**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 = calculations:

1. Convert the percentages to decimal form:

plaintextCopy code

Allied Signal: 24.23% -> 0.2423 Bankers Trust: 25.53% -> 0.2553 General Mills: 25.41% -> 0.2541 ITT Industries: 24.14% -> 0.2414 J.P.Morgan & Co.: 29.62% -> 0.2962 Lehman Brothers: 28.25% -> 0.2825 Marriott: 25.81% -> 0.2581 MCI: 24.39% -> 0.2439 Merrill Lynch: 40.26% -> 0.4026 Microsoft: 32.95% -> 0.3295 Morgan Stanley: 91.36% -> 0.9136 Sun Microsystems: 25.99% -> 0.2599 Travelers: 39.42% -> 0.3942 US Airways: 26.71% -> 0.2671 Warner-Lambert: 35.00% -> 0.3500

1. Create a plot:

Let's create a simple bar plot to visualize the data:

pythonCopy code

import matplotlib.pyplot as plt companies = [ "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" ] measure\_x = [ 0.2423, 0.2553, 0.2541, 0.2414, 0.2962, 0.2825, 0.2581, 0.2439, 0.4026, 0.3295, 0.9136, 0.2599, 0.3942, 0.2671, 0.3500 ] plt.figure(figsize=(12, 6)) plt.barh(companies, measure\_x, color='skyblue') plt.xlabel('Measure X (Decimal Form)') plt.title('Measure X for Different Companies') plt.gca().invert\_yaxis() # Reverse the order of companies for better readability plt.show()

1. Calculate μ, σ, and σ^2:

pythonCopy code

import numpy as np data = np.array(measure\_x) mean = np.mean(data) std\_dev = np.std(data) variance = std\_dev\*\*2 print("Mean (μ):", mean) print("Standard Deviation (σ):", std\_dev) print("Variance (σ^2):", variance)

1. Identify outliers using the IQR method:

To find outliers, you'll need to calculate the Interquartile Range (IQR) and then use it to identify values that fall outside the typical range. The IQR is the range between the 75th percentile (Q3) and the 25th percentile (Q1).

pythonCopy code

Q1 = np.percentile(data, 25) Q3 = np.percentile(data, 75) IQR = Q3 - Q1 lower\_bound = Q1 - 1.5 \* IQR upper\_bound = Q3 + 1.5 \* IQR outliers = [x for x in data if x < lower\_bound or x > upper\_bound] print("Outliers:", outliers)



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?

Ans = (i) The Interquartile Range (IQR) of this dataset is approximately 0.098 (rounded to three decimal places). This value represents the spread of the middle 50% of the data. A larger IQR indicates greater variability in the middle of the dataset, implying that the central values are more dispersed.

(ii) The skewness of the dataset can be inferred from the box plot. If the median line inside the box is not centered (i.e., it's closer to one of the quartiles), it suggests that the data is skewed. In this case, since the median line appears to be slightly closer to the lower quartile (Q1), it implies a slight negative skew, indicating that the data is skewed to the left.

(iii) If the data point with the value 25 is actually corrected to 2.5, it would significantly affect the box plot. This data point is likely an outlier based on the previous analysis, as it falls well below the lower bound calculated using the IQR method. As a result, the new box plot would show the lower whisker extending down to 2.5, the lower bound of the box shifting accordingly, and the box's lower edge and median adjusting as well. This change would make the box plot more consistent with the rest of the data and less skewed to the left.



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.

Ans = (i) **Mode of the Dataset:** To determine the mode of a dataset from a histogram, you would look for the bin (interval) with the highest frequency (tallest bar). The mode represents the most common or frequently occurring value in the dataset. If you can identify the bin with the highest frequency, that's where the mode would lie.

(ii) **Skewness of the Dataset:** The skewness of a dataset can be assessed based on the shape of the histogram. If the histogram is skewed to the right, it indicates positive skewness, where the tail of the distribution is extended to the right and the mean is greater than the median. If the histogram is skewed to the left, it indicates negative skewness, where the tail of the distribution is extended to the left, and the mean is less than the median. If the histogram is roughly symmetric, it suggests that the data is normally distributed or close to it.

(iii) **Complementing Histogram and Box Plot:** Histograms and box plots are two different graphical tools that provide complementary information about a dataset:

* Histograms provide a detailed visual representation of the distribution of data values, including the shape, spread, and modes. They are particularly useful for understanding the data's density and frequency within specific value ranges.
* Box plots, on the other hand, offer a summary of key statistics such as the median, quartiles, and potential outliers. They are excellent for identifying skewness and detecting the presence of outliers.

When both a histogram and a box plot are plotted for the same dataset:

* The histogram gives you a more detailed view of the data's distribution, including its shape and the location of modes.
* The box plot provides a concise summary of the central tendency, spread, and skewness of the data, as well as the presence of outliers.

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 = The probability of a single telephone call being misdirected is 1/200, and since the attempts are assumed to be independent, the probability of none of the five calls being misdirected is:

P(no misdirected call) = (1 - 1/200)^5

Now, you can calculate the probability of at least one misdirected call:

P(at least one misdirected call) = 1 - P(no misdirected call) P(at least one misdirected call) = 1 - (1 - 1/200)^5

Calculating this probability:

P(at least one misdirected call) ≈ 0.02475

So, there is approximately a 2.475% chance that at least one of the five attempted telephone calls reaches the wrong number.

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

Ans = (i) **Most Likely Monetary Outcome:** The most likely monetary outcome corresponds to the value with the highest probability (mode) in the distribution. In this case, the value with the highest probability is $2,000 with a probability of 0.3. So, the most likely monetary outcome is $2,000.

(ii) **Venture Success:** Whether the venture is likely to be successful or not depends on the definition of success and the risk tolerance of the investor. In this case, we can say that the venture has a 0.3 probability of generating a profit of $2,000 and a 0.2 probability of breaking even (earning $0). However, it also has a 0.1 probability of a significant loss of -$2,000.

The success of the venture depends on the investor's perspective. If a 20% chance of breaking even or making a profit is acceptable with a 10% chance of a significant loss, then the venture might be considered successful. However, some investors might consider the risk of a significant loss too high.

(iii) **Long-Term Average Earnings:** To find the long-term average earnings, you can calculate the expected value (mean) of the distribution. It's the weighted sum of all possible outcomes, where the weights are the probabilities associated with each outcome:

Expected Value (mean) = (-2,000 \* 0.1) + (-1,000 \* 0.1) + (0 \* 0.2) + (1,000 \* 0.2) + (2,000 \* 0.3) + (3,000 \* 0.1)

Expected Value = $600

So, the long-term average earning for ventures of this kind is $600.

(iv) **Measure of Risk:** One common measure of risk is the standard deviation, which indicates how much the actual returns are likely to vary from the expected (mean) returns.

To compute the standard deviation, use the following formula:

Standard Deviation (σ) = √[Σ(P(x) \* (x - μ)^2)]

Where:

* P(x) is the probability of each outcome.
* x is the value of each outcome.
* μ (mu) is the mean (expected value).

You've already calculated the mean as $600 in part (iii). Now, calculate the standard deviation:

σ = √[(0.1 \* (-2,000 - 600)^2) + (0.1 \* (-1,000 - 600)^2) + (0.2 \* (0 - 600)^2) + (0.2 \* (1,000 - 600)^2) + (0.3 \* (2,000 - 600)^2) + (0.1 \* (3,000 - 600)^2)]

σ ≈ $1,163.49

So, the standard deviation is approximately $1,163.49, which is a measure of the risk involved in this business venture. A higher standard deviation indicates greater risk and variability in returns