homens (H), uma pessoa será sorteada para presidir uma reunião. Queremos saber as probabilidades de o presidente ser do sexo masculino ou feminino.

1) Espaço amostral: { M, H }

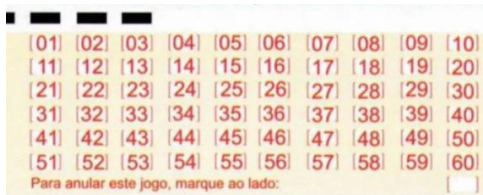
2) Modelo teórico:

Sexo	M	H	Total
Frequência teórica	2/5	3/5	1

Fonte: Exemplo extraído do livro Estatística Básica página 108

Noções de Probabilidade

Qual a probabilidade de ganhar o prêmio máximo da Mega Sena com um único jogo de seis dezenas?



O jogo da Mega-sena consiste em escolher seis dezenas dentre as 60 dezenas.

Noções de Probabilidade

- Qual a probabilidade de ganhar o prêmio máximo da Mega Sena com um único jogo de seis dezenas?

Lembrando que o jogo da Mega-sena consiste em escolher seis dezenas dentre as 60 dezenas.

Espaço amostral consiste da enumeração de todos os resultados possíveis do jogo.

$$\Omega = \{ \underset{1}{(1,2,3,4,5,6)}, \underset{2}{(1,2,3,4,5,7)}, \underset{3}{(1,2,3,4,5,8)}, \underset{?}{\dots} \}$$

Noções de Probabilidade

- Evento A = ganhar o prêmio máximo da Mega Sena com um único jogo de seis dezenas

Análise Combinatória Simples : $C_{(n,p)} = \binom{n}{p} = \frac{n!}{(n-p)!p!}$

n=60 e p= 6

$$\Omega = \binom{60}{6} = \frac{60!}{(60-6)!6!} = \frac{60!}{54!6!} = \frac{60.59.58....3.2.1}{(54.533...3.2.1).6.5.4.3.2.1} = 50.063.860$$

$$\text{Probabilidade}(A) = \frac{A}{(\Omega)} = \frac{1}{\binom{60}{6}} = \frac{1}{50.063.860}$$

Noções de Probabilidade

Análise Combinatória e Probabilidade

Fatorial : $n! = n \cdot (n - 1) \cdot (n - 2) \dots 3 \cdot 2 \cdot 1$

Permutação: $P = n!$

Arranjo: $C_{(n,p)} = \frac{n!}{(n - p)!}$

Análise Combinatória Simples : $C_{(n,p)} = \frac{n!}{(n - p)! p!}$

PROBABILIDADE

Exemplo 2:

O jogo da Mega-sena consiste em escolher seis dezenas dentre as 60 dezenas (01, 02, 03, ..., 60). O jogador pode marcar num cartão de 6 a 15 dezenas. Os custos em reais (R\$) de cada jogo estão relacionados abaixo.

Dezenas	Custo (R\$)
6	4,50
7	31,50
8	126,00
9	378,00
10	945,00
11	2.079,00
12	4.158,00
13	7.722,00
14	13.513,50
15	22.522,50

$$\binom{60}{6} = \frac{60!}{(60-6)! \cdot 6!} = 50.063.860 \text{ possibilidades}$$

Com um único jogo a probabilidade de ganhar o prêmio

máximo é $\frac{1}{\binom{60}{6}}$, isto é, aproximadamente uma chance em 50 milhões.

Fonte: Exemplo extraído do livro Estatística Básica página 109. O custo e a probabilidade foram atualizadas.

PROBABILIDADE

Exemplo 2:

O jogo da Mega-sena consiste em escolher seis dezenas dentre as 60 dezenas (01, 02, 03, ..., 60). O jogador pode marcar num cartão de 6 a 15 dezenas. Os custos em reais (R\$) de cada jogo estão relacionados abaixo.

Dezenas	Custo (R\$)	Probabilidade de acerto (1 em) Sena
6	4,50	50.063.860
7	31,50	7.151.980
8	126,00	1.787.995
9	378,00	595.998
10	945,00	238.399
11	2.079,00	108.363
12	4.158,00	54.182
13	7.722,00	29.175
14	13.513,50	16.671
15	22.522,50	10.003

Fonte: Exemplo extraído do livro Estatística Básica página 109. O custo e a probabilidade foram atualizadas.

Noções de Probabilidade

Artigos:

<https://mundoeducacao.uol.com.br/matematica/anagrama.htm>

<https://mundoeducacao.uol.com.br/matematica/combinacao-simples.htm>

<https://mundoeducacao.uol.com.br/matematica/permutacao-envolvendo-elementos-repetidos.htm>

<https://mundoeducacao.uol.com.br/matematica/principio-fundamental-contagem-fatorial.htm>

<https://mundoeducacao.uol.com.br/matematica/arranjos-simples.htm>

DISTRIBUIÇÃO DE PROBABILIDADE

Distribuição de Probabilidade

Exemplo:

- Construir o modelo preditivo a fim de prever o resultado de partidas do campeonato brasileiro.
- Considerando que a quantidade de gols marcados (K) em uma partida de futebol do Campeonato Brasileiro de Futebol (Brasileirão) em 2018 seja uma variável aleatória que segue a distribuição de Poisson com média de gols igual a λ .
- Calcule a probabilidade de ocorrer $k = 0, 1, 2, 3, 4, 5, 6$ e 7

Distribuição de Probabilidade

Distribuição de Poisson

A probabilidade de existam exatamente k ocorrências é:

$$P(K = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$k = 0, 1, 2, 3 \dots$ (número inteiro positivo)

$e = 2.71828$

$k! = k \cdot (k-1) \cdot (k-2) \dots 3 \cdot 2 \cdot 1$

λ = média do valor esperado de k

Distribuição de Probabilidade

- Dados históricos

<https://github.com/henriquegomide/caRtola/tree/master/data>

game	round	date	home_team	score	away_team	gol_m	gol_v	gols	arena	Lcal
1	1	14/04/2018 - 16:00	Cruzeiro - MG	0 x 1	0	1	1	1	Grêmio - RS	Mineirão - Belo Horizonte - MG
2	1	15/04/2018 - 19:00	Atlético - PR	5 x 1	5	1	6	6	Chapecoense - SC	Arena da Baixada - Curitiba - PR
3	1	15/04/2018 - 11:00	América - MG	3 x 0	3	0	3	3	Sport - PE	Independência - Belo Horizonte - MG
4	1	14/04/2018 - 19:00	Vitória - BA	2 x 2	2	2	4	4	Flamengo - RJ	Manoel Barradas - Salvador - BA
5	1	15/04/2018 - 16:00	Vasco da Gama - RJ	2 x 1	2	1	3	3	Atlético - MG	São Januário - Rio de Janeiro - RJ
6	1	16/04/2018 - 20:00	Botafogo - RJ	1 x 1	1	1	2	2	Palmeiras - SP	Nilton Santos - Rio de Janeiro - RJ
7	1	16/04/2018 - 20:00	São Paulo - SP	1 x 0	1	0	1	1	Paraná - PR	Morumbi - Sao Paulo - SP

λ = média do número de gols em uma partida de futebol do Brasileirão.

Distribuição de Probabilidade

- Dados históricos

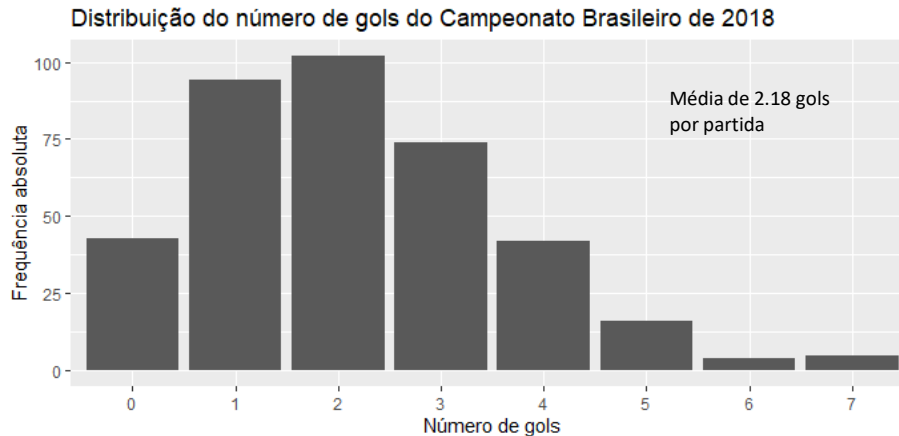
<https://github.com/henriquegomide/caRtola/tree/master/data>

game	round	date	home_team	score	away_team	gol_m	gol_v	gols	arena	Lcal
1	1	14/04/2018 - 16:00	Cruzeiro - MG	0 x 1	0	1	1	1	Grêmio - RS	Mineirão - Belo Horizonte - MG
2	1	15/04/2018 - 19:00	Atlético - PR	5 x 1	5	1	6	6	Chapecoense - SC	Arena da Baixada - Curitiba - PR
3	1	15/04/2018 - 11:00	América - MG	3 x 0	3	0	3	3	Sport - PE	Independência - Belo Horizonte - MG
4	1	14/04/2018 - 19:00	Vitória - BA	2 x 2	2	2	4	4	Flamengo - RJ	Manoel Barradas - Salvador - BA
5	1	15/04/2018 - 16:00	Vasco da Gama - RJ	2 x 1	2	1	3	3	Atlético - MG	São Januário - Rio de Janeiro - RJ
6	1	16/04/2018 - 20:00	Botafogo - RJ	1 x 1	1	1	2	2	Palmeiras - SP	Nilton Santos - Rio de Janeiro - RJ
7	1	16/04/2018 - 20:00	São Paulo - SP	1 x 0	1	0	1	1	Paraná - PR	Morumbi - Sao Paulo - SP

λ = média do número de gols em uma partida de futebol do Brasileirão.

Distribuição de Probabilidade

- Dados históricos



```
> mean(cartola2018$gols)
```

```
> [1] 2.18
```

Distribuição de Probabilidade

Distribuição de Poisson

A probabilidade de existam exatamente k ocorrências é:

$$P(K = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

λ = média de 2.18 gols

Qual a probabilidade de ocorrer 0 gol?

$$P(K = 0) = \frac{e^{-\lambda} \lambda^k}{k!} = \frac{e^{-2.18} 2.18^0}{0!} = 0.135$$

Qual a probabilidade de ocorrer 1 gol?

$$P(K = 1) = \frac{e^{-\lambda} \lambda^k}{k!} = \frac{e^{-2.18} 2.18^1}{1!} = 0.271$$

Distribuição de Probabilidade

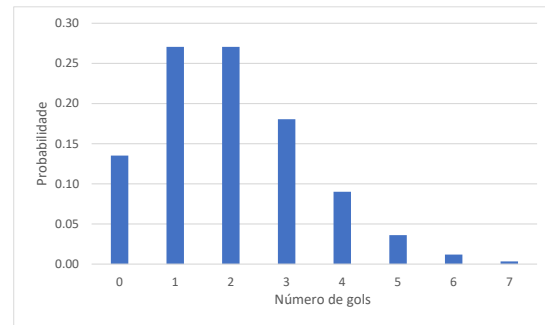
Distribuição de Poisson

A probabilidade de existam exatamente k ocorrências é:

$$P(K = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

λ = média de 2.18 gols

k (gols)	Probabilidade
0	0.14
1	0.27
2	0.27
3	0.18
4	0.09
5	0.04
6	0.01
7	0.00
Soma	1.00





TABELAS ESTATÍSTICAS



Distribuição de Probabilidade

Distribuição de Poisson

K (gols)	Probabilidade
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

Considerando que a quantidade de gols marcados (K) em uma partida de futebol do Campeonato Brasileiro de Futebol (Brasileirão) em 2018 seja uma variável aleatória que segue a distribuição de Poisson com média de gols igual a λ .

$$P(K = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Use a Tabela de Poisson para identificar a probabilidade para cada k ocorrência, dada que a média de gols é igual a 2.5 (λ = média de 2.5 gols).

DISTRIBUIÇÃO POISSON

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

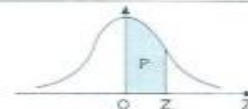
Tabela II — Distribuição de Poisson

 $X \sim \text{Pois}(\lambda)$ Corpo da tabela dá as probabilidades $P(X = j), j = 0, 1, 2, \dots$

$X \backslash \lambda$	0,001	0,005	0,010	0,015	0,020	0,025	0,030	0,035	0,040	0,045	0,050	0,055	0,060	0,065	0,070	0,075	$\lambda \backslash X$
0	9990	9950	9900	9851	9802	9753	9704	9656	9608	9560	9512	9465	9418	9371	9324	9277	0
1	0010	0050	0099	0148	0196	0244	0291	0338	0384	0430	0476	0521	0565	0609	0653	0696	1
2	0*	0*	0*	0001	0002	0003	0004	0006	0008	0010	0012	0014	0017	0020	0022	0026	2
3	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0001	3
≥ 4	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	≥ 4
$X \backslash \lambda$	0,080	0,085	0,090	0,095	0,100	0,200	0,300	0,400	0,500	0,600	0,700	0,800	0,900	1,000	1,200	1,400	$\lambda \backslash X$
0	9231	9185	9139	9094	9048	8187	7408	6703	6065	5488	4966	4493	4066	3679	3012	2466	0
1	0739	0781	0823	0864	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679	3614	3452	1
2	0030	0033	0037	0041	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839	2169	2417	2
3	0001	0001	0001	0001	0002	0011	0033	0071	0126	0198	0284	0383	0494	0613	0867	1128	3
4	0*	0*	0*	0*	0*	0001	0003	0007	0016	0030	0050	0077	0111	0153	0260	0395	4
5	0*	0*	0*	0*	0*	0*	0*	0001	0002	0004	0007	0012	0020	0031	0062	0111	5
6	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0002	0003	0005	0012	0026	6
7	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0002	0005	7
8	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	8
≥ 9	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	≥ 9
$X \backslash \lambda$	1,600	1,800	2,000	2,500	3,000	3,500	4,000	4,500	5,000	5,500	6,000	6,500	7,000	8,000	9,000	10,000	$\lambda \backslash X$
0	2019	1653	1353	0821	0498	0302	0183	0111	0067	0041	0025	0015	0009	0003	0001	0*	0
1	3230	2975	2707	2052	1494	1057	0733	0500	0337	0225	0149	0098	0064	0027	0011	0005	1
2	2584	2678	2707	2565	2240	1850	1465	1125	0842	0618	0446	0318	0223	0107	0050	0023	2
3	1378	1607	1804	2138	2240	2158	1954	1687	1404	1133	0892	0688	0521	0286	0150	0076	3
4	0551	0723	0902	1336	1680	1888	1954	1898	1755	1558	1339	1118	0912	0573	0337	0189	4
5	0176	0260	0361	0668	1008	1322	1563	1708	1755	1714	1606	1454	1277	0916	0607	0378	5
6	0047	0078	0120	0278	0504	0771	1042	1281	1462	1571	1606	1575	1490	1221	0911	0631	6
7	0011	0020	0034	0100	0216	0385	0595	0824	1044	1234	1377	1462	1490	1396	1171	0901	7
8	0002	0005	0009	0031	0081	0169	0298	0463	0653	0849	1033	1188	1304	1396	1318	1126	8
9	0*	0001	0002	0009	0027	0066	0132	0232	0363	0519	0688	0858	1014	1241	1318	1251	9
10	0*	0*	0*	0002	0008	0023	0053	0104	0181	0285	0413	0558	0710	0993	1186	1251	10
11	0*	0*	0*	0*	0002	0007	0019	0043	0082	0143	0225	0330	0452	0722	0970	1137	11
12	0*	0*	0*	0*	0*	0002	0006	0016	0034	0065	0113	0179	0264	0481	0728	0948	12
13	0*	0*	0*	0*	0*	0*	0002	0006	0013	0028	0052	0089	0142	0296	0504	0729	13
14	0*	0*	0*	0*	0*	0*	0*	0002	0005	0011	0022	0041	0071	0169	0324	0521	14
15	0*	0*	0*	0*	0*	0*	0*	0001	0002	0004	0009	0018	0033	0090	0194	0347	15
16	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0003	0007	0014	0045	0109	0217	16
17	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0003	0006	0021	0058	0128	17
18	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0002	0009	0029	0071	18
19	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0004	0014	0037	19
20	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0002	0006	0019	20
21	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*	0001	0003	0009	21

DISTRIBUIÇÃO NORMAL

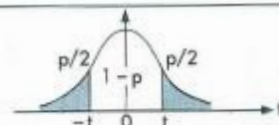
Tabela III — Distribuição Normal Padrão
 $Z \sim N(0, 1)$
 Corpo da tabela dá a probabilidade p , tal que $p = P(0 < Z < Z_c)$



$$P(Z > 0) = ?$$

parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
0,0	p = 0	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4
3,5	49977	49978	49979	49979	49980	49981	49981	49982	49983	49983	3,5
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989	3,6
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992	3,7

Tabela V — Distribuição t de Student
 Corpo da tabela dá os valores t_c tais que $P(-t_c < t < t_c) = 1 - p$.
 Para $v > 120$, usar a aproximação normal.



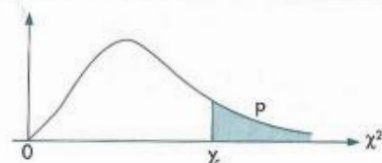
Graus de liberdade v	Tabela V — Distribuição t de Student															Graus de liberdade v
	$p = 90\%$	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	15,894	31,821	63,657	318,309	636,619	1
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	4,849	6,965	9,925	22,327	31,598	2
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	3,482	4,541	5,841	10,214	12,924	3
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	2,998	3,747	4,604	7,173	8,610	4
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	2,756	3,365	4,032	5,893	6,869	5
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	2,612	3,143	3,707	5,208	5,959	6
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,517	2,998	3,499	4,785	5,408	7
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,449	2,896	3,355	4,501	5,041	8
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,398	2,821	3,250	4,297	4,781	9
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,359	2,764	3,169	4,144	4,587	10
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,328	2,718	3,106	3,025	4,437	11
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,303	2,681	3,055	3,930	4,318	12
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,282	2,650	3,012	3,852	4,221	13
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,264	2,624	2,977	3,787	4,140	14
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,248	2,602	2,947	3,733	4,073	15
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,235	2,583	2,921	3,686	4,015	16
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,224	2,567	2,898	3,646	3,965	17
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,214	2,552	2,878	3,610	3,922	18
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,205	2,539	2,861	3,579	3,883	19
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,197	2,528	2,845	3,552	3,850	20
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,189	2,518	2,831	3,527	3,819	21
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,183	2,508	2,819	3,505	3,792	22
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,177	2,500	2,807	3,485	3,768	23
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,172	2,492	2,797	3,467	3,745	24
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,166	2,485	2,787	3,450	3,725	25
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,162	2,479	2,779	3,435	3,707	26
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,158	2,473	2,771	3,421	3,690	27
28	0,127	0,256	0,389	0,530	0,684	0,855	1,056	1,313	1,701	2,048	2,154	2,467	2,763	3,408	3,674	28
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,150	2,462	2,756	3,396	3,659	29
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,147	2,457	2,750	3,385	3,646	30
35	0,126	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	2,030	2,133	2,438	2,724	3,340	3,591	35
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,123	2,423	2,704	3,307	3,551	40
50	0,126	0,254	0,387	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,109	2,403	2,678	3,261	3,496	50
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,099	2,390	2,660	3,232	3,460	60
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,076	2,358	2,617	3,160	3,373	120
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,054	2,326	2,576	3,090	3,291	∞
	$p = 90\%$	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	

Tabela IV – Distribuição Qui-quadrado

$$Y \sim \chi^2 (v)$$

Corpo da tabela dá os valores y_c tais que $P(Y > y_c) = p$.

Para valores $v > 30$, use a aproximação normal dada no texto.



Graus de liberdade v	p = 99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	Graus de liberdade v
1	0,016	0,063	0,001	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	4,218	5,024	5,412	6,635	9,550	10,827	1
2	0,020	0,040	0,051	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	6,438	7,378	7,824	9,210	12,429	13,815	2
3	0,115	0,185	0,216	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	8,311	9,348	9,837	11,345	14,796	16,266	3
4	0,297	0,429	0,484	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	10,026	11,143	11,668	13,277	16,924	18,467	4
5	0,554	0,752	0,831	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	11,644	12,832	13,388	15,086	18,907	20,515	5
6	0,872	1,134	1,237	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	13,198	14,449	15,033	16,812	20,791	22,457	6
7	1,239	1,564	1,690	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	14,703	16,013	16,622	18,475	22,601	24,322	7
8	1,646	2,032	2,180	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	16,171	17,534	18,168	20,090	24,352	26,125	8
9	2,088	2,532	2,700	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	17,608	19,023	19,679	21,666	26,056	27,877	9
10	2,558	3,059	3,247	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	19,021	20,483	21,161	23,209	27,722	29,588	10
11	3,053	3,609	3,816	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	20,412	21,920	22,618	24,725	29,354	31,264	11
12	3,571	4,178	4,404	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	21,785	23,337	24,054	26,217	30,957	32,909	12
13	4,107	4,765	5,009	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	23,142	24,736	25,472	27,688	32,535	34,528	13
14	4,660	5,368	5,629	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	24,485	26,119	26,873	29,141	34,091	36,123	14
15	5,229	5,985	6,262	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	25,816	27,488	28,259	30,578	35,628	37,697	15
16	5,812	6,614	6,908	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	27,136	28,845	29,633	32,000	37,146	39,252	16
17	6,408	7,255	7,564	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	28,445	30,191	30,995	33,409	38,648	40,790	17
18	7,015	7,906	8,231	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	29,745	31,526	32,346	34,805	40,136	42,312	18
19	7,633	8,567	8,906	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	31,037	32,852	33,687	36,191	41,610	43,820	19
20	8,260	9,237	9,591	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	32,321	34,170	35,020	37,566	43,072	45,310	20
21	8,897	9,915	10,283	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	33,597	35,479	36,343	38,932	44,522	46,997	21
22	9,542	10,600	10,982	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	34,867	36,781	37,659	40,289	45,962	48,268	22
23	10,196	11,293	11,688	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	36,131	38,076	38,968	41,638	47,391	49,728	23
24	10,856	11,992	12,401	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	37,389	39,364	40,270	42,980	48,812	51,179	24
25	11,524	12,697	13,120	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	38,642	40,646	41,566	44,314	50,223	52,620	25
26	12,198	13,409	13,844	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	39,889	41,923	42,856	45,642	51,627	54,052	26
27	12,879	14,125	14,573	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	41,132	43,194	44,140	46,963	53,022	55,476	27
28	13,565	14,847	15,308	16,928	18,939	21,588	23,647	27,336	31,319	34,027	37,916	41,337	42,370	44,461	45,419	48,278	54,411	56,893	28
29	14,258	15,574	16,047	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	43,604	45,722	46,693	49,588	55,792	58,302	29
30	14,953	16,306	16,791	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	44,834	46,979	47,962	50,892	57,167	59,703	30
p = 99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%		

Tabelas Estatísticas

Afinal, o que são as tabelas estatísticas?

Probabilidade é a base dos modelos teóricos para buscar determinar a chance de eventos acontecerem, sejam eventos simples ou compostos.

Por que elas são úteis?

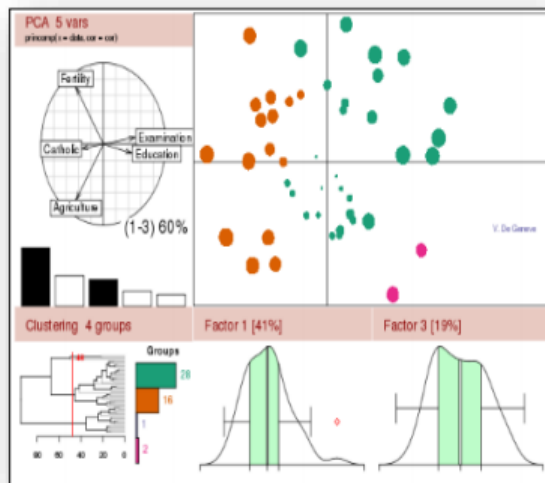
Os valores das probabilidades encontram-se em tabelas estatísticas que podem ser facilmente utilizadas para análise de teste de hipóteses, análise de associação e saídas de modelos preditivos.

Exemplo: Qual a probabilidade de um determinado time ganhar? Use a distribuição de Poisson.

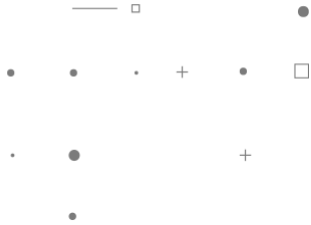
<https://www.goal.com/br/not%C3%ADcias/como-calcular-a-probabilidade-de-gols-marcados-para-apostas/bhonn6lceb171ohkvmh2rn4xa>

<https://www.youtube.com/watch?v=a6dRG3V5l6s>

DISTRIBUIÇÃO DE POISSON



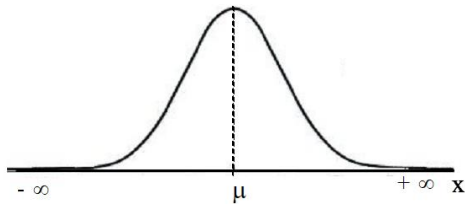
Exercícios



DISTRIBUIÇÃO NORMAL



DISTRIBUIÇÃO NORMAL



CARACTERÍSTICAS:

- A) A variável pode assumir qualquer valor no conjunto real.
- B) O gráfico da distribuição é uma curva em forma de sino, simétrica em torno da média μ , que é igual à mediana e à moda.
- C) A área sob a curva é igual a 1, e corresponde à probabilidade de a variável assumir valores entre $[-\infty \dots +\infty]$.
- D) $\mu; \sigma$ (Mi e Sigma) representam os parâmetros de posição e dispersão da distribuição.
- E) Os pontos de inflexão da curva ocorrem nos valores definidos por $(\mu - \sigma$ e $\mu + \sigma)$.
- F) A expressão da função densidade de probabilidade é:

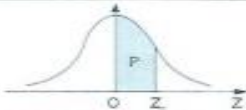
$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-1/2[(X-\mu)/\sigma]^2}$$

Qual a probabilidade de ocorrer valores entre 0 e 1?

$$P(0 < Z < 1.00) = ?$$

DISTRIBUIÇÃO NORMAL

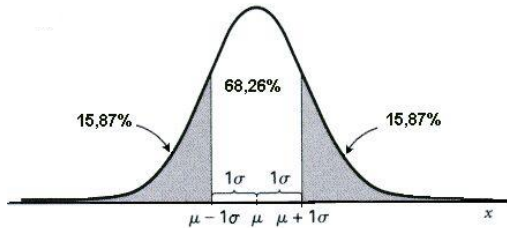
Tabela III — Distribuição Normal Padrão
 $Z \sim N(0, 1)$
 Corpo da tabela dá a probabilidade p , tal que $p = P(0 < Z < Z_c)$



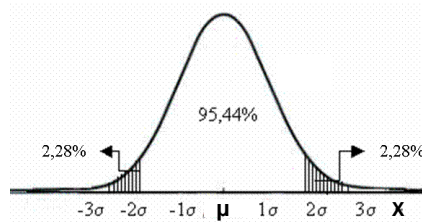
parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39797	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4

Distribuição Normal Padronizada

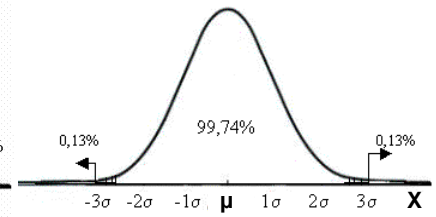
$$Z \sim N(0,1)$$



$$P[(\mu - \sigma) < X < (\mu + \sigma)] = 68.25\%$$



$$P[(\mu - 2\sigma) < X < (\mu + 2\sigma)] = 95.44\%$$



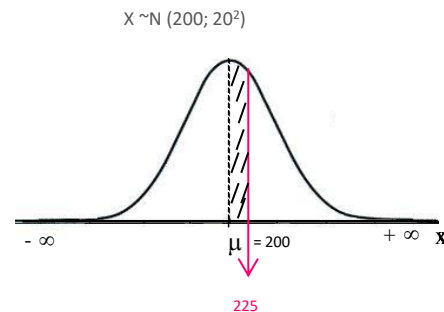
$$P[(\mu - 3\sigma) < X < (\mu + 3\sigma)] = 99.74\%$$

DISTRIBUIÇÃO NORMAL

Exemplo:

X = Faturamento anual da empresa

- Probabilidade de um resultado entre 200 e 225?

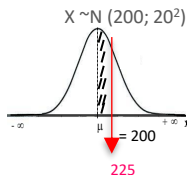


$$P(200 < X < 225) = ? \quad \rightarrow \quad f(X) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-1/2[(X-\mu)/\sigma]^2}$$

DISTRIBUIÇÃO NORMAL

Distribuição Normal

$$X \sim N(\mu, \sigma^2)$$



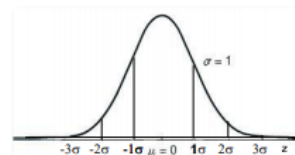
X = Faturamento anual da empresa

$$P(200 < X < 225) = ?$$

Distribuição Normal
Padronizada

$$Z = \frac{X - \mu}{\sigma}$$

$$Z \sim N(0,1)$$



$$P(200 < X < 225) = P(? < Z < ?)$$

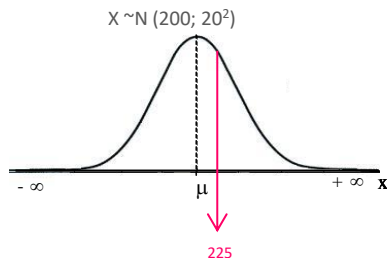
DISTRIBUIÇÃO NORMAL

- Distribuição Normal Padronizada

Exemplo:

X = Faturamento anual da empresa

- Probabilidade de um resultado entre 200 e 225 ?



$$P(200 < X < 225) = ?$$

$$P(200 < X < 225) = P(? < Z < ?)$$

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 200}{20} = 0$$

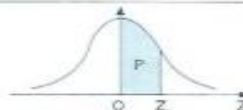
$$Z = \frac{X - \mu}{\sigma} = \frac{225 - 200}{20} = 1,25$$

$$P(200 < X < 225) = P(0 < Z < 1,25)$$

$$P(200 < X < 225) = 0,39$$

DISTRIBUIÇÃO NORMAL

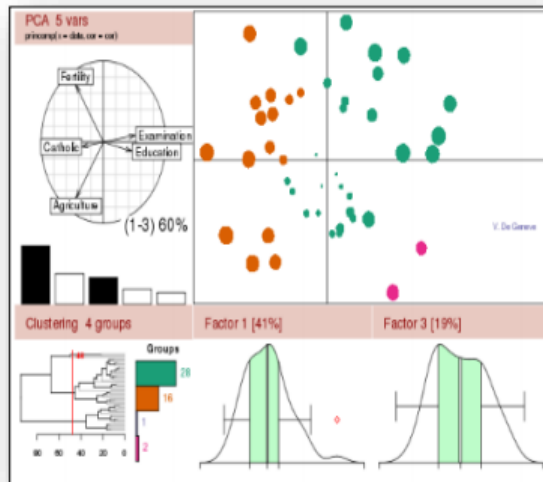
Tabela III — Distribuição Normal Padrão

 $Z \sim N(0, 1)$ Corpo da tabela dá a probabilidade p , tal que $p = P(0 < Z < Z_c)$ 

parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4

$$P(0 < Z < 1,25) = 0,39$$

DISTRIBUIÇÃO NORMAL



Exercícios

NORMALIZAÇÃO DOS DADOS

- Distribuição Normal Padronizada

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

```
from sklearn.preprocessing import StandardScaler
```

```
# padronizar a variável para Normal com média igual a zero e desvio padrão igual a 1
scaler = StandardScaler()
df_padrao = scaler.fit_transform(df)
```

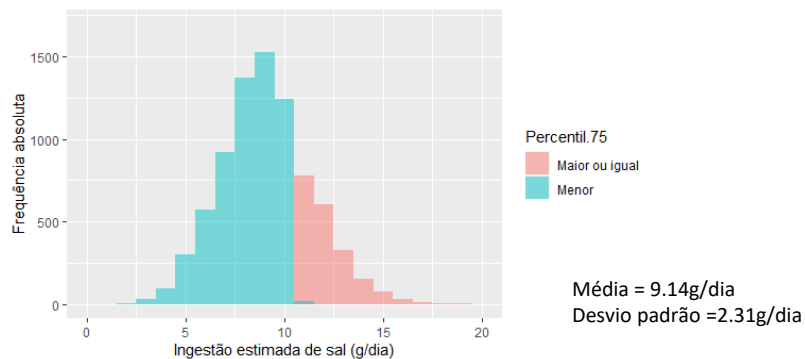
EXEMPLO

Exemplos de aplicações da normalização dos dados:

Convolutional Neural Networks (CNNs) e Algoritmos de Machine Learning (Regressão, SVM, Cluster e outros)

NORMALIZAÇÃO DOS DADOS

Distribuição Normal

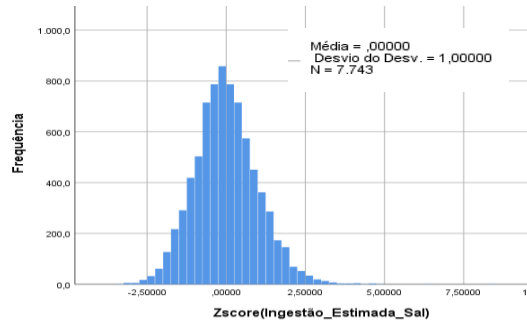
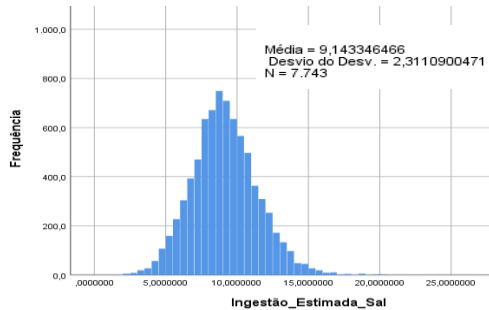


Fonte: Pesquisa Nacional de Saúde 2013 – População adulta

NORMALIZAÇÃO DOS DADOS

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

Exemplo: Ingestão de sal estimada na população adulta. PNS, 2013



DISTRIBUIÇÃO NORMAL PADRONIZADA

EXEMPLO

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

Exemplo: Ingestão de sal estimada na população adulta. PNS, 2013

Ingestão de sal (X)

8.56

$$Z = \frac{8.56 - 9.14}{2.31} = -0.25$$

4.54

$$Z = (4.54 - 9.14) / 2.31 = -1.99 \sim -2 \text{dp}$$

13.7

$$Z = (13.7 - 9.14) / 2.31 = +1.97 \sim +2 \text{ dp}$$

NORMALIZAÇÃO DOS DADOS

- Distribuição Normal Padronizada

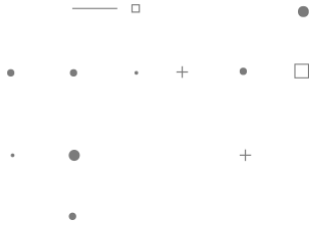
$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

- Máximo e Mínimo

$$X_p = \frac{X - X_{\text{mínimo}}}{X_{\text{máximo}} - X_{\text{mínimo}}}$$

Exemplos de aplicações da normalização dos dados:

Convolutional Neural Networks (CNNs) e Algoritmos de Machine Learning (Regressão, SVM, Cluster e outros)



PROBABILIDADE CONDICIONAL



PROBABILIDADE CONDICIONAL



Exemplo 3:

Considere o evento A = chover em SP no dia 12 de janeiro do próximo ano.

Suponha que uma pessoa morando em Fortaleza tenha que calcular essa probabilidade. Se ela não tiver informação sobre o tempo em São Paulo, poderá atribuir a probabilidade de $\frac{1}{2}$.

Já o morador de São Paulo tem informações adicionais, como por exemplo, ele saberá que janeiro, fevereiro e março são os meses mais chuvosos e poderá arriscar uma probabilidade de $\frac{2}{3}$ de ocorrer o evento A.

Fonte: Exemplo extraído do livro Estatística Básica página 121.

PROBABILIDADE CONDICIONAL



O fenômeno aleatório pode ser separado em etapas. A informação que ocorreu em uma determinada etapa pode influenciar nas probabilidades de ocorrências das etapas sucessivas.

Definição:

Dados dois eventos A e B, a probabilidade condicional de A dado que ocorreu B é representado por $P(A/B)$ e dada por:

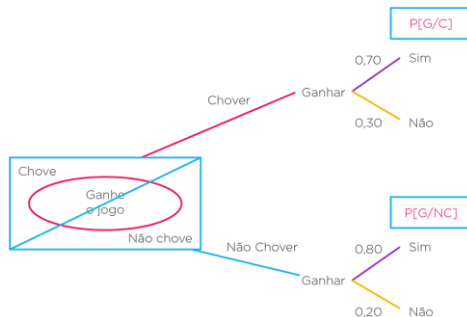
$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

Fonte: Exemplo extraído do livro Noções de Probabilidade e Estatística página 41.

PROBABILIDADE CONDICIONAL

Exemplo 4:

O São Paulo Futebol Clube ganha com probabilidade 0,7 se chove e com 0,8 se não chove. Em Setembro a probabilidade de chuva é de 0,3. O São Paulo ganhou uma partida em Setembro, qual a probabilidade de ter chovido nesse dia?



$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

Eventos:

C: Chover

G: São Paulo ganhar um jogo

$$P(C/G) = \frac{P(C) * P(G/C)}{P(C) * P(G/C) + P(NC) * P(G/NC)}$$

$$P(C/G) = \frac{0,30 * 0,70}{0,30 * 0,70 + 0,70 * 0,80} = \frac{0,21}{0,21 + 0,56} = 0,273$$

Fonte: Exemplo extraído e adaptado do livro Noções de Probabilidade e Estatística página 41.

PROBABILIDADE CONDICIONAL

A técnica de Basket Analysis utiliza a probabilidade condicional para encontrar cestas de produtos.

Um Exemplo de Sucesso!



- ✓ Descobriu-se que homens entre trinta e quarenta e cinco anos, que compram cervejas, nas sextas-feiras, após as dezesseis horas, também compram fraldas!
- ✓ Resultado: apenas mudando os produtos de lugar, colocando as fraldas ao lado de cervejas nos pontos de venda, obteve-se um aumento de mais de quarenta por cento nas vendas de fraldas.
- ✓ O que acha de possuir uma informação como essa?

A Wall-Mart soube tirar bom proveito dela!

Medidas estatísticas

- Support (frequência)

$$(A \cap B \Rightarrow C) = \%$$

- Confidence (probabilidade condicional)

$$(A \cap B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)}$$

- Lift(associação)

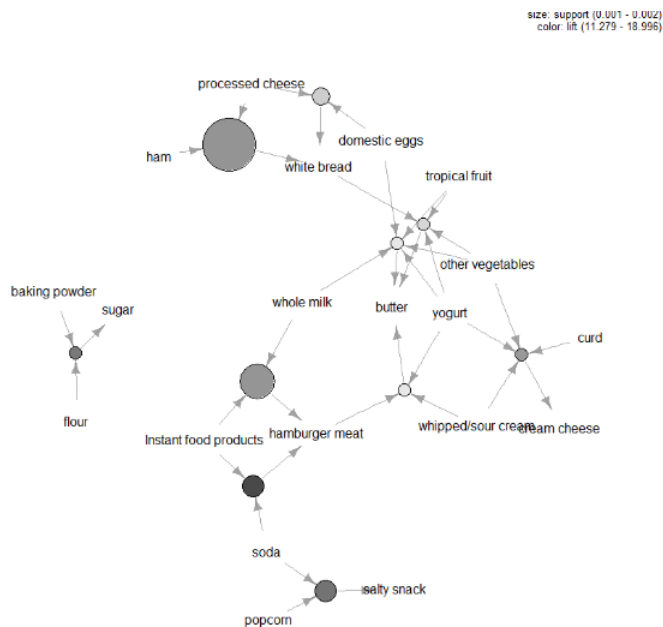
$$(A \& B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)P(C)}$$

MARKET BASKET ANALYSIS

É uma técnica estatística para identificar cestas de produtos, por meio de regras de associação, a qual relaciona todos os produtos adquiridos em uma mesma transação/ticket disponível na base de dados.

Essa técnica é muito utilizada na área de Marketing para identificar hábitos de compra de clientes. O exemplo clássico, citado na literatura, é a associação encontrada entre fralda e cerveja pela rede de supermercados americana WalMart.

Graph-based visualization of the top ten rules in terms of lift.

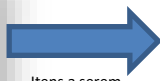


Exemplo

Técnica Descoberta de Sequências

Quais associações são significativas ?

Item comprado anteriormente



Itens a serem sugeridos de acordo com a força

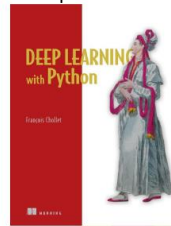
1.o produto



2.o produto

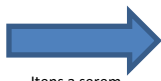


3.o produto



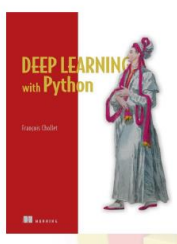
EXEMPLO

Item comprado anteriormente



Itens a serem sugeridos de acordo com a força

1.o produto



2.o produto



3.o produto



+

+

.

□

.

.

.



.

.

.

.

Exemplo

Market Basket Analysis

- A pizzaria XPTO vendeu 2000 pizzas:
 - 100 de cogumelos, 150 de pepperoni , 200 com extra queijo
 - 400 de cogumelos e pepperoni, 300 de cogumelos e extra queijo, 200 de pepperoni e extra queijo
 - 100 de cogumelos, pepperoni e extra queijo
 - 500 outros
- Cálculo da probabilidade das combinações dos itens:

Pepperoni



$$150 + 400 + 200 + 100 = 850 \text{ pizzas}$$

$$frequência = \frac{850}{2000} * 100 = 42.5\%$$

Market Basket Analysis

- Cálculo da probabilidade das combinações dos itens:

Cogumelo



$$100 + 400 + 300 + 100 = 900 \text{ pizzas}$$

$$frequência = \frac{900}{2000} * 100 = 45\%$$

Pepperoni



$$150 + 400 + 200 + 100 = 850 \text{ pizzas}$$

$$frequência = \frac{850}{2000} * 100 = 42.5\%$$

Queijo



$$200 + 300 + 200 + 100 = 800 \text{ pizzas}$$

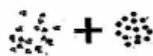
$$frequência = \frac{800}{2000} * 100 = 40\%$$

Exemplo

Market Basket Analysis

- Cálculo da probabilidade das combinações dos itens:

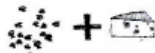
Cogumelo e Pepperoni



$$400 + 100 = 500 \text{ pizzas}$$

$$frequência = \frac{500}{2000} * 100 = 25\%$$

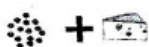
Cogumelo e Queijo



$$300 + 100 = 400 \text{ pizzas}$$

$$frequência = \frac{400}{2000} * 100 = 20\%$$

Pepperoni e Queijo



$$200 + 100 = 300 \text{ pizzas}$$

$$frequência = \frac{300}{2000} * 100 = 15\%$$

Exemplo

Market Basket Analysis

- Cálculo da probabilidade das combinações dos itens:

Cogumelo, Pepperoni e Queijo



100 pizzas

$$frequência = \frac{100}{2000} * 100 = 5\%$$

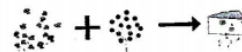
Exemplo


Market Basket Analysis

- Regra 1 com os três itens:

Exemplo

Se pizza de Cogumelo e Pepperoni então Queijo



Estatísticas	Valor
Support = $(A \cap B \Rightarrow C)$	 = 5% ou 0.05
Confidence = $(A \cap B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)}$	$\frac{\text{mushrooms + pepperoni + cheese}}{\text{mushrooms + pepperoni}} = \frac{5\%}{25\%} = 0.20$
Improvement/lift $(A \& B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)P(C)}$	$\frac{\text{mushrooms + pepperoni + cheese}}{\text{mushrooms + pepperoni} \times \text{cheese}} = \frac{5\%}{25\% \times 40\%} = 0.5$

Market Basket Analysis

- Três regras com os três itens:

Se pizza de Cogumelo e Pepperoni então Queijo



Support = 0.05
Confidence = 0.20
Improvement/lift = 0.5

Se pizza de Cogumelo e Queijo então Pepperoni



Support = 0.05
Confidence = 0.25
Improvement/lift = 0.588

Se pizza de Queijo e Pepperoni então Cogumelo



Support = 0.05
Confidence = 0.33
Improvement/lift = 0.74





Exemplo

Data Mining

Market Basket Analysis

Aplicação na área da saúde

Exemplo de arquivo de entrada na área da saúde

	 COD_SERVICO	 DIAGNOSTICO		 DT_ATENDIMENTO
1	20010028	HAS		50440926 12DEC2008:00:00:00
2	20010010	HAS		50440926 16APR2009:00:00:00
3	20010141	HAS		50440926 05JUN2009:00:00:00
4	20010010	HAS		50440926 30NOV2009:00:00:00
5	20010010	HAS		50440926 15APR2010:00:00:00
6	20010010	HAS		50440926 14JUL2010:00:00:00
7	20010010	HAS		50440926 03NOV2010:00:00:00
8	20010010	HAS		50440926 14FEB2011:00:00:00
9	20010010	HAS		50440926 01JUL2011:00:00:00
10	20010010	HAS		50440926 14OCT2011:00:00:00
11	20010141	HAS		50440926 19OCT2011:00:00:00
12	20010010	HAS		50440926 08FEB2012:00:00:00
13	20010010	HAS		50440926 02JUL2012:00:00:00
14	20010141	HAS		50440926 07JUL2012:00:00:00
15	20010010	HAS		50440926 11DEC2012:00:00:00
16	20010141	HAS		50440926 23JAN2013:00:00:00
17	20010010	HAS		50440926 16MAY2013:00:00:00
18	20010010	HAS		50440926 15AUG2013:00:00:00
19	20010028			50440945 05FEB2009:00:00:00
20	20010010			50440945 12MAR2009:00:00:00
21	20010010			50440945 09JUN2009:00:00:00
22	20020058			50440945 16JUL2009:00:00:00
23	20010010			50440945 10DEC2009:00:00:00
24	20010010			50440945 05MAY2010:00:00:00
25	20010028			50440945 24MAY2010:00:00:00
26	20010028			50440945 04APR2011:00:00:00
27	20010010			50440945 13JUN2012:00:00:00
28	20010028			50440945 25JUN2012:00:00:00
29	20020058			50440945 20SEP2013:00:00:00
30	20010010			50440959 08MAR2010:00:00:00
31	20010133			50440959 29MAR2010:00:00:00
32	20020058			50440959 27SEP2010:00:00:00

Exemplo na área da saúde

Análise considerando o tempo (seqüência)

ID	INTERVAL	1
SEQUENCE	INTERVAL	1
TARGET	NOMINAL	1

Sequence Report

Chain Length	Transaction Count	Support (%)	Confidence (%)	Rule
2	1979	55.95	63.21	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG
2	1396	39.47	44.59	ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES
3	1272	35.96	64.27	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG
2	1255	35.48	58.43	ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG
3	859	24.29	61.53	ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG
4	837	23.66	65.80	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG
3	791	22.36	39.97	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES
2	754	21.32	61.55	TESTE ERGOMETRICO (TE) - EM BICICLETA OU EM ESTEIRA ==> ELETROCARDIOGRAMA - ECG
3	745	21.06	59.36	ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG
2	744	21.03	23.76	ELETROCARDIOGRAMA - ECG ==> TESTE ERGOMETRICO (TE) - EM BICICLETA OU EM ESTEIRA
2	729	20.61	23.28	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG & ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES
2	717	20.27	33.38	ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES
2	711	20.10	57.02	ELETROCARDIOGRAMA - ECG & ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG
4	530	14.98	61.70	ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG
4	524	14.81	66.25	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG
2	498	14.08	15.91	ELETROCARDIOGRAMA - ECG ==> SISTEMA HOLTER - 24 HORAS - 2 CANAIS
4	498	14.08	39.15	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO
3	497	14.05	66.80	ELETROCARDIOGRAMA - ECG ==> TESTE ERGOMETRICO (TE) - EM BICICLETA OU EM ESTEIRA ==> ELETROCARDIOGRAMA - ECG
2	463	13.09	62.99	SISTEMA HOLTER - 24 HORAS - 2 CANAIS ==> ELETROCARDIOGRAMA - ECG
3	454	12.84	32.52	ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO
4	452	12.78	60.67	ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG
3	451	12.75	35.94	ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA - ECG ==> ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO
3	439	12.41	22.18	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG & ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A
3	437	12.36	57.96	TESTE ERGOMETRICO (TE) - EM BICICLETA OU EM ESTEIRA ==> ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG
3	435	12.30	59.67	ELETROCARDIOGRAMA - ECG ==> ELETROCARDIOGRAMA - ECG & ECOCARDIOGRAMA BIDIMENSIONAL COM MAPEAMENTO DE FLUXO A CORES ==> ELETROCARDIOGRAMA

Análise de Coluna

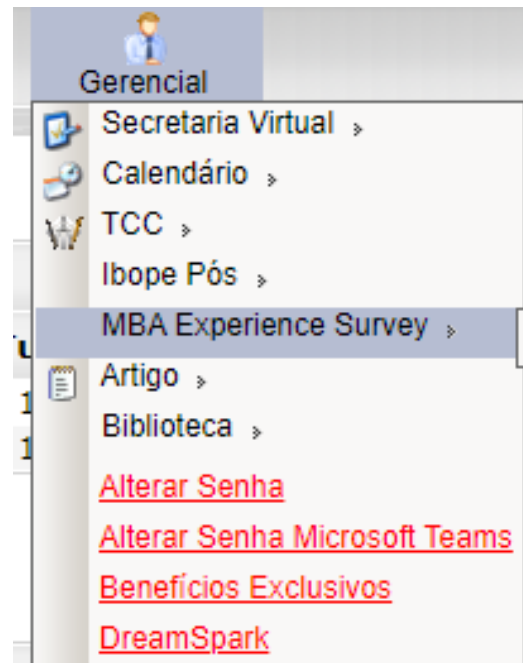
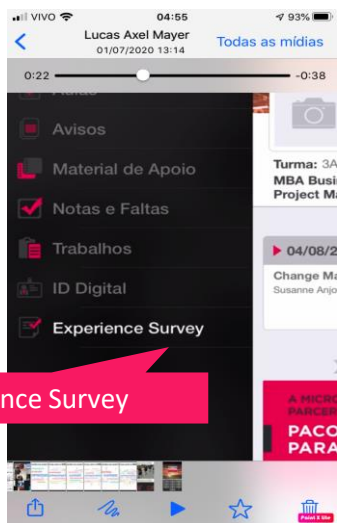
Association Report

Relations	Expected			Transaction			Rule
	Confidence (%)	Confidence (%)	Support (%)	Lift	Count	Rule	
3	6.99	80.00	2.15	11.45	4.00	COLUNA LOMBO-SACRA & ARTICULACAO TIBIO-TARSICA ==> PE OU PODODACTILO	
3	2.69	30.77	2.15	11.45	4.00	PE OU PODODACTILO ==> COLUNA LOMBO-SACRA & ARTICULACAO TIBIO-TARSICA	
2	6.99	77.78	3.76	11.13	7.00	ARTICULACAO TIBIO-TARSICA ==> PE OU PODODACTILO	
2	4.84	53.85	3.76	11.13	7.00	PE OU PODODACTILO ==> ARTICULACAO TIBIO-TARSICA	
3	4.84	50.00	2.15	10.33	4.00	PE OU PODODACTILO & COLUNA LOMBO-SACRA ==> ARTICULACAO TIBIO-TARSICA	
3	4.30	44.44	2.15	10.33	4.00	ARTICULACAO TIBIO-TARSICA ==> PE OU PODODACTILO & COLUNA LOMBO-SACRA	
3	5.91	46.15	3.23	7.80	6.00	COLUNA TOTAL OU ESCOLIOSE PANORAMICA & COLUNA LOMBO-SACRA ==> COLUNA PARA ESCOLIOSE: TA-LATERAL	
3	6.99	54.55	3.23	7.80	6.00	COLUNA PARA ESCOLIOSE: TA-LATERAL ==> COLUNA TOTAL OU ESCOLIOSE PANORAMICA & COLUNA LOMBO-SACRA	
3	5.38	36.36	2.15	6.76	4.00	TORAX: PA & COLUNA CERVICAL: AP-LAT-TO OU FLEXAO ==> ARTICULACAO ESCAPULO-UMERAL	
3	5.91	40.00	2.15	6.76	4.00	ARTICULACAO ESCAPULO-UMERAL ==> TORAX: PA & COLUNA CERVICAL: AP-LAT-TO OU FLEXAO	
3	3.76	23.53	2.15	6.25	4.00	COLUNA LOMBO-SACRA & BACIA ==> ARTICULACAO COXO-FEMORAL (CADA LADO)	
3	9.14	57.14	2.15	6.25	4.00	ARTICULACAO COXO-FEMORAL (CADA LADO) ==> COLUNA LOMBO-SACRA & BACIA	
3	5.91	36.36	2.15	6.15	4.00	COLUNA TOTAL OU ESCOLIOSE PANORAMICA & COLUNA DORSAL: AP-LATERAL ==> COLUNA PARA ESCOLIOSE: TA-LATERAL	
3	5.91	36.36	2.15	6.15	4.00	COLUNA PARA ESCOLIOSE: TA-LATERAL ==> COLUNA TOTAL OU ESCOLIOSE PANORAMICA & COLUNA DORSAL: AP-LATERAL	
3	3.23	17.39	2.15	5.39	4.00	RADIOSCOPIA PARA ACOMPANHAMENTO DE PROCEDIMENTO CIRURGICO & COLUNA LOMBO-SACRA ==> JOELHO OU ROTULA: AP.-LAT - AXIAL	
3	12.37	66.67	2.15	5.39	4.00	JOELHO OU ROTULA: AP.-LAT - AXIAL ==> RADIOSCOPIA PARA ACOMPANHAMENTO DE PROCEDIMENTO CIRURGICO & COLUNA LOMBO-SACRA	
2	15.05	80.00	2.15	5.31	4.00	ESCANOMETRIA ==> COLUNA TOTAL OU ESCOLIOSE PANORAMICA	
2	2.69	14.29	2.15	5.31	4.00	COLUNA TOTAL OU ESCOLIOSE PANORAMICA ==> ESCANOMETRIA	
3	15.05	75.00	3.23	4.98	6.00	COLUNA PARA ESCOLIOSE: TA-LATERAL & COLUNA LOMBO-SACRA ==> COLUNA TOTAL OU ESCOLIOSE PANORAMICA	
3	4.30	21.43	3.23	4.98	6.00	COLUNA TOTAL OU ESCOLIOSE PANORAMICA ==> COLUNA PARA ESCOLIOSE: TA-LATERAL & COLUNA LOMBO-SACRA	
4	12.37	60.00	3.23	4.85	6.00	TORAX :PA - LAT & BACIA ==> RADIOSCOPIA PARA ACOMPANHAMENTO DE PROCEDIMENTO CIRURGICO & COLUNA LOMBO-SACRA	
4	5.38	26.09	3.23	4.85	6.00	RADIOSCOPIA PARA ACOMPANHAMENTO DE PROCEDIMENTO CIRURGICO & COLUNA LOMBO-SACRA ==> TORAX :PA - LAT & BACIA	*
2	15.05	72.73	4.30	4.83	8.00	COLUNA PARA ESCOLIOSE: TA-LATERAL ==> COLUNA TOTAL OU ESCOLIOSE PANORAMICA	
2	5.91	28.57	4.30	4.83	8.00	COLUNA TOTAL OU ESCOLIOSE PANORAMICA ==> COLUNA PARA ESCOLIOSE: TA-LATERAL	
4	4.30	20.69	3.23	4.81	6.00	TORAX :PA - LAT & COLUNA LOMBO-SACRA ==> RADIOSCOPIA PARA ACOMPANHAMENTO DE PROCEDIMENTO CIRURGICO & BACIA	

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)



OBRIGADA



/ Regina T. I. Bernal

FIAP

Copyright © 2023 | Professora Dra. Regina Tomie Ivata Bernal
Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente
proibido sem consentimento formal, por escrito, do professor/autor.

FIAP