**Topics: Descriptive Statistics and Probability**

1. Look at the data given below. Plot the data, find the outliers and find out

|  |  |
| --- | --- |
| **Name of company** | **Measure X** |
| Allied Signal | 24.23% |
| Bankers Trust | 25.53% |
| General Mills | 25.41% |
| ITT Industries | 24.14% |
| J.P.Morgan & Co. | 29.62% |
| Lehman Brothers | 28.25% |
| Marriott | 25.81% |
| MCI | 24.39% |
| Merrill Lynch | 40.26% |
| Microsoft | 32.95% |
| Morgan Stanley | 91.36% |
| Sun Microsystems | 25.99% |
| Travelers | 39.42% |
| US Airways | 26.71% |
| Warner-Lambert | 35.00% |

Ans:

import numpy as np

import pandas as pd

import seaborn as sns

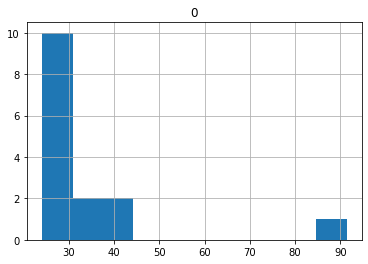
df = np.array([[24.23],[25.53],[25.41],[24.14],[29.62],[28.25],[25.81],[24.39],[40.26],[32.95],[91.36],[25.99],[39.42],[26.71],[35.00]])

df = pd.DataFrame(df)

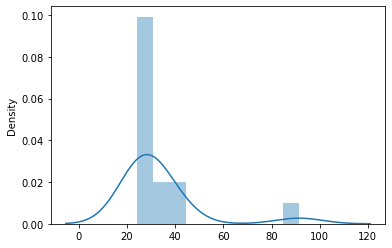
df

df.describe()

df.hist()



sns.distplot(df)



df.skew()

df.kurt()

df

# Plot the data

import matplotlib.pyplot as plt

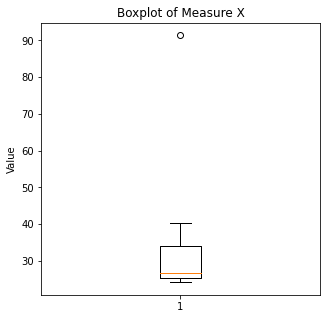
plt.figure(figsize=(5, 5))

plt.boxplot(df)

plt.title("Boxplot of Measure X")

plt.ylabel("Value")

plt.show()



# Calculating mean

mean = np.mean(df)

# Calculating standard deviation

std\_dev = np.std(df)

# Calculating variance

variance = np.var(df)

print("Mean :", mean) #33.271333 (MEAN)

print("Standard Deviation :", std\_dev) #16.370813 (Standard deviation)

print("Variance :", variance) #268.003505 (Variance)



Answer the following three questions based on the box-plot above.

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

Ans: Inter-quartile range (IQR) = Q3 – Q1

From figure , we can say that => Q1(25th percentile) = 5

Q3 (75th percentile) = 12

IQR = 12-5

IQR =7

IQR wiil tells us the range of half of the data

1. What can we say about the skewness of this dataset?

Ans:

It is Positively Skewed , (mean>median>mode) Tail is found extending towards right side of the curve

1. If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

Ans:

If the data point with the value 25 is corrected to 2.5, it would significantly affect the dataset's distribution and the box plot. This change could also effect the mean, median, and the skewness



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Ans:

The mode of the dataset from the histogram would lie between the interval 5 to 8 because that's where the histogram has the highest bar, indicating the most frequent data values.

1. Comment on the skewness of the dataset.

Ans:

It is Positively Skewed , (mean>median>mode) Tail is found extending towards right side of the curve

1. Suppose that the above histogram and the box-plot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

Ans:

They both are right-skewed and both have outliers, the median can be easily visualized in box plot where as in histogram mode is more visible

1. AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that “could happen.” Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

Ans: Given that, one in 200 long-distance telephone calls is misdirected.

probability that at least one in five attempted telephone calls reaches the wrong number

=> probability of call misdirecting p = 1/200

=0.005

=> Probability of call not Misdirecting = 1 - 1/200

= 199/200

=0.995

Number of Calls = 5

P(x) = ⁿCₓpˣqⁿ⁻ˣ ,n = 5 , p = 1/200 , q = 199/200

at least one in five attempted telephone calls reaches the wrong number

= 1 - none of the call reaches the wrong number

= 1 - P(0)

= 1 - ⁵C₀(1/200)⁰(199/200)⁵⁻⁰

= 1 - (199/200)⁵

= 0.02475

probability that at least one in five attempted telephone calls reaches the wrong number = 0.02475

1. Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution

|  |  |
| --- | --- |
| x | P(x) |
| -2,000 | 0.1 |
| -1,000 | 0.1 |
| 0 | 0.2 |
| 1000 | 0.2 |
| 2000 | 0.3 |
| 3000 | 0.1 |

1. What is the most likely monetary outcome of the business venture?

Ans:

In this case, the value of 'x' with the highest probability is 2000, with a probability of 0.3. So, the most likely monetary outcome of the business venture is $2,000.

1. Is the venture likely to be successful? Explain

Ans:

Long term average = \sum{P(xi)\*Xi} = (-2000\*0.1) +(-1000\*0.1) +(0) +(1000\*0.2) +(2000\*0.3) +(3000\*0.1) = 800$

As the long-term average gives positive numbers the Business venture likely to be successful

1. What is the long-term average earning of business ventures of this kind? Explain

Ans:

The expected value represents the long-term average return.

The expected value (mean) can be calculated as follows:

E(x) = Σ [x \* P(x)]

E(x) = (-2000 \* 0.1) + (-1000 \* 0.1) + (0 \* 0.2) + (1000 \* 0.2) + (2000 \* 0.3) + (3000 \* 0.1)

E(x) = -200 - 100 + 0 + 200 + 600 + 300

E(x) = 800

The expected value (mean) is $800. Since the expected value is positive, the

venture is likely to be successful on average. However, this does not consider the level of risk.

1. What is the good measure of the risk involved in a venture of this kind? Compute this measure

Ans: X P(X) E(X)= X . P(X) E(X²) = X² . P(X)

-2000 0.1 -200 400000

-1000 0.1 -100 100000

0 0.2 0 0

1000 0.2 200 200000

2000 0.3 600 1200000

3000 0.1 300 900000

Total 800 2800000

The good measure of the risk involved in a venture of this kind depends on the Variability in the distribution. Higher Variance means more chances of risk

Var (X) = E(X^2) –(E(X))^2

= 2800000 – 800^2

= 2160000