応用データ解析中間レポート

03230966 河田顕帆

2024 年 2 月 3 日

本レポート作成に用いたソースコードは github にて公開している。

1 課題 1

期待値・中央値・標本分散・不偏分散・歪度・尖度は以下のようにして計算した。

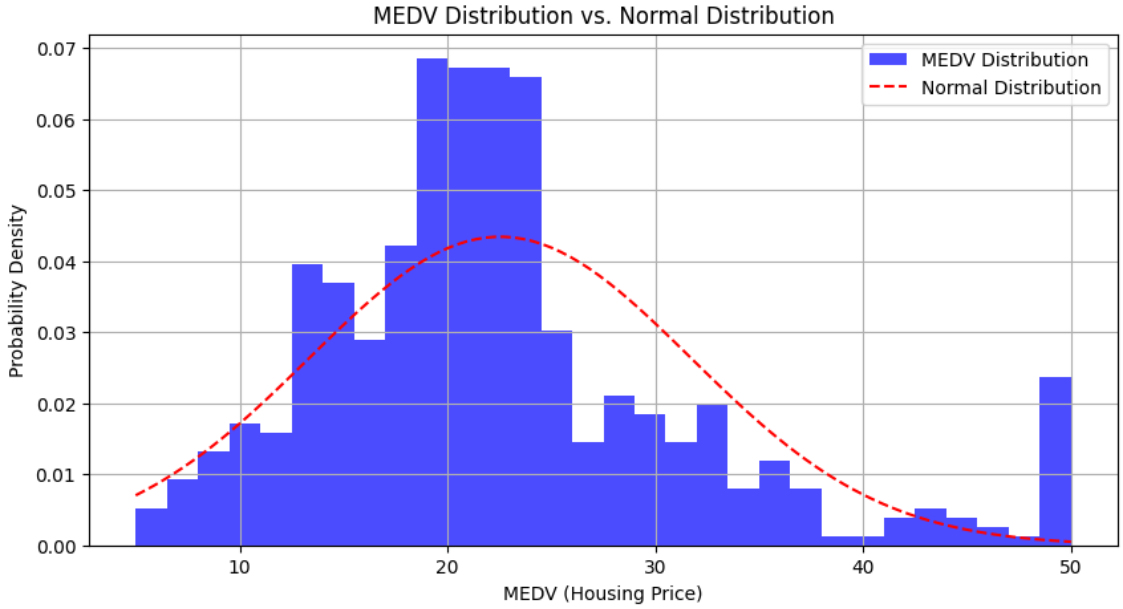
```
import numpy as np
import pandas as pd

df = pd.read_csv('BostonHousing.csv')

medv_data = df['MEDV']

# 期待値（平均値）
mean_medv = np.mean（medv_data）

# 中央値
median_medv = np.median（medv_data）

# 標本分散
sample_variance_medv = np.var（medv_data, ddof=0）

# 不偏分散
unbiased_variance_medv = np.var（medv_data, ddof=1）

# 歪度
skewness_medv = medv_data.skew（）

# 尖度
kurtosis_medv = medv_data.kurtosis（）
```

結果は以下のようになった。

- 期待値（平均値）: 22.532806324110677

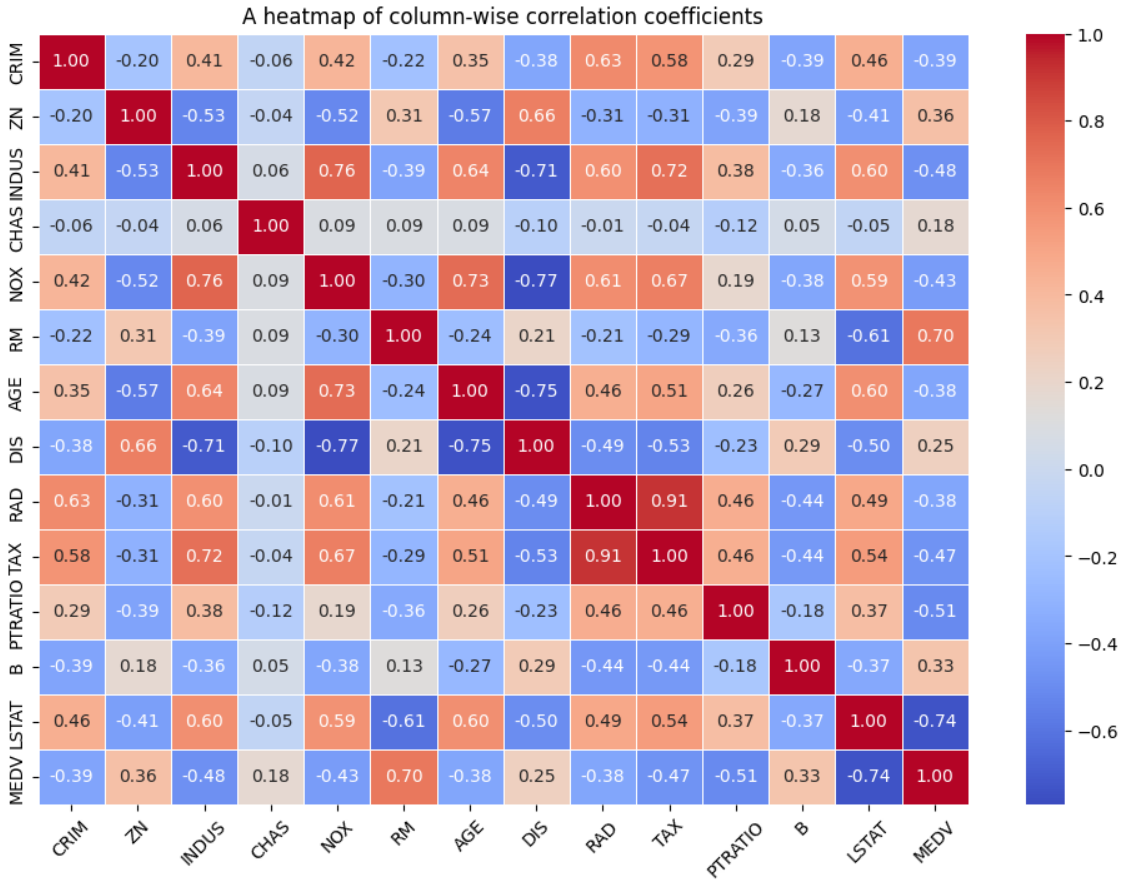
- 中央値: 21.2
- 標本分散: 84.41955615616556
- 不偏分散: 84.58672359409856
- 歪度: 1.1080984082549072
- 尖度: 1.495196944165818

平均値よりも中央値の方が小さい値を取っているため、データは左に歪んでいると言える。歪度に関しては、正の歪度があることを示しており、データは右裾が長いと言える。尖度に関しては 1.5 となっているため、尖度が 0 の正規分布よりも尖っていて、正規分布よりも価格のピークが集中していることが分かる。実際にヒストグラムを描画してみると、以下のようになり、上記の事柄が確かめられる。



2 課題 2

データ内の各変数の共分散行列を heatmap にして図示すると以下のようになる。
MEDV と相関係数の絶対値が比較的大きい変数は LSTAT (相関係数:-0.74) と RM (相関係数:0.70) であることが分かる。逆に、相関係数の絶対値が比較的小さい変数は CHAS (相関係数:0.18) と DIS (相関係数:0.25) であることが分かる。



- 相関係数の絶対値が大きい変数
  - LSTAT

    人口の低所得者の割合と価格の負の相関がある。これは、低所得者の割合が高い地域ほど価格が安いということを示している。
  - RM

    1 戸あたりの平均部屋数と価格の正の相関がある。これは、部屋数が多いほど価格が高いということを示している。
- 相関係数の絶対値が小さい変数
  - CHAS

    チャールズ川沿いかどうかと価格の相関が小さい。これは、チャールズ川沿いかどうかと価格にはあまり関係がないということを示している。
  - DIS

    ボストンの主な雇用センターまでの重み付き距離と価格の相関が小さい。これは、雇用センターまでの距離と価格にはあまり関係がないということを示している。