**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% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name of company** | **Measure X** |  | **Find the Outlier** |  | **Define Outlier Boundaries** |
| Allied Signal | 24.23% |  | Q1(25th percentile) |  | Lower Bound |
| Bankers Trust | 25.53% |  | 0.2547 |  | 12.71% |
| General Mills | 25.41% |  |  |  |  |
| ITT Industries | 24.14% |  | Q3(75th percentile) |  | Upper Bound |
| J.P.Morgan & Co. | 29.62% |  | 0.33975 |  | 46.73% |
| Lehman Brothers | 28.25% |  |  |  |  |
| Marriott | 25.81% |  | IQR |  |  |
| MCI | 24.39% |  | 0.08505 |  |  |
| Merrill Lynch | 40.26% |  |  |  |  |
| Microsoft | 32.95% |  |  |  |  |
| Morgan Stanley | 91.36% |  | **Mean** |  |  |
| Sun Microsystems | 25.99% |  | 33.27% |  |  |
| Travelers | 39.42% |  | **Variance** |  |  |
| US Airways | 26.71% |  | 2.68% |  |  |
| Warner-Lambert | 35.00% |  | **Standard Deviation** |  | Conclusion : |
| . | . |  | 16.37% |  | **The only outlier is 91.36% (Morgan Stanley).** |



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?

**(i) Inter-Quartile Range (IQR):**

The IQR is the difference between the third quartile (Q3) and the first quartile (Q1).

* From the box-plot:
  + **Q3 (Upper edge of the box)** ≈ 12
  + **Q1 (Lower edge of the box)** ≈ 6

Thus, IQR=Q3−Q1=12−6=6.

**Interpretation**: This value implies that the middle 50% of the data points lie within a range of 6 units.  
  
**(ii) Skewness:**

The box plot suggests slight positive skewness (right skew) because:

* The whisker on the right side is slightly longer than the whisker on the left.
* There is a potential outlier on the right (value 25).

**(iii) Effect of correcting the data point from 25 to 2.5:**

* The current potential outlier (25) would no longer exist.
* The maximum value would likely reduce to the next largest data point (possibly around 15).
* The right whisker would shorten significantly, and the box-plot would appear more symmetric.
* The dataset might lose its skewness and become approximately symmetrical.



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.

**(i) Mode of the dataset:**

The mode is the value or range with the highest frequency.

* From the histogram, the highest bar lies within the interval 5−7.55 - 7.55−7.5.
* Hence, the mode of this dataset lies approximately **between 5 and 7.5**.

**(ii) Skewness of the dataset:**

The histogram shows a clear **right skew (positive skew)** because:

* The right tail (values greater than 15) is longer and has a smaller frequency compared to the left.
* Most of the data is concentrated towards lower values, with a few higher values stretching the distribution to the right.

**(iii) How the histogram and box-plot complement each other:**

1. **Histogram**:
   * It provides a detailed view of the frequency distribution of the data.
   * It allows for identifying the shape (e.g., skewness, modality) and clustering of the data.
2. **Box-Plot**:
   * It provides a summary of key statistical measures such as median, quartiles, and IQR.
   * It identifies outliers and visually represents the spread of the data.
3. 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.)

**Given:**

1. Probability of a misdirected call P(misdirected)=1200=0.005.
2. Probability of a correctly directed call P(correct)=1−0.005=0.995.
3. Number of attempts n=5n = 5n=5.

We want the probability that at least one call is misdirected. This is the complement of the probability that **none of the calls** are misdirected.

P(none misdirected)=(0.995)5≈0.9753

P(at least one misdirected)=1−0.9753≈0.0247

The probability that at least one of the five attempted telephone calls is misdirected is approximately **0.0247**, or **2.47%**.

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

**(i) Most likely monetary outcome:**

The most likely outcome corresponds to the highest probability P(x)P(x)P(x).  
From the table:

* P(x=2000)=0.3P(x = 2000) = 0.3P(x=2000)=0.3, the highest probability.

Thus, the most likely monetary outcome is **$2,000**.

**(ii) Is the venture likely to be successful?**

* A business venture is likely to be successful if the expected returns (long-term average earnings) are **positive**.
* We calculate the expected value E(X)E(X)E(X) in part (iii). If E(X)>0E(X) > 0E(X)>0, the venture is considered likely to succeed.

**(iii) Long-term average earning (expected value):**

The expected value E(X) is calculated as:

E(X)=∑(x⋅P(x))

Substituting the given values:

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=800

Thus, the **long-term average earning** is **$800**. This indicates a positive return, meaning the venture is likely to be **successful** on average.

**(iv) Good measure of risk (variance and standard deviation):**

The **variance** Var(X) is a good measure of risk, as it quantifies the spread of outcomes. Standard deviation, σ(X) , is the square root of variance and provides an interpretable measure of variability in the same units as the outcomes.

1. **Formula for variance:**

Var(X)=∑(P(x)⋅(x−E(X))2)

1. **Steps:**
2.  First, calculate (x−E(X))2 for each x.
3.  Multiply P(x)⋅(x−E(X))2 , then sum up.

Var(X)=∑P(x)⋅(x−E(X))2=1660000

1. **Standard deviation**:

σ(X)=Var(X)​=1660000​≈1289.9

**Final Answers:**

1. **Most likely outcome:** $2,000.
2. **Venture success:** Likely, as the long-term average earning is positive ($800).
3. **Long-term average earning:** $800.
4. **Risk measure:** The standard deviation of returns is approximately **$1,289.90**, indicating a moderate level of variability in outcomes.