DATA SCIENCE

AULA 3 - ESTATÍSTICA

PROFª. ANA CAROLINA B. ALBERTON

INTRODUÇÃO

- Montgomery (2004) em seu livro Estatística Aplicada a Engenharia descreve que: "Engenheiros resolvem problemas de interesse da sociedade pela aplicação eficiente de princípios científicos".

- O campo da estatística lida com a coleta, a apresentação, a análise e o uso de dados para tomar decisões e resolver problemas.

- Métodos estatísticos são usados para nos ajudar a entender a **variabilidade**.

- Todos nós encontramos variabilidade em nosso dia-a-dia e o julgamento estatístico pode nos dar uma maneira útil para incorporar essa variabilidade em nossos processos de tomada de decisão.

EDA - Exploratory Data Analisys

Exploratory Data Analisys - Análise de Dados Exploratória

Busca obter informações ocultas sobre os dados:

- Variação
- Anomalias
- Distribuição
- Tendências
- Padrões
- Relações

A EDA faz parte do pipeline de qualquer processo de análise de dados, mesmo que informal

EDA - Exploratory Data Analisys

Elementos que compõem a EDA:

ESTATÍSTICA APLICADA

- Montgomery (2004) em seu livro Estatística Aplicada a Engenharia descreve que: "Engenheiros resolvem problemas de interesse da sociedade pela aplicação eficiente de princípios científicos".

- O campo da estatística lida com a coleta, a apresentação, a análise e o uso de dados para tomar decisões e resolver problemas.

- Métodos estatísticos são usados para nos ajudar a entender a **variabilidade**.

- Todos nós encontramos variabilidade em nosso dia-a-dia e o julgamento estatístico pode nos dar uma maneira útil para incorporar essa variabilidade em nossos processos de tomada de decisão.

ESTATÍSTICA APLICADA

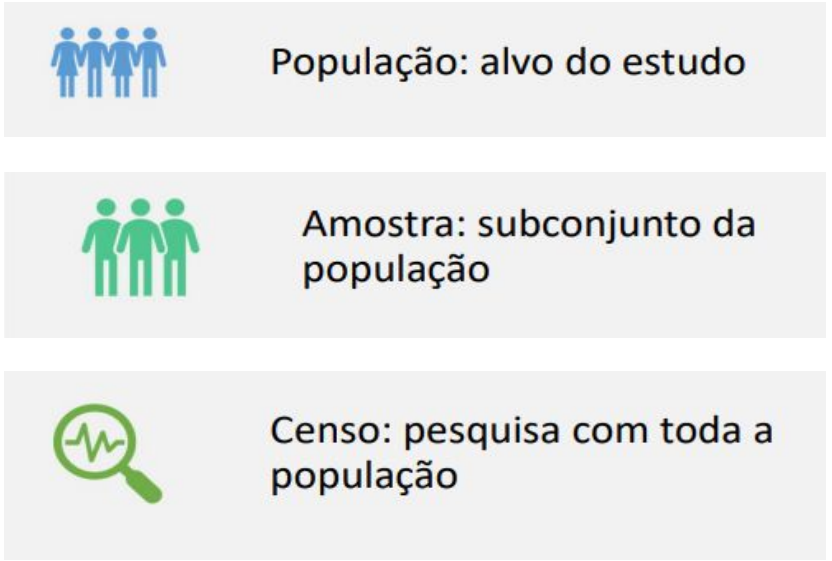
- Principais Divisões
  - Descritiva
  - Probabilística
  - Inferencial

ESTATÍSTICA APLICADA

- Descritiva
  - Organizar demonstrar e resumir dados
- Probabilidade
  - Analisar situações sujeitas ao acaso
- Inferência
  - Obter respostas sobre um fenômeno com dados representativos

AMOSTRAGEM

- Amplamente utilizada em pesquisas, estudos etc.
- Faz parte do dia a dia de um Cientista de Dados



PEQUENAS AMOSTRAS

Se eu jogar um dado 6x, qual a possivel média de resultados:

1+2+3+4+5+6 = 21/6 = 3,5

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
2.1666666666666665
```

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
3.3333333333333335
```

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
3.1666666666666665
```

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
3.0
```

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
4.333333333333333
```

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
3.6666666666666665
```

```
#Exemplo de pequenas amostras:
x = random.choices(range(1, 7), k=6)
print(np.mean(x))
```

```
3.6666666666666665
```

AMOSTRAGEM

Porque Amostra?

Pode ser caro ou impossível inferir sobre toda a população (censo)

AMOSTRAGEM

- É possível inferir sobre uma amostra
- Uma amostra feita corretamente deve representar as mesmas características da população de onde foi retirada.
- Se ela não representa a população, dizemos que ele é enviesada

AMOSTRAGEM

Enviesamento

Você subestima ou superestima o parâmetro da população

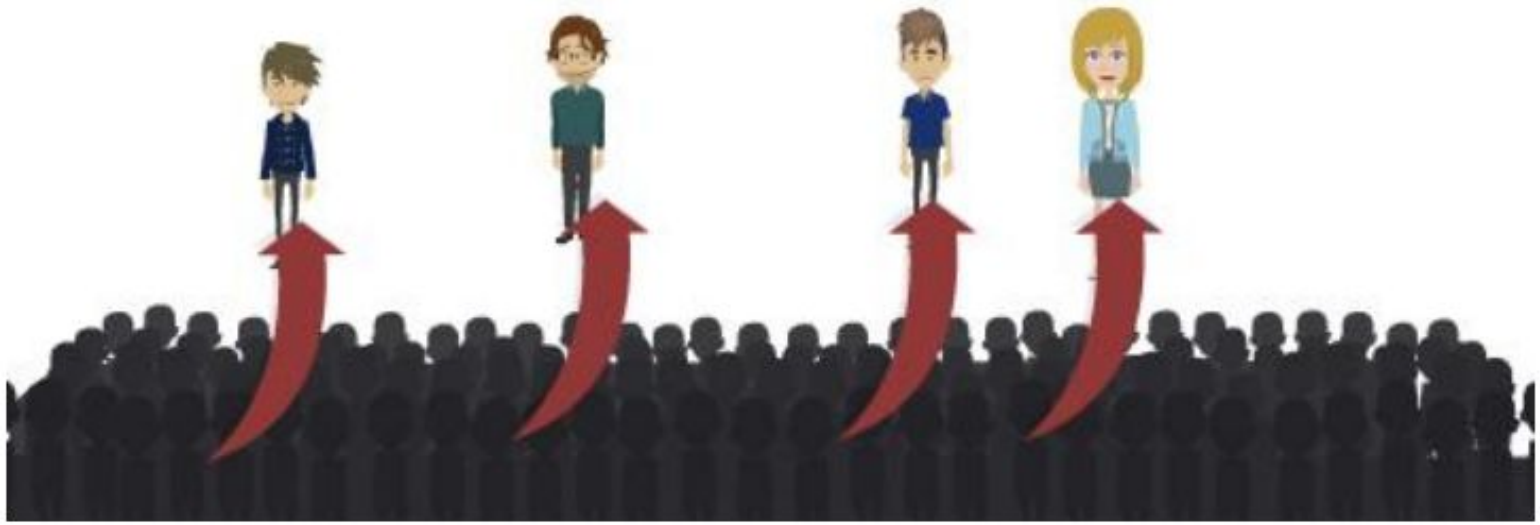
Causas:

- Pesquisa de pessoas próximas ou de fácil acesso
- Pesquisas pela Internet
- Sem uso de mecanismo de seleção aleatório

AMOSTRAGEM

- Principais tipos de amostras
  - Aleatória Simples
  - Estratificada
  - Sistemática

AMOSTRA



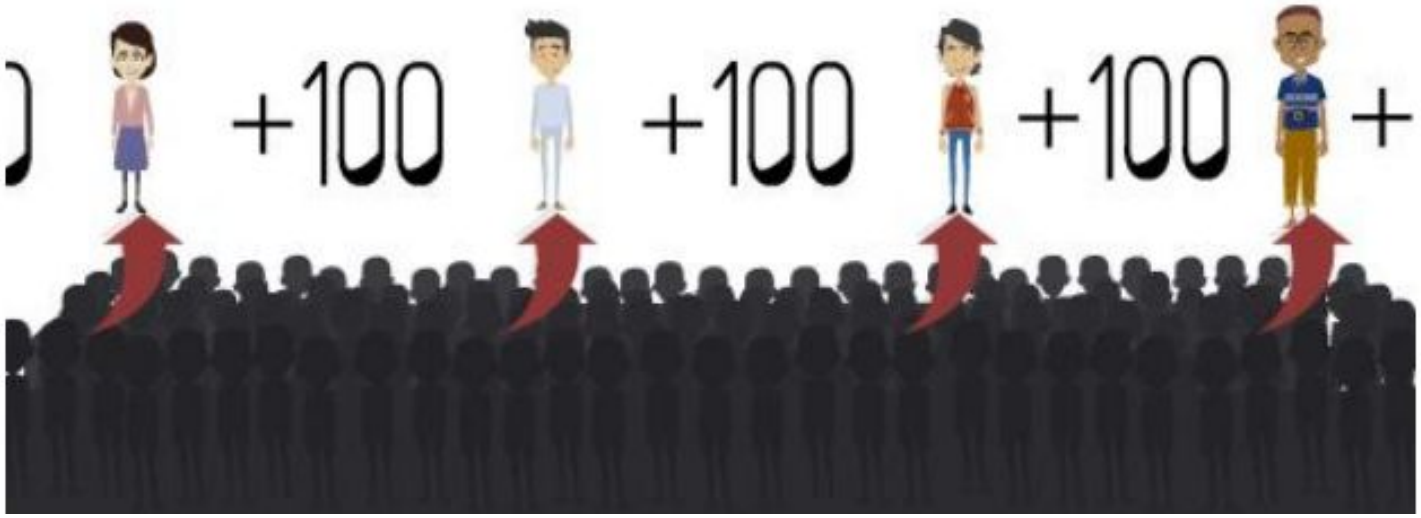
Amostras Aleatórias Simples

- Um determinado número de elementos é retirado da população de forma aleatória
- Todos os elementos da população alvo do processo de amostragem, devem ter as mesmas chances de serem selecionados para fazer parte da amostra

AMOSTRA

- As vezes as populações estão divididas nos chamados estratos.

AMOSTRA



- Amostra sistemática
  - Nesse tipo de amostragem é escolhido um elemento aleatório e a partir daí a cada N elementos um novo membro é escolhido

MÉDIA

- Quando falamos de variáveis quantitativas, os conjuntos de dados possuem medidas estatísticas que podem nos auxiliar a tomar decisões básicas sobre um comportamento ou uma tendência sobre os dados que possuímos;

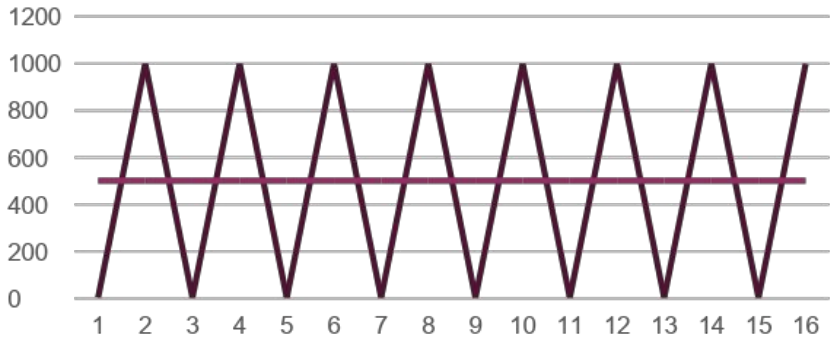
- A partir desse momento vamos chamar essas medidas de características dos seus conjuntos de dados e vamos falar sobre as características mais básicas da estatística;

- A característica mais básica de um conjunto de dados é a sua **média**. Porém estatisticamente o seu uso é mais raro, apesar de que no dia-a-dia é comum que utilizamos a média para parâmetros, como notas de escola ou até decisões estratégicas.

MÉDIA

- O seu uso é raro no planejamento de experimentos porque a média pode esconder grandes alterações entre as observações;

- Num exemplo hipotético, em um observação que sempre repete o padrão valor 1000 no tempo X e 2 no tempo Y e assim sucessivamente, traria ao seu observador uma média 501, que é muito longe de qualquer observação que ele vai ter ao longo do seu experimento.



MÉDIA

- Isso porque a média é um valor calculado, que pode não existir dentro da sua distribuição, um outro exemplo são as próprias notas de classe, é muito provável que a sua média não seja igual a nenhuma das notas que você tenha tirado ao longo do semestre;

- Portanto, se o seu objetivo é identificar como um fenômeno se comporta que sentido faz utilizar um recurso matemático, cujo o resultado irá retornar um valor que existe uma probabilidade alta de não acontecer;

- Talvez seja novidade para alguns, mas mesmo a média pode se subdividida, temos a média aritmética, a média ponderada, a média geométrica, entre outras.

MEDIANA

- Uma outra característica de dados muito importante é a **Mediana**. Que, segundo a literatura, é a medida de tendência central, que divide os dados em duas partes iguais, metade abaixo da mediana e metade acima.

- O seu cálculo é bem simples: No conjunto de dados conta-se a quantidade de registros apresentados, e caso a quantidade seja impar, pega-se exatamente o registro que divide o conjunto no meio para a ser a mediana:
  - Ex: 45 Registros, a mediana será o registro de número 23. Pois assim teremos 22 dois registros para um lado e para o outro.

MEDIANA

- No caso do conjunto de dados possuir uma quantidade par, pega-se o número do registro que corresponde a metade do conjunto e o registro superior para calcular a mediana.
- Com esses dois registros selecionados, eles são somados e divididos por 2, o resultado então é a mediana do conjunto.
    - Ex: Um conjunto com 44 registros, o registro de número 22 e 23 serão somados e divididos por dois, o seu resultado será a mediana do conjunto.

- Aqui pode parecer um contrassenso apontar que para achar a mediana, calcule-se a média de algo, sendo que falamos que a média pode nos levar à decisões erradas;
- Porém, para o cálculo da mediana (e de todas as medidas estatísticas que veremos ao longo dessa aula), se faz necessário que os dados numéricos da distribuição sejam ordenados em ordem crescente.

- Isso nos mostrará que mesmo utilizando uma média entre dois valores, nesse caso, o cálculo da media nos traz muito mais próximo da realidade da nossa distribuição, que é o nosso principal foco.

- Vejamos o seguinte exemplo ao lado, temos uma distribuição de 20 números que estão aparentemente desordenados;

- A sua média é 65.45, o que se a gente olhar atentamente para a distribuição podemos ver que ela não se relaciona muito com os dados;

- Porém agora vamos calcular a sua mediana.

130
53
300
32
50
74
5
200
50
52
4
32
75
46
48
56
10
43
0
49

65.45
Média

- Ordenamos nosso dado, só nessa situação já conseguimos visualizar um padrão;
- 8 dos nossos 20 casos ficam entre valores de 43 e 53, o que nos dá 40% de chance que, se estivermos em processo de repetição, no futuro teremos o mesmo padrão;
- Voltemos a mediana, no caso pegamos os 10° e 11° valores da distribuição (49, 50) e calculamos o seu valor;
- Agora eu lhes pergunto quais das duas características está mais próxima da realidade da distribuição?

130
53
300
32
50
74
5
200
50
52
4
32
75
46
48
56
10
43
0
49

65.45
Média

0
4
5
10
32
32
43
46
48
49
50
50
52
53
56
74
75
130
200
300

49.5
Mediana

MODA

- Ok, já sabemos que entre a média e a mediana, já saberíamos dizer o que mais nos seria útil se a gente fosse tentar encontrar uma característica para tentar emular futuros comportamentos do nosso fenômeno;

- Uma outra característica bem básica do conjunto de dados é a **moda**, que pode ser representada com o termo que aparece o maior número de vezes dentro do conjunto de dados.
- Exemplo, qual a moda?
  - {1,2,3,5,7,8,9}
  - {1,2,2,5,7,8,8}

EXEMPLO 1

- Apesar da semelhança entre moda, mediana e média, a apresentação desses valores podem variar consideravelmente dependendo do conjunto de dados. Vejamos o exemplo a seguir.

- Dado o conjunto ao lado, calcule:

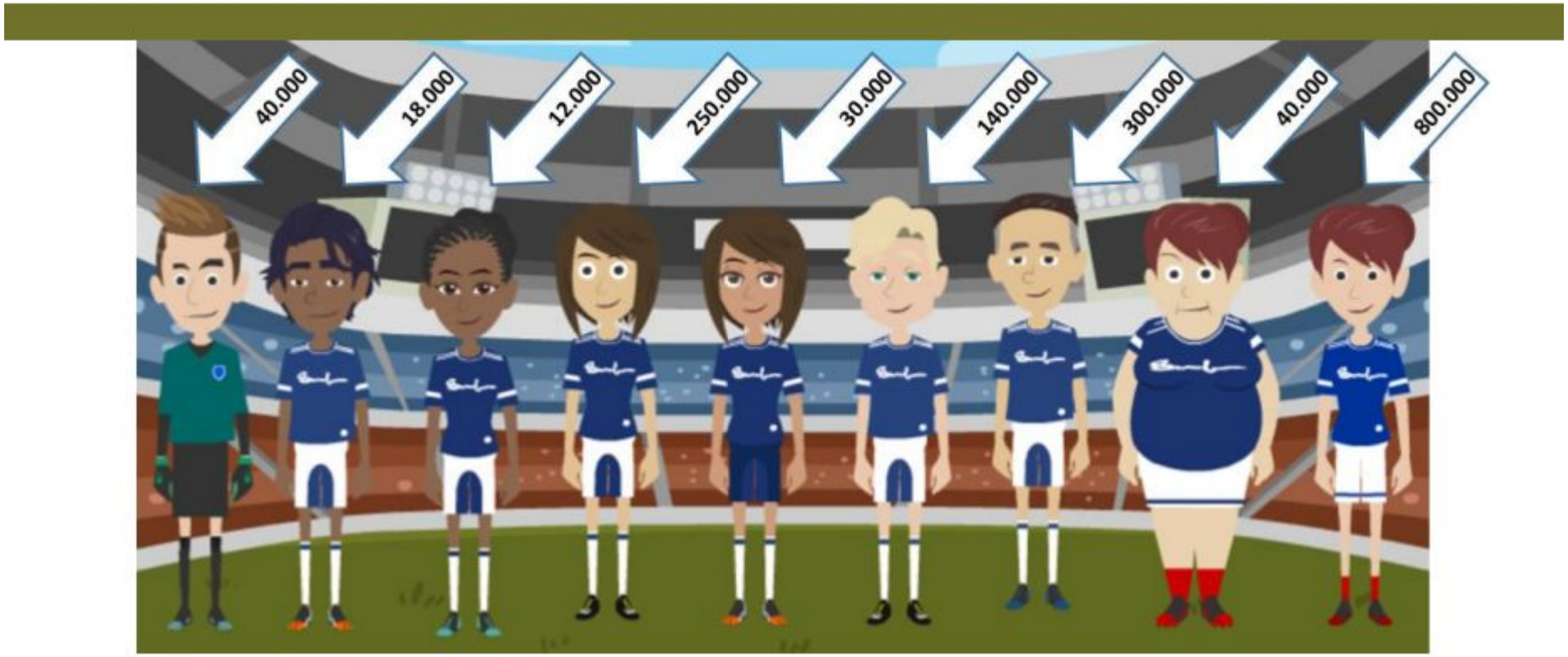
  - A média (Aritmética);
  - A moda;
  - A mediana;

- Apesar de ser as vezes bem sedutora a ideia de calibrar um modelo simplesmente ao utilizar o número que mais ocorre;

- Podemos ver pelo exemplo que nem sempre, isso é o mais correto a ser utilizado.

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

EXEMPLO 2: EM PYTHON



ESTATÍSTICA APLICADA

- Duas outras medidas de dispersão, que são mais usadas no contexto estatístico, são elas:
  - Variância;
  - Desvio Padrão.

- Ambas as medidas são utilizadas para medir a mesma coisa: O quanto as medidas do conjunto de dados se afastam da média aritmética. Porém uma acaba sendo meio que interpretada como uma evolução da outra.

- Começaremos pela **variância**. É uma medida mais básica nesse contexto. A fórmula para o cálculo da variância pode ser escrita com a fórmula a seguir.

VARIÂNCIA

- Onde:
  - V – Variância.
  - n – Número de registros
  - Xi – Registro de índice i
  - Ma – Média aritmética.

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

- Então vejamos, ainda utilizando o conjunto anterior (exemplo 1), calculamos a variância.

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

VARIÂNCIA

- Quanto maior for a variância, mais distante os pontos do conjunto de dados estão da média. O mesmo acontece com o inverso, quanto menor a variância, mais próximos serão os dados da sua média.

- Não existe uma escala de variância, o que pode ser considerado muito alto ou muito baixo e isso vai variar sempre do conjunto de dados. E o seu conhecimento sobre os dados (contexto), vai fazer a diferença quando formos determinar o que é uma variância alta ou baixa.

- No nosso resultado, tivemos uma variância de 266.36, o que alinhado com o nosso conhecimento dados, podemos caracterizar como alta, pois no nosso conjunto temos muitos pontos longe da média do conjunto.

DESVIO PADRÃO

- Porém, temos uma deficiência na variância que pode nos atrapalhar quando estamos realizando alguma análise estatística.
- Como as diferenças entre o valor representado e a média é elevada ao quadrado, os pontos mais longe da média acabam por elevar de forma as vezes tendenciosa a variância.
- Isso pode ser perigoso, principalmente quando temos outliers extremos, comprometendo então as nossas decisões.
- E ai é que entra o **desvio padrão**.

DESVIO PADRÃO

- 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.

- Sendo assim, a fórmula para o cálculo do desvio padrão é a seguinte:

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

- Basicamente, o cálculo do desvio padrão é a raiz quadrada da variância. Vamos ao cálculo então.

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

EXEMPLO 2

CALCULAR A VARIÂNCIA E O DESVIO PADRÃO DO SEGUINTE CONJUNTO:

[446, 458, 470, 482, 494, 506, 518, 530, 542, 554]

PROBABILIDADE

- Probabilidade (P): 0 <= P <= 1
- P=1 evento certo
- P=0 evento impossível
- Probabilidade: 0,5 ~½
- Impossível: -0,5 ou -20% ou 2/1

PROBABILIDADE - CONCEITOS

- Experimento: o que está sendo estudado
- Espaço Amostral: todas as possibilidades de ocorrência do evento
- Evento: resultados ocorridos

EXEMPLO

Jogar uma moeda

Cara ou coroa

Deu coroa

PROBABILIDADE

Eventos Excludentes:

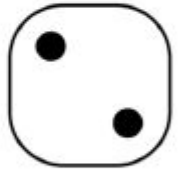
Quando não podem ocorrer ao mesmo tempo.

Exemplo: Jogar um dado e ser **1** e **par.**

PROBABILIDADE

Eventos Não Excludentes:

Quando podem ocorrer ao mesmo tempo.

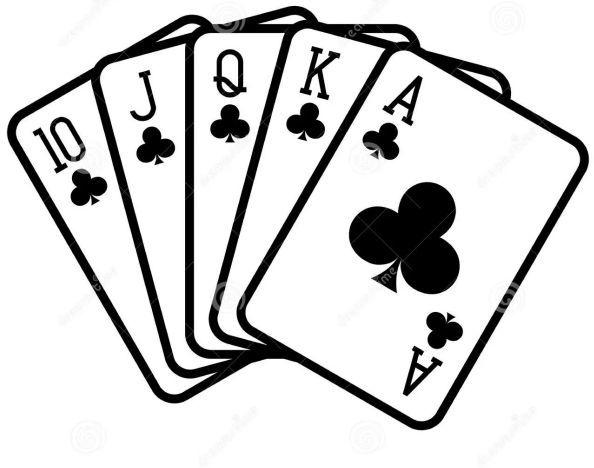


Exemplo: Jogar um dado e ser **2** e **par.**

PROBABILIDADE

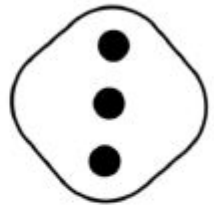
Eventos Dependentes:

A ocorrência de um evento afeta o outro.
Um tem que ocorrer para que depois o outro ocorra.

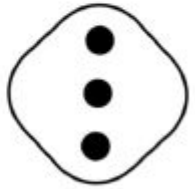


PROBABILIDADE

Eventos Não Dependentes:



A ocorrência de um evento não afeta o outro.



PROBABILIDADE

Um único evento

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo 1: Jogar uma moeda e dar cara:

P = ½

P = 0,5

PROBABILIDADE

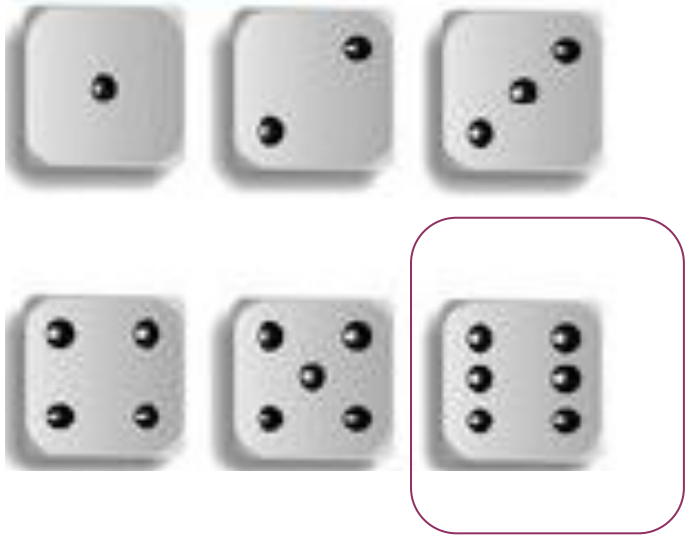
Um único evento

P = Ocorrência Esperada/Número de Eventos Possíveis

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

P = ⅙

P = 0,16



PROBABILIDADE

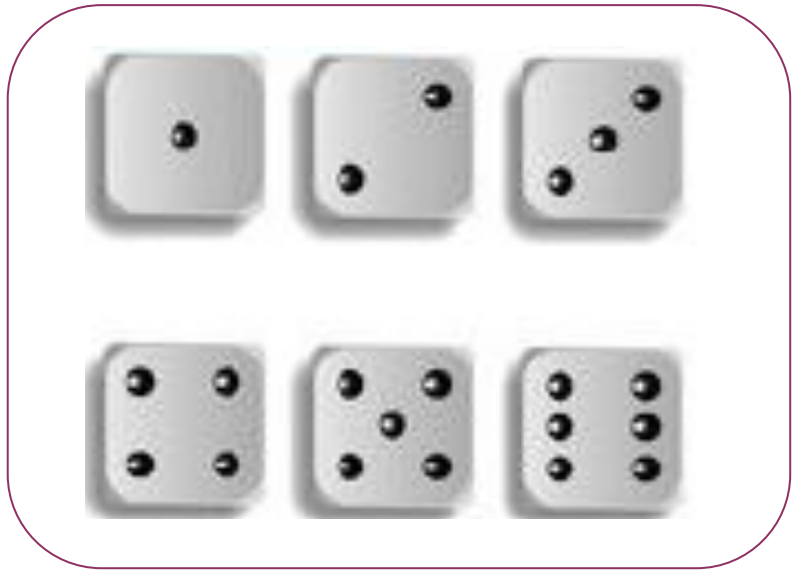
Um único evento

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo: Jogar um dado e dar **1, 2 ,3, 4, 5 ou 6**:

P = 6/6

P=1



PROBABILIDADE

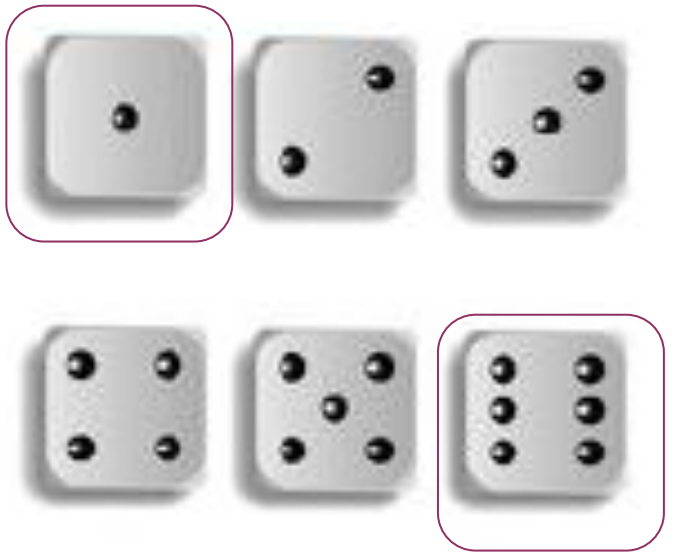
Um único evento

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo: Jogar um dado e dar **1 ou 6**:

P = 2/6

P = 0,33



PROBABILIDADE

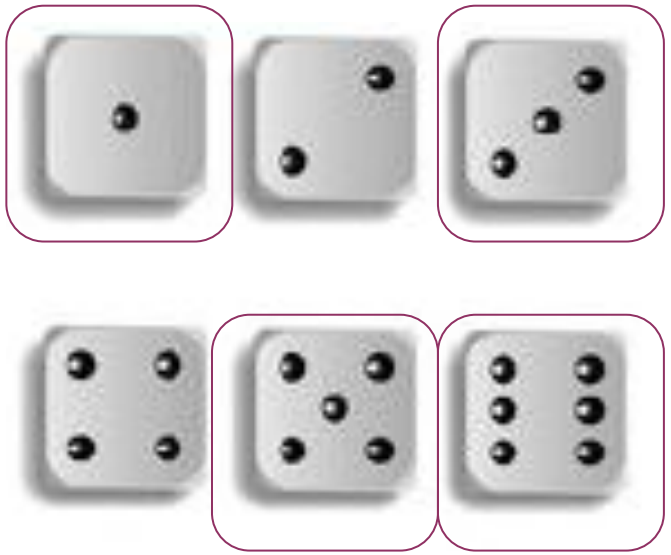
Um único evento

P = Ocorrência Esperada/Número de Eventos Possíveis

Exemplo: Jogar um dado e dar **ímpar ou maior que 4**:

P = 4/6

P = 0,67



PROBABILIDADE

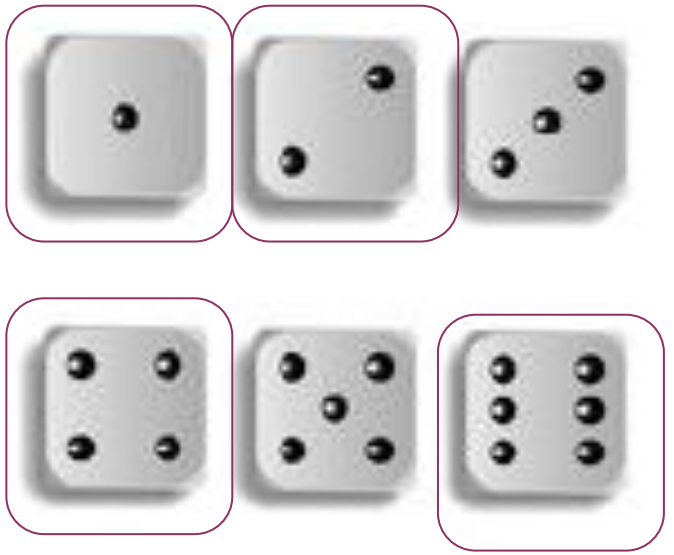
Eventos Excludentes

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



PROBABILIDADE

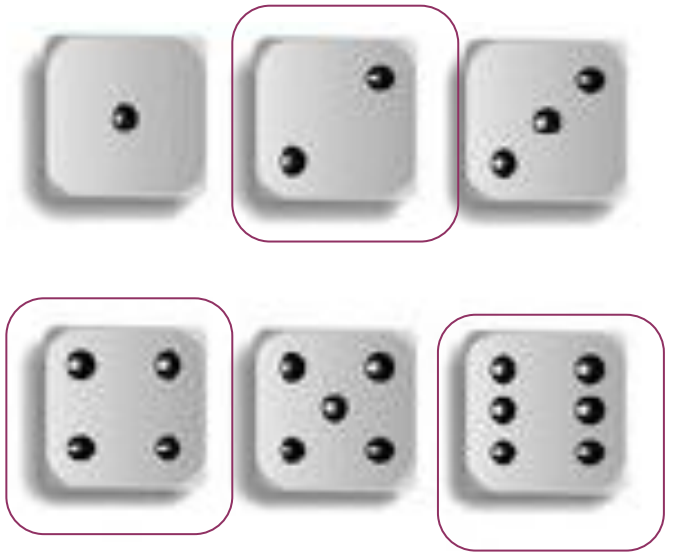
Eventos Não Excludentes

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



PROBABILIDADE

Eventos Independentes

Mais de um evento, eles se relacionam com Multiplicação

Exemplo: Qual a probabilidade de jogar dois dados e dar **1 e 6**?

(Dois eventos independentes)

$P = \frac{1}{6} * \frac{1}{6} = 0{,}028$

PROBABILIDADE

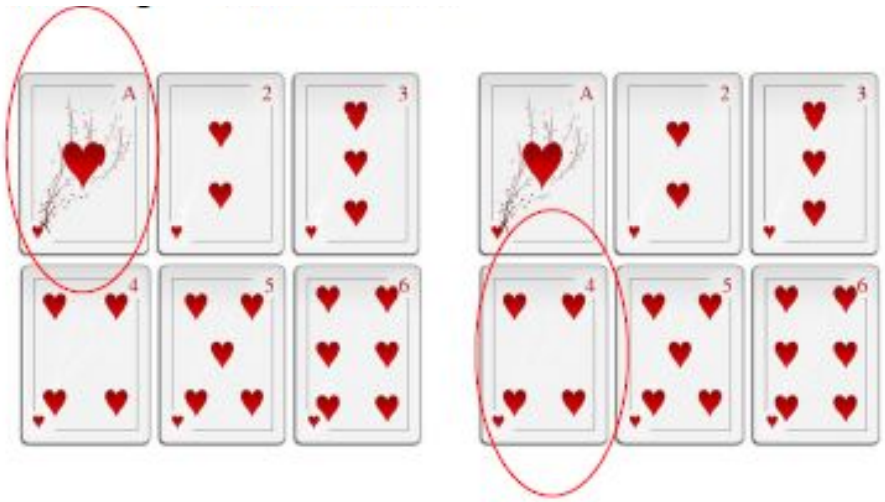
Eventos Dependentes

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



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.