**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: See Set\_1.ipynb file



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:

1. First Quartile, Q1 = 5 (approx.)

Third Quartile, Q3 = 12 (approx.)

Hence IQR = Q3 - Q1 = 12 - 5 = 7

This also implies the middle value i.e. the Second Quartile, Q2 = 7

1. We can clearly say that the dataset is Positively skewed since the median is towards left of the Normal distribution position(center)
2. Since the datapoint is an outlier at the upper extreme, reducing it to 2.5 will significantly reduce the spread of the data and it will change the Quartile values also.



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:

1. Mode is the datapoint with the highest frequency. If we need to estimate it without knowing the exact datapoints, we can say that mode will lie between Y=4 to Y=8
2. Since there is a long tail to the right of the dataset, we can determine that it is a positive skewed dataset.
3. The histogram clearly depicts that the data is skewed and there are potential outliers and the box plot highlights the specific values of the quartiles and median, which helps in identifying the center and spread of the data. The box plot also identifies the exact values of the outliers which can only be estimated by the histogram.
4. 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 that a call is misdirected is 1/200.

The probability of call not being misdirected = 1 - 1/200 = 199/200

The probability of at least 1 call being misdirected = 1 - (Probability of no calls being misdirected)

= 1 - (199/200)^5

= 1 - 0.9752 (approx.)

= 0.0247

So, there is a 2.47% chance that at least one in five calls goes to 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:

1. Since there is a high probability at x=2000, we can say that most likely there would be a profit of $2000 in this business venture.
2. If we sum up the probabilities of Profit and Loss, we can clearly see that there is a greater probability we would get a profit in this business venture since P(Profit)>P(Loss). Hence, we can determine that it will most likely be successful.
3. We can determine the long-time average earning by multiplying the expected profit/loss with its respective probability and summing these values.  
   (-2000)\*0.1 + (-1000)\*0.1 + (0)\*0.2 + (1000)\*0.2 + (2000)\*0.3 + (3000)\*0.1 = 800  
   Hence, we can get an average earning of $800 in the long term.
4. For the above data, we can calculate standard deviation to know the difference between expected returns and actual returns.

Variance = ∑(x-μ)2/(n-1) = 3500000

Standard Deviation = √Variance = 1870.82

So expected returns are $800 and standard deviation is $1870.82 which means there is high risk involved in this business venture.