ESTATÍSTICA EM ANÁLISE DE DADOS

Estatística Descritiva como um segmento da Matemática que objetiva descrever, sumarizar e representar graficamente o comportamento dos dados oriundos de ***variáveis*** de diversas áreas de aplicação.

- **POPULAÇÃO**: conjunto amplo de indivíduos de uma classe com suas características descritas por variáveis.

- **AMOSTRAGEM**: subconjunto da POPULAÇÃO

- **VARIÁVEIS**: atributos que descrevem o indivíduo da classe a ser analisada.
Por conta da natureza dos dados assumidos por uma variável, esta pode ser:

1. **QUANTITATIVA** (dados numéricos)
2. **QUALITATIVA** (não numéricos, literais ou categóricos)

- Para variáveis **quantitativas** temos:
a. **DISCRETAS** (para valores inteiros. Ex: idade, artigos publicados, falhas de acesso, e-mails em spam,...)
b. **CONTINUAS** (para valores reais. Ex: renda salarial, peso, taxa de juros,...)

- Para variáveis **qualitativas** temos:
c. **NOMINAIS** (quando descreve categorias sem ordenação qualquer. Ex: nome, sexo, UF, cidade,...)
d. **ORDINAIS** (quando a categoria pode ser ordenada. Ex: pouco, moderado, muito; Junior, Pleno, Senior,...)

Ilustração:
Classificar as variáveis da tabela seguinte:

Out[1]:

	nome	idade	sexo	peso	altura	salario	e_civil
0	Joao	20	M	50	1.70	1400	S
1	Maria	35	F	62	1.80	3000	C
2	Pedro	92	M	60	1.55	4800	S
3	Alice	20	F	50	1.49	1240	S
4	Amanda	38	F	65	1.70	2400	C
5	Sandro	27	M	57	1.63	2140	S
6	Clara	29	F	54	1.67	2000	C
7	Roberta	65	F	60	1.72	1500	C
8	Marcos	40	M	58	1.76	3500	C
9	Carol	45	F	70	1.85	1800	C
10	Cintia	20	F	64	1.58	1600	C
11	Jonas	19	M	70	1.95	1450	S
12	Silvia	40	F	57	1.68	2100	C
13	Ana	29	F	55	1.60	1300	C
14	Jonata	37	M	59	1.68	2300	C

Análise Univariada: consiste em classificar e graficar cada variável individualmente de modo a obter a informação desejada sobre a mesma. A **Distribuição de Frequência** é um dos recursos mais utilizados nesta etapa de conhecimento sobre a variável.

Análises sobre uma Variável Qualitativa Nominal ou Ordinal

Tomamos como exemplo a variável: *e_civil*. Com ela podemos:

- Obter uma tabela de frequências (absoluta e/ou relativa)

- Elaborar um gráfico (barras/setores). Para variáveis Ordinais, gráfico de setores não é interessante.

- Obter a **MODA**, ou seja, o valor que ocorre com a maior frequência.

e_civil	***f***	***fr***
S	4	0.2666...
C	11	0.7333...
∑ =	**15**	**1**

Notas:

$$fr_i = \frac{fi}{\sum fi}$$

Python: Contagem de ocorrências de um elemento em Lista

Counter(<lista>)

```
In [13]: idades = [19,12,21,45,67,19,30,51,17,19]
         print("Frequencia de cada idade = ",Counter(idades))
```

Análises sobre uma Variável Quantitativa Discreta:

Tomamos como exemplo a variável: *idade*. Com ela podemos:

- Obter as frequências (como no caso anterior)

Obs. Para um conjunto pequeno de dados podemos tratar cada idade, nesse caso, como uma categoria específica, assim como se fosse uma variável Qualitativa Ordinal, no entanto, quando se tornar inviável pelo volume de dados então essa categoria será sumarizada como uma varável Qualitativa Contínua (a seguir...)

- Elaborar gráficos de frequência

Análises sobre uma Variável Quantitativa Contínua:

Variáveis contínuas precisam ser organizadas em intervalos (CLASSES). Fica a critério do analista a definição dos intervalos ou pode-se recorrer a uma forma usual:

(AT) Amplitude Total da Classe = $\boldsymbol{max\ (x_i) - min\ (x_i)}$

após a ordenação dos dados brutos da classe.

Peso = {50, 50, 54, 55, 57, 57, 58, 59, 60, 60, 62, 64, 65, 70, 70}

Agora, a Distribuição de Frequência para a variável *Peso* será organizada na forma:

AT = (70 – 50) / k

k = √ *n* onde, *n = número de dados observados na variável*

logo, k = √ 15 ≈ 3,8729

Assim,

AT = 20 / 3,8729 ≈ 5,164 (Podemos arredondar AT para 5, nesse caso)

Nota: Também é possível utilizar a *Regra de Sturges* para determinar o número de classes numa Distribuição.

Número de Classes = 1 + 3,322 log n

Notação de Intervalos:

a |-- b **ou** **[a,b)**

a --|b **ou** **(a,b]**

Distribuição de Frequências: *Peso*

Classes	*f*	*fr*	*F*	*Fr*
50 |-- 55	3	0.2	3	0.2
55 |-- 60	5	0.3333	8	0.5333
60 |-- 65	4	0.2666	12	0.7999
65 |-- 70	1	0.0666	13	0.8665
70 |-- 75	2	0.1333	15	1
∑	**15**	**1**		

Distribuição de Frequências em Python

- análise de apenas um atributo

```
import pandas as pd
pessoas=pd.read_csv('pessoas.csv')
print(pessoas)
sorted(pessoas['idade'].unique())
#print("Menor Idade = ",pessoas['idade'].min()) e print("Maior Idade = ",pessoas['idade'].max())
pessoas['sexo'].value_counts()
pessoas['sexo'].value_counts(normalize=True)*100
freq=pessoas['sexo'].value_counts()
perc=pessoas['sexo'].value_counts(normalize=True)*100
tabela_freq=pd.DataFrame({'fi':freq,'%':perc})
print(tabela_freq)
```

```
       nome  idade sexo  peso  altura  salario e_civil
0      Joao     20    M    50    1.70     1400       S
1     Maria     35    F    62    1.80     3000       C
2     Pedro     92    M    60    1.55     4800       S
3     Alice     20    F    50    1.49     1240       S
4    Amanda     38    F    65    1.70     2400       C
5    Sandro     27    M    57    1.63     2140       S
6     Clara     29    F    54    1.67     2000       C
7   Roberta     65    F    60    1.72     1500       C
8    Marcos     40    M    58    1.76     3500       C
9     Carol     45    F    70    1.85     1800       C
10   Cintia     20    F    64    1.58     1600       C
11    Jonas     19    M    70    1.95     1450       S
12   Silvia     40    F    57    1.68     2100       C
13      Ana     29    F    55    1.60     1300       C
14   Jonata     37    M    59    1.68     2300       C
   fi     %
F   9  60.0
M   6  40.0
```

- análise cruzada (sexo X estado civil)

```
import pandas as pd
pessoas=pd.read_csv('pessoas.csv')
#print(pessoas)
sorted(pessoas['idade'].unique())
#print("Menor Idade = ",pessoas['idade'].min()) e print("Maior Idade = ",pessoas['idade'].max())
pessoas['sexo'].value_counts()
pessoas['sexo'].value_counts(normalize=True)*100
freq=pessoas['sexo'].value_counts()
perc=pessoas['sexo'].value_counts(normalize=True)*100
tabela_freq=pd.DataFrame({'fi':freq,'%':perc})
print(tabela_freq)
sexo={'F':'Feminino','M':'Masculino'}
estado_civil={'S':'Solteiro','C':'Casado'}
freq_sexo_estado_civil=pd.crosstab(pessoas.sexo,pessoas.e_civil)
freq_sexo_estado_civil.rename(index = sexo, inplace = True)
freq_sexo_estado_civil.rename(columns = estado_civil, inplace = True)
# print(freq_sexo_estado_civil)
perc_sexo_estado_civil=pd.crosstab(pessoas.sexo,pessoas.e_civil,normalize=True)*100
perc_sexo_estado_civil.rename(index = sexo, inplace = True)
perc_sexo_estado_civil.rename(columns = estado_civil, inplace = True)
print(perc_sexo_estado_civil)
```

```
   fi     %
F   9  60.0
M   6  40.0
e_civil       Casado   Solteiro
sexo
Feminino   53.333333   6.666667
Masculino  13.333333  26.666667
```

- análise cruzada (Média Salarial: sexo X estado civil)

```
import pandas as pd
import numpy as np
pessoas=pd.read_csv('pessoas.csv')
sexo={'F':'Feminino','M':'Masculino'}
estado_civil={'S':'Solteiro','C':'Casado'}
dados_cruzados=pd.crosstab(pessoas.sexo,pessoas.e_civil,aggfunc='mean',values=pessoas.salario)
dados_cruzados.rename(index = sexo, inplace = True)
dados_cruzados.rename(columns = estado_civil, inplace = True)
print(dados_cruzados)
```

```
e_civil    Casado  Solteiro
sexo
Feminino   1962.5    1240.0
Masculino  2900.0    2447.5
```

Outros parâmetros funcionais de Agregação:

- min
- max
- sum
- count

Possíveis Bibliotecas (*import*) para Python

```
from collections import Counter
import numpy
import statistics
import math
from scipy import stats
```

Listas em Python como Recurso para Funções Estatísticas

Ilustração:

idades = [19,12,21,45,67,32,30,51,17,26]

municipios_alto_vale = [‘Rio do Sul’,’Lontras’,’Imbuia’,’Salete’,’Aurora’,’Ituporanga’,’Ibirama’]

Algumas medidas de Tendência Central

Média Aritmética Simples

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Python com a biblioteca ***statistics*** podemos aplicar:

```
m = statistics.mean(idades)
print("Média aritmética das Idades é ", m)
```

- Python com ***pandas*** podemos aplicar:

```
import pandas as pd

df = pd.read_csv('pessoas.csv')
print(df['idade'].mean())
```

Média Harmônica

Tipo de média utilizada para dados com grandezas inversamente proporcionais (velocidade, vazão, densidade,...)

$$\bar{H} = \frac{n}{\sum_{i=1}^{n} \left(\frac{1}{x_i}\right)}$$

Em Python:

m_harmonica = statistics.harmonic_mean(lista_de_numeros)
print("Média harmônica é ", m_harmonica)

```
In [15]: import statistics
         A = [4, 7, 9, 10, 15]
         print("Média Harmônica é ",(statistics.harmonic_mean(A)))

         Média Harmônica é  7.455621301775148
```

Mediana

Valor que está no centro do conjunto de dados. Quando o número de observações na distribuição for ímpar, a mediana é o valor central, caso contrário será a média das duas observações mais centrais.

- Python com biblioteca ***statistics*** podemos aplicar:

```
In [1]: import statistics
        idades = [19,12,21,45,67,32,30,51,17,26]
        mediana = statistics.median(idades)
        print("Mediana é ", mediana)

        Mediana é  28.0
```

- Python com ***pandas*** podemos aplicar:

```
In [16]: import pandas as pd
         df=pd.read_csv('pessoas.csv')
         print(df['idade'].median())

         35.0
```

Quantil

É uma generalização da mediana. Consiste no valor abaixo do qual está um certo percentual dos dados. Para a mediana tipicamente, esse percentual é de 50% (q=0.5). Podemos estimar outro valor para *q*.

```
In [17]: import pandas as pd
         lista_idades =  pd.Series([19,12,21,45,67,32,30,51,17,26,90])
         print(lista_idades.quantile())

         30.0
```

```
In [20]: import pandas as pd
         lista_idades =  pd.Series([19,12,21,45,67,32,30,51,17,26,90])
         print(lista_idades.quantile(q=0.25))

         20.0
```

Moda

É o valor que possui maior frequência no conjunto de dados.

- Python com biblioteca ***statistics*** podemos aplicar:

```
import statistics
temperaturas_registradas = [10.5,12,13,9.5,12,15,15,8,5,-1,4,12,15]
print("Moda é ",statistics.mode(temperaturas_registradas))

Moda é  12
```

Problema: retorna a primeira ocorrência modal.

Então trocamos para a função ***multimod()***

```
In [33]: import statistics
         temperaturas_registradas = [10.5,12,13,9.5,12,15,15,8,5,-1,4,12,15]
         print("Moda é ",statistics.multimode(temperaturas_registradas))

         Moda é  [12, 15]
```

- Python com ***pandas*** podemos aplicar:

```
In [28]: import pandas as pd
         temperaturas_registradas = pd.Series([10.5,12,13,9.5,12,15,15,8,5,-1,4,12,15])
         print("Moda é ",temperaturas_registradas.mode())

         Moda é  0    12.0
         1    15.0
         dtype: float64
```

Medidas de Dispersão

Variância

Um indicador estatístico que mostra o quão os dados de um conjunto analisado estão afastados em relação à média.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

onde temos:

σ^2 = medida de variância ou variabilidade
x_i = *i-ésimo* elemento do conjunto
x = média dos elementos do conjunto
n = número de elementos do conjunto

Desvio Padrão

$$\sigma = \sqrt{\sigma^2}$$

Equação óbvia, no entanto, σ indica variação para dados em um conjunto com distribuição normal. Um valor baixo para σ mostra que os valores se aproximam mais da média, o contrário, indica maior dispersão.

Nota: Variância e Viés são medidas importantes em processos de *Learning Machine*, tendo em vista que ambas são consideradas como erros em relação a expectativa e aos resultados gerados. Métodos de Regressão, por exemplo, tendem a estimar uma função a partir da minimização dos erros de predição e aqueles alcançados de fato.

Ilustração:

	Venda Mensal 1o. Semestre (em R$ mil)						Indicadores				
Vendedor	*Jan*	*Fev*	*Mar*	*Abr*	*Mai*	*Jun*	Média	Variância	DP	Melhor Projeção	Pior Projeção
A	20	70	40	35	60	50	45,83	324,1667	18,00463	63,84	27,83
B	35	40	40	45	42	46	41,33	15,86667	3,983298	45,32	37,35
C	30	32	60	25	15	50	35,33	276,6667	16,6333	51,97	18,70
D	35	45	55	60	55	45	49,17	84,16667	9,174239	58,34	39,99
E	60	40	50	55	60	65	55,00	80	8,944272	63,94	46,06

Qual o vendedor com perfil mais confiável em termos de cumprimento de uma meta estimada?

Variabilidade e o Diagrama de Dispersão (*vamos a planilha!!!*)

Variância e Desvio Padrão em Python

```
import numpy as np
df=pd.read_csv('pessoas.csv')
print(df)
print("Variância do atributo IDADE = ",np.var(df['idade']))
print("Desvio Padrão do atributo IDADE = ",np.std(df['idade']))
fem = df[df.sexo == 'F']
mas = df[df.sexo == 'M']
peso_fem = fem['peso']
peso_mas = mas['peso']
print("Desvio Padrão do atributo PESO para Homens = ",np.std(peso_mas))
print("Desvio Padrão do atributo PESO para Mulheres = ",np.std(peso_fem))
```

```
       nome  idade sexo  peso  altura  salario e_civil
0      Joao     20    M    50    1.70     1400       S
1     Maria     35    F    62    1.80     3000       C
2     Pedro     92    M    60    1.55     4800       S
3     Alice     20    F    50    1.49     1240       S
4    Amanda     38    F    65    1.70     2400       C
5    Sandro     27    M    57    1.63     2140       S
6     Clara     29    F    54    1.67     2000       C
7   Roberta     65    F    60    1.72     1500       C
8    Marcos     40    M    58    1.76     3500       C
9     Carol     45    F    70    1.85     1800       C
10   Cintia     20    F    64    1.58     1600       C
11    Jonas     19    M    70    1.95     1450       S
12   Silvia     40    F    57    1.68     2100       C
13      Ana     29    F    55    1.60     1300       C
14   Jonata     37    M    59    1.68     2300       C
Variância do atributo IDADE =  354.3288888888889
Desvio Padrão do atributo IDADE =  18.823625816746596
Desvio Padrão do atributo PESO para Homens =  5.887840577551898
Desvio Padrão do atributo PESO para Mulheres =  5.90668171555645
```

Na sequência:

- Visualização

    - Histogramas

    - Gráficos de Barras / outros

    - Mapas

Exercício para hoje:

O arquivo ***COMBUSTÍVEIS.CSV*** registra os dados de uma pesquisa que coletou o preço da Gasolina, Etanol e Diesel em três cidades e com uma amostra de 5 postos em cada cidade. Você precisa analisar os dados para discutir sobre a Conformidade/Variabilidade dos preços dos combustíveis em cada cidade. O foco da análise é perceber se há certo acordo entre proprietários dos postos numa política de uniformidade ou se trabalham com autonomia e competitividade. Considerar a partir da amostra o quão vale a pena os usuários pesquisarem por preços nas referidas cidades. Verifique também, qual combustível tem maior discrepância de preço. Sugestão: visualize o arquivo numa forma mais estruturada para identificar os dados. Aplique consultas por meio das funções estatísticas e discuta os resultados.