untitled6

September 8, 2024

APPLYING STATISTICAL METHODS TO DATASETS

1.DESCRIPTIVE STATICS

Installing required libraries:

1.Scipy library

We use the *SciPy* library to apply statistical methods to datasets because it provides a comprehensive set of tools for performing scientific and technical computing. Here's why SciPy is particularly useful for statistics:

1. *Rich Collection of Statistical Functions*: These functions are essential for analyzing and interpreting data.
2. *High-Level Abstractions*: SciPy abstracts complex mathematical operations, making it easier to implement advanced statistical methods without diving into the details of the algorithms.
3. *Integration with NumPy*: SciPy is built on top of NumPy, which provides efficient array operations. This means that you can handle large datasets and perform statistical computations efficiently.
4. *Probability Distributions*: SciPy has an extensive library of continuous and discrete probability distributions, which are useful for modeling and understanding data distributions.
5. *Hypothesis Testing*: SciPy provides a variety of statistical tests, such as t-tests, chi-square tests, and ANOVA, making it convenient for hypothesis testing.
6. *Optimization and Fitting*: SciPy includes tools for curve fitting and optimization, which are essential for statistical modeling and finding best-fit parameters in data.
7. *Interoperability*: SciPy works well with other scientific computing libraries like pandas and matplotlib, allowing you to integrate statistical analysis with data manipulation and visuali

[1]:
```
pip install scipy
```

```
Requirement already satisfied: scipy in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages
(1.14.0)
Requirement already satisfied: numpy<2.3,>=1.23.5 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
scipy) (2.0.1)
Note: you may need to restart the kernel to use updated packages.
```

2.Scikit - learn statsmodels:

We use *scikit-learn* for machine learning tasks like classification, regression, and clustering with easy-to-use APIs and performance-focused algorithms. *Statsmodels* is ideal for traditional statistical modeling, providing in-depth outputs like p-values and confidence intervals. Scikit-learn emphasizes predictive modeling, while statsmodels offers detailed statistical analysis. Both integrate well with NumPy and pandas, complementing each other for comprehensive data analysis.

[3]:
```
pip install scikit-learn statsmodels
```

```
Requirement already satisfied: scikit-learn in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (1.5.1)
Requirement already satisfied: statsmodels in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages
(0.14.2)
Requirement already satisfied: numpy>=1.19.5 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
scikit-learn) (2.0.1)
Requirement already satisfied: scipy>=1.6.0 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
scikit-learn) (1.14.0)
Requirement already satisfied: joblib>=1.2.0 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
scikit-learn) (3.5.0)
Requirement already satisfied: pandas!=2.1.0,>=1.4 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
statsmodels) (2.2.2)
Requirement already satisfied: patsy>=0.5.6 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
statsmodels) (0.5.6)
Requirement already satisfied: packaging>=21.3 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
statsmodels) (24.1)
Requirement already satisfied: python-dateutil>=2.8.2 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
pandas!=2.1.0,>=1.4->statsmodels) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
pandas!=2.1.0,>=1.4->statsmodels) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
pandas!=2.1.0,>=1.4->statsmodels) (2024.1)
Requirement already satisfied: six in
c:\users\ruppa\appdata\local\programs\python\python312\lib\site-packages (from
patsy>=0.5.6->statsmodels) (1.16.0)
```

```
Note: you may need to restart the kernel to use updated packages.
```

[ ]:

2.DESCRIPTIVE STATISTICS:

Inferential Statistics in Python

Inferential statistics involve drawing conclusions about a population based on a sample. Here are some key techniques:

Hypothesis Testing: Used to determine whether there is enough evidence to reject a null hypothesis. Confidence Intervals: Provide a range of values that likely contain the population parameter. Regression Analysis: Examines relationships between variables.

Exercises for Applying Statistical Methods Exercise 1: Analyzing a Health-Related Dataset

Choose any health-related dataset from Kaggle or the UCI Machine Learning Repository. Once you have selected a dataset:

Load the dataset into Python using pandas. Calculate the mean, median, mode, standard deviation, and variance for all the relevant features. Conduct a hypothesis test to determine if a specific feature (e.g., average blood pressure, cholesterol levels, etc.) is significantly different from a chosen value. Compute a 95% confidence interval for the mean of a selected feature. 5.2 Exercise 2: Exploring Regression Analysis on a New Dataset

Using the same dataset you selected from Kaggle or the UCI Machine Learning Repository:

Perform a linear regression analysis to determine the relationship between two or more variables (e.g., how BMI affects disease progression or how age impacts blood pressure). Interpret the coefficients, p-values, and R-squared value from the regression model summary. Create visualizations to illustrate the relationships between variables and the regression line.

2.1 Loading a dataset from kaggle into python using pandas

[5]:
```
import pandas as pd
data=pd.read_csv('fetal_health.csv')
df=pd.DataFrame(data)
print(df.head())
```

```
   baseline value  accelerations  fetal_movement  uterine_contractions  \
0           120.0          0.000             0.0                 0.000
1           132.0          0.006             0.0                 0.006
2           133.0          0.003             0.0                 0.008
3           134.0          0.003             0.0                 0.008
4           132.0          0.007             0.0                 0.008

   light_decelerations  severe_decelerations  prolongued_decelerations  \
0                0.000                   0.0                       0.0
1                0.003                   0.0                       0.0
2                0.003                   0.0                       0.0
3                0.003                   0.0                       0.0
4                0.000                   0.0                       0.0
```

```
   abnormal_short_term_variability  mean_value_of_short_term_variability  \
0                             73.0                                   0.5
1                             17.0                                   2.1
2                             16.0                                   2.1
3                             16.0                                   2.4
4                             16.0                                   2.4

   percentage_of_time_with_abnormal_long_term_variability  …  histogram_min  \
0                                                 43.0     …           62.0
1                                                  0.0     …           68.0
2                                                  0.0     …           68.0
3                                                  0.0     …           53.0
4                                                  0.0     …           53.0

   histogram_max  histogram_number_of_peaks  histogram_number_of_zeroes  \
0          126.0                        2.0                         0.0
1          198.0                        6.0                         1.0
2          198.0                        5.0                         1.0
3          170.0                       11.0                         0.0
4          170.0                        9.0                         0.0

   histogram_mode  histogram_mean  histogram_median  histogram_variance  \
0           120.0           137.0             121.0                73.0
1           141.0           136.0             140.0                12.0
2           141.0           135.0             138.0                13.0
3           137.0           134.0             137.0                13.0
4           137.0           136.0             138.0                11.0

   histogram_tendency  fetal_health
0                 1.0           2.0
1                 0.0           1.0
2                 0.0           1.0
3                 1.0           1.0
4                 1.0           1.0

[5 rows x 22 columns]
```

2.2 Performing Descriptive Statistics

[6]:
```
#calculating basic descriptive statistics
print("Mean:\n", df.mean())
```

```
Mean:
 baseline value                                          133.303857
accelerations                                             0.003178
fetal_movement                                            0.009481
uterine_contractions                                      0.004366
```

```
light_decelerations                                          0.001889
severe_decelerations                                         0.000003
prolongued_decelerations                                     0.000159
abnormal_short_term_variability                             46.990122
mean_value_of_short_term_variability                         1.332785
percentage_of_time_with_abnormal_long_term_variability       9.846660
mean_value_of_long_term_variability                          8.187629
histogram_width                                             70.445908
histogram_min                                               93.579492
histogram_max                                              164.025400
histogram_number_of_peaks                                    4.068203
histogram_number_of_zeroes                                   0.323612
histogram_mode                                             137.452023
histogram_mean                                             134.610536
histogram_median                                           138.090310
histogram_variance                                          18.808090
histogram_tendency                                           0.320320
fetal_health                                                 1.304327
dtype: float64
```

[7]:
```
#calculate median
print("\nMedian:\n", df.median())
```

```
Median:
 baseline value                                            133.000
accelerations                                               0.002
fetal_movement                                              0.000
uterine_contractions                                        0.004
light_decelerations                                         0.000
severe_decelerations                                        0.000
prolongued_decelerations                                    0.000
abnormal_short_term_variability                            49.000
mean_value_of_short_term_variability                        1.200
percentage_of_time_with_abnormal_long_term_variability      0.000
mean_value_of_long_term_variability                         7.400
histogram_width                                            67.500
histogram_min                                              93.000
histogram_max                                             162.000
histogram_number_of_peaks                                   3.000
histogram_number_of_zeroes                                  0.000
histogram_mode                                            139.000
histogram_mean                                            136.000
histogram_median                                          139.000
histogram_variance                                          7.000
histogram_tendency                                          0.000
fetal_health                                                1.000
dtype: float64
```

[8]:
```
#calculating mode
print("\nMode:\n", df.mode().iloc[0])
```

```
Mode:
 baseline value                                                133.0
accelerations                                                  0.0
fetal_movement                                                 0.0
uterine_contractions                                           0.0
light_decelerations                                            0.0
severe_decelerations                                           0.0
prolongued_decelerations                                       0.0
abnormal_short_term_variability                               60.0
mean_value_of_short_term_variability                           0.8
percentage_of_time_with_abnormal_long_term_variability         0.0
mean_value_of_long_term_variability                            0.0
histogram_width                                               39.0
histogram_min                                                 50.0
histogram_max                                                157.0
histogram_number_of_peaks                                      1.0
histogram_number_of_zeroes                                     0.0
histogram_mode                                               133.0
histogram_mean                                               143.0
histogram_median                                             146.0
histogram_variance                                             1.0
histogram_tendency                                             0.0
fetal_health                                                   1.0
Name: 0, dtype: float64
```

[9]:
```
#calculating standard deviation
print("\nStandard Deviation:\n", df.std())
```

```
Standard Deviation:
 baseline value                                                9.840844
accelerations                                                 0.003866
fetal_movement                                                0.046666
uterine_contractions                                          0.002946
light_decelerations                                           0.002960
severe_decelerations                                          0.000057
prolongued_decelerations                                      0.000590
abnormal_short_term_variability                              17.192814
mean_value_of_short_term_variability                          0.883241
percentage_of_time_with_abnormal_long_term_variability       18.396880
mean_value_of_long_term_variability                           5.628247
histogram_width                                              38.955693
histogram_min                                                29.560212
histogram_max                                                17.944183
```

```
histogram_number_of_peaks                                 2.949386
histogram_number_of_zeroes                                0.706059
histogram_mode                                           16.381289
histogram_mean                                           15.593596
histogram_median                                         14.466589
histogram_variance                                       28.977636
histogram_tendency                                        0.610829
fetal_health                                              0.614377
dtype: float64
```

[10]:
```
#calculating variance
print("\nVariance:\n", df.var())
```

```
Variance:
 baseline value                                          9.684222e+01
accelerations                                           1.494279e-05
fetal_movement                                          2.177701e-03
uterine_contractions                                    8.679323e-06
light_decelerations                                     8.762835e-06
severe_decelerations                                    3.283272e-09
prolongued_decelerations                                3.480381e-07
abnormal_short_term_variability                         2.955928e+02
mean_value_of_short_term_variability                    7.801153e-01
percentage_of_time_with_abnormal_long_term_variability  3.384452e+02
mean_value_of_long_term_variability                     3.167716e+01
histogram_width                                         1.517546e+03
histogram_min                                           8.738061e+02
histogram_max                                           3.219937e+02
histogram_number_of_peaks                               8.698876e+00
histogram_number_of_zeroes                              4.985198e-01
histogram_mode                                          2.683466e+02
histogram_mean                                          2.431602e+02
histogram_median                                        2.092822e+02
histogram_variance                                      8.397034e+02
histogram_tendency                                      3.731116e-01
fetal_health                                            3.774589e-01
dtype: float64
```

[13]:
```
# Additional descriptive statistics
print("\nRange:\n", df.max() - df.min())
print("\nSkewness:\n", df.skew())
print("\nKurtosis:\n", df.kurt())
```

```
Range:
 baseline value                                          54.000
accelerations                                            0.019
```

```
fetal_movement                                            0.481
uterine_contractions                                      0.015
light_decelerations                                       0.015
severe_decelerations                                      0.001
prolongued_decelerations                                  0.005
abnormal_short_term_variability                          75.000
mean_value_of_short_term_variability                      6.800
percentage_of_time_with_abnormal_long_term_variability   91.000
mean_value_of_long_term_variability                      50.700
histogram_width                                         177.000
histogram_min                                           109.000
histogram_max                                           116.000
histogram_number_of_peaks                                18.000
histogram_number_of_zeroes                               10.000
histogram_mode                                          127.000
histogram_mean                                          109.000
histogram_median                                        109.000
histogram_variance                                      269.000
histogram_tendency                                        2.000
fetal_health                                              2.000
dtype: float64

Skewness:
 baseline value                                           0.020312
accelerations                                             1.204392
fetal_movement                                            7.811477
uterine_contractions                                      0.159315
light_decelerations                                       1.718437
severe_decelerations                                     17.353457
prolongued_decelerations                                  4.323965
abnormal_short_term_variability                          -0.011829
mean_value_of_short_term_variability                      1.657339
percentage_of_time_with_abnormal_long_term_variability    2.195075
mean_value_of_long_term_variability                       1.331998
histogram_width                                           0.314235
histogram_min                                             0.115784
histogram_max                                             0.577862
histogram_number_of_peaks                                 0.892886
histogram_number_of_zeroes                                3.920287
histogram_mode                                           -0.995178
histogram_mean                                           -0.651019
histogram_median                                         -0.478414
histogram_variance                                        3.219974
histogram_tendency                                       -0.311632
fetal_health                                              1.849934
dtype: float64

Kurtosis:
```

```
 baseline value                                            -0.292943
accelerations                                               0.767648
fetal_movement                                             64.260821
uterine_contractions                                       -0.635071
light_decelerations                                         2.517461
severe_decelerations                                      299.424142
prolongued_decelerations                                   20.515918
abnormal_short_term_variability                            -1.051030
mean_value_of_short_term_variability                        4.700756
percentage_of_time_with_abnormal_long_term_variability      4.252998
mean_value_of_long_term_variability                         4.131254
histogram_width                                            -0.902287
histogram_min                                              -1.290422
histogram_max                                               0.632769
histogram_number_of_peaks                                   0.504211
histogram_number_of_zeroes                                 30.365084
histogram_mode                                              3.009531
histogram_mean                                              0.933427
histogram_median                                            0.667259
histogram_variance                                         15.131589
histogram_tendency                                         -0.652639
fetal_health                                                2.091215
dtype: float64
```

[14]:
```
df.columns
```

[14]:
```
Index(['baseline value', 'accelerations', 'fetal_movement',
       'uterine_contractions', 'light_decelerations', 'severe_decelerations',
       'prolongued_decelerations', 'abnormal_short_term_variability',
       'mean_value_of_short_term_variability',
       'percentage_of_time_with_abnormal_long_term_variability',
       'mean_value_of_long_term_variability', 'histogram_width',
       'histogram_min', 'histogram_max', 'histogram_number_of_peaks',
       'histogram_number_of_zeroes', 'histogram_mode', 'histogram_mean',
       'histogram_median', 'histogram_variance', 'histogram_tendency',
       'fetal_health'],
      dtype='object')
```

[15]:
```
df.describe()
```

[15]:

	baseline value	accelerations	fetal_movement	uterine_contractions	\
count	2126.000000	2126.000000	2126.000000	2126.000000	
mean	133.303857	0.003178	0.009481	0.004366	
std	9.840844	0.003866	0.046666	0.002946	
min	106.000000	0.000000	0.000000	0.000000	
25%	126.000000	0.000000	0.000000	0.002000	
50%	133.000000	0.002000	0.000000	0.004000	

```
75%        140.000000       0.006000        0.003000              0.007000
max        160.000000       0.019000        0.481000              0.015000

       light_decelerations  severe_decelerations  prolongued_decelerations  \
count          2126.000000           2126.000000               2126.000000
mean              0.001889              0.000003                  0.000159
std               0.002960              0.000057                  0.000590
min               0.000000              0.000000                  0.000000
25%               0.000000              0.000000                  0.000000
50%               0.000000              0.000000                  0.000000
75%               0.003000              0.000000                  0.000000
max               0.015000              0.001000                  0.005000

       abnormal_short_term_variability  mean_value_of_short_term_variability  \
count                      2126.000000                           2126.000000
mean                         46.990122                              1.332785
std                          17.192814                              0.883241
min                          12.000000                              0.200000
25%                          32.000000                              0.700000
50%                          49.000000                              1.200000
75%                          61.000000                              1.700000
max                          87.000000                              7.000000

       percentage_of_time_with_abnormal_long_term_variability  …  \
count                                            2126.00000       …
mean                                               9.84666       …
std                                               18.39688       …
min                                                0.00000       …
25%                                                0.00000       …
50%                                                0.00000       …
75%                                               11.00000       …
max                                               91.00000       …

       histogram_min  histogram_max  histogram_number_of_peaks  \
count    2126.000000    2126.000000                2126.000000
mean       93.579492     164.025400                   4.068203
std        29.560212      17.944183                   2.949386
min        50.000000     122.000000                   0.000000
25%        67.000000     152.000000                   2.000000
50%        93.000000     162.000000                   3.000000
75%       120.000000     174.000000                   6.000000
max       159.000000     238.000000                  18.000000

       histogram_number_of_zeroes  histogram_mode  histogram_mean  \
count                 2126.000000     2126.000000     2126.000000
mean                     0.323612      137.452023      134.610536
std                      0.706059       16.381289       15.593596
```

```
min                    0.000000       60.000000       73.000000
25%                    0.000000      129.000000      125.000000
50%                    0.000000      139.000000      136.000000
75%                    0.000000      148.000000      145.000000
max                   10.000000      187.000000      182.000000

       histogram_median  histogram_variance  histogram_tendency  fetal_health
count       2126.000000         2126.000000         2126.000000   2126.000000
mean         138.090310           18.808090            0.320320      1.304327
std           14.466589           28.977636            0.610829      0.614377
min           77.000000            0.000000           -1.000000      1.000000
25%          129.000000            2.000000            0.000000      1.000000
50%          139.000000            7.000000            0.000000      1.000000
75%          148.000000           24.000000            1.000000      1.000000
max          186.000000          269.000000            1.000000      3.000000

[8 rows x 22 columns]
```

2.3 Performing inferential statistics

[22]:
```
from scipy import stats
import pandas as pd
data=pd.read_csv('fetal_health.csv')
df=pd.DataFrame(data)
fetal_health=df['fetal_health']

# Hypothetical population mean for fetal_health
population_mean = 20

# Perform one-sample t-test
t_stat, p_value = stats.ttest_1samp(fetal_health, population_mean)

print(f"T-Statistic: {t_stat}")
print(f"P-Value: {p_value}")
```

```
T-Statistic: -1403.0976033589639
P-Value: 0.0
```

REGRESSION ANALYSIS

[27]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score,␣
 ↪confusion_matrix
```

[32]:
```
#splitting the features i.e, seperating target variable
X = df.drop(columns=['fetal_movement'])
```

```
y = df['fetal_health']
```

[33]:
```
X
```

[33]:

	baseline value	accelerations	uterine_contractions	\
0	120.0	0.000	0.000	
1	132.0	0.006	0.006	
2	133.0	0.003	0.008	
3	134.0	0.003	0.008	
4	132.0	0.007	0.008	
…	…	…	…	
2121	140.0	0.000	0.007	
2122	140.0	0.001	0.007	
2123	140.0	0.001	0.007	
2124	140.0	0.001	0.006	
2125	142.0	0.002	0.008	

	light_decelerations	severe_decelerations	prolongued_decelerations	\
0	0.000	0.0	0.0	
1	0.003	0.0	0.0	
2	0.003	0.0	0.0	
3	0.003	0.0	0.0	
4	0.000	0.0	0.0	
…	…	…	…	
2121	0.000	0.0	0.0	
2122	0.000	0.0	0.0	
2123	0.000	0.0	0.0	
2124	0.000	0.0	0.0	
2125	0.000	0.0	0.0	

	abnormal_short_term_variability	mean_value_of_short_term_variability	\
0	73.0	0.5	
1	17.0	2.1	
2	16.0	2.1	
3	16.0	2.4	
4	16.0	2.4	
…	…	…	
2121	79.0	0.2	
2122	78.0	0.4	
2123	79.0	0.4	
2124	78.0	0.4	
2125	74.0	0.4	

	percentage_of_time_with_abnormal_long_term_variability	\
0	43.0	
1	0.0	
2	0.0	

```
3                                                 0.0
4                                                 0.0
…                                                 …
2121                                             25.0
2122                                             22.0
2123                                             20.0
2124                                             27.0
2125                                             36.0

      mean_value_of_long_term_variability  …  histogram_min  histogram_max  \
0                                     2.4  …           62.0          126.0
1                                    10.4  …           68.0          198.0
2                                    13.4  …           68.0          198.0
3                                    23.0  …           53.0          170.0
4                                    19.9  …           53.0          170.0
…                                       …  …              …              …
2121                                  7.2  …          137.0          177.0
2122                                  7.1  …          103.0          169.0
2123                                  6.1  …          103.0          170.0
2124                                  7.0  …          103.0          169.0
2125                                  5.0  …          117.0          159.0

      histogram_number_of_peaks  histogram_number_of_zeroes  histogram_mode  \
0                           2.0                         0.0           120.0
1                           6.0                         1.0           141.0
2                           5.0                         1.0           141.0
3                          11.0                         0.0           137.0
4                           9.0                         0.0           137.0
…                             …                         …             …
2121                        4.0                         0.0           153.0
2122                        6.0                         0.0           152.0
2123                        5.0                         0.0           153.0
2124                        6.0                         0.0           152.0
2125                        2.0                         1.0           145.0

      histogram_mean  histogram_median  histogram_variance  \
0              137.0             121.0                73.0
1              136.0             140.0                12.0
2              135.0             138.0                13.0
3              134.0             137.0                13.0
4              136.0             138.0                11.0
…                  …                 …                   …
2121           150.0             152.0                 2.0
2122           148.0             151.0                 3.0
2123           148.0             152.0                 4.0
2124           147.0             151.0                 4.0
2125           143.0             145.0                 1.0
```

```
      histogram_tendency  fetal_health
0                    1.0           2.0
1                    0.0           1.0
2                    0.0           1.0
3                    1.0           1.0
4                    1.0           1.0
…                    …             …
2121                 0.0           2.0
2122                 1.0           2.0
2123                 1.0           2.0
2124                 1.0           2.0
2125                 0.0           1.0

[2126 rows x 21 columns]
```

[34]:
```
print(y)
```

```
0       2.0
1       1.0
2       1.0
3       1.0
4       1.0
       …
2121    2.0
2122    2.0
2123    2.0
2124    2.0
2125    1.0
Name: fetal_health, Length: 2126, dtype: float64
```

[ ]:
```

```