FIAP

MBA

MBA em DATA SCIENCE

STATISTICS FUNDAMENTALS

Dra. Regina Tomie Ivata Bernal
Cientista de Dados na área da Saúde

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

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

Aula 2
Noções de Probabilidade

NOÇÕES DE **PROBABILIDADE**

PROBABILIDADE

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

Exemplos:

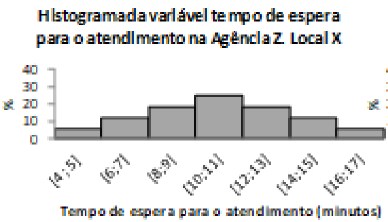
- Condições climáticas no próximo domingo.
- Faturamento da empresa no próximo mês.
- Quantidade de clientes cancelados nos próximos seis meses.
- Taxa de inflação no próximo mês.

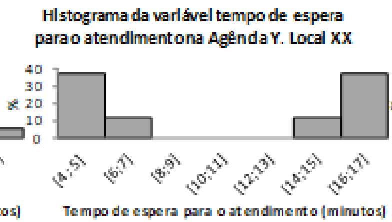
Modelos podem ser estabelecidos para quantificar as INCERTEZAS. → **MODELOS PROBABILÍSTICOS**

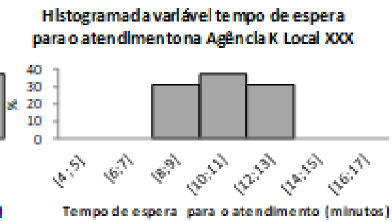
PROBABILIDADE

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. $P(S)$ = 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

FIAP MBA+

PROBABILIDADE

Definição 2:

"As frequências relativas são estimativas de probabilidade de ocorrência de certos eventos de interesse. Com suposições adequadas, e sem observamos 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?

01	02	03	04	05	06	07	08	09	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60

Para anular este jogo, marque ao lado:

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.

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

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 : $n! = n.(n-1).(n-2)\dots 3.2.1$

Permutação: $P = n!$

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

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

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

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

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 ... (número inteiro positivo)

e = 2.71828

k! = k.(k-1).(k-2) ... 3.2.1

λ = média do valor esperado de k

Distribuição de Probabilidade

- Dados históricos

https://github.com/henriquepgomide/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	Grêmio - RS	Mineirão - Belo Horizonte - MG
	2	1	15/04/2018 - 19:00	Atlético - PR	5 x 1	5	1	6	Chapecoense - SC	Arena da Baixada - Curitiba - PR
	3	1	15/04/2018 - 11:00	América - MG	3 x 0	3	0	3	Sport - PE	Independência - Belo Horizonte - MG
	4	1	14/04/2018 - 19:00	Vitória - BA	2 x 2	2	2	4	Flamengo - RJ	Manoel Barradas - Salvador - BA
	5	1	15/04/2018 - 16:00	Vasco da Gama - RJ	2 x 1	2	1	3	Atlético - MG	São Januário - Rio de Janeiro - RJ
	6	1	16/04/2018 - 20:00	Botafogo - RJ	1 x 1	1	1	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	Paraná - PR	Morumbi - Sao Paulo - SP

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

Distribuição de Probabilidade

- Dados históricos

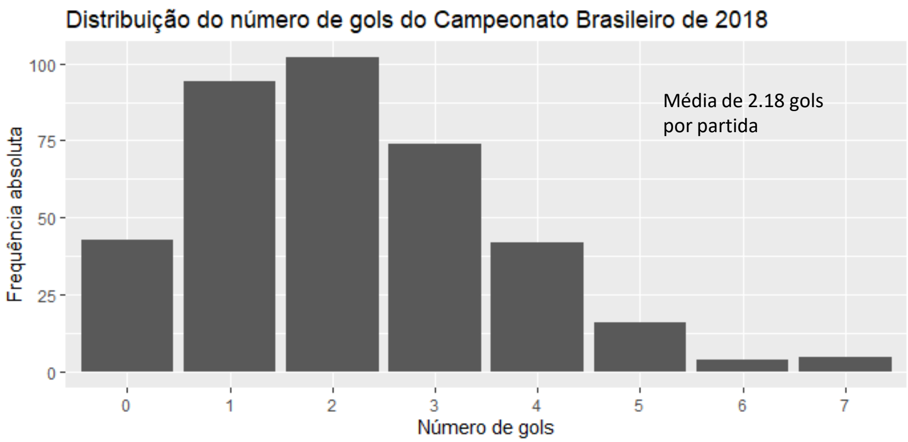
https://github.com/henriquepgomide/caRtola/tree/master/data

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	Grêmio - RS	Mineirão - Belo Horizonte - MG
	2	1	15/04/2018 - 19:00	Atlético - PR	5 x 1	5	1	6	Chapecoense - SC	Arena da Baixada - Curitiba - PR
	3	1	15/04/2018 - 11:00	América - MG	3 x 0	3	0	3	Sport - PE	Independência - Belo Horizonte - MG
	4	1	14/04/2018 - 19:00	Vitória - BA	2 x 2	2	2	4	Flamengo - RJ	Manoel Barradas - Salvador - BA
	5	1	15/04/2018 - 16:00	Vasco da Gama - RJ	2 x 1	2	1	3	Atlético - MG	São Januário - Rio de Janeiro - RJ
	6	1	16/04/2018 - 20:00	Botafogo - RJ	1 x 1	1	1	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	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

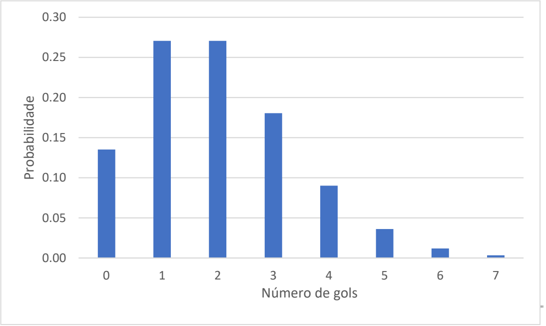
- Distribuição de Poisson

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

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 uteis?

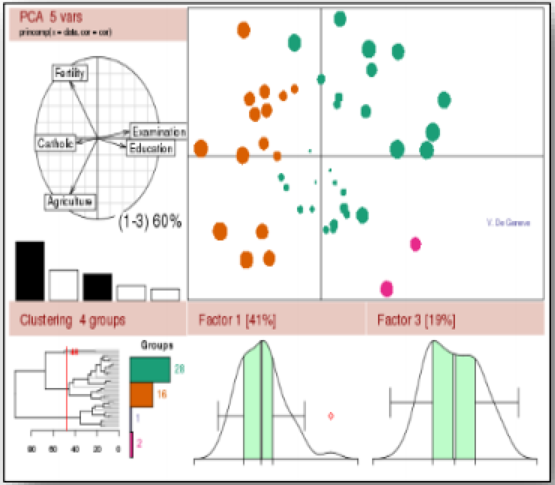
Os valores das probabilidades encontram-se em tabelas estatísticas que podem ser facilmente utilizadas para análise de teste de hipóteses, análise de associação e saídas de modelos preditivos.

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

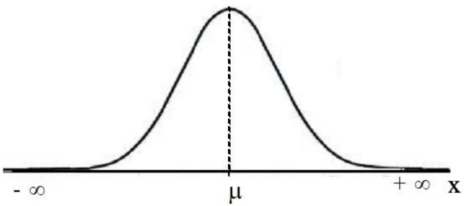
https://www.youtube.com/watch?v=a6dRG3V5l6s

DISTRIBUIÇÃO **DE POISSON**



Exercícios

DISTRIBUIÇÃO **NORMAL**



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 $[-\infty \ldots\ldots +\infty]$.

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:

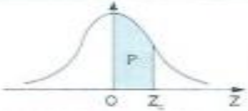
$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}.e^{-1/2[(X-\mu)/\sigma]^2}$$

DISTRIBUIÇÃO **NORMAL**

Qual a probabilidade de ocorrer valores entre 0 e 1?

P (0< Z < 1.00) = ?

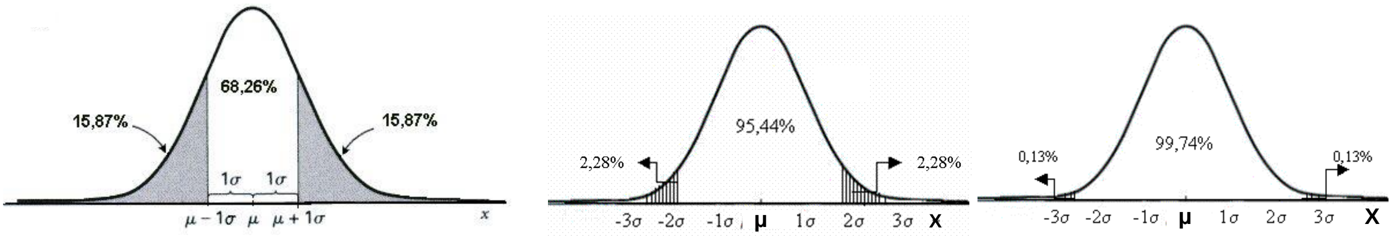
Tabela III — Distribuição Normal Padrão
$Z \sim N(0, 1)$
Corpo da tabela dá a probabilidade p, tal que $p = P(0 < Z < Z_c)$



parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c 0	1	2	3	4	5	6	7	8	9	parte inteira e primeira decimal de Z_c
	p = 0										
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

Distribuição Normal Padronizada

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

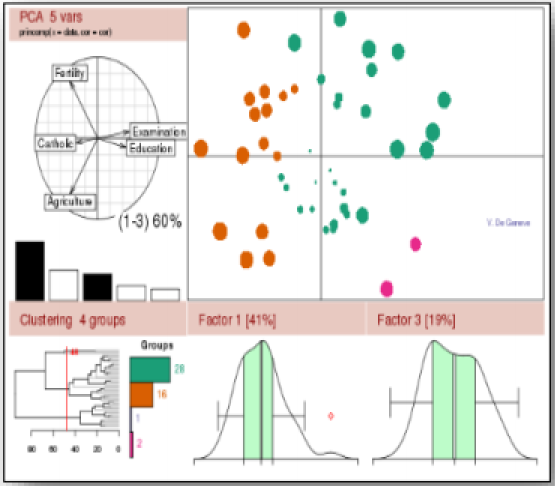


$$P[(\mu-\sigma) < X < (\mu+\sigma)] = 68.25\%$$

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

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

DISTRIBUIÇÃO **NORMAL**



Exercícios

NORMALIZAÇÃO **DOS DADOS**

- Distribuição Normal Padronizada

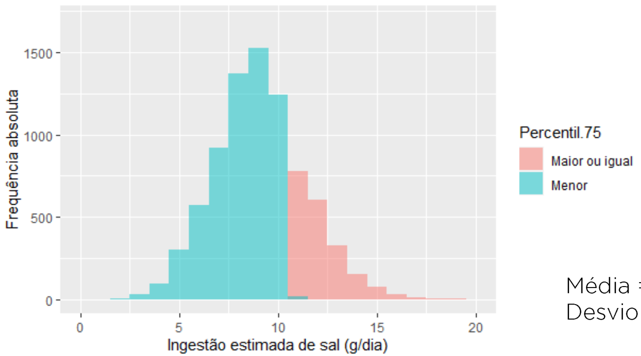
$$X \sim N(\mu, \sigma^2) \Rightarrow \boxed{Z = \frac{X - \mu}{\sigma}} \Rightarrow Z \sim N(0,1)$$

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



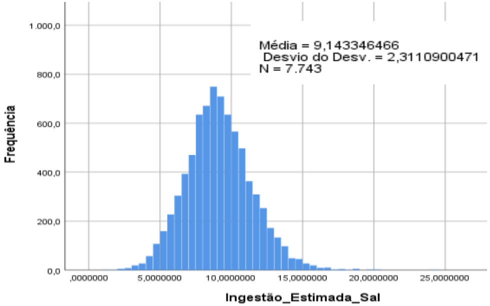
Média = 9.14g/dia
Desvio padrão =2.31g/dia

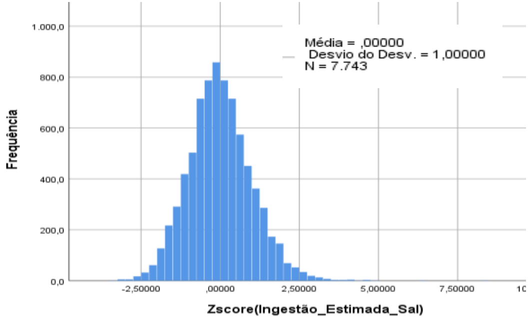
Fonte: Pesquisa Nacional de Saúde 2013 – População adulta

NORMALIZAÇÃO **DOS DADOS**

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

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





FIAP MBA+

DISTRIBUIÇÃO **NORMAL PADRONIZADA**

EXEMPLO

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

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

Ingestão de sal (X)
8.56
4.54
13.7

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

Z= (4.54-9.14)/2.31 = -1.99 ~ -2dp

Z= (13.7 – 9.14)/2.31 = +1.97 ~+2 dp

FIAP MBA+

NORMALIZAÇÃO **DOS DADOS**

- Distribuição Normal Padronizada

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

- Máximo e Mínimo

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

PROBABILIDADE **CONDICIONAL**

PROBABILIDADE **CONDICIONAL**

INFERÊNCIA BAYESIANA → PROBABILIDADES SUBJETIVAS

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 ½. 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 2/3 de ocorrer o evento A.

Fonte: Exemplo extraído do livro Estatística Básica página 121.

PROBABILIDADE **CONDICIONAL**

INFERÊNCIA BAYESIANA → PROBABILIDADES SUBJETIVAS

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:

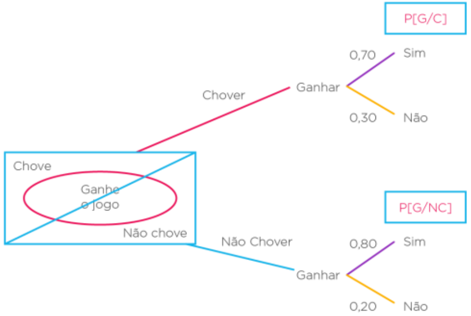
$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \; P(B) > 0$$

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?



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

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

EXEMPLO

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)

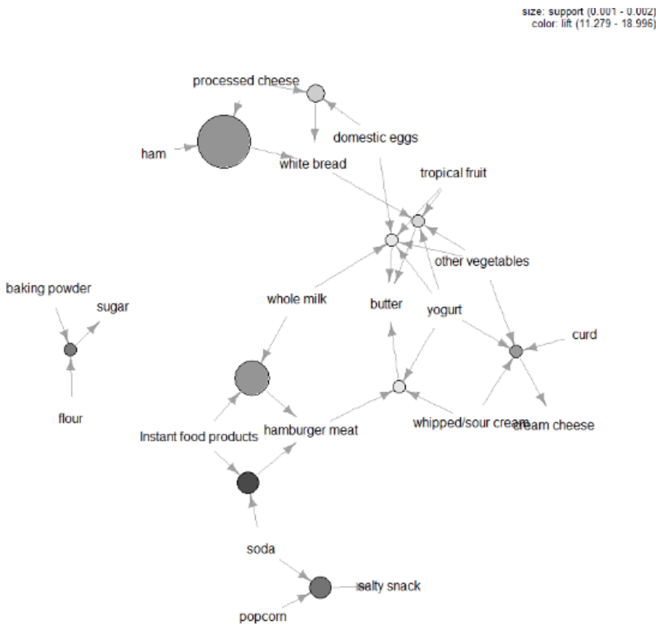
$$(A \& B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)P(C)}$$

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.



Técnica Descoberta de Sequências

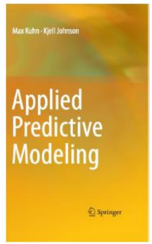
Quais associações são significativas ?

EXEMPLO

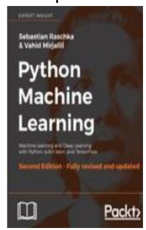
Item comprado anteriormente

Itens a serem sugeridos de acordo coma força

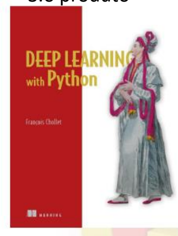
1.o produto



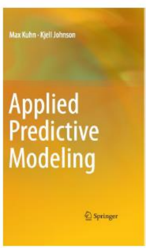
2.o produto



3.o produto

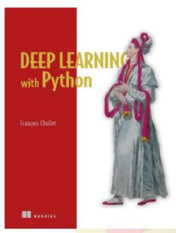


Item comprado anteriormente



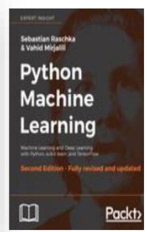
Itens a serem sugeridos de acordo coma força

1.o produto



2.o produto

3.o produto



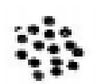
Market Basket Analysis

Exemplo

- 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 pizzas

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

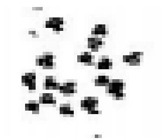
Fonte: BERRY,M.J.A., LINOFF,G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011, página 146.

Market Basket Analysis

- Cálculo da probabilidade das combinações dos itens:

Exemplo

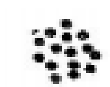
Cogumelo



100 + 400 + 300 + 100 = 900 pizzas

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

Pepperoni



150 + 400 + 200 + 100 = 850 pizzas

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

Queijo

200 + 300 + 200 + 100 = 800 pizzas

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

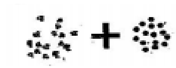
Fonte: BERRY,M.J.A., LINOFF,G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011, página 146.

Market Basket Analysis

Exemplo

- Cálculo da probabilidade das combinações dos itens:

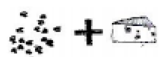
Cogumelo e Pepperoni



400 + 100 = 500 pizzas

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

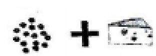
Cogumelo e Queijo



300 + 100 = 400 pizzas

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

Pepperoni e Queijo



200 + 100 = 300 pizzas

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

Fonte: BERRY,M.J.A., LINOFF,G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011, página 146.

Market Basket Analysis

- Cálculo da probabilidade das combinações dos itens:

Exemplo

Cogumelo, Pepperoni e Queijo

100 pizzas

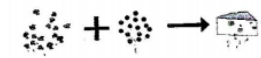
$$frequência = \frac{100}{2000} * 100 = 5\%$$

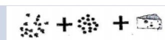
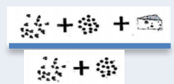
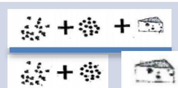
Fonte: BERRY,M.J.A., LINOFF,G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011, página 146.

Market Basket Analysis

- Regra 1 com os três itens:

Exemplo

Se pizza de Cogumelo e Pepperoni então Queijo 

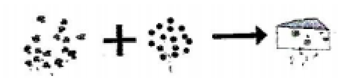
Estatísticas	Valor
Support = $(A \cap B \Rightarrow C)$	 = 5% ou 0.05
Confidence $=(A \cap B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)}$	 $= \frac{5\%}{25\%} = 0.20$
Improvement/lift $(A \& B \Rightarrow C) = \frac{P(A \cap B \cap C)}{P(A \cap B)P(C)}$	 $= \frac{5\%}{25\% x 40\%} = 0.5$

Market Basket Analysis

Exemplo

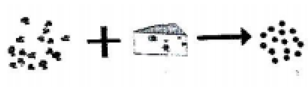
- Três regras com os três itens:

Se pizza de Cogumelo e Pepperoni então Queijo



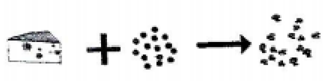
Support = 0.05
Confidence = 0.20
Improvement/lift = 0.5

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



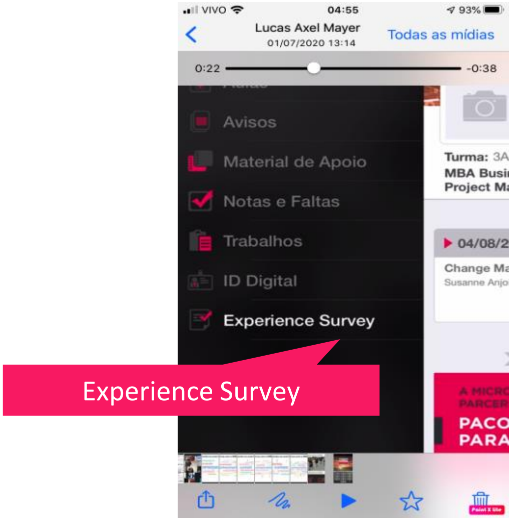
Support = 0.05
Confidence = 0.33
Improvement/lift = 0.

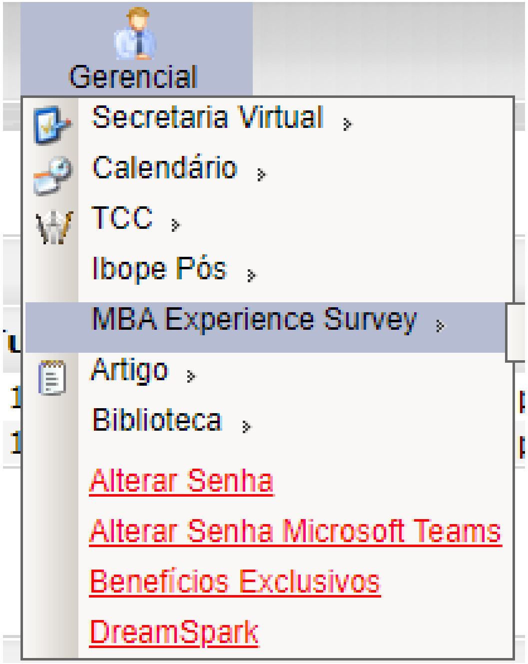
Fonte: BERRY,M.J.A., LINOFF,G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011, página 146.

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

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





A grande finalidade do conhecimento não é conhecer, mas agir.

T. Huxley

OBRIGADO

/ Regina T. I. Bernal

FIAP

FIAP