Introdução à Ciência dos Dados
Ciência da Computação – UFV/Florestal

Aula Prática 02
Estatística Descritiva

Objetivo: avaliar dados numéricos com base em cálculos estatísticos
Pré-requisitos: linguagem de programação Python, Linux, estatística
Meta: ao final da prática, o aluno será capaz de utilizar ferramentas de análise de dados para calcular indicadores estatísticos e comparar valores.

Roteiro:

Ler dados de arquivo (arquivo *series.csv*)

```
1 data = pd.read_csv('series.csv', index_col=False, header=None, squeeze=True);
2 print(data)
```

```
0      13
1       3
2       5
3       6
4       7
5       9
6       5
7      33
8      67
9     432
10      5
11      7
12     35
13     67
14     83
15     57
16     88
Name: 0, dtype: int64
```

Explorar os dados com base em estatísticas descritivas

```
1 # Mínimo
2 data.min()
```

```
3
```

```
1 # Máximo
2 data.max()
```

```
432
```

```
1 # Média
2 data.mean()
```

```
54.235294117647058
```

```
1 # Desvio Padrão
2 data.std()
```

```
101.93780543287454
```

```
1 # Mediana
2 data.median()
```

```
13.0
```

```
# Moda
data.mode()
```

```
0    5
```

Visualização formatada das estatísticas (usar função *round* se quiser limitar casas decimais)

```
print('MIN: {}'.format(data.min()))
print('MAX: {}'.format(data.max()))
print('MÉDIA: {}'.format(data.mean()))
print('DESVIO PADRÃO: {}'.format(data.std()))
```

```
MIN: 3
MAX: 432
MÉDIA: 54.2352941176
DESVIO PADRÃO: 101.937805433
```

Calcular os percentis

```
# 25o percentil (1o quartil)
data.quantile(.25)
```

```
6.0
```

```
# 50o percentil (2o quartil)
data.quantile(.50)
```

```
13.0
```

```
# 75o percentil (3o quartil)
data.quantile(.75)
```

```
67.0
```

```
# 95o percentil
data.quantile(.95)
```

```
156.79999999999976
```

Calcular a tabela de frequências

```
1 # Tabela de Frequências
2 data.value_counts()
```

```
5      3
67     2
7      2
57     1
88     1
83     1
432    1
13     1
9      1
6      1
3      1
33     1
35     1
Name: 0, dtype: int64
```

Atividade (Entregar via PVANet o código fonte Python):

1. Faça um código para ler os arquivos *altura_homens.csv* e *altura_mulheres.csv*. Esses arquivos contém as alturas (em cm) de 1000 homens e 1000 mulheres, respectivamente. Em seguida, responda às seguintes perguntas:

a) Qual a altura mínima e máxima dos homens e das mulheres dessas amostras?

b) Qual a média de altura dos homens e das mulheres? E qual a mediana dessas alturas?

c) Qual o desvio padrão da altura dos homens e das mulheres?

d) Qual o percentual de homens com altura menor que 160cm?

e) Qual o percentual de mulheres com altura maior que 180cm?

f) Um homem com altura 185cm está em qual percentil? (pesquise sobre a função *percentileofscore* do pacote *scipy*)

g) Uma mulher com altura 150cm está em qual percentil?

h) Quais as três alturas de homens que são as mais frequentes? Quantos homens possuem essas alturas?

i) Quais as três alturas de mulheres que são as mais frequentes? Quantas mulheres possuem essas alturas?

j) Um homem com altura 185cm está distante quantos desvios padrões da média dos homens?

k) Um homem com altura 145cm está distante quantos desvios padrões da média dos homens?
l) Uma mulher com altura 185cm está distante quantos desvios padrões da média das mulheres?
m) Uma mulher com altura 145cm está distante quantos desvios padrões da média das mulheres?
n) É possível afirmar com determinado grau de confiança que uma pessoa com altura 150cm é um homem ou uma mulher?
o) E uma pessoa com altura 190cm?
p) E uma pessoa com altura 165cm?
q) As alturas dos homens e mulheres seguem uma distribuição Normal?

2. Para que serve a função "*describe()*" de uma Series?
3. Para que serve a função "*unique()*" de uma Series?

Referência:

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.html
(Seção Computations / Descriptive Stats)