**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%** |

Answer : import pandas as pd

import numpy as np

import matplotlib.pyplot.plt

import seaborn as sns

%matplot inline

X=pd.Series([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])

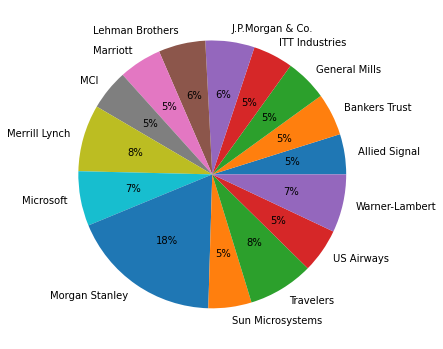
names=[‘Allied Signal’, ’Bankers Trust’, ‘General Mills’, ‘ITT Industries’, ‘J.P. Morgan & co’, ‘Lehman Brothers’, ‘Marriott’, ‘MCI’, ‘Merrill Lynch’, ‘Microsoft’, ‘Morgan Stanley’, ‘Sun Microsystems’, ‘Travelers’, ‘US Airways’, ‘Warner-Lambert’]

#Pie Plot

plt.figure(figsize=(6,8))

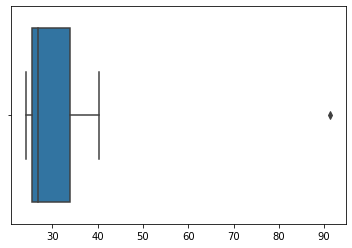
plt.pie(x, labels=name, autopct= ’%1.0f%%’)

plt.show()



#Box Plot to find outliers

Sns.boxplot(X)



#mean

X.mean()

33.27133333

#Varience

X.var()

287.1466123

#Standard Deviation

X.std()

16.9454009

****

**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.**
2. **What can we say about the skewness of this dataset?**
3. **If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?**

Answer : (i) The data set has a first quartile (Q1) of approximately 5, a third quartile (Q3) of approximately 12, and a median (Q2) of 7. The interquartile range (IQR) is calculated as Q3 minus Q1, which results in 7. The second quartile range represents the median value.

(ii) The data set exhibits a right-skewed distribution where the median is shifted towards the left side. It is not a normal distribution.

(iii) If there were no outliers present in the given dataset, the positive skewness observed in the data would diminish, leading to a more normal distribution.

****

**Answer the following three questions based on the histogram above.**

1. **Where would the mode of this dataset lie?**
2. **Comment on the skewness of the dataset.**
3. **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.**

Answer :

(i) The most frequently occurring values in this data set are concentrated between the ranges of 5 to 10 and approximately 4 to 8.

(ii) The distribution of the data is skewed to the right, which means that the tail of the distribution extends more towards higher values. The average value (mean) is greater than the middle value (median), and the mode is also smaller than the mean and median.

(iii) Both distributions exhibit a right-skewed shape, indicating that the majority of the data points are concentrated towards the lower end, with a long tail stretching towards higher values. Additionally, both distributions contain outliers, which are extreme values that deviate significantly from the rest of the data. In a box plot, the median can be easily identified as the line dividing the box, while in a histogram, the mode (most frequent value) stands out as the highest peak.

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.)**

Answer : Given that one in 200 long-distance telephone calls is misdirected, the probability of a call being misdirected is 1/200. Therefore, the probability of a call not being misdirected is 199/200.

We want to find the probability that at least one in five attempted telephone calls reaches the wrong number. This is equivalent to finding the probability that none of the five calls are misdirected and subtracting it from 1.

Using the formula for the binomial probability, where n is the number of trials (5), p is the probability of success (1/200), and q is the probability of failure (199/200), we can calculate the probability as follows:

P(x) = ⁵Cₓ \* (1/200)ˣ \* (199/200)⁵⁻ˣ

To find the probability that none of the calls are misdirected (P(0)), we substitute x = 0 into the formula:

P(0) = ⁵C₀ \* (1/200)⁰ \* (199/200)⁵ = (199/200)⁵

Finally, we subtract P(0) from 1 to obtain the probability that at least one in five attempted telephone calls reaches the wrong number:

Probability = 1 - P(0) = 1 - (199/200)⁵ = 0.02475

Therefore, the probability that at least one in five attempted telephone calls reaches the wrong number is approximately 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?**
2. **Is the venture likely to be successful? Explain**
3. **What is the long-term average earning of business ventures of this kind? Explain**
4. **What is the good measure of the risk involved in a venture of this kind? Compute this measure**

Answer :

(i) The most likely monetary outcome of the business venture is $2000, which corresponds to the maximum probability of 0.3.

(ii) To determine if the venture is successful, we need to consider the monetary outcome (X) being positive. Let's denote success as S.

The probability of success is calculated as:

P(S) = P(X = 1000) + P(X = 2000) + P(X = 3000)

= 0.2 + 0.3 + 0.1

= 0.6

Since the probability of success (0.6) is greater than 0.5, the venture is likely to be successful.

(iii) The long-term average earnings of the business ventures can be represented by the expected value, denoted as E(X). The calculation for E(X) involves summing up the products of each possible outcome (X) and its corresponding probability (P(X)):

E(X) = ∑ X \* P(X)

= (1000 \* 0.2) + (2000 \* 0.3) + (3000 \* 0.1)

= 800

Therefore, the long-term average earnings of the business ventures is $800.

(iv) The risk involved in a venture can be quantified using the variance (Var(X)) and standard deviation (SD). The variance is calculated as the difference between the expected value of X squared (E(X²)) and the square of the expected value (E(X)):

Var(X) = E(X²) - [E(X)]²

= 2800000 - 800²

= 2160000

The standard deviation is the square root of the variance:

SD = √Var(X) ≈ √2160000 ≈ $1470

The high variance and standard deviation indicate a higher level of variability and hence a higher level of risk in the venture.