DATA SCIENCE

AULA 4 - ESTATÍSTICA

PROFª. ANA CAROLINA B. ALBERTON

AULA PASSADA

- O desvio padrão é uma medida que indica o quando o conjunto de dados é uniforme, considerando o quanto os dados se afastam da sua média.

$$D_P = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M_A)^2}{n}}$$

- O cálculo do desvio padrão é a raiz quadrada da variância. Vamos ao cálculo então.

AULA PASSADA - PROBABILIDADE

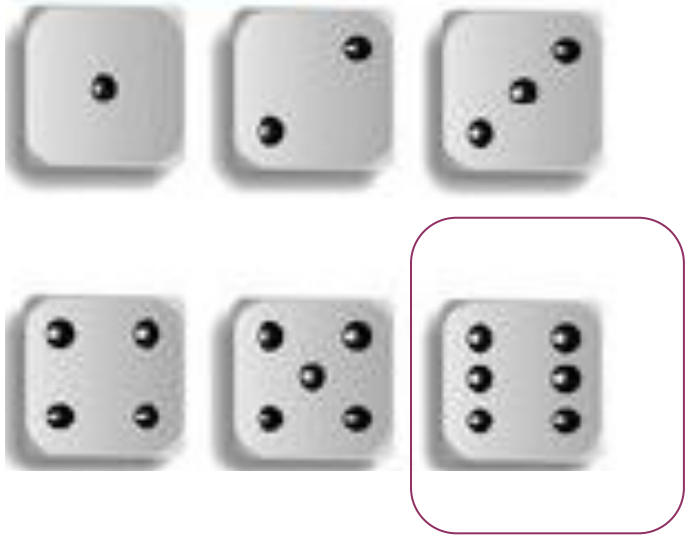
Um único evento

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo: Jogar um dado e **dar 6**:

P = ⅙

P = 0,16



AULA PASSADA - PROBABILIDADE

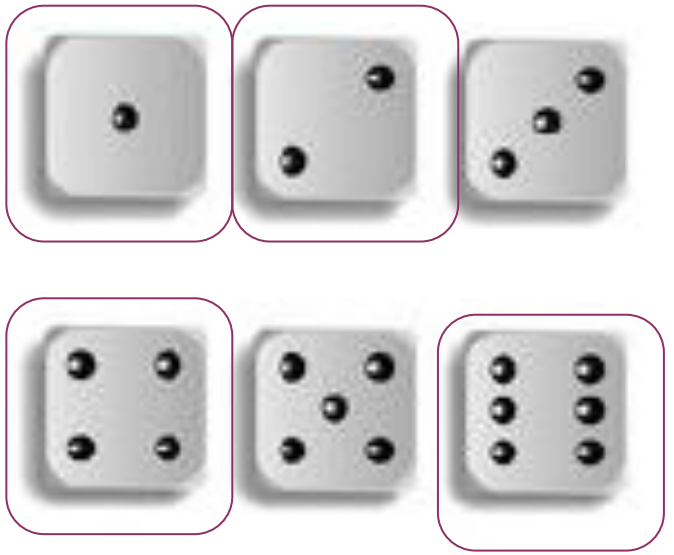
Eventos Excludentes

Soma-se as probabilidades

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo: Jogar um dado e ser **1 ou par**:
P = ⅙ + 3/6

P = 0,67



AULA PASSADA - PROBABILIDADE

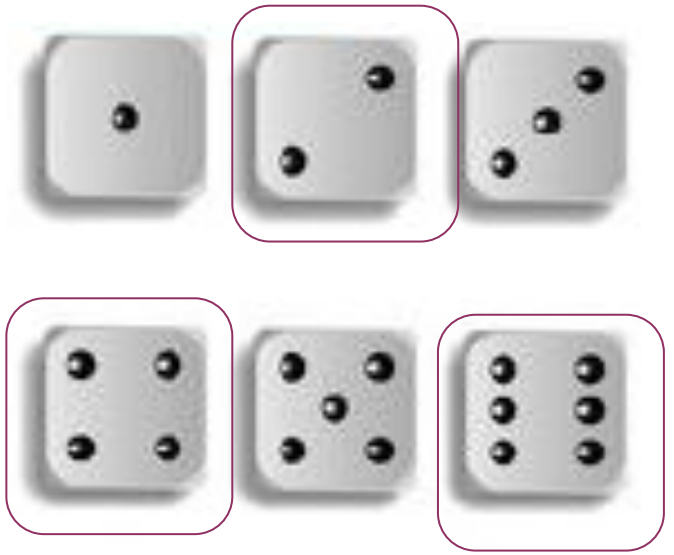
Eventos Não Excludentes

Soma-se as probabilidades, diminui-se as sobreposições

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo: Jogar um dado e ser **2 ou par**:

P = ⅙ + 3/6 - ⅙
P = 0,5



AULA PASSADA - PROBABILIDADE

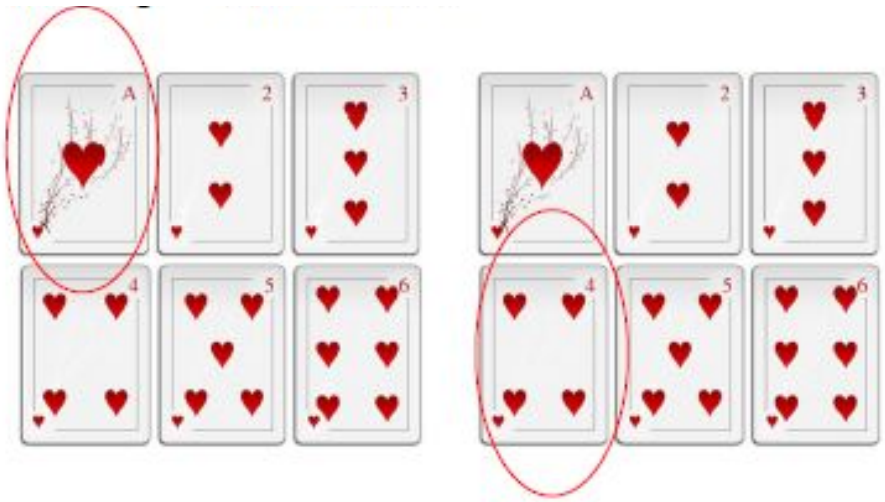
Eventos Dependentes

Mais de um evento, eles se relacionam com Multiplicação

Exemplo: Com 6 cartas na mão, qual a probabilidade de tirar primeiro um **A** e depois um **4**?

(Dois eventos dependentes)

P = ⅙ * ⅕ = 0,028



PASSEIO ALEATÓRIO

- Passeio aleatório é um modelo que descreve uma sequência de passos aleatórios que formam um caminho.
- O passeio aleatório pode ser usado para modelar vários cenários do mundo real que envolvem aleatoriedade
- Sucessão de Etapas Aleatórias e Independentes (uma não influencia a outra);
- Em finanças é usado para estudar os preços de ações;
- É usado para estudar vários fenômenos naturais e empresariais;

PASSEIO ALEATÓRIO

- Não é "totalmente aleatório"
- Existe uma distribuição de probabilidades

PASSEIO ALEATÓRIO

$n \setminus x$	-5	-4	-3	-2	-1	0	1	2	3	4	5
0						1					
1					$\frac{1}{2}$	0	$\frac{1}{2}$				
2				$\frac{1}{4}$	0	$\frac{2}{4}$	0	$\frac{1}{4}$			
3			$\frac{1}{8}$	0	$\frac{3}{8}$	0	$\frac{3}{8}$	0	$\frac{1}{8}$		
4		$\frac{1}{16}$	0	$\frac{4}{16}$	0	$\frac{6}{16}$	0	$\frac{4}{16}$	0	$\frac{1}{16}$	
5	$\frac{1}{32}$	0	$\frac{5}{32}$	0	$\frac{10}{32}$	0	$\frac{10}{32}$	0	$\frac{5}{32}$	0	$\frac{1}{32}$

PASSEIO ALEATÓRIO EM AÇÕES/INVESTIMENTO

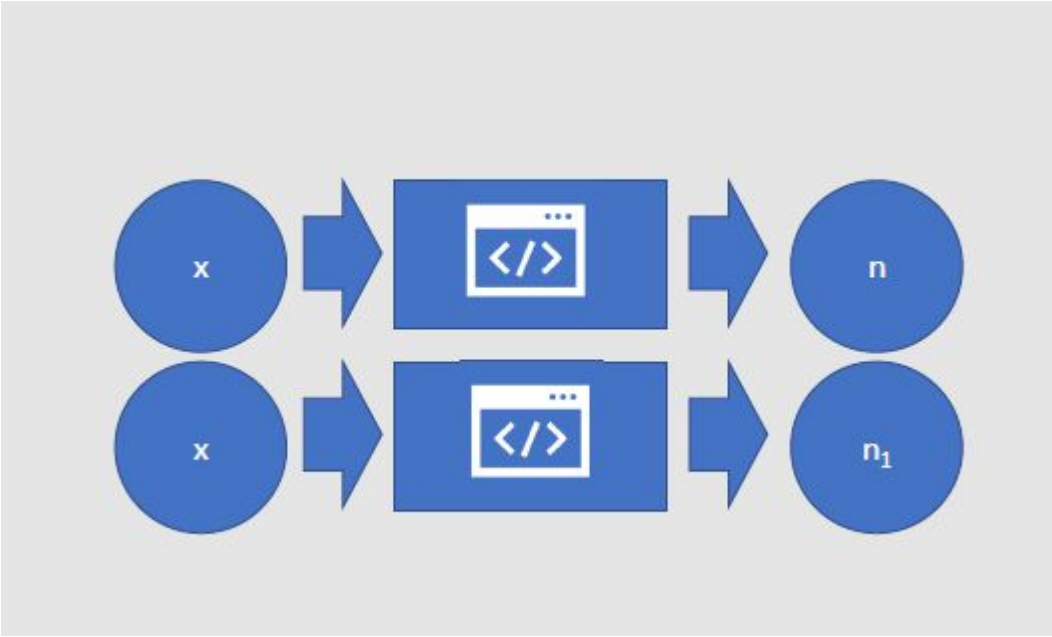
Duas correntes:

- Comportamento de ações são um passeio aleatório, portanto imprevisíveis com Análise Técnica
  - Effcient Market Hypothesis (EMH): única maneira de ganhar é com alto risco
- Comportamento de ações mantém padrões e tendências ao longo do tempo.
- Ver exemplo PYTHON

MODELO ESTOCÁSTICO E DETERMINÍSTICO

Estocástico: dada uma mesma entrada, a saída pode variar

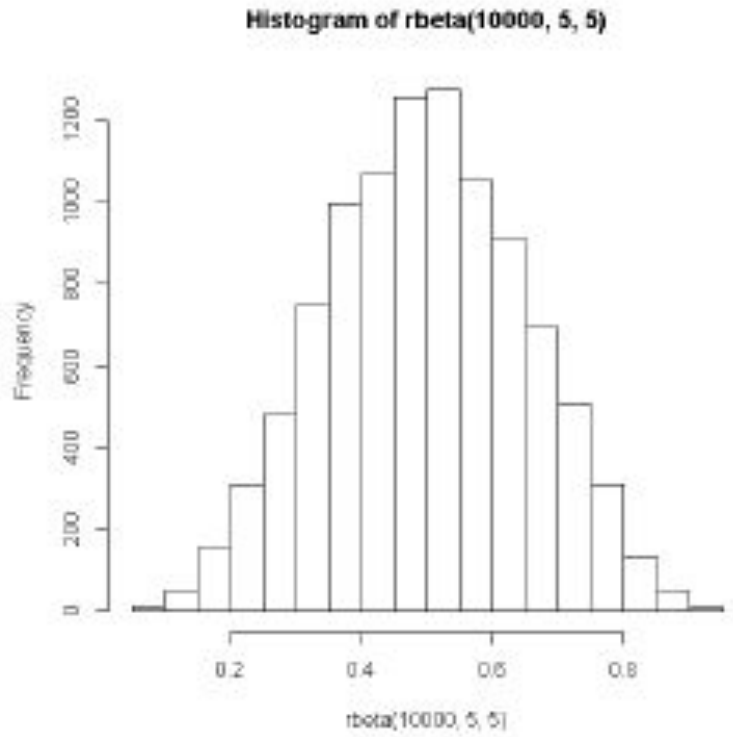
Determinístico: dada uma mesma entrada, apresenta sempre a mesma saída



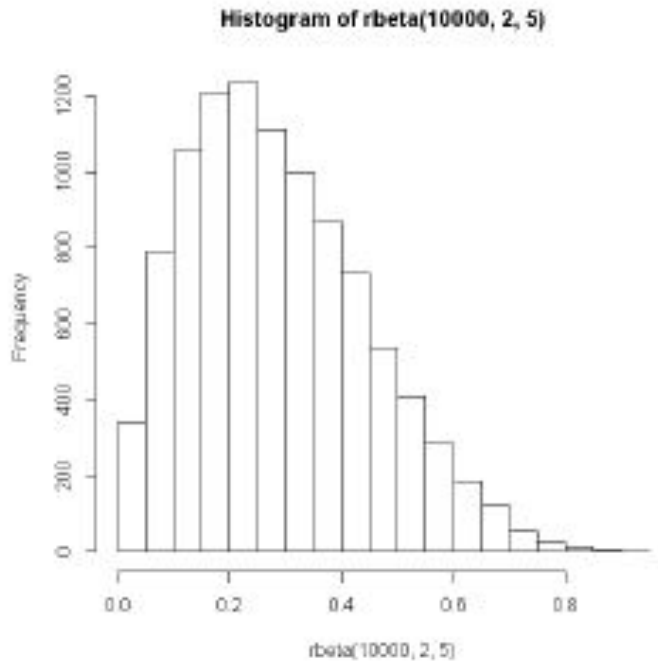
EX: Algoritmos como Random Forest e k-means são Estocásticos

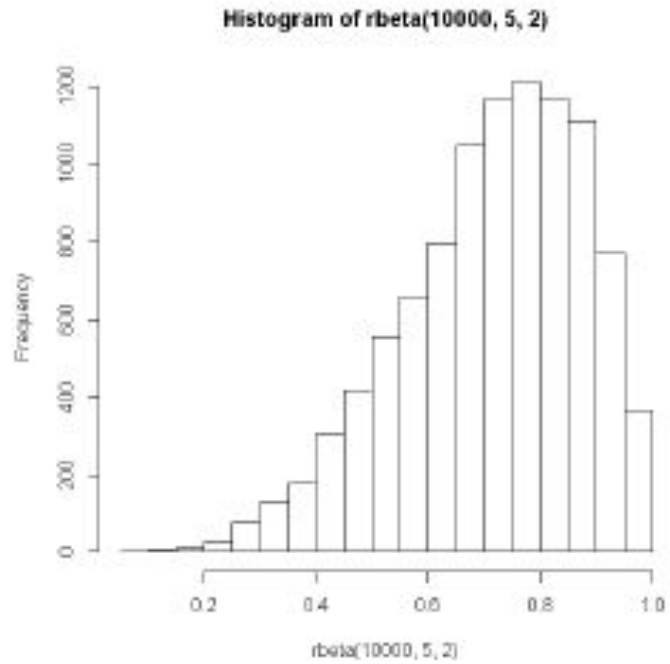
DISTRIBUIÇÃO

- Usado principalmente na teoria da probabilidade
- Comportamento de dados aleatórios



DISTRIBUIÇÃO



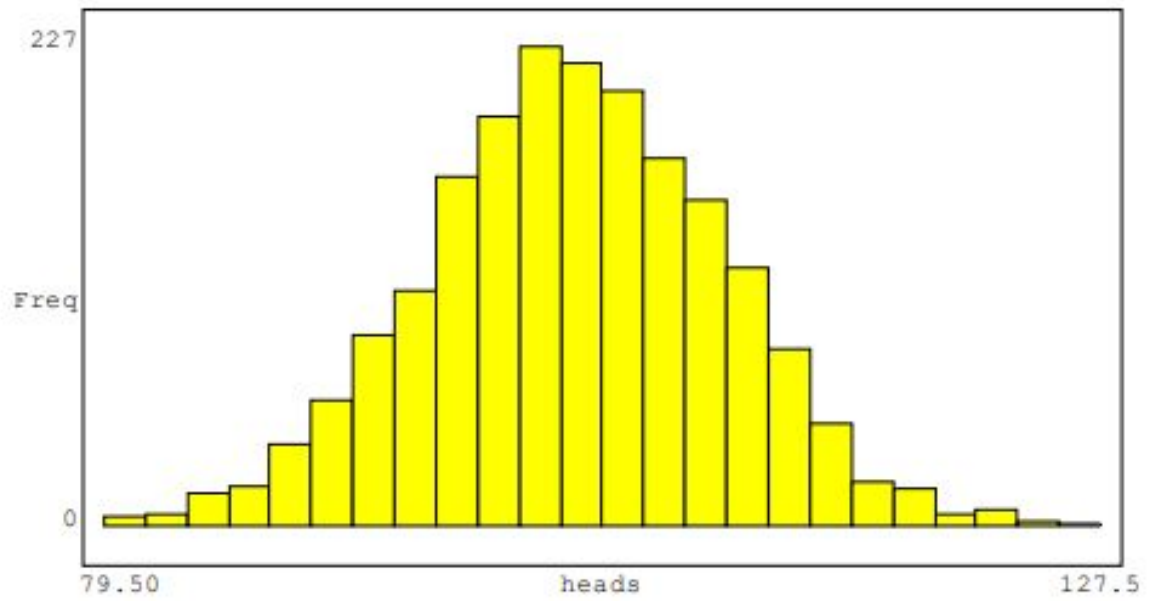


DISTRIBUIÇÃO NORMAL - O HISTOGRAMA

- Por exemplo, suponhamos 2000 lançamentos de 200 moedas. Vamos calcular quantas 'caras' vou encontrar. Utilizando um software de simulação, obtivemos os seguintes resultados:

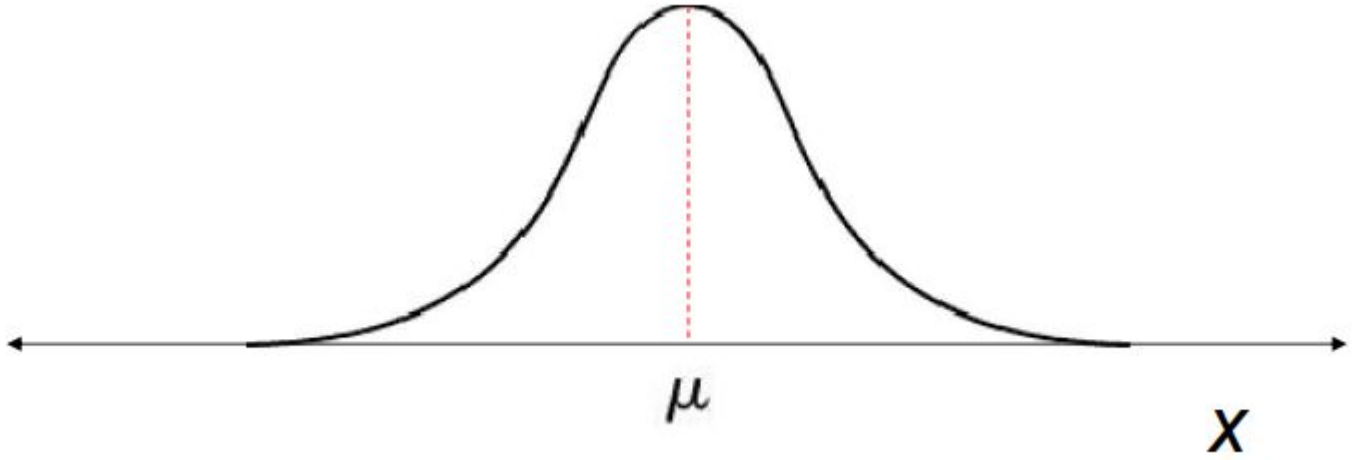
[80 ... 81]	5
[82 ... 83]	6
[84 ... 85]	16
[86 ... 87]	19
[88 ... 89]	39
[90 ... 91]	60
[92 ... 93]	91
[94 ... 95]	111
[96 ... 97]	165
[98 ... 99]	194
[100 ... 101]	227
[102 ... 103]	220
[104 ... 105]	206
[106 ... 107]	174
[108 ... 109]	155
[110 ... 111]	123
[112 ... 113]	84
[114 ... 115]	49
[116 ... 117]	21
[118 ... 119]	18
[120 ... 121]	6
[122 ... 123]	8
[124 ... 125]	2
[126 ... 127]	1

sample mean = 102.132
sample st dev = 7.238

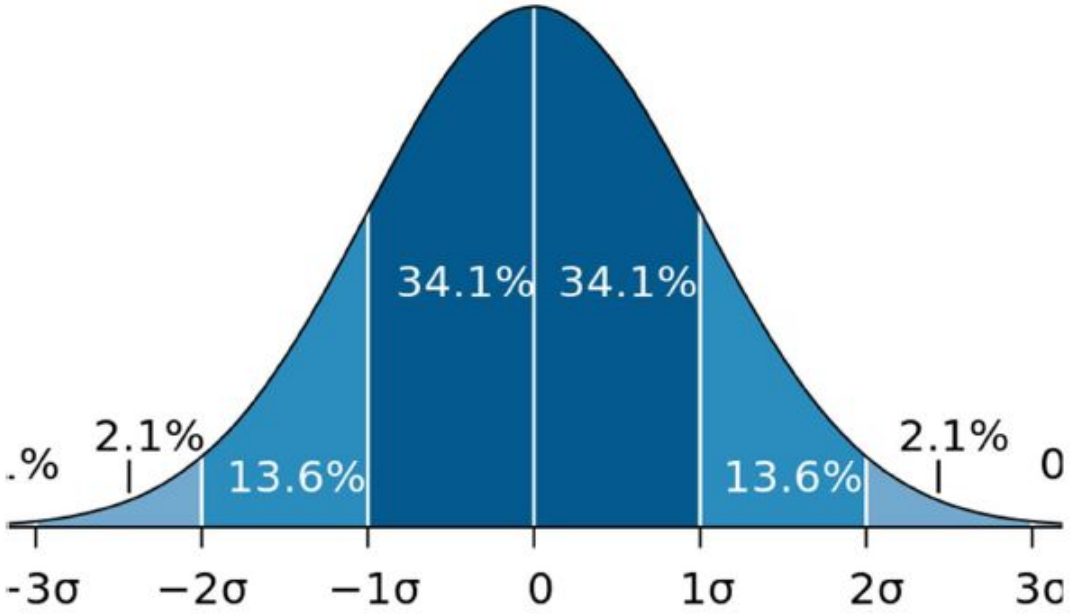


DISTRIBUIÇÃO NORMAL

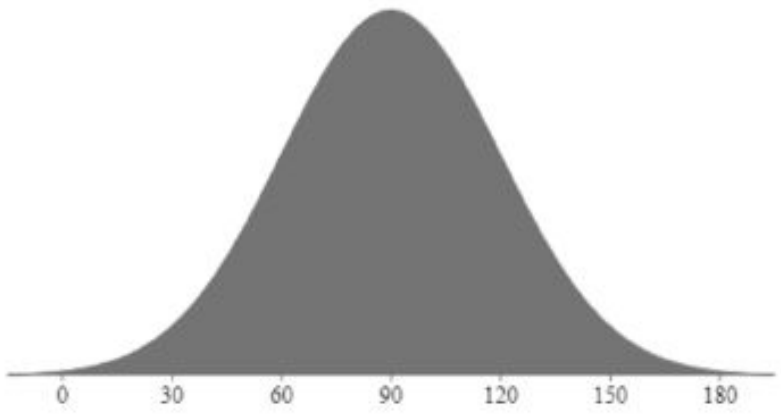
- O mundo é Normal! Muitos dos fenômenos aleatórios que encontramos na prática apresentam uma distribuição muito peculiar, chamada Normal.
- Tem forma de sino e é simétrica em torno da média.
- A área total sob a curva é de 1.



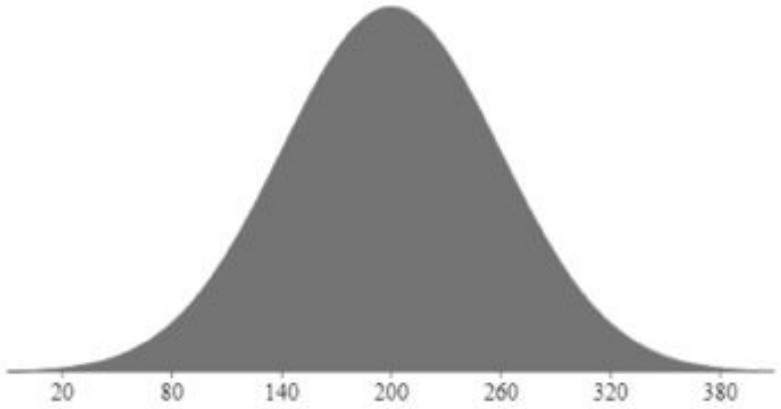
DISTRIBUIÇÃO NORMAL



DISTRIBUIÇÃO NORMAL



$\mu = 90$
$\sigma = 30$



$\mu = 200$
$\sigma = 60$

DISTRIBUIÇÃO NORMAL - EXEMPLO

7.57	6.72	5.59	9.56	4.79	4.84	5.87	10.23	9.53	6.99
9.51	9.21	5.78	6.72	8.96	7.32	7.64	8.53	5.90	7.93
8.82	8.45	7.99	5.77	4.76	4.49	8.97	6.60	8.55	6.30
6.54	5.98	10.88	8.92	7.01	7.58	9.47	6.34	6.17	7.46
8.78	7.13	7.71	8.06	7.67	7.05	9.66	4.37	15.08	9.20
7.64	5.89	11.16	5.35	5.75	8.98	8.74	8.20	8.79	5.80
11.70	5.53	7.75	6.54	9.79	7.43	9.14	5.78	10.31	10.12
9.68	8.11	5.54	10.41	8.83	10.00	5.54	10.32	6.96	7.93
10.14	9.66	10.67	8.17	8.86	8.40	5.15	6.98	8.19	8.72
8.76	8.02	8.93	8.54	3.26	10.06	8.18	2.43	9.17	12.00

Média = 8

Desvio Padrão = 2

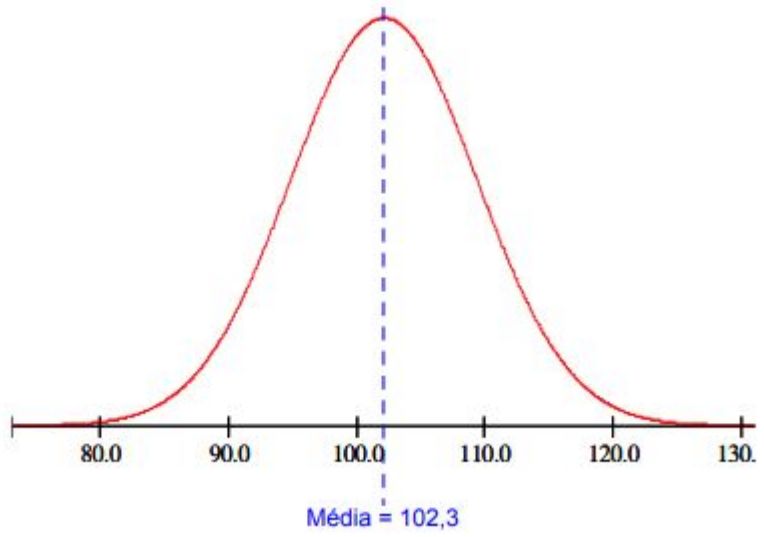
DISTRIBUIÇÃO NORMAL

- Essa curva que chamamos de sino recebe um nome especial: Curva Normal.
- A Curva Normal é a representante do modelo normal e é obtida a partir da função densidade que nada mais é do que uma função que origina o gráfico anterior. Assim, se X é uma variável aleatória com distribuição Normal com média μ e variância σ^2 então a sua densidade é dada por

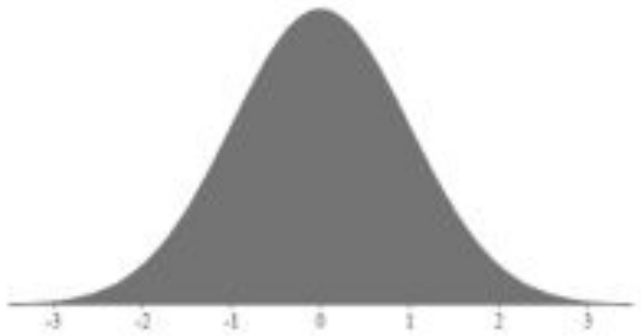
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

- A Normal apresenta as seguintes propriedades:
  - é simétrica ao redor da média;
  - a área sobre a curva é igual a 1;
  - para valores muito grandes de x, tendendo a infinito

(ou muito pequenos, tendendo a menos infinito), a curva tende a zero.



DISTRIBUIÇÃO NORMAL PADRÃO (Z)



- Distribuição de Referência para outras Distribuições Normais
- Média Zero e Desvio Padrão =1

DISTRIBUIÇÃO NORMAL PADRÃO (Z)

Mostra o número de desvios padrões que o valor está acima ou abaixo da média (score z ou valor z)

- Ex: score z = -2 quer dizer que os dados estão a dois desvios padrão abaixo da média

- Usa-se uma fórmula para calcular a probabilidade de seus dados com relação a tabela Z

DISTRIBUIÇÃO NORMAL PADRÃO (Z) E PROBABILIDADE

Para encontrar a probabilidade, utiliza-se a Tabela Z para facilitar.

Com a fórmula abaixo você transforma a probabilidade da sua distribuição na probabilidade da tabela Z

Então você olha a probabilidade na tabela!

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

onde X = seu valor

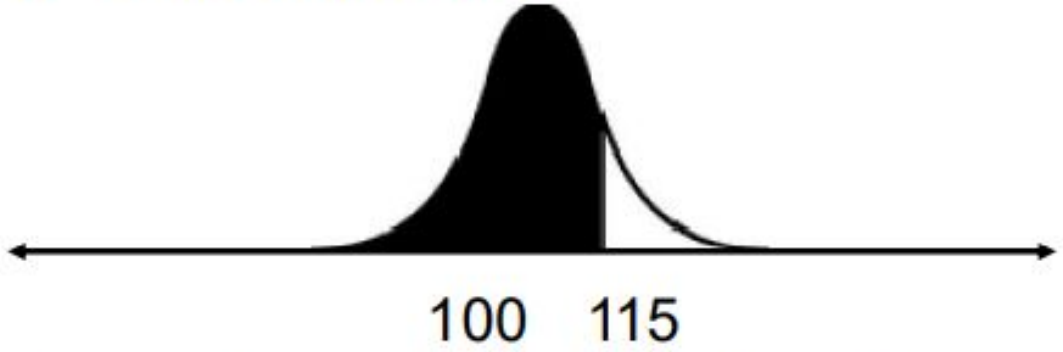
μ = média

σ = desvio padrão

PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

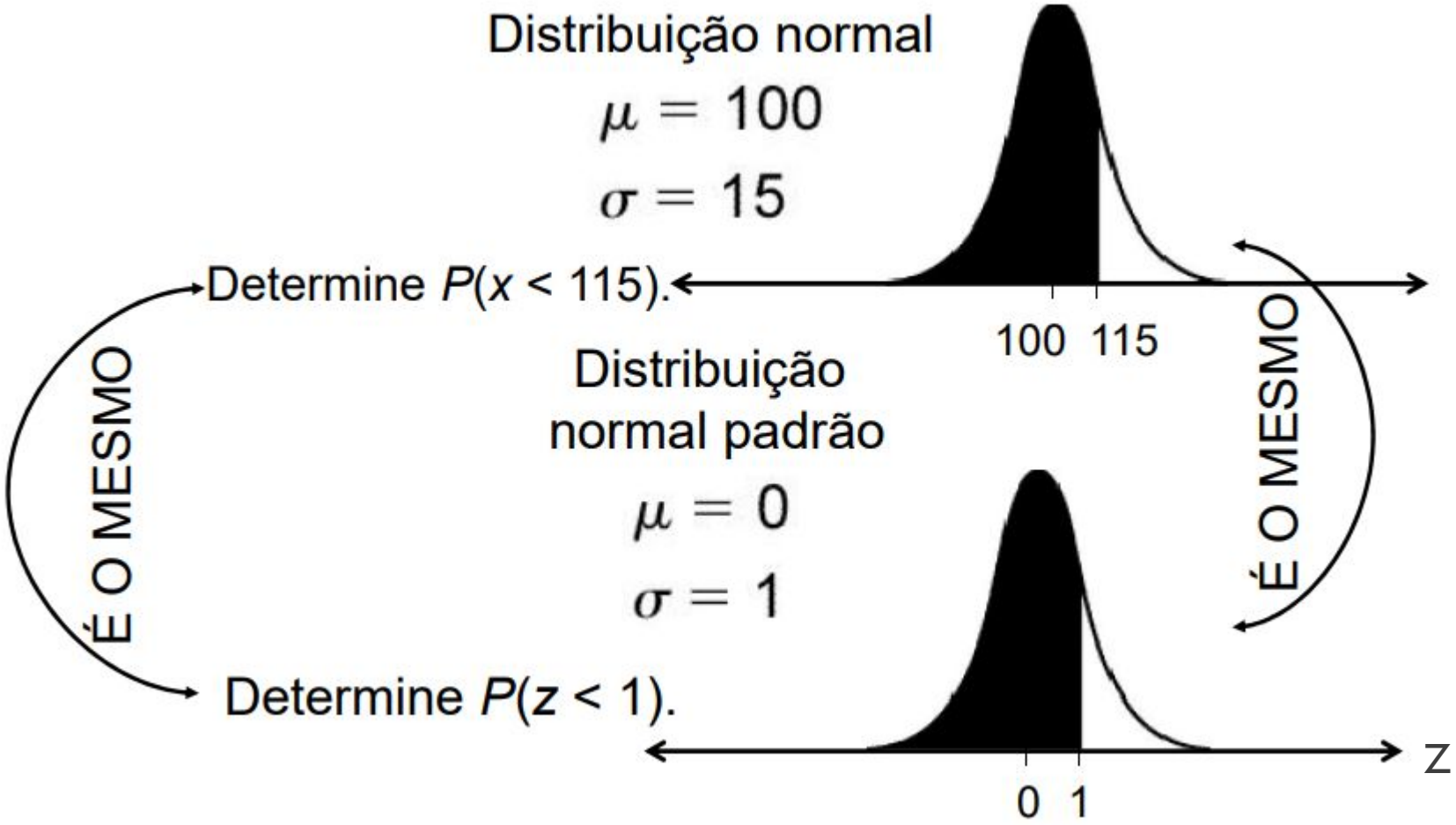
- Se uma variável aleatória x é normalmente distribuída, a probabilidade de que ela esteja dentro de dado intervalo é igual à área sob a curva nesse intervalo.

- **Ex:** Pontuações de QI são normalmente distribuídas, com uma média de 100 e um desvio padrão de 15. Determine a probabilidade de que uma pessoa selecionada aleatoriamente tenha uma pontuação de QI inferior a 115.
- Para determinar a área nesse intervalo, primeiro encontre o escore z correspondente a x = 115.



$$z = \frac{150 - 100}{15} = 1$$

PROBABILIDADE E DISTRIBUIÇÕES NORMAIS



Z=I

NORMAIS

Distribuição Normal: Valores de p tais que $P(0 \leq Z \leq z) = p$

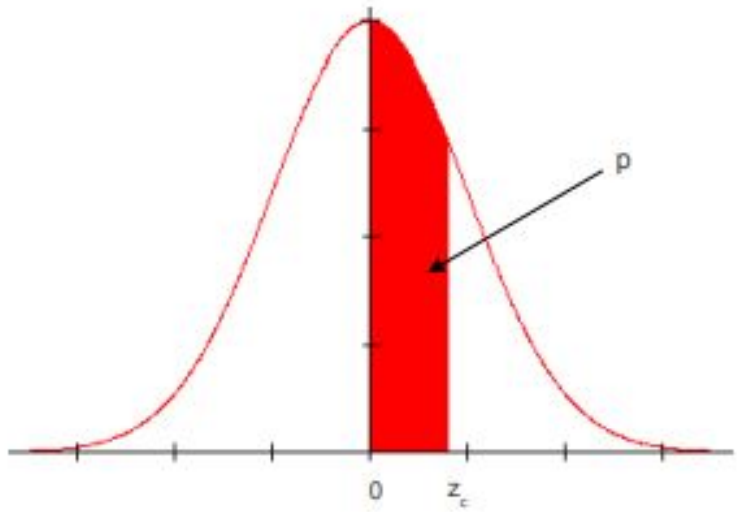
Parte inteira e primeira decimal de z	Segunda casa decimal de z									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

$P(z < 1) = 0,3413,$

logo $P(x < 115) = 0,5 + 0,3413$

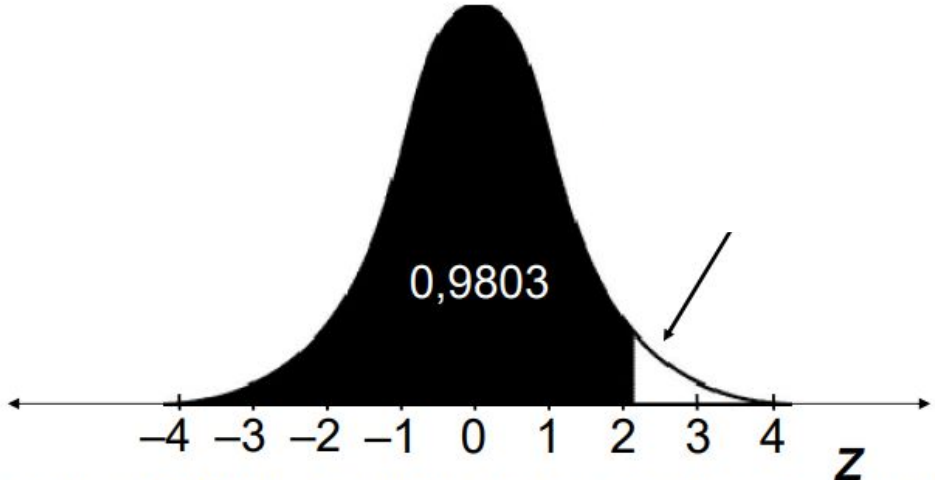
$P(x < 115) = 0,8413$

$= 84,13\%$



PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

- Exemplo: Determine o escore z correspondente a uma área acumulada de 0,9803.



- Resp: Localize o 0,9083 – 0,5 na tabela. Leia os valores no início da linha e no alto da coluna correspondente.

O Escore Z será de 2,06

PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

- Exemplo 1: Existe um conjunto de objetos em uma cesta, cujos pesos são normalmente distribuídos com
  média = 8 e desvio padrão igual a 2.
  Qual a chance de se tirar um objeto pesando menos de 6 quilos?

PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

- Exemplo 1: Existe um conjunto de objetos em uma cesta, cujos pesos são normalmente distribuídos com média = 8 e desvio padrão igual a 2.
  Qual a chance de se tirar um objeto pesando menos de 6 quilos?

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

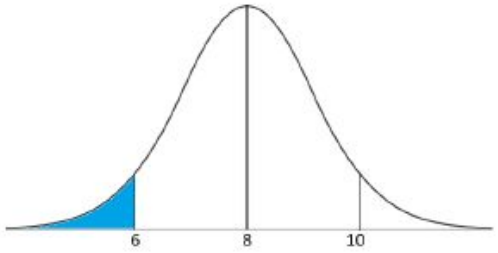
$$X = ?\,(6)$$
$$\mu = média$$
$$\sigma = desvio\ padrão$$

$$z = \frac{6 - 8}{2}$$

$$z = -1$$

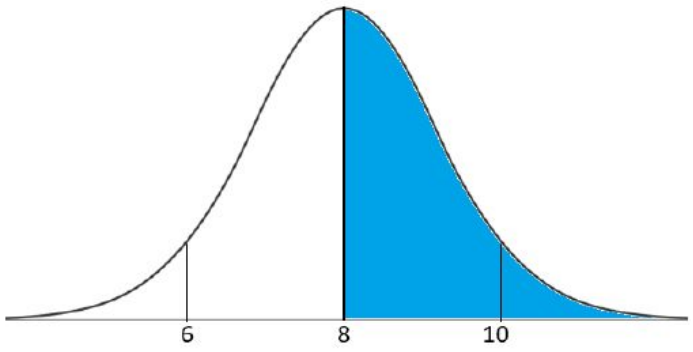
$$P = 0{,}1587$$

z	.00	.01	.02	.03	.04	.05	.06
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446



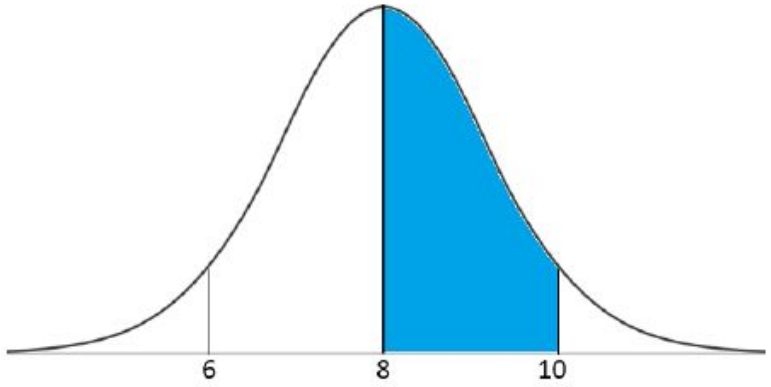
PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

- Exemplo 2: Existe um conjunto de objetos em uma cesta, cujos pesos são normalmente distribuídos com média = 8 e desvio padrão igual a 2.
  Qual a chance de se tirar um objeto pesando mais de 8 quilos?



PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

- Exemplo 3: Existe um conjunto de objetos em uma cesta, cujos pesos são normalmente distribuídos com média = 8 e desvio padrão igual a 2.
  Qual a chance de se tirar um objeto pesando mais de 8 e menos de 10 quilos?



DISTRIBUIÇÃO NORMAL PADRÃO (Z)

- Exemplo 4: Determinado atacadista efetua suas vendas por telefone. Após alguns meses, verificou-se que os pedidos se distribuem normalmente com média de 3.000 pedidos e desvio-padrão de 180 pedidos. Qual a probabilidade de que um mês selecionado ao acaso esta empresa venda menos de 2700 pedidos

PROBABILIDADE E DISTRIBUIÇÕES NORMAIS

ex 5: Considerando a distribuição da nossa última aula:

- [20, 32, 32, 36, 39, 43, 46, 48, 49, 50, 52, 53, 56, 57, 63, 64, 65, 74, 75, 90]

Se você escolhe um número, qual seria a chance de você escolher um número entre 39 e 63?

lembrando que:

μ = média = 52.2,

σ = desvio padrão = 16.32

REFERÊNCIAS

- FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de Análise de Dados:** Estatística e Modelagem Multivariada com Excel, SPSS e Stata. Rio de Janeiro: Ltc, 2020.
- CIFERRI, Cristina Dutra de Aguiar; CIFERRI, Ricardo Rodrigues. **Modelagem Multidimensional.** São Paulo: Usp, 2020. 14 slides, color. Disponível em: <http://wiki.icmc.usp.br/images/6/6a/SCC5911-02-ModelagemMultidimensional.pdf>. Acesso em: 10 jan. 2020.
- RESENDE, Tânia. **Modelagem multidimensional conceitos básicos.** São Paulo: Slideshare, 2016. 28 slides, color. Disponível em: <https://pt.slideshare.net/TANIARESENDE/modelagem-multidimensional-conceitos-bsicos>. Acesso em: 18 fev. 2020.
- JARDIM, Edgar Silveira; OLIVEIRA, Marcus Vinícius Abreu de; MORAVIA, Rodrigo Vitorino. Diferença Entre Banco de Dados Relacional e Banco de Dados Dimensional. **Revista Pensar Tecnologia**, Belo Horizonte, v. 2, n. 4, p.1-17, julho 2015. Mensal. Disponível em: <http://revistapensar.com.br/tecnologia/pasta_upload/artigos/a122.pdf>. Acesso em: 18 fev. 2020.
- SERGENTI, Alexsandro. **Modelagem Relacional e Multidimensional:** uma análise envolvendo Sistemas de Apoio a decisão. 2015. Disponível em: <https://www.linkedin.com/pulse/modelagem-relacional-e-multidimensional-uma-an%C3%A1lise-de-sergenti/>. Acesso em: 18 fev. 2020.