**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

df = pd.read\_csv('/content/set1Q1.csv')

df

# converting the Measure X column from string to float by replacing the % and dividing with 100

df['Measure X']= df['Measure X'].str.replace('%', '').astype(float) / 100

df

# Using boxplot to detect the outliers

import matplotlib.pyplot as plt

plt.boxplot(df['Measure X'], vert = False)

plt.title("Detecting outliers using Boxplot")

plt.xlabel('Measure X')

plt.show()

A graph with a line and a square

Description automatically generated with medium confidence

def colname(column):

    Q1 = column.quantile(0.25)

    Q3 = column.quantile(0.75)

    IQR = Q3 - Q1

    lower = Q1 - 1.5\*IQR

    upper = Q3 + 1.5\*IQR

    col1 = column[(column < lower) | (column > upper)]

    return list(col1)

out =colname(df['Measure X'])

print(out)

**output : [0.9136]**

# Mean

m = df['Measure X'].mean().round(3)

# Standard deviation

s = df['Measure X'].std().round(3)

# Variance

v = df['Measure X'].var().round(3)

print('Mean :', m,' Standard deviation :', s ,' Variance :', v)

**Output : Mean : 0.333, Standard deviation : 0.169, Variance : 0.029**



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: It is the quartile range where the maximum data points lies.

Mathematically, It is represented as the IQR = 3rd quartile – 1st quartile

From the box plot above IQR = 12 – 5

**= 7**

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

**Ans:**  The line inside the box represents the 2nd quartile which is also called median. Since, the line is towards the left side i.e., Median << mean. We can say the above box plot is right skewed.

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:**  The box plot will have no outliers and the 1st quartile of the box will move towards the left side.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

**Ans: “** The value which repeatedly occurs in a data set are called Mode”. From the graph it is approximately 4 to 8 of the variable ‘Y’.

1. Comment on the skewness of the dataset.

**Ans:** from the histogram the maximum frequency is on the left side and the it has the positive skewness.

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:** Both the plots show the outliers and the skewness. However the histogram shows the mode.

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:**  let us say 1 in the 200 long-distance telephone calls are misdirected. Then the probability of misdirected calls p = 1/200

the probability of not misdirected calls q = 1 – (1/200)

the number of attempts n = 5

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

P(x) = ⁿCₓ pˣ qⁿ⁻ˣ,

ⁿCₓ = !

x = 1 then after applying the values in the formula we get

P (1) = 0.0245037

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:** From the table we can say that the monetary outcome of the business venture is 2000 because the probability is high i.e., 0.3

1. Is the venture likely to be successful? Explain

**Ans:** Yes, the business will be in profitability because the sum of probability of returns are higher i.e., 0.2+0.2+0.1+0.3 = 0.8. that means there is 80% of chance to be successful.

1. What is the long-term average earning of business ventures of this kind? Explain
2. **Ans:** Average earning of business venture = Σ(x\*P(x))

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

**= 800**

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

**Ans:** import numpy as np

x = [-2000,-1000,0,1000,2000,3000]

prob = [0.1,0.1,0.2,0.2,0.3,0.1]

v = np.array([x,prob])

variance = np.var(v)

print('Variance: ',variance.round(3))

**Output :** 1520791.676

The variance is high, therefore the chances of risk is very high.