**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 :

Mean (µ) : 33.27133%

SD (σ) : 16.9454%

Var (σ²) : 02.8714%

Outlier : 91.36% (Morgan Stanley)

***What is an outlier?***

An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

**What are the criteria to identify an outlier?**

* Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile and below the 1st quartile
* Data point that falls outside of 3 standard deviations. we can use a z score and if the z score falls outside of 2 standard deviation

**What is the reason for an outlier to exists in a dataset?**

An outlier could exist in a dataset due to

Variability in the data

An experimental measurement error

**What is the impact of an outlier?**

causes serious issues for statistical analysis

skew the data,

significant impact on mean

significant impact on standard deviation.

**How can we identify an outlier?**

using scatter plots

using Z score

using the IQR interquartile range

Using Scatter Plot

We can see the scatter plot and it shows us if a data point lies outside the overall distribution of the dataset

Using Z score

**Formula for Z score = (Observation — Mean)/Standard Deviation**

**z = (X — μ) / σ**

EX:-

import numpy as np

import pandas as pd

data\_1= [10,12,12, 13,12,11,14,13,15,10,10, 10, 100,12, 14,13, 12,10, 10,11,12,15,12,13,12,11,14,13,15,10, 15,12,10,14,13,15,10]

len(data\_1)

o/p:- 37

we write a function that takes numeric data as an input argument.

we find the mean and standard deviation of the all the data points

We find the z score for each of the data point in the dataset and if the z score is greater than 3 than we can classify that point as an outlier. Any point outside of 3 standard deviations would be an outlier.

import numpy as np

import pandas as pd

outliers=[]

def detect\_outlier(data\_1):

threshold=3

mean\_1 = np.mean(data\_1)

std\_1 =np.std(data\_1)

for y in data\_1:

z\_score= (y - mean\_1)/std\_1

if np.abs(z\_score) > threshold:

outliers.append(y)

return outliers

we now pass dataset that we created earlier and pass that as an input argument to the detect\_outlier function

outlier\_datapoints = detect\_outlier(dataset)

print(outlier\_datapoints)

o/p:- [100]

output of the outlier\_datapoints

Using IQR

IQR tells how spread the middle values are. It can be used to tell when a value is too far from the middle.

An outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile.

we will use the same dataset

step 1:

Arrange the data in increasing order

Calculate first(q1) and third quartile(q3)

Find interquartile range (q3-q1)

Find lower bound q1\*1.5

Find upper bound q3\*1.5

Anything that lies outside of lower and upper bound is an outlier

Fist sorting the dataset

sorted(dataset)

Finding first quartile and third quartile

q1, q3= np.percentile(dataset,[25,75])

q1 is 11 and q3 is 14

Find the IQR which is the difference between third and first quartile

iqr = q3 - q1

iqr is 3

Find lower and upper bound

lower\_bound = q1 -(1.5 \* iqr)   
upper\_bound = q3 +(1.5 \* iqr)

lower\_bound is 6.5 and upper bound is 18.5, so anything outside of 6.5 and 18.5 is an outlier.

--------------------------------------------------------------------

import numpy as np

import pandas as pd

data\_1= [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]

len(data\_1)

print(np.mean(data\_1)) #33.27133333333333

print(np.median(data\_1)) #26.71

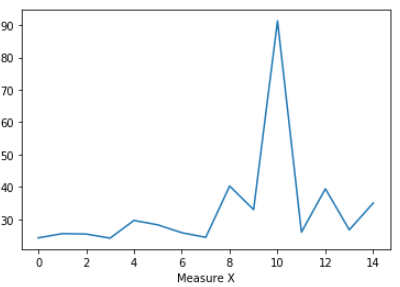
print(np.var(data\_1)) # 268.00350488888887

print(np.std(data\_1)) #16.370812590976932

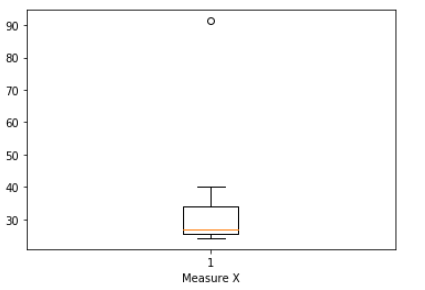
data\_1= [24.23,25.53,25.41,24.14,29.62,28.25,25.81,24.39,40.26,32.95,91.36,25.99,39.42,26.71,35.00]

plt.plot(data\_1)

plt.xlabel('Measure X')



plt.boxplot(data\_1)





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: IQR = (12-5) = 7 i.e. 50% of the whole data is present inside this range.

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

Ans : The data is positively 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 : May be the boxplot is lacking its symmetry due to the outlier 25. So replacing 25 by 2.5 , we may get changed mean and median value over there. May be mean approximately equal to the median value in this case.

The interquartile range is a measure of where the “[middle fifty](https://www.statisticshowto.datasciencecentral.com/middle-fifty/)” is in a data set. Where a [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/)is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That’s why it’s preferred over many other [measures of spread](https://www.statisticshowto.datasciencecentral.com/measures-of-spread/) (i.e. the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)or [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)) when reporting things like school performance or SAT scores.

The interquartile range formula is the first [quartile](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/)subtracted from the third [quartile](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/):

IQR = Q3 – Q1.

* Step 1: Put the numbers in order.  
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
* Step 2: Find the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median).  
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
* Step 3: Place parentheses around the numbers above and below the median.  
  Not necessary statistically, but it makes Q1 and Q3 easier to spot.  
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
* Step 4: Find Q1 and Q3  
  Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.  
  (1, 2, 5, 6, 7),  9, ( 12, 15, 18, 19, 27). Q1 = 5 and Q3 = 18.
* Step 5: Subtract Q1 from Q3 to find the interquartile range.  
  18 – 5 = 13.

What if I Have an Even Set of Numbers?

Sample question: Find the IQR for the following data set: 3, 5, 7, 8, 9, 11, 15, 16, 20, 21.

Step 1: Put the numbers in order.  
3, 5, 7, 8, 9, 11, 15, 16, 20, 21.

Step 2: Make a mark in the center of the data:  
3, 5, 7, 8, 9, | 11, 15, 16, 20, 21.

Step 3: Place parentheses around the numbers above and below the mark you made in Step 2–it makes Q1 and Q3 easier to spot.  
(3, 5, 7, 8, 9), | (11, 15, 16, 20, 21).

Step 4: Find Q1 and Q3  
Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.  
(3, 5, 7, 8, 9), | (11, 15, 16, 20, 21). Q1 = 7 and Q3 = 16.

Step 5: Subtract Q1 from Q3.  
16 – 7 = 9.  
This is your IQR.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Ans : The Mode lies in between 5 to 7

1. Comment on the skewness of the dataset.

Ans : Data is positively skewed

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:In Histogram we are able to get the modal value bot not exactly the median value. But incase of the Boxplot we can obtain the median value. Looking either of the plot we can comment on the distribution of the data. In histogram its little bit difficult to comment outlier, whereas in boxplot we can see the outlier in dot or oval shape.

Looking at the histogram we can’t comment the Interquartile range (major 50% of the data) but incase of boxplot we can define the Inter quartile Range.

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: We are reaching one wrong number in 200 calls. So, prob. of reaching a wrong no is 1/200.

Hence in cycle of 40 attempts of 5 calls we are reaching one wrong number.

So, we conclude that probability of reaching one wrong number in 5 calls is 1/40 = 0.025.

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: $2,000

1. Is the venture likely to be successful? Explain

Ans: Probability ofGetting profit is (0.2+0.3+0.1) = 0.6. So, we can say the venture is likely to be successful.

1. What is the long-term average earning of business ventures of this kind? Explain

Ans: As we know E(X) = Σ ( Pi Xi )

E(X) = 800 which is positive.

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