|  |  |
| --- | --- |
| 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 | ORDINAL |
| Fahrenheit Temperature | INTERVAL |
| Height | RATIO |
| Type of living accommodation | ORDINAL |
| Level of Agreement | ORDINAL |
| IQ(Intelligence Scale) | INTERVAL |
| Sales Figures | RATIO |
| Blood Group | NOMINAL |
| Time Of Day | RATIO |
| Time on a Clock with Hands | INTERVAL |
| Number of Children | NOMINAL |
| Religious Preference | ORDINAL |
| 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?

1. Possible outcomes when three coins are tossed: HHH,HHT,HTH,HTT,THH,THT,TTH,TTT

Total possible outcomes:8

Favorable outcomes: HHT,HTH,THH,

Probability=No of favorable outcomes/Total no of Outcomes

=**3/8**

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
4. When two dice are rolled there are 36 possible outcomes:
5. Outcomes whose sum equal to 1: NILL

Probability=0/36=**0**

1. Outcomes whose sum is less than or equal to 4: (1,1) ,(1,2),(1,3), (2,1) ,(2,2), (3,1)

Probability=**6/36=1/6**

1. Outcomes whose sum is divisible by 2 and 3: (1,5),(2,4),(3,3),(4,2),(5,1),(6,6)

Probability=**6/36=1/6**

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?

1. **None of the balls drawn is blue, this can only happen when the two balls drawn at random are either red and green or both. Total number of balls = 2 + 3 + 2 = 7 ⇒ Number of ways of drawing 2 balls out of 7 = 7C2 = (7 × 6) / (2 × 1) = 42/2 = 21**

**Number of balls other than blue = 5 ⇒ Number of ways of drawing 2 balls out of 5 = 5C2 = (5 × 4) / (2 × 1) = 20/2 = 10**

**∴ Required Probability = 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

1. **The expected number of candies for a randomly selected child is:**

**Expected value = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \*0.120)**

**Expected value = 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24.**

**Expected value = 3.09.**

**Therefore, the expected number of candies for a randomly selected child is 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.

**Use Q7.csv file**

**A.**

**Mean: Points = 3.59, Score = 3.21 and Weigh = 17.84.**

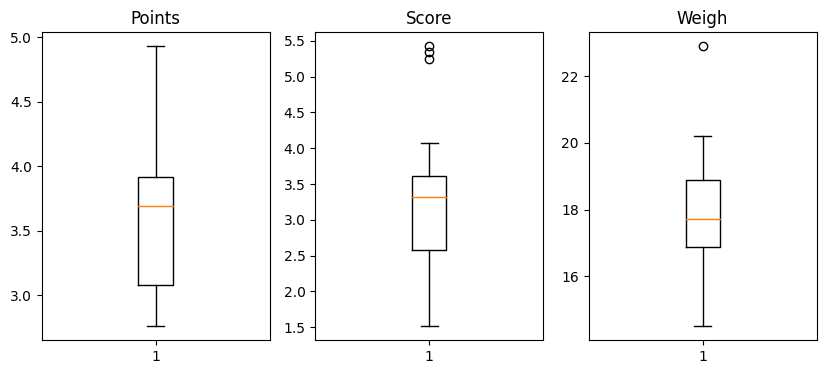
**Median: Points = 3.69, Score = 3.32 and Weigh = 17.71.**

**Mode: Points = 3.07, Score = 3.44 and Weigh = 17.02.**

**Variance: Points = 0.28, Score = 0.95, Weigh = 3.19.**

**Standard Deviation: Points = 0.53, Score = 0.97, Weigh = 1.78.**

**Range: points=2.17, Score=3.911, Weigh=8.4.**

**Inferences:**

**There are some no outliers in points attribute but in score and weigh attributes there are some outliers**.

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?

1. **Given weights (X) of patients:108, 110, 123, 134, 135, 145, 167, 187, 199Step 1: Calculate the sum of the weights.Sum = 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199 = 1308Step 2: Calculate the total number of patients, which is the number of data points.Total number of patients = 9Step 3: Calculate the expected value (mean) using the formula:Expected Value (E) = Sum of weights / Total number of patientsE = 1308/ 9 ≈ 145.33So, the expected value (mean) of the weight of a randomly selected patient is approximately 145 pounds.**

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

Cars speed and distance

Use Q9\_a.csv

**A.a.** **For Cars Speed Skewness value= -0.1139**

**-The data is negatively skewed, meaning it has a long left tail**

**ii. Kurtosis value= -0.577**

**-It signifies that the distribution has more values in the tails compared to a normal distribution**

**\* For Cars Distance**

**i. Skewness value = 0.78**

**- It signifies the data is rightly skewed, meaning it has a long right tail**

**ii. Kurtosis value = 0.24**

**- It means that the distribution has fatter tails and is more concentrated around the mean, resulting in a sharper, more pointed peak**.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**a.** **1. For SP Skewness = 1.58 , kurtosis = 2.70.**

**b. For WT Skewness = -0.60, Kurtosis = 0.81.**

Q10) Draw inferences about the following boxplot & histogram



**The histograms peak has right skew and tail is on right. Mean>Median. We**

**have outliers on the higher side.**



**The boxplot has outliers on the maximum 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?

1. **Given Average sample Weight Xbar=200pounds std=30 pounds**
2. **Confidence Interval for 94% level of confidence**

**conf\_94 =stats.t. interval (0.94,df=1999, loc=200, scale=30/np.sqrt (2000))**

**print (np. round(conf\_94,0))**

**print(conf\_94)**

**For 94% confidence interval Range is [ 198.73 – 201.26]**

**For 98% confidence interval range is [198.43 – 201.56]**

**For 96% confidence interval range is [198.62 – 201.37]**

**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?

1. **Mean=738/18=41**

**Median=40+41/2=40.5**

**Mode=41(4 times repeated)**

**Variance=25.67**

**Standard deviation=5.07**

B. **We don’t have outliers and the data is slightly skewed towards right**

**because mean is greater than median**.

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

1. **When the mean and median of a data set are equal, it typically indicates that the data is symmetrically distributed. In this case, the skewness is close to zero, suggesting that there is little or no skewness in the data. The distribution is approximately symmetric**.

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

A. **When the mean is greater than the median, it indicates positive skewness. This means that the distribution is right-skewed, with a longer tail on the right side of the distribution. In a positively skewed distribution, the majority of the data points are concentrated on the left side, with some higher values on the rig**ht.

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

A. **When the median is greater than the mean, it indicates negative skewness. This means that the distribution is left-skewed, with a longer tail on the left side of the distribution. In a negatively skewed distribution, the majority of the data points are concentrated on the right side, with some lower values on the left.**

Q16) What does a positive kurtosis value indicate for data?A. **A positive kurtosis value indicates that the data has heavy tails and is leptokurtic. In a leptokurtic distribution, the data has more extreme values (outliers) and a higher peak in the center compared to a normal distribution. It means that the distribution has fatter tails and is more concentrated around the mean, resulting in a sharper, more pointed peak.**

Q17) What does a negative kurtosis value indicate for data?A. **A negative kurtosis value indicates that the data has light tails and is platykurtic. In a platykurtic distribution, the data has fewer extreme values and a flatter peak in the center compared to a normal distribution. It means that the distribution has thinner tails and is more spread out, resulting in a flatter peak.**

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



What can we say about the distribution of the data?

A. **The above Boxplot is not normally distributed the median is towards the**

**higher value.**

What is nature of skewness of the data?

A. **The data is a skewed towards left. The whisker range of minimum value is**

**greater than maximum**

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

A.**IQR=UpperQuartile-LowerQuartile=18-10=8**

Q19) Comment on the below Boxplot visualizations?



**Here there is a representation of 2 box plots in which box plot 2) is highly**

**distributed across the plane and 1) is slightly less distributed. (variance)**

**Whiskers in these diagrams also show this.100% of the data is spread across**

**values from 350 in 2 whereas its spread in range 250-290 app x in 1**

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

A. **Here when we compare box plot 1 with box plot 2, we can say that the data in boxplot 1 is widely spread. Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2. The median in the 2 box plots is equal. And the data spread is both of them are symmetrical**.

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)

A. a. **1-stats.norm.cdf(38,Cars.MPG.mean(),Cars.MPG.std())**

**P(MPG>38) = 0.3475939**

b. **stats.norm.cdf(40,Cars.MPG.mean(),Cars.MPG.std())**

**P(MPG<40) = 0.7293498762**

**c. stats.norm.cdf(50,Cars.MPG.mean(),Cars.MPG.std())- stats.norm.cdf(40,Cars.MPG.mean(),Cars.MPG.std())**

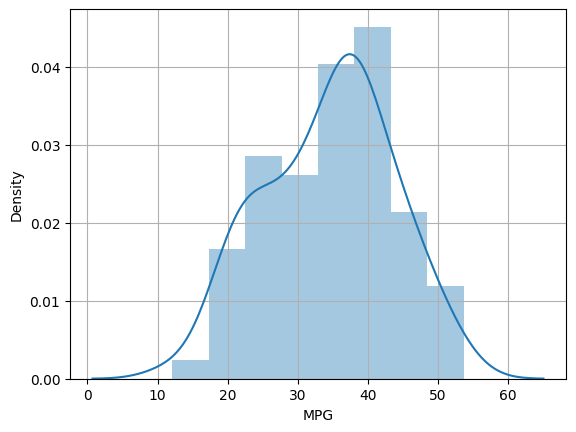
**P (20<MPG<50)=0.898868916**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

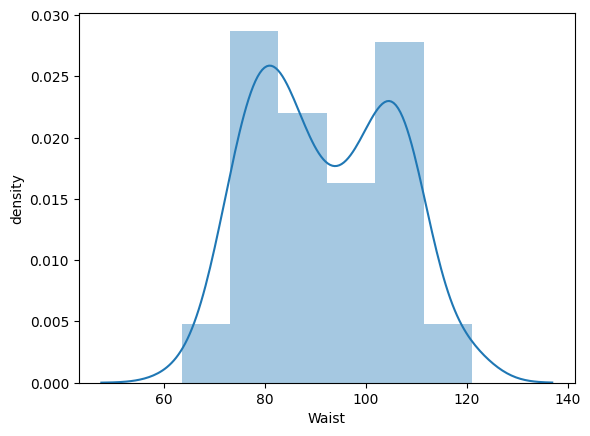
A.**The MPG of cars appears to be normally distributed**

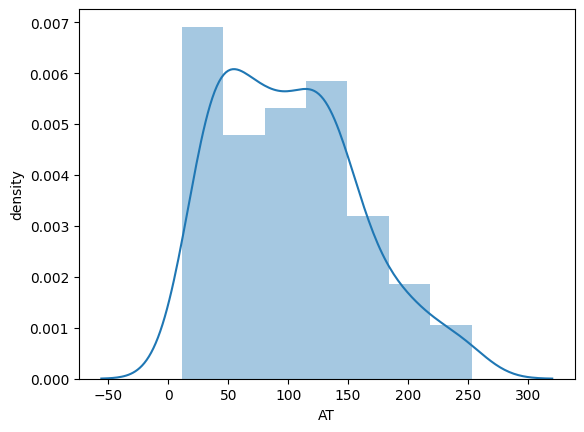


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

1. **The AT and Waist values from the wc-at.csv dataset appear to be approximately normally distributed**





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

A.  **a. Z scores of 90% confidence interval:**

**x=stats.norm.ppf(0.90)**

**= 1.281551.**

**b. Z scores of 90% confidence interval:**

**x=stats.norm.ppf(0.94)**

**= 1.55477.**

**c. Z scores of 60% co nfidence interval:**

**x=stats.norm.ppf(0.60)**

**= 0.253347**

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

A. **t score of 95% confidence interval:**

**stats.t.ppf(0.975,24)**

**= 2.0638985616280205**

**b.  96% confidence interval:**

**stats.t.ppf(0.98,24)**

**= 2.1715446760080677**

**c.99% confidence interval:**

**stats.t.ppf(0.995,24)**

**= 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

1. **μ = 270, N=18, x bar=260, s=90**

**We need to use t-test here because std of population is not known only std of sample is known**

**T score = t= (260-270)/ (90/18\*\*0.5).**

**= -0.47.**

**Required probability = 0.32**

**p\_value=1-stats.t.cdf(abs(-0.4714),df=17)**