|  |  |
| --- | --- |
| 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 | Category |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Category |

# **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 | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval:::(Measure that the difference between an IQ of 80 and 90 is the same as an IQ of 90 and 100. If IQ were ordinal, their designations could not be meaningful and outcome will be if someone is more or less smart than someone else, there would be no numerical measure.) |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal::::(Assumption: time of day is morning, evening, dawn dusk etc..).However, if we consider the time of day as continuous numerical value then the data type will be Interval.) |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval(Similar to IQ score) |
| Years of Education | Interval(determines distance between the values) |

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

*ANS: When 3 coins are tossed, the sample space is 8, and the possibilities of two head and a tail is 3 times. {HHT,HTH,THH} Hence the probability is P(tow heads & a tail) = 3/8= .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

*ANS:*

1. *0 probability of getting sum of 1 froma 2 dice roll.*
2. *Sample space={(1,1),(1,2),(1,3),(2,1),(3,1),(2,2)}*

*required probability=6/36=1/6=****0.167 (less than equal to 4)***

1. *sample space={(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}, divisible by both 2 and 3.*

*P(sum divisible by 2 and 3)=6/36=1/6=****.167***

# **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?**

*ANS: P (first ball drawn not blue) = 5/7 P (second ball drawn not blue) = 4/6 (as one ball is out, then total remaining ball will be 6) P (none of the balls drawn is blue) = (5/7) \*(4/6) = 10/21=.476*

**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

*ANS::*

*expected number of candies for a randomly selected child*

*= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.1*

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

*= 3.09 ~ 3*

***Expected number of candies for randomly selected child is 3***

**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.

**Use Q7.csv file**

***Mean:***

*Points 3.596563*

*Score 3.217250*

*Weigh 17.848750*

*dtype: float64*

*--------------------------------*

***Median:***

*Points 3.695*

*Score 3.325*

*Weigh 17.710*

*dtype: float64*

*--------------------------------*

***Standard deviation:***

*Points 0.534679*

*Score 0.978457*

*Weigh 1.786943*

*dtype: float64*

*--------------------------------*

***Variance:***

*Points 0.285881*

*Score 0.957379*

*Weigh 3.193166*

*dtype: float64*

*--------------------------------*

***Mode\_Point****::: ModeResult(mode=array([3.07]), count=array([3]))*

*--------------------------------*

***Mode\_Score****::: ModeResult(mode=array([3.44]), count=array([3]))*

*--------------------------------*

***Mode\_Weigh****::: ModeResult(mode=array([17.02]), count=array([2]))*

***Points\_Range****: 2.17*

***Score\_Range****: 3.91*

***Weigh\_Range****: 8.4*

***Inferences:***

1. *The Points column has the lowest deviation while the Weigh column has the highest deviation.*
2. *All "Points, score and weigh" are +ve scewed and doesn't follow emperical rule*
3. *None of the data is nomally distributed*

**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?*

*Expected value = Sum( Probalility \* value)*

*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)19*

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

*= (1/9) ( 1308)*

*= 145.33*

##### **Expected weight of a randomly chosen patient =145.33 pounds**

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

**Cars speed and distance**

Use Q9\_a.csv

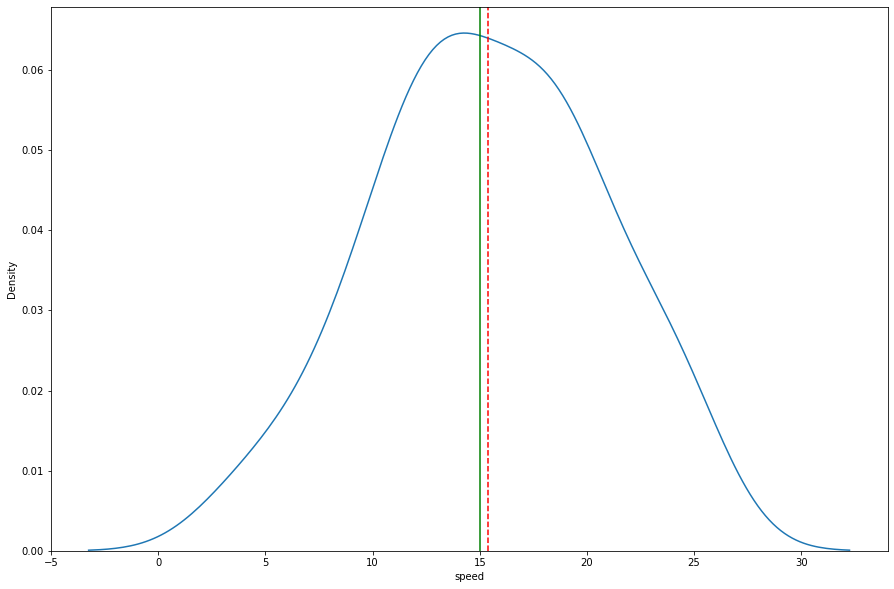
SP and Weight(WT)

Use Q9\_b.csv

### ***Inference SPEED***

### *"A general guideline for skewness is that if the number is greater than +1 or lower than –1, this is an indication of a substantially skewed distribution. For kurtosis, the general guideline is that if the number is greater than +1, the distribution is too peaked.*

### *Hence for case1 the distribution is left skewed and not peaked*

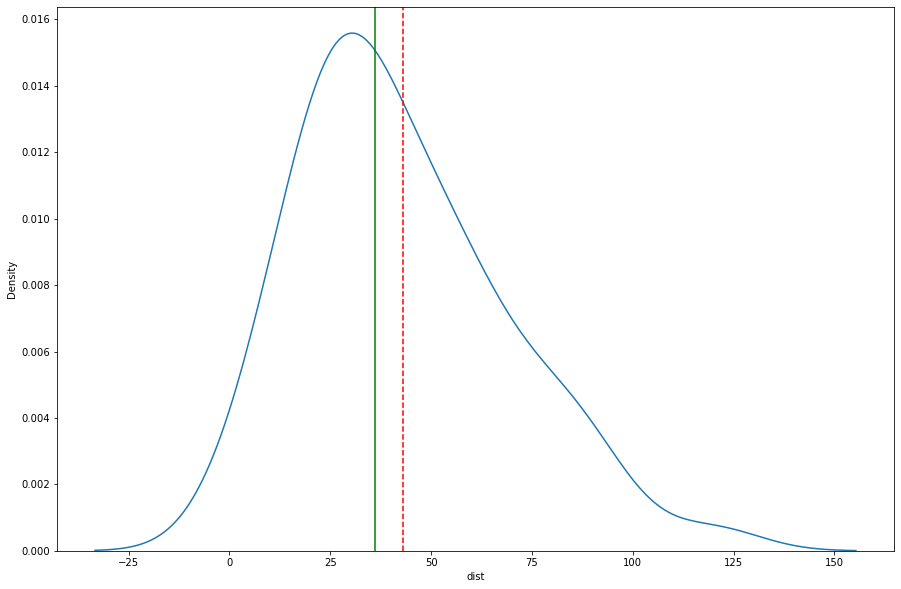
****

### ***Inference DIST***

###### *"A general guideline for skewness is that if the number is greater than +1 or lower than –1, this is an indication of a substantially skewed distribution. For kurtosis, the general guideline is that if the number is greater than +1, the distribution is too peaked.*

#### for dist, the distribution is right skewed and peaked near to the median.

#### kurtosis f is less than 3 so distribution is platykurtic.

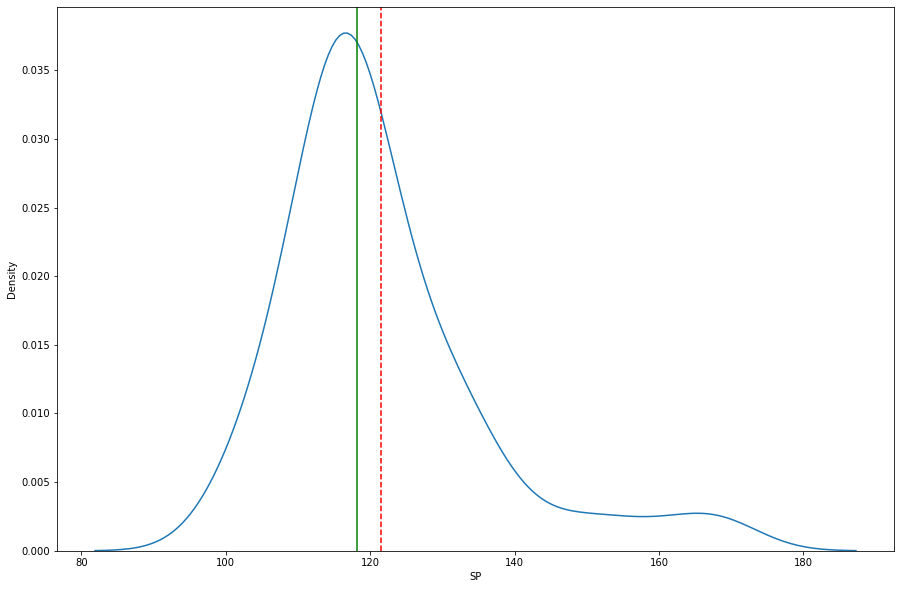
****

### ***Inference SP***

###### *"A general guideline for skewness is that if the number is greater than +1 or lower than –1, this is an indication of a substantially skewed distribution. For kurtosis, the general guideline is that if the number is greater than +1, the distribution is too peaked.*

#### for SP, the distribution is right skewed and too peaked.

#### kurtosis is less than 3 so distribution is platykurtic.

****

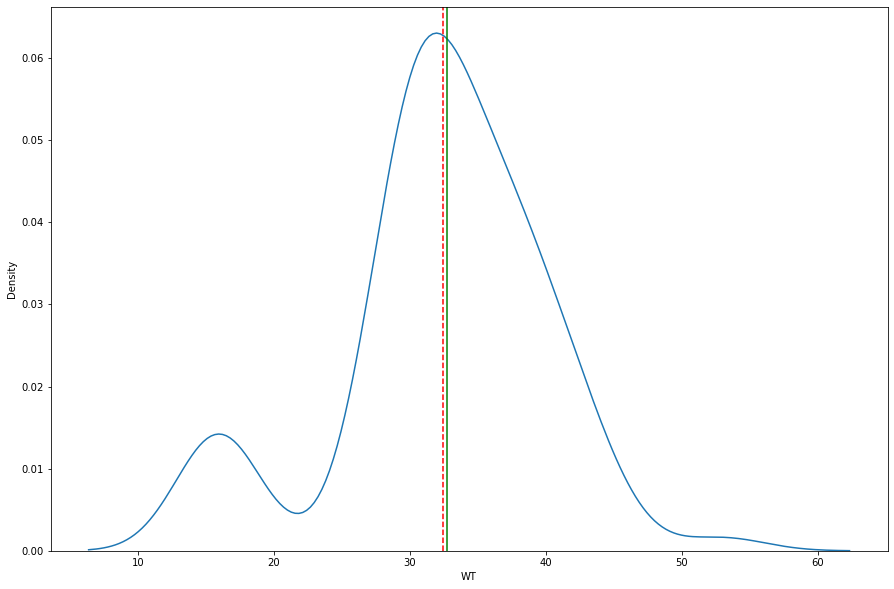
### ***Inference WT***

###### *"A general guideline for skewness is that if the number is greater than +1 or lower than –1, this is an indication of a substantially skewed distribution. For kurtosis, the general guideline is that if the number is greater than +1, the distribution is too peaked.*

#### for WT, the distribution is left skewed and peaked.

*From the histogram we can see that:*

1. *mode is between 50-100*
2. *median is approx 200*
3. *As mode is less than median, the above distribution is positively skewed*

****

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



#### ANS(Histogram)

*1.+ve skewed*

*2.most chickweight is between 50-100*

*3.around 50% are between 50-150 weight range*

*4. less than 15% are between 200-400 weight range (higher end)*



### *ANS*

*1.Outlier on the upperside*

*2.the distribution is positively skewed or skewed to the right.*

*3. The top whisker is much longer than the bottom whisker and the line is gravitating towards the bottom of the box.*

**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?**

#### n=2000 , mean=200 , std= 30 C.I=94% , alpha=0.06 C.I=98% , alpha=0.02 C.I=96% , alpha=0.04

###### as the population variance/std is unknown we will use the sample space std(s).

1. We are 94% confident that average age of an adult male is between: (198.73986828253865, 201.26013171746135)
2. 96% confident that average age of an adult male is between: (198.6239882298767, 201.3760117701233)
3. 98% confident that average is between: (198.6239882298767, 201.3760117701233)

**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.
2. What can we say about the student marks?

*ANS*

*mean is: 41.0*

*median is: 40.5*

*Mode is: ModeResult(mode=array([41]), count=array([4]))*

*Variance is: 24.11111111111111*

*std deviation is: 4.910306620885412*

## *Analysis*

1. *+ve skewed (skew value = 1.54 )*
2. *Average marks obtained was 41 with the spread around 5*
3. *arnd 50% of the tests have marks between 39-41*
4. 95% test have pass mark (assuming pass is 35 marks)

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

*Ans:skewness is zero when mean an dmedian are equal.*

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

*Ans: Distribution is +ve skewed*

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

*Ans: Distribution is -ve skewed*

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

*Ans: +ve kurtosis means distribution is peaked*

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

*Ans: -ve kurtosis means distribution is flat*

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



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

*Ans:*

*distribution is -ve skewed*

*There are no outliers present in the data.*

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

*Ans:IQR = Q3-Q1 => 19-10= 9*

**Q19) Comment on the below Boxplot visualizations?**



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

# *ANS*

1. *the median of both box plots are same is same arnd 263.*
2. *Both datasets are symmetrically distributed as evident from the nature of their boxplots.*
3. *box plot 1 has the min wbs value*
4. *the variation for box1 is less, that means it is more consistent as oppose to box 2.*
5. *the highest/lowest WBS value is more than the entire boxplot 2 values.*

**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)
  3. P (20<MPG<50)

*ANS:*

1. *the probability the MPG is <40 = 0.7293498762151616*
2. *the probability the MPG is >38 = 0.3475939251582705*
3. *the probability the MPG is between 20 and 50= 0.8988689169682046*

**Q 21) Check whether the data follows normal distribution**

1. Check whether the MPG of Cars follows Normal Distribution

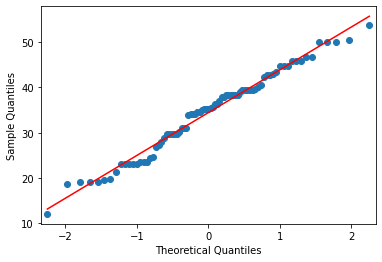
Dataset: Cars.csv

# *ANS*

1. *MPG(SKEW)= -0.177947, so -ve skewed, not completely symmetric*
2. *Also, the std =9.13 with mean of 34.3. Applying empirical rule*

*61% of the data is in 1 standard deviation, which means the data doesn't follow the emprical rule*

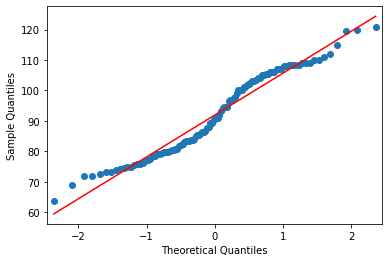
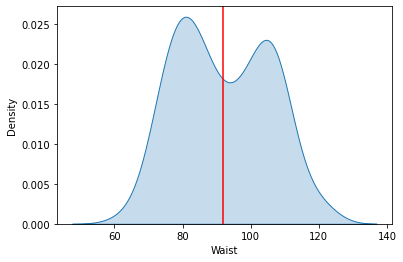
1. *from the QQ plot as below, it can derived that the data doeant follow normal distribution.*



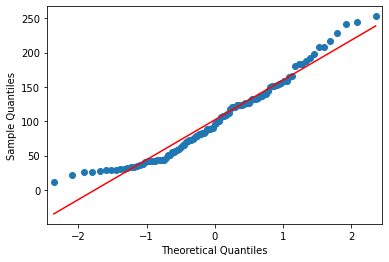
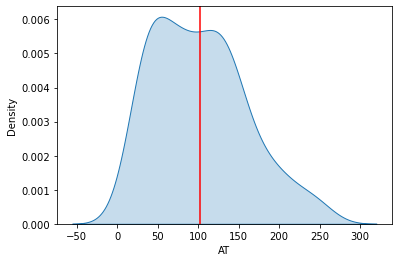
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

### ***Waist data does not follow normal distribution***

### ***AT +ve skewed , AT doesnt follow normal dist***

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

* 1. *Z-score of 90% CI is: 1.6448536269514722*
  2. *Z-score of 94% CI is: 1.8807936081512509*
  3. *Z-score of 60% CI is: 0.8416212335729143*

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

1. *T-score of 95% CI is: 2.0638985616280205*
2. *T-score of 96% CI is: 2.1715446760080677*
3. *T-score of 99% CI is: 2.796939504772804*

**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

*Ans:*

*The P\_Value = 0.3218140331685075*