Lab#1: Introduction

What is Statistical Data Analysis?

Statistical data analysis is the process of collecting, examining, and interpreting data to uncover patterns, trends, relationships, and insights. It involves the application of statistical methods and techniques to analyze data sets and draw meaningful conclusions. This process is fundamental in various fields, including business, science, engineering, healthcare, and social sciences, to make informed decisions based on empirical data.

Types of Statistical Data Analysis

Statistics data analysis is a class of analysis that includes different techniques and methods for collection, data analysis, interpretation and presentation of data. Knowing the approach to data analysis is one of the crucial aspects that allows drawing a meaningful conclusion. In this article, the most fundamental types of statistical data analysis will be described. The authors will explain all the terms and concepts easily.

Descriptive Statistics

Descriptive statistics are intended to make a basic summary of data and variables in the sample while providing point measures as the main features of a dataset.

Measures of Central Tendency

- **Mean:** The arithmetical mean, which is the mean value of all observation points. It is done by summing up all the values and dividing them by the total number of values.
- **Median:** The central value in a dataset when the values are linearly sorted in order of rise. The median is computed as the average of the two middle numbers if the dataset is even where there is an equal number of observations.
- **Mode:** This is the most recurring value in a set of data. It can be used as the only mode, multiple modes, or it can have no modes at all.

Measures of Variability

- **Range**: Such as Minimum/Maximum. The difference between the highest and lowest data points in a dataset. It delineates the values by generating a primary method of their spread.
- **Variance**: The averaging of the squared deviations away from the mean. This helps to find out how much the data values in the data set are uniformity with the mean.

- **Standard Deviation:** The square root of the variance which gives us the standard deviation. It is a more intuitive measure of variability as the scale of the estimate is matched the data units.
- **Interquartile Range (IQR):** It scores the distance stretching from the 25th percentile (IQR: 1) to 75th percentile (IQR:3). It monitors at what tier the middle 50th percentile of the distribution lies.

Frequency Distribution

- **Histogram:** Trend line representing the frequency distribution of a dataset. The horizontal axis (X-axis) is divided into blocks of length, and the height of each bar represents the frequency of points within each interval.
- **Bar Chart:** A chart with rectangular bars as the frequency or count of the categorical data in the form of bars below each other. Bar sizes reflect the amount they are worth.
- **Pie Chart:** A circular graph of dedicated spaces to various sectors, each having a proportion of the total area. It is a useful tool for making proportional relations between the whole and its parts.

Example: X-Y plot

```
import matplotlib.pyplot as plt

import numpy as np

#x = np.array([0, 1, 2, 3, 4, 5])

x = np.arange(0, 6, 1)

y = np.array([3, 8, 1, 10, 12,13])

plt.plot(x,y, linestyle = 'dashed', color = 'r', linewidth = '1', marker = 'o')

plt.show()

print(x)
```

Example: Pi chart plot

```
# Import libraries

from matplotlib import pyplot as plt

import numpy as np

# Creating dataset

cars = ['AUDI', 'BMW', 'FORD','TESLA', 'JAGUAR', 'MERCEDES']

data = [23, 17, 35, 29, 12, 41]

# Creating plot

fig = plt.figure(figsize=(10, 7))

plt.pie(data, labels=cars)

# show plot

plt.show()
```

Example: Range Calculation

```
# Sample Data

arr = [1, 2, 3, 4, 5]

#Finding Max

Maximum = max(arr)

# Finding Min

Minimum = min(arr)

# Difference Of Max and Min

Range = Maximum-Minimum

print("Maximum = {}, Minimum = {} and Range = {}".format( Maximum, Minimum, Range))
```

Example: Histogram Construction

The data in the following code are generated by using one of the random generated functions:

1) **np.random.randn** is a function in the NumPy library used in Python to generate random numbers from a standard normal distribution, which has a mean of 0 and a standard deviation of 1.
2) **np.random.normal** is a function in the NumPy library used in Python to generate random numbers from a normal (Gaussian) distribution with specified mean and standard deviation.
3) **np.random.randint** is a function in the NumPy library used in Python to generate random integers within a specified range.

```
import matplotlib.pyplot as plt

import numpy as np

# Generate 100 random numbers from a standard normal distribution

#data = np.random.randn(100)

#data = np.random.normal(80,0,1000)

data = np.random.randint(0, 10,100000 )

# create histogram

plt.hist(data, bins=5,  color='red', edgecolor='black', rwidth=1)

 plt.xlabel('Values')

plt.ylabel('Frequency')

plt.title('Basic Histogram')

# display histogram

plt.show()

print(data)
```