| Activity | Data Type |
| --- | --- |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continuous |
| Weight of Gold | Continuous |
| Distance between two places | Continuous |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

| Data | Data Type |
| --- | --- |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Nominal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Discrete |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | INterval |
| Years of Education | Ratio |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

solution:

sample space={(H,H,H) (H,H,T) (H,T,H) (H,T,T) (T,H,H) (T,H,T) (T,T,H) (T,T,T)}

A=two heads and one tail

={(H,H,T) (H,T,H) (T,H,H)}

P(A)=⅜

P(A)=0.375

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

Solution: sample space ={(1,1)(1,2)(1,3)(1,4)(1,5)(1,6)

(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)

(3,1)(3,2)(3,3)(3,4)(3,5)(3,6)

(4,1)(4,2)(4,3)(4,4)(4,5)(4,6)

(5,1)(5,2)(5,3)(5,4)(5,5)(5,6)

(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)}

A: sum is Equal to 1

B:sum is Less than or equal to 4

C:Sum is divisible by 2 and 3

P(A)=0/36

P(A) =0

P(B)=6/36

P(B)=1/6

P(B)= 0.1666

P(C)=6/36

P(C)=0.1666

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

Solution:

total no of chances to draw ***2*** balls at random from ***7*** coloured balls,**n(s)=7C2=21**

let **E** be an event to draw ***2*** balls other than **blue** .

no of chances to draw two balls other than blue are,**n(E)=2C2+3C2+2C1.3C1=1+3+6=10**

the probability that none of balls drawn is blue is,**p(E)=n(E)/n(s)=10/21**

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

| CHILD | Candies count | Probability |
| --- | --- | --- |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Solution:

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

= 3.09

**Expected number of candies for a randomly selected child = 3.09**

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Solution:** Basic Statistics\_Level-1\_Q7.ipynb file

**Use Q7.csv file**

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Solution:**

Expected Value = ∑ (probability \* Value) = ∑ P(x). E(x)

There are 9 patients,

Probability of selecting each patient = 1/9

Ex: 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x):  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

Expected Value = (1/9)\*108 + (1/9)\*110  + (1/9)\*123 + (1/9)\*134 + (1/9)\*135 + (1/9)\*145 + (1/9)\*167 + (1/9)\*187 + (1/9)\*199

= (1/9) (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9)\*(1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

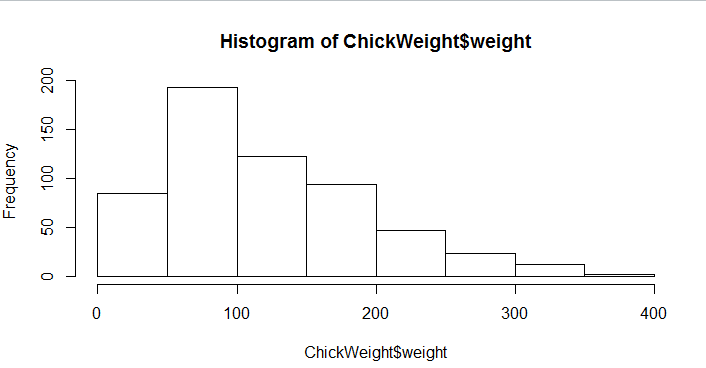
**Solution:** Basic Statistics\_Level-1\_Q9\_a.ipynb file

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Solution: Basic Statistics\_Level-1\_Q9\_b.ipynb file**

**Q10) Draw inferences about the following boxplot & histogram**

**Histogram:**

1.ChickWeight data is right skewed or positively skewed.

2.More than 50% ChickWeight is between 50 to 150.

3.Most of the ChickWeight is between 50 to 100.



**Boxplot:**

1. The data is right skewed.

2. There are outliers at upper side.

**Q11)** 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?

**Solution:**  
The information given is:

* Sample **mean**of .
* Sample **standard deviation** of .
* Sample **size**of .

The **interval** is:



* In which **t** is the critical value for the two-tailed confidence interval.

Considering a **94%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 1.8916**, hence:





The **94%** confidence interval is **(198.73, 201.27).**

Considering a **96%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.0673**, hence:





The **96%** confidence interval is **(198.61, 201.39).**

Considering a **98%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.3452**, hence:





The **98%** confidence interval is **(198.43, 201.57).**

**Q12)** 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.

Solution:

**Mean:**

As we know,

Mean= Sum of the terms/ Number of the terms.

Mean= (34+36+36+38+38+39+39+40+40+41+41+41+41+42+45+49+56)/18= 738/18=**41**

**Median:**

For median first I have to arrange in ascending order but here scores are already in ascending order so I directly calculated the median.

Median = (9th+ 10th Term)/2 = (40+41)/2 = **40.5**

**Variance:**

Mean(m)=41

Scores(s) s-m (s-m)^2

34 -7 49

36 -5 25

36 -5 25

38 -3 9

38 -3 9

39 -2 4

39 -2 4

40 -1 1

40 -1 1

41 0 0

41 0 0

41 0 0

41 0 0

42 1 1

45 4 16

49 8 64

56 15 225

Sum 0 433

Variance = 433/17= **25.47**

standard deviation:

As We know

standard deviation= (Variance)^(1/2)

So standard deviation= (24.05) ^ (1/2) = **5.05**

1. What can we say about the student marks?

**Solution:** Basic Statistics\_Level-1\_Q12.ipynb file

Q13) What is the nature of skewness when mean, median of data are equal?

**Solution**:

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero.

When mean, median and mode is equal then it is a normal distribution.

So, the nature of skewness is **zero** when mean and median of data is equal.

Q14) What is the nature of skewness when mean > median ?

**Solution**:

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

A positively skewed distribution is the distribution with the tail on its right side. The value of skewness for a positively skewed distribution is greater than zero.

When the value of mean is greater than median and mode then it is called positive skewed.

So the nature of skewness is **positive** when mean > median.

Q15) What is the nature of skewness when median > mean?

**Solution:**

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

A negatively skewed distribution is the distribution with the tail on its left side. The value of skewness for a negatively skewed distribution is less than zero.

When the value of mean is less than median and mode then it is called negative skewed.

So the nature of skewness is **negative** when median > mean.

Q16) What does positive kurtosis value indicates for a data ?

**Solution:**

Kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. It can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution).

Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

 Excess kurtosis = Kurt – 3

Leptokurtic is having very long and skinny tails, which means there are more chances of outliers.

**Positive values of kurtosis** indicate that distribution is **peaked and possesses thick tails**. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.

Q17) What does negative kurtosis value indicates for a data?

**Solution:**

Kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. It can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution).

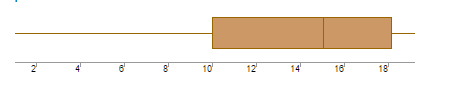
Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

 Excess kurtosis = Kurt – 3

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

A distribution with a **negative kurtosis** value indicates that **the distribution has lighter tails than the normal distribution**

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

**Solution:**

a) first 25% of data value is less than 10, next 25% of data value lies between 10 to 15.4, next 25% of data value lies between 15.4 to 18.2 and last 25% of data value is greater than 18.

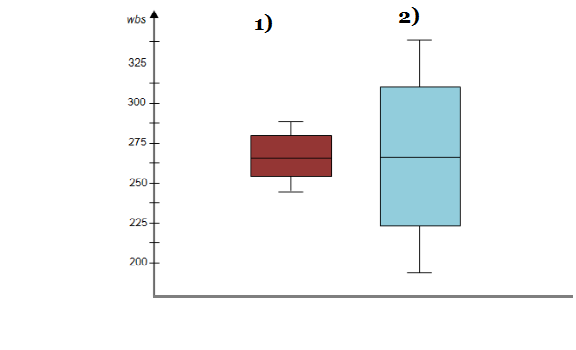
b) According to boxplot, median > mean.

So, the nature of skewness is negatively skewness.

c) Here Q1=10, Q2=15.4 and Q3=18.2

So, IQR is 18.2-10=8.2

Q19) Comment on the below Boxplot visualizations?



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

**Solution:** According to 1st boxplot:

The range of value is from 245 to 290 in which first 25%of data value lies between 245 to 255 and next 25% data value lies between 255 to 265 and next 25% data value is lies between 265 to 280 and last 25% of data values is lies between 280-290.

According to 2nd Boxplot:

The range of data value is from 190 to 340 in which first 25%of data value lies between 190 to 225 and next 25% data value lies between 225 to 255 and next 25% data value is lies between 255 to 305 and last 25% of data values is lies between 305-340.

Therefore, by comparing of both boxplot, I can say that the median value of both distribution is same but in 1st distribution most of value is similar or nearly comparison of 2nd distribution.

The IQR for 1st distribution is 280-255=25

The IQR for 2nd distribution is 305-225=80

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

**Solution:** Basic Statistics\_Level-1\_Q20.ipynb file

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Solution:** Basic Statistics\_Level-1\_Q21\_a.ipynb file

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**Solution:** Basic Statistics\_Level-1\_Q21\_b.ipynb file

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Solution:** Basic Statistics\_Level-1\_Q22.ipynb file

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**Solution:** Basic Statistics\_Level-1\_Q23.ipynb file

a)The sample size is n=25,

So the degrees of freedom is n−1=25−1=24

Thus, we are interested in the quantity

t(α/2)=t(0.05/2)=t(0.025)

for a t-distribution with 24 degrees of freedom.

Upon using a t-table, we see that the critical t-value for this 95% confidence interval is t(α/2)=2.064.

b)Upon using a t-table, we see that the critical t-value for this 96% confidence interval is t(α/2)=2.171.

c)Upon using a t-table, we see that the critical t-value for this 99% confidence interval is t(α/2)=2.797.

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

**Solution:** Basic Statistics\_Level-1\_Q24.ipynb file