Estatística Descritiva

Medidas Resumo

- Muitas vezes os dados a serem analisados serão dispostos em tabelas, sendo uma linha para cada observação;

- Para facilitar a uma visualização/generalização das informações, utilizam-se medidas que possam resumir as informações dispostas em colunas

- Dentre as medidas mais utilizadas estão as de centralidade (também chamadas de medidas de posição) e medidas de dispersão.

idp

Medidas de centralidade (de posição)

- Como o nome sugere, estas medidas expressam a ideia de valores centrais de um determinado conjunto. Destacam-se **moda, mediana e média.**

- A **moda** é o valor que mais se repete em um conjunto de valores observados. Um conjunto pode possuir mais de uma moda.

- A **mediana** é o valor que ocupa a posição central do conjunto, uma vez organizado em ordem crescente

Média

A **média aritmética** é a soma dos valores observados dividida pela quantidade de observações do conjunto. Para variáveis discretas:

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

lê se x barra.

No Python...

- Na biblioteca **Pandas:** nome_variável**.mean( )**
- Na biblioteca **Numpy: np.average(**nome_variável**)**
- Na biblioteca **Scipy: sc.mean(**nome_variável**)**
- Na biblioteca **Statistics: st.mean(**nome_variável**)**

idp

Média Aparada

Uma variação da média é a chamada **média aparada**

Neste caso são excluídos uma certa quantidade de valores em cada um dos extremos dos dados (inferior e superior), diminuindo sua influência nos resultados

$$\bar{x}_{aparada} = \frac{1}{n-2p} \sum_{i=p+1}^{n-p} x_i$$

No Python:

- Biblioteca **Scipy: trim_mean**(nome_variável, % dos dados retirados)

Média Ponderada

É o cálculo da média no qual cada valor do conjunto é multiplicado por um peso.

Em geral, o peso é um valor entre 0 e 1, que somado dá 1.

Essa média é importante para reduzir a variabilidade dos dados (em dados altamente variáveis os pesos são menores) e/ou para dar corrigir diferenças entre dados de grupos de dados diferentes.

$$\bar{x}_{geométrica} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

idp

Mediana

Como dito anteriormente, a **mediana** é o valor central em um conjunto de dados ordenados:

1. Se o número de observações foi ímpar, então a mediana será o valor que está no meio dos dados.
2. Se o número de observações for par, toma-se a média dois valores centrais em dois.

Vantagem em relação à média: como ela não usa todos os valores do conjunto de dados em seu cálculo, então ela não é influenciada por valores extremos.

$$\tilde{X}\begin{cases} \left(\frac{X_n+1}{2}\right), & \text{se } n \text{ é ímpar} \\ \left(\frac{X_n}{2}\right)+\left(\frac{X_n+1}{2}\right), & \text{se } n \text{ é par} \end{cases}$$

idp

Mediana

No Python:

- Na biblioteca **Pandas:** nome_variável**.median( )**
- Na biblioteca **Numpy: np.median(**nome_variável**)**
- Na biblioteca **Scipy: sc.median(**nome_variável**)**
- Na biblioteca **Statistics: st.median(**nome_variável**)**

idp

Moda

A moda é o valor que mais se repete em um conjunto de dados.

Quando existe apenas um valor que se repete com frequência diz-se que os dados são **unimodais**.

Entretanto isso não significa que outros valores não se repitam frequentemente, podem haver outros e dizemos que os dados são ou **bimodais**, ou **trimodais**, ou **multimodais**.

Moda

No Python:

- Na biblioteca **Pandas:** nome_variável**.mode( )[0]**
- Na biblioteca **Statistics: st.mode(**nome_variável**)**
- Na biblioteca **Scipy: mode(**nome_variável**)[0][0]**

Quartis, decis e percentis

É possível agrupar ainda os dados em partes iguais e verificar a distribuição dos dados nessas partes:

1. **Quartis ($Q_i, com\ i = 1, 2, 3, 4$):** divide o conjunto de dados em 4 partes iguais. É bastante utilizado no conceito do boxplot de dados. Nessa estratégia os dados são agrupados em 4 categorias 1º Quartil (1º quarto dos dados e assim sucessivamente), 2º quartil, 3º Quartil e 4º Quartil. Esse conceito é utilizado na plotagem do boxplot;
2. **Decil ($D_i, com\ i = 1, 2, \dots, 9, 10$) :** nesta abordagem os dados se dividem em 10 grupos
3. **Percentil ($P_i, com\ i = 1, 2, \dots, 99, 100$) :** nesta abordagem os dados se dividem em 100 grupos

idp

Calculando os quantis

No Python:

- Na biblioteca **Pandas:** nome_variável**.quantile(quartis/decis/percentis)**
- Na biblioteca **Scipy: mquantiles(**nome_variável**,quartis/decis/percentis)**
- Na biblioteca **Numpy: np.percentile(**nome_variável**,np.multiply(quartis))**

idp

Exemplo de medidas de centralidade

Para entendermos essas medidas, sejam os dados referentes a notas de diferentes agrupamentos de estudantes em uma disciplina:

1. Grupo A: 3, 4, 5, 6, 7
2. Grupo B: 1, 3, 5, 7, 9
3. Grupo C: 5, 5, 5, 5, 5
4. Grupo D: 3, 5, 5, 5, 7
5. Grupo E: 3, 5, 5, 6, 6

Determine a média, a moda e a mediana de cada grupo.

Exemplo de medidas de centralidade Solução

	Média	Moda	Mediana
Grupo A	5	-	5
Grupo B	5	-	5
Grupo C	5	5	5
Grupo D	5	5	5
Grupo E	5	6 e 5	5

Pergunta-se: a média, ou outra medida central é um bom descritor para os conjuntos, mesmo com todas as diferenças entre os valores apresentados?

Medidas de dispersão

- Uma vez que utilizamos medidas de centralidade para representar os dados, é fácil ver que os valores observados do conjunto ficarão ora acima, ora abaixo da média.

- Essa dispersão é chamada de **variação dos dados** ou **dispersão dos dados**.

- As medidas de dispersão mais comum são: amplitude total, desvio-médio, desvio padrão e variância

Medidas de dispersão
Amplitude total

Definição: a amplitude total de um conjunto é a diferença entre o valor mais alto e mais baixo do conjunto.

Esta medida dá uma ideia, ainda que muito vaga, sobre o quão espalhados estão os dados.

Uma vez que se conheça a média e os limites que geraram a amplitude total, pode-se ter uma ideia de valores extremos (outliers)

Medidas de dispersão
Desvio Médio (ou desvio absoluto médio)

Da mesma forma que utiliza-se uma medida de centralidade para exprimir o conjunto, utiliza-se também uma medida de centralidade para verificar o quão dispersos estão os dados.

Assim a média torna-se a referência para estudar a variabilidade do conjunto de dados.

Definição: o desvio médio de um conjunto de N observações é dado por

$$dm = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

Onde $\bar{x}$ é a média aritmética das observações

Medidas de dispersão
Desvio Médio – Um exemplo

Voltemos ao exemplo das notas

1. Grupo A: 3, 4, 5, 6, 7
2. Grupo B: 1, 3, 5, 7, 9
3. Grupo C: 5, 5, 5, 5, 5
4. Grupo D: 3, 5, 5, 5, 7
5. Grupo E: 3, 5, 5, 6, 6

a) Calcule o somatório dos desvios de cada nota, sem considerar o módulo.

b) Calcule o desvio médio dos conjuntos

Medidas de dispersão
Variância

A variância considera ao invés do módulo a soma dos quadrados dos desvios dividido pelo total de observações.

Definição: a variância de um conjunto de N observações é dado por

$$var = s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

Onde $\bar{x}$ é a média aritmética das observações

Percebam que tal qual o desvio médio, que utiliza o módulo, a utilização da diferença ao quadrado evita que o somatório dos desvios fique igual a zero.

Alguns autores indicam ao desvio padrão por $\boldsymbol{\sigma}^2$.

Medidas de dispersão
Desvio Padrão

Percebam que a variância não apresenta a mesma unidade de medida da variável estudada. Por exemplo, se a unidade de medida da variável fosse metros, a variância estaria em metros quadrados...

A fim de evitar essa discrepância, utiliza-se o desvio-padrão, que nada mais é do que a raiz quadrada da variância.

Definição: o desvio padrão de um conjunto de N observações é dado por

$$dp = \sqrt{var} = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Onde $\bar{x}$ é a média aritmética das observações. Alguns autores indicam ao desvio padrão por $\boldsymbol{\sigma}$.

Coeficiente de Variação

- Este coeficiente mede o grau de dispersão dos dados em relação à média;
- Uma vantagem desse medida é ser adimensional, podendo ser expressa como porcentagem;
- Sua fórmula é:

$$CV = \frac{\sigma}{\bar{X}} = \frac{desvio\ padrão}{média}$$

No Scipy: variation(nome_da_variável [‘coluna_da_variável’])

idp

Medidas de dispersão
Variância e Desvio Padrão – Um exemplo

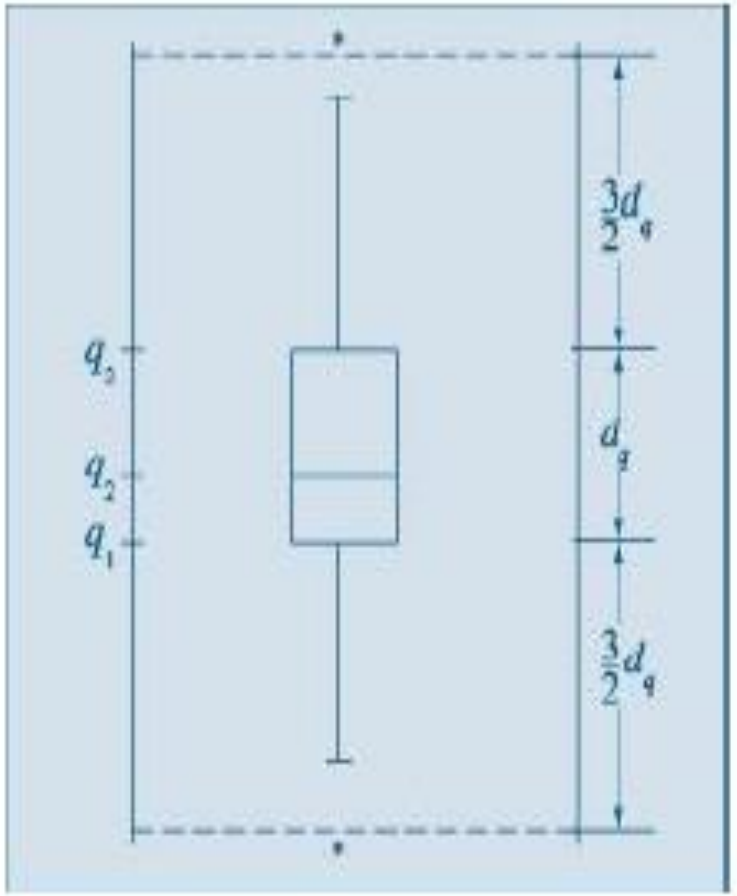
Voltemos ao exemplo das notas. Calcule a variância e o desvio padrão dos conjuntos

1. Grupo A: 3, 4, 5, 6, 7
2. Grupo B: 1, 3, 5, 7, 9
3. Grupo C: 5, 5, 5, 5, 5
4. Grupo D: 3, 5, 5, 5, 7
5. Grupo E: 3, 5, 5, 6, 6

Boxplot

A construção de um boxplot deve considerar a mediana e os quartis.

- A diferença interquartílica (d_q) é dada por $Q_3 - Q_1$.
- O limite inferior é dado por $LI = Q_1 - (1{,}5)d_q$
- O limite superior é dado por $LS = Q_3 + (1{,}5)d_q$



Intervalo Interquartil

O **intervalo interquartílico** ou **Amplitude interquartílica** é a diferença entre o 3º quartil (75º percentil) e o 1º quartil (25º percentil).

Essa medida difere da variância e do desvio padrão, pois ela leva em consideração o ordenamento dos dados.

Boxplot

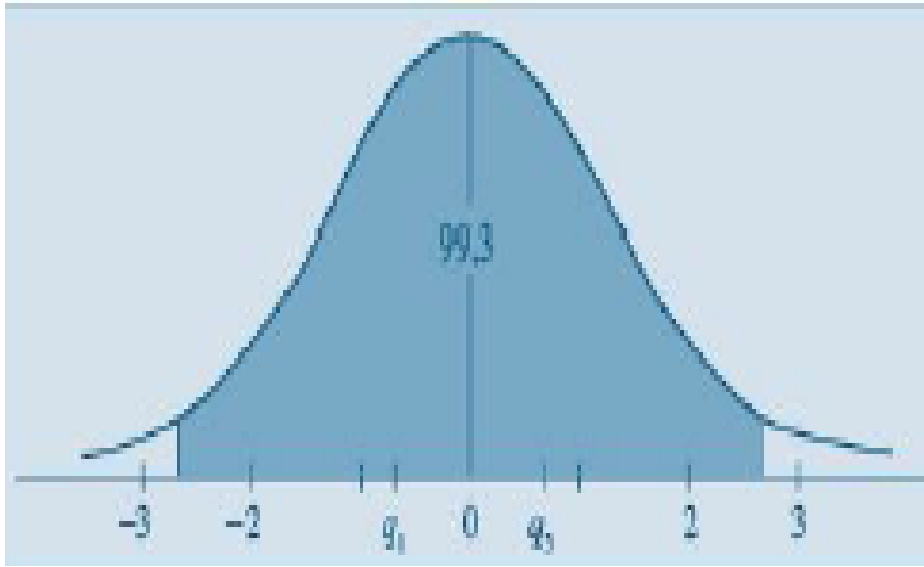
O boxplot dá a ideia de:

- Posição dos dados;
- Dispersão dos dados;
- Assimetria;
- Cauda (dados nas extremidades)
- Dados Discrepantes.

Boxplot

A justificativa para os limites do boxplot está no entendimento da distribuição de probabilidades pela curva normal:

- No exemplo ao lado, com média e mediana zero;
- Assim, $Q_1 = -0{,}6745$, $Q_3 = 0{,}6745$ e $d_q = 1{,}349$.
- Nesse caso $LI = -2{,}698$ e $LS = 2{,}698$, com área sob a curva de 0,993, ou seja, 99,3%.



Gráficos de Assimetria e Curtose

Essas medidas estão relacionadas a distribuição dos dados de uma variável.

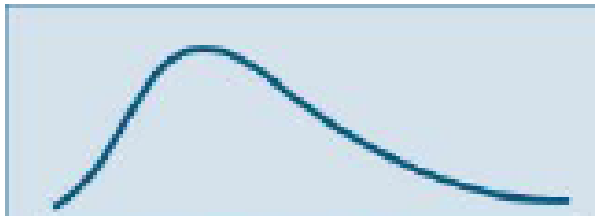
O estudo de distribuições são importantes para realizar a modelagem de populações.

Alguns coeficientes expressam o comportamento das distribuições.

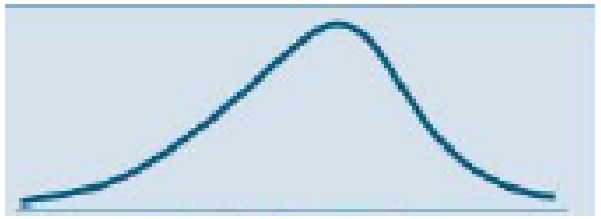
Coeficiente de Assimetria

Essa medida verifica onde os dados estão se concentrando em relação à média:

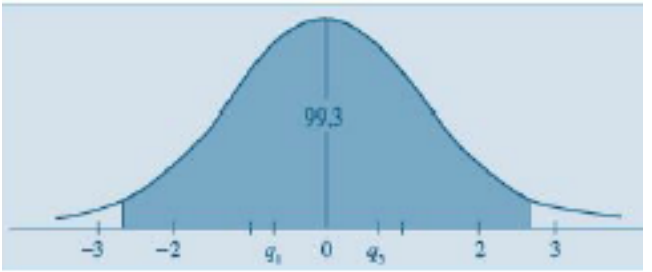
- Se os dados **estão à esquerda da média**, então há **assimetria negativa**;



- Se os dados concentram-se a **direita da média**, então há **assimetria positiva**;



- Se os dados estiverem concentrados junto a média, dizemos que a **distribuição dos dados é simétrica**.



Coeficiente de assimetria

A fórmula do coeficiente de assimetria é

$$S = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3 / n}{\sigma^3}$$

- Se S for menor que zero $(S < 0)$, a distribuição do conjunto de dados é assimétrica negativa;
- Se S for maior que zero $(S > 0)$, a distribuição do conjunto de dados é assimétrica positiva;
- Se S for igual a zero $(S = 0)$, a distribuição do conjunto de dados é assimétrica.

Coeficiente de assimetria

No Python:

- **Pandas**

  nome_da_variável [‘nome_da_coluna’].skew()

- **Scipy**

  Skew(nome_da_variável[‘nome_da_coluna’])