Estatística Descritiva

Simulação Discreta

Filipe Saraiva

Conteúdo

Conteúdo

Introdução

O que é Estatística?

Estatística é um conjunto de técnicas que permitem, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos e experimentos.

Introdução

Algumas áreas da Estatística são:

- Probabilidade;
- Inferência Estatística;
- Estatística Descritiva.

Introdução

O que é Estatística Descritiva?

Estatística Descritiva é aquela que visa sumarizar e descrever conjuntos de dados.

O objetivo, portanto, é organizar dados e aprender o que eles apresentam sobre uma população. Não há expectativa de uso de indução ou inferência.

Introdução

A Estatística Descritiva trabalha com 2 conjuntos de medidas principais:

- Medidas de Centralidade - informações que, em certa medida, representam todo o conjunto;
- Medidas de Dispersão - para se obter o grau de variabilidade de um elemento em relação ao conjunto.

Introdução

As principais medidas de cada conjunto são:

- Medidas de Centralidade - média, mediana, moda, mínimo, máximo, quartis.
- Medidas de Dispersão - amplitude, variância, desvio padrão, coeficiente de variação, coeficiente de assimetria.

Introdução

Tomando as medidas de centralidade e de dispersão como ponto de partida, trateremos sobre as principais medidas de cada conjunto nessa aula.

Ao final, veremos como realizar os cálculos desses valores utilizando Python.

Introdução

Para os exemplos, faremos:

- X é o conjunto de elementos;
- x_i é o i-ésimo elemento do conjunto X;
- n é $|X|$, ou o tamanho/módulo do conjunto.

Conteúdo

Medidas de Centralidade

Como já comentado, Medidas de Centralidade apresentam informações que, em certa medida, representam o conjunto como um todo.

Veremos as seguintes medidas de centralidade nessa apresentação:

- Média
- Moda
- Mediana
- Máximo
- Mínimo
- Quartis

Medidas de Centralidade

Média

Média ou **Média Aritmética** é a soma dos valores de elementos dividido pela quantidade de elementos.
A média identifica o comportamento médio ou a média dos valores presentes no conjunto.

A fórmula matemática da média (*media*) é, portanto:

$$media = \frac{\sum_{i=0}^{n} x_i}{n}$$

Medidas de Centralidade

Exemplo da Média:

Fazendo $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, o cálculo da média seria:

$$media = \frac{\sum_{i=0}^{n} x_i}{n} = \frac{0+3+1+6+3+4+1+3}{8} = \frac{21}{8} = 2,625$$

Portanto, $media = 2,625$

Medidas de Centralidade

Moda

Moda é o valor que aparece com mais frequência no conjunto.

Imaginando $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, a *moda* desse conjunto seria $moda = 3$.

Medidas de Centralidade

Mediana

Mediana é o valor que divide o conjunto X pela metade, criando 2 conjuntos com exatos 50% dos elementos do conjunto original para cada um.

Quando o n é ímpar, o valor da mediana será exatamente o elemento cujo valor está na metade do conjunto. Para o caso de n par, a mediana será a média aritmética entre o último elemento do primeiro conjunto, e o primeiro elemento do segundo conjunto.

Medidas de Centralidade

Exemplo da Mediana:

Exemplo ímpar:
Fazendo $X = \{0, 3, 1, 6, 3, 4, 1\}$, ordenando-o teríamos $X = \{0, 1, 1, 3, 3, 4, 6\}$. Nesse caso, o valor 3 dividiria X em $X_1 = \{0, 1, 1\}$ e $X_2 = \{3, 4, 6\}$.
Assim, $mediana = 3$.

Exemplo par:
Fazendo $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, ordenando-o teríamos $X = \{0, 1, 1, 3, 3, 3, 4, 6\}$. Nesse caso, os valores médios serão 3e3. A média aritmética destes será $\frac{3+3}{2} = 3$. Esse valor permitirá a criação dos conjuntos $X_1 = \{0, 1, 1, 3\}$ e $X_2 = \{3, 3, 4, 6\}$
Assim, $mediana = 3$.

Medidas de Centralidade

Máximo e Mínimo

Máximo é o maior valor pertencente ao conjunto, enquanto **Mínimo** é o menor.

Retomando $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, tem-se:
$maximo(X) = 6$
$minimo(X) = 0$

Medidas de Centralidade

Quartis

Quartis são valores que dividem o conjunto em subconjuntos com 25% dos elementos, cada.

Há, no geral, 3 quartis, sendo eles:

- Quartil 1 (Q1) cujo valor indica que 25% do conjunto está à esquerda;
- Quartil 2 (Q2) que coloca 50% do conjunto à esquerda (mediana);
- Quartil 3 (Q3) onde 75% do conjunto está à esquerda.

Medidas de Centralidade

Exemplo de Quartis:

Fazendo $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, e ordenando-o para $X = \{0, 1, 1, 3, 3, 3, 4, 6\}$ temos:

- Quartil 1 (Q1) = 1
- Quartil 2 (Q2) = *mediana* = 3
- Quartil 3 (Q3) = 3,5

Conteúdo

Medidas de Dispersão

Como já falado anteriormente, Medidas de Dispersão apresentam o grau de variabilidade entre um elemento e o conjunto.

Veremos as seguintes medidas de dispersão nessa apresentação:

- Amplitude;
- Variância;
- Desvio Padrão;
- Coeficiente de Variação;
- Coeficiente de Assimetria.

Medidas de Dispersão

Amplitude

Amplitude é a diferença entre o valor máximo e o valor mínimo do conjunto. Ela identifica o quão distante estão os elementos.

Retomando $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, temos:

- $maximo = 6$
- $minimo = 0$
- $amplitude = maximo - minimo = 6 - 0 = 6$

Medidas de Dispersão

Variância

Variância é uma medida de dispersão que indica o quão longe os valores estão da média do conjunto. A fórmula matemática é dada pela soma do quadrado da diferença de cada valor para a média, dividido pelo número de elementos, conforme abaixo:

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

Para o exemplo $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, cuja $media = 2,625$, temos:

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

$\frac{(0-2,625)^2+(3-2,625)^2+(1-2,625)^2+(6-2,625)^2+(3-2,625)^2+(4-2,625)^2+(1-2,625)^2+(3-2,625)^2}{8}$

$= 3,234375$

Medidas de Dispersão

Desvio Padrão

Desvio Padrão identifica o quão próximo ou longe da média estão os elementos do conjunto. Quanto mais baixo(/alto), o desvio padrão identifica que os elementos estão mais próximos(/distantes) da média. O cálculo do desvio padrão é dado pela raiz quadrada da variância, conforme:

$$desvio\_padrao = \sqrt{variancia}$$

Sabendo que a $variancia(X) = 3,234375$, teremos:

$$desvio\_padrao = \sqrt{variancia} = \sqrt{3,234375} = 1,798436$$

Medidas de Dispersão

Coeficiente de Variação

Coeficiente de Variação é a razão entre o desvio padrão e a variância, normalmente calculado em termos de porcentagem.

Ele apresenta a extensão da variabilidade da população em termos da média. A fórmula matemática é dada por:

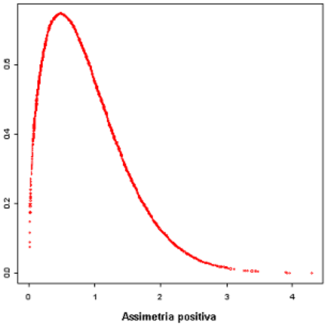
$$coeficiente_v ariacao = \frac{desvio\_padrao}{variancia} * 100$$

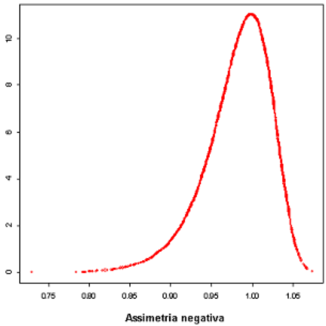
Seguindo com o exemplo $X = \{0, 3, 1, 6, 3, 4, 1, 3\}$, onde $desvio\_padrao = 1,798436$ e $variancia = 3,234375$, temos: $coeficiente_v ariacao = \frac{desvio\_padrao}{variancia} * 100 = \frac{1,798436}{3,234375} * 100 = 55.6038$

Medidas de Dispersão

Coeficiente de Assimetria

Coeficiente de Assimetria identifica assimetrias no gráfico da função de densidade do conjunto. Um valor positivo indica uma elevação na esquerda, enquanto um negativo indica uma elevação na direita.





Medidas de Dispersão

Coeficiente de Assimetria

A fórmula do **Coeficiente de Assimetria** é dada por:

$$coeficiente\_assimetria = \frac{Q3}{mediana^{\frac{3}{2}}}$$

Para o exemplo, relembrando que $Q3(X) = 3,5$ e $mediana = 3$:

$$coeficiente\_assimetria = \frac{Q3}{mediana^{\frac{3}{2}}} = \frac{3{,}5}{3^{\frac{3}{2}}} = 0,6735$$

O que indica que a curva terá uma assimetria para a direita.

Conteúdo

Estatística Descritiva em Python

Em Python há vários pacotes matemáticos e estatísticos com funções pré-definidas para os cálculos das medições de Estatística Descritiva descritas nessa aula, como o **scipy** e o **numpy**.

A partir do Python 3.4, foi disponibilizado um pacote padrão para esse fim chamado **statistics**, que contém funções para os cálculos mais básicos apresentados. Utilizaremos esta biblioteca para os cálculos que estudamos nessa apresentação.

Estatística Descritiva em Python

Para utilizar o pacote, importamo-os com:

import statistics

Para não ficar com um nome muito grande de biblioteca, recomendamos a criação de um alias:

import statistics as st

Estatística Descritiva em Python

Medidas de Centralidade:

Média	**st.mean(X)**
Moda	**st.mode(X)**
Mediana	**st.median(X)**
Máximo	**max(X)**
Mínimo	**min(X)**
Quartis	*

Não há fórmula pré-definida para o Quartis, que deve ser implementada do zero.

Estatística Descritiva em Python

Medidas de Centralidade:

Cálculo dos Quartis:
Pré-requisito:

1. Faz-se a ordenação da lista X = **sorted(X)**;

- Q1 = **st.median(X[:len(X)//2])**
- Q2 = **st.median(X)**
- Q3 = **st.median(X[len(X)//2:])**

Lembrando que $X[:i]$ é a sublista do primeiro índice até o índice i, enquanto $X[i:]$ é a sublista do índice i até o fim da lista.

Estatística Descritiva em Python

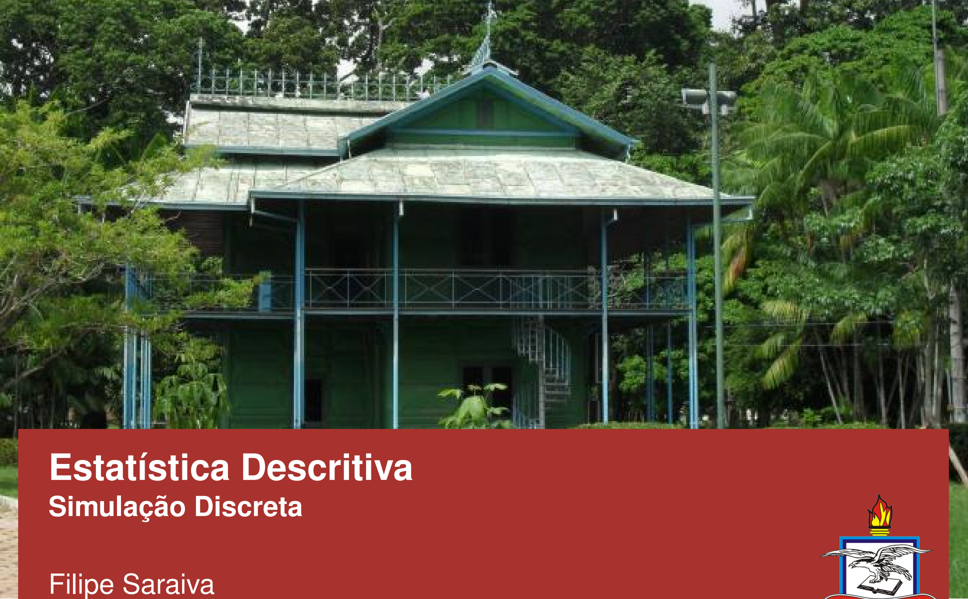
Medidas de Dispersão:

Amplitude	**max(X) - min(X)**
Variância	**st.pvariance(X)**
Desvio Padrão	**st.pstdev(X)**
Coeficiente de Variação	**st.pstdev(X) / st.pvariance(X) * 100**
Coeficiente de Assimetria	**st.median(X[len(X)//2:]) / st.median() ** 3/2**

Conteúdo

Conclusões

- Estatística Descritiva é aquela voltada a descrever conjuntos de dados;
- Os dois principais conjuntos de medições são de centralidade e dispersão;
- O pacote **statistics** do Python 3 provê várias funções pré-definidas para cálculos de estatística descritiva.



Estatística Descritiva

Simulação Discreta

Filipe Saraiva