Q1) Calculate Skewness, Kurtosis & draw inferences on the following data

a. Cars speed and distance

****

**skewness**

speed -0.117510 – negative skewness – data are skewed to the left – left tail is longer

dist 0.806895 – positive skewness – data are skewed to the right – right tail is longer

**kurtosis**

speed -0.508994 negative kurtosis – less peaked – lighter and thinner tails

dist 0.405053 – positive kurtosis – highly peaked – heavier tails

kurtosis(speed) - 2.422853

kurtosis(dist)- 3.248019 – positive kurtosis > 3 - leptokurtic

b. Top Speed (SP) and Weight (WT)

****

cartopspeed.skew()

SP 1.611450 – positive skewness – data are skewed to the right

WT -0.614753 – negative skewness – data are skewed to the right

cartopspeed.kurtosis()

SP 2.977329 – positive kurtosis – highly peaked – heavier tails

WT 0.950291- positive kurtosis – highly peaked – heavier tails

> skewness(SP) - 1.581454 - positive skewness – data are skewed to the right

> skewness(WT) - -0.6033099 – negative skewness – data are skewed to the right

> kurtosis(SP) - 5.723521- positive kurtosis >3 - leptokurtic

> kurtosis(WT) - 3.819466 - positive kurtosis >3 - leptokurtic

Q2) Draw inferences about the following boxplot & histogram

**Inferences from Histogram**

1. **Frequency of ChickWeight is maximum in the range 50-100 which is 200.**
2. **Frequency of ChickWeight started decreasing gradually as weight increases and it is almost zero when weight = 400**
3. **Distribution of Chickweight is skewed to the right**





Inferences from boxplot

1. There are outliers in data
2. Data is skewed to the right

**Q3)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

Sample size n = 2000

Average weight of the sample = 200 pounds

Standard deviation(sd) = 30 pounds

Standard error = sd/sqrt(n) = 30/sqrt(2000) = 0.67082

**94% - confidence interval**

Alpha = 1-confidencelevel/100 = 1-94/100 = 0.06

Critical probability = 1-alpha/2 = 1-0.06/2 = 0.97

Critical value = 1. 5554

Margin of Error = critical value \* standard error = 1.56\*0.67082 = 1.0464792

94% Confidence Interval is 200+/-1.046

**98% - confidence interval**

Alpha = 1-confidenceinterval/100 = 1-98/100 = 0.02

Critical probability = 1-alpha/2 = 1-0.02/2 = 0.99

Critical value = 2.055

Margin of Error = critical value \* standard error = 2.055 \* 0.67082 = 1.3785351

98% confidence interval is 200+/-1.379

**96% confidence interval**

Alpha = 1-confidenceinterval/100 = 1-0.96/100 = 0.04

Critical probability = 1-alpha/2 = 1-0.04/2 = 0.98

Critical value = 1.7515

Margin of Error = critical value \* standard error = 1.7515 \* 0.67082 = .17494

96% confidence interval is 200+/-1.749

**Q4)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.

print(np.mean(marks))

print(np.median(marks))

print(max(marks,key=marks.count))

print(np.std(marks))

print(np.var(marks))

Mean = 41.0

Median = 40.5

Mode = 41

Standard deviation = 4.910306620885412

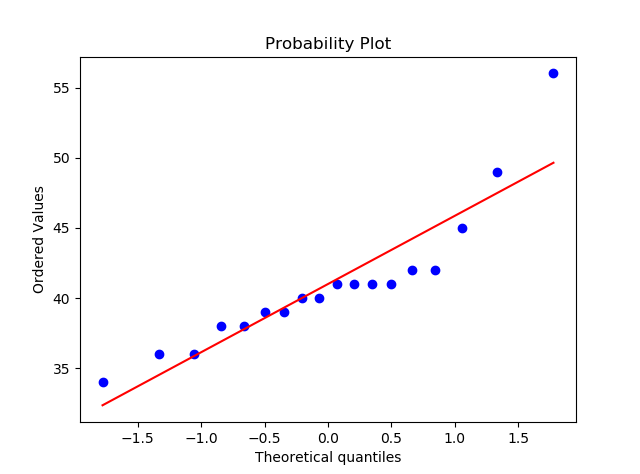
Variance = 24.11111111111111

1. What can we say about the student marks?

import scipy.stats as stats

import pylab

stats.probplot(marks,dist="norm",plot=pylab)



As we can see from the graph,

1. Marks forms a curve over the straight line. Hence, it does not follow a normal distribution. 56 is an outlier
2. Mean, median and mode are almost equal, hence marks is almost symmetrical

Q5) What is the nature of skewness when mean, median of data are equal?

Skewness = 0. It is a symmetric distribution

Q6) What is the nature of skewness when mean > median?

Data are skewed to the right

Q7) What is the nature of skewness when median > mean?

Data are skewed to the left

Q8) What does positive kurtosis value indicates for a data?

1. It indicates that data distribution is highly peaked and it has thicker tails than a normal distribution.
2. Leptokurtic distributions have positive kurtosis value

Q9) What does negative kurtosis value indicates for a data?

1. It indicates that data distribution is flatter(less peaked) and it has thinner tails than a normal distribution.
2. Platykurtic distributions have positive kurtosis value

Q10) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

1. Range is (1-20)
2. Median is approximately 15.5
3. Minimum value = 1, Lower quartile (Q1) = 10, Median = 15.5, Q3 = 18.2(approx.)

What is nature of skewness of the data?

1. Skewed left as median is in the upper side of the box and the whisker is shorter on the upper side of the box

What will be the IQR of the data (approximately)?

IQR = Q3-Q1 = 18.2 – 10 = 8.2

Q11) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

|  |  |  |
| --- | --- | --- |
| Parameter | Boxplot1 | Boxplot2 |
| Median | 262.5 | 262.5 |
| Range | 237.5 to 287.5 | 175 to 350 |
| Q1 | 250 | 225 |
| Q3 | 280 | 325 |
| IQR(Q3-Q1) | 30 | 100 |
| Skewness | Symmetric | Symmetric |

Inferences :

1. Data in boxplot 2 is wide spread than data in box plot 1
2. No outliers in both data

Q12)



Answer the following three questions based on the boxplot above.

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

(Q3-Q1) = 12-5 = 7

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

Median is 7.5 and lower whisker is 0 , upper whisker is 19(approx)

Median is closer to lower whisker and hence data is positively skewed(i.e. skewed to the right)

1. If it was found that the data point with the value 25 is actually 2.5, how would the new boxplot be affected?

There would not be any outliers

Q13)



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?

Between 4 and 8 (approx.). it is bi-modal as there are two ranges with maximum frequencies

1. Comment on the skewness of the dataset.

Data are skewed to the right

1. Suppose that the above histogram and the boxplot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.
2. Histogram cannot identify the outliers in the dataset where as boxplot will identify them
3. Histogram can be used to identify mode of the distribution whereas boxplot cannot identify mode
4. Skewness of data can be found accurately from boxplot rather using histogram alone