FIAP

Statistics for Machine Learning

Aula 14: Medidas de posição e variabilidade

Prof. Jones Egydio

profjones.egydio@fiap.com.br

2024

Objetivos

- Introduzir os conceitos de medidas de posição e variabilidade
- Formas de representação;
- Exemplos e exercícios;
- Conclusão;
- Perguntas.

Medidas

- São as medidas mais utilizadas:
  - Mínimo – min()
  - Máximo – max()
  - Média – mean()

```
dados_paises.groupby('pais') \
            .agg(min_idh = pd.NamedAgg('idh', 'min'),
                 max_idh = pd.NamedAgg('idh', 'max'),
                 media_idh = pd.NamedAgg('idh', 'mean')) \
            .reset_index()
```

	pais	min_idh	max_idh	media_idh
0	Alemanha	0.8466	0.9050	0.882880
1	Austrália	0.8958	0.9290	0.914507
2	Brasil	0.6464	0.7180	0.684653

Moda

- É a observação mais frequente.
- No Python, a melhor estratégia seria o seguinte:

```
dados_paises.groupby('pais')['idh'] \
            .apply(lambda x: x.mode().iloc[0]) \
            .to_frame() \
            .reset_index()
```

	pais	idh
0	Alemanha	0.8466
1	Austrália	0.8958
2	Brasil	0.6464

O comando mode retorna um dataframe. Diante disto, precisamos considerar somente a coluna da modelo, no caso, utiliza-se o comando .iloc[0]

Mediana

- Ocupa a posição central numa sequência de valores ordenados.
- 50% das observações são menores de que a MEDIANA e 50% são maiores.

Posição	Valor
1	80
2	91
3	100
4	105
5	125

O valor 100 é a mediana, uma vez que divide os menores 50% dos maiores 50%.

Mediana

- Para estimar a mediana utiliza-se a função median.

```
dados_paises.groupby('pais') \
            .agg(median_idh = pd.NamedAgg('idh', 'median')) \
            .reset_index()
```

	pais	median_idh	max_idh	media_idh
0	Alemanha	0.8888	0.9050	0.882880
1	Austrália	0.9156	0.9290	0.914507
2	Brasil	0.6866	0.7180	0.684653

FIAP

Percentil

- O percentil divide os dados ordenados em 100 partes.

- O percentil p% é obtido de tal forma que p% das observações sejam menores que ele.

Exemplo:

  - Se calcularmos o percentil 5% da renda e obter o valor de R$1.000,00 então 5% da amostra possui renda menor de que R$1.000,00

Percentil

Posição	Valor	Percentil
1	1	10
2	1	20
3	2	30
4	2	40
5	3	50
6	4	60
7	4	70
8	7	80
9	8	90
10	9	100

Dividimos os dados em 100 partes.

O percentil 20% tem valor 1.

O percentil 50% tem valor 3.

O percentil 90% tem valor 8. Ou seja, 90% dos dados tem valor abaixo de 8.

Percentil

- Em alguns casos não conseguimos dividir as 100 partes diretamente nos dados.
- Diante disto, podemos realizar algumas interpolações.
- Neste caso, podemos utilizar a seguinte fórmula:
- $k = \frac{p \cdot (n+1)}{100}$, *k* é a posição do respectivo percentil *p*.
- Exemplo: o percentil 35% nos dados do slide anterior seria:
- $k = \frac{35 \cdot (10+1)}{100} = 3{,}85$. Neste caso, o percentil 35 está entre as posições 3 e 4.
- Deste modo, fazemos a média ponderada: $\frac{(2*3)+(2*0{,}85)}{3{,}85} = 2$.
- Assim, percentil 35% seria igual a 2.

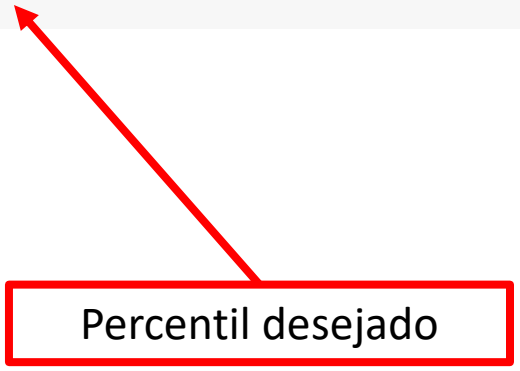
Percentil

- No python temos a função quantile do módulo numpy.

```
dados = [1, 1, 1, 2, 2, 3, 4, 4, 7, 8, 9]
```

```
np.quantile(dados, 0.75)
```

```
5.5
```



Percentil

- Para estimar o percentil em um pandas dataframe pode-se utilizar a função quantile.

```
dados_paises.groupby('pais')['idh'] \
            .apply(lambda x: x.quantile([0.05, 0.95])) \
            .to_frame() \
            .reset_index() \
            .rename(columns={'level_1': 'percentil'})
```

	pais	percentil	idh
0	Alemanha	0.05	0.85066
1	Alemanha	0.95	0.90360
2	Austrália	0.05	0.89818
3	Austrália	0.95	0.92760
4	Brasil	0.05	0.65074
5	Brasil	0.95	0.71590

Quartil

- Os quartis são pontos específicos dos percentis.
- O primeiro quartil é o percentil 25%
- O segundo quartil é o percentil 50%
- O terceiro quartil é o percentil 75%

Exemplo:

- Se calcularmos o primeiro quartil da renda e obter o valor de R$2.000,00 então 25% da amostra possui renda menor de que R$2.000,00

Quartil

Para estimar os quartis também utiliza-se a função quantile, porém utilizando os respectivos valores de 25%, 50% e 75%.

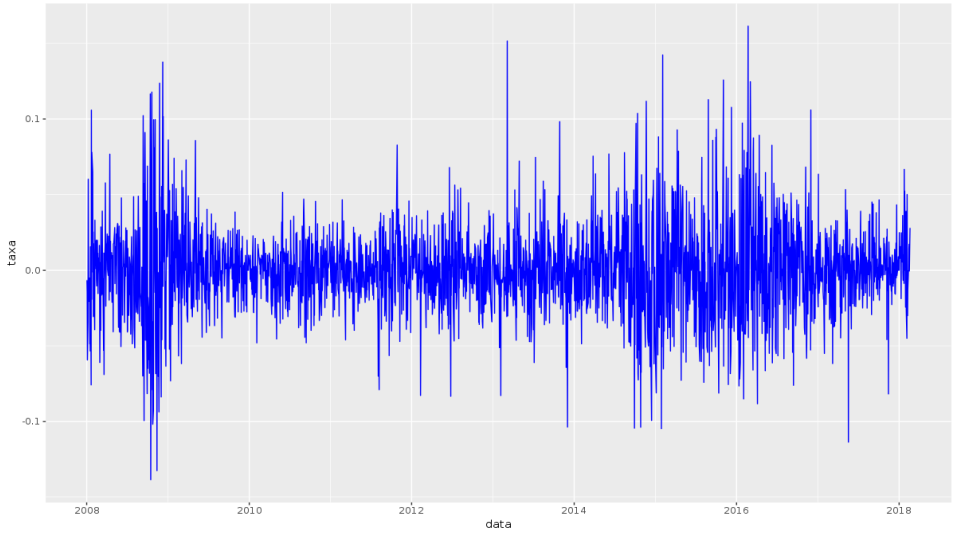
```
dados_paises.groupby('pais')['idh'] \
            .apply(lambda x: x.quantile([0.25, 0.5, 0.75])) \
            .to_frame() \
            .reset_index() \
            .rename(columns={'level_1': 'percentil'})
```

	pais	percentil	idh
0	Alemanha	0.25	0.8671
1	Alemanha	0.50	0.8888
2	Alemanha	0.75	0.9005
3	Austrália	0.25	0.9072
4	Austrália	0.50	0.9156
5	Austrália	0.75	0.9230

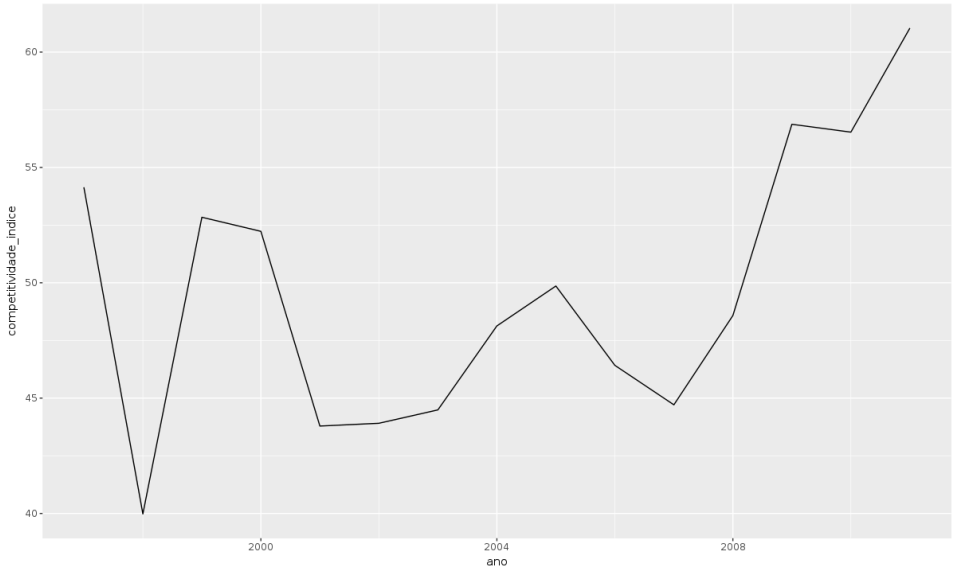
Conceito de variabilidade

A variabilidade pode ser considerada como a diferença entre o que é esperado e o que é realmente ocorre

Variação das ações da Petrobras



Índice competitividade Brasil



Medidas de variabilidade

- Medidas de variabilidade:
  - Desvio
  - Desvio Médio
  - Desvio Médio Absoluto
  - Variância
  - Desvio Padrão

Desvio

- O desvio é simplesmente a diferença entre um determinado valor e seu valor esperado.
- No exemplo abaixo, o valor esperado foi a média:

```
media_paises = dados_paises.groupby('pais') \
                           .agg(media_idh = pd.NamedAgg('idh', 'mean')) \
                           .reset_index()
```

```
dados_paises = dados_paises.merge(media_paises, on = 'pais', how = 'left')
```

```
dados_paises['desvio_idh'] = dados_paises['idh'] - dados_paises['media_idh']
```

Cálculo do desvio

```
dados_paises[['pais', 'ano', 'idh', 'desvio_idh']]
```

	pais	ano	idh	desvio_idh
0	África do Sul	1997	0.6328	0.02056
1	África do Sul	1998	0.6272	0.01496
2	África do Sul	1999	0.6216	0.00936

Desvio Médio

- O desvio médio é simplesmente a média dos desvios.
- Porém, esta métrica não é utilizada pelo problema abaixo:
  - Devido a este problema utilizam-se as próximas métricas

```
dados_paises.groupby('pais') \
            .agg(desvio_medio_idh = pd.NamedAgg('desvio_idh', 'mean')) \
            .reset_index()
```

	pais	desvio_medio_idh
0	Alemanha	1.258253e-16
1	Austrália	1.110223e-16
2	Brasil	-2.220446e-17

Valores de desvio médio são todos aproximadamente iguais a **zero**.

Desvio Médio Absoluto

- Uma possível métrica para medir variabilidade é o desvio médio absoluto.
- A métrica calcula a média dos desvios desconsiderando os sinais.
- Desvio médio absoluto:

$$MAD = \frac{\sum_{i=1}^{n} |x - \bar{x}|}{n}$$

em que $\bar{x}$ é a média.

Desvio Médio Absoluto

- No pandas utilizamos a função mad.

```
dados_paises.groupby('pais') \
            .agg(desvio_medio_abs_idh = pd.NamedAgg('idh', 'mad')) \
            .reset_index()
```

	pais	desvio_medio_abs_idh
0	Alemanha	0.017301
1	Austrália	0.008740
2	Brasil	0.018983

```
def desvio_medio_absoluto(x):
    return np.mean(np.abs(x - np.mean(x)))
```

```
[43]
import numpy as np

# Aplicar o GroupBy e calcular o desvio médio absoluto
desvio_medio_abs = dados_paises.groupby('pais') \
                               .agg(desvio_medio_abs_idh=('idh', desvio_medio_absoluto)) \
                               .reset_index()

# Exibir o resultado
print(desvio_medio_abs)
```

FIAP

Variância e Desvio-Padrão

- São as principais métricas para mensurar variabilidade.
- Ambas medem os desvios médios em relação a média.
- Variância: $s^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n-1}$
- Desvio-Padrão: $s = \sqrt{s^2}$
- O desvio-padrão é somente a raiz da variância.

FIAP

Variância e Desvio-Padrão

- As funções para variância e desvio-padrão no python

```
dados_paises.groupby('pais') \
            .agg(variancia_idh = pd.NamedAgg('idh', 'var'),
                 dp_idh = pd.NamedAgg('idh', 'std')) \
            .reset_index()
```

	pais	variancia_idh	dp_idh
0	Alemanha	0.000408	0.020187
1	Austrália	0.000110	0.010485
2	Brasil	0.000517	0.022727

FIAP

Exercício Complementar

- Aplicar as métricas apresentadas nas variáveis PIB, Corrupção índice, Competitividade índice e Globalização índice.

- Resolver as perguntas presentes no arquivo “analise_projetos.ipynb”

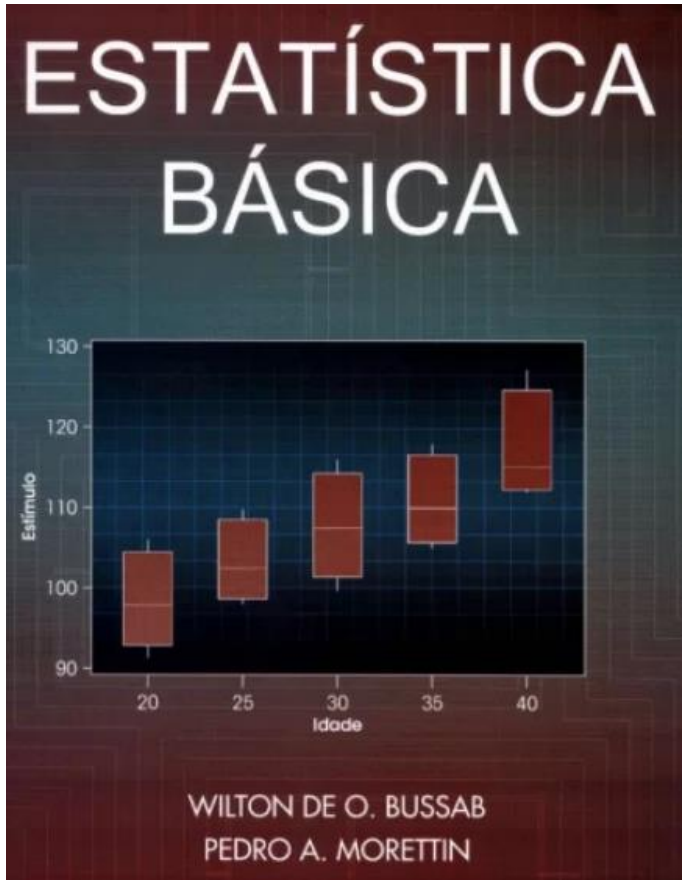
Material extra

Média e mediana

https://pt.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/measuring-center-quantitative/v/statistics-intro-mean-median-and-mode

Percentil e quartil

https://pt.khanacademy.org/math/ap-statistics/density-curves-normal-distribution-ap/percentiles-cumulative-relative-frequency/v/calculating-percentile



Referências bibliográficas

- BUSSAB, W., MORETITIN, P., Estatística Básica, Editora Saravia, 2014.

Obrigado!