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

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

Ans) **Three coins are tossed the total number of possible combinations are = 8. The possible outcomes are HHH, HHT, HTH, THH, TTH, THT, HHT, TTT.**

**The probability of two heads and one tail is 3/8 = 0.375.**

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

**Ans)**

* **(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)**
* **Total outcome=6×6=36**

1. Equal to 1

Ans. **0/36 = 0**

1. Less than or equal to 4

Ans. **6/36 = 0.1666**

1. Sum is divisible by 2 and 3

Ans. **6/36 = 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?

Ans) **Total number of balls = (2 + 3 + 2) = 7  
Let S be the sample space.  
Then, n(S) = Number of ways of drawing 2 balls out of 7  
=7C2​  
=(2×1)(7×6)​  
=21  
Let E = Event of drawing 2 balls, none of which is blue.  
∴n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.  
=5C2​  
=(2×1)(5×4)​  
=10  
∴P(E)=n(S)n(E)​=10/21**

**The probability that none of the balls drawn are blue is** **10/21=0.47619**

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 \* 0.12**

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

**= 3.090**

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**

**Ans)**

|  |  |  |
| --- | --- | --- |
| **Column1** | **Formula** | **Points** |
| **Mean** | **Mean(Q7$Points)** | **3.5965625** |
| **Median** | **Median(Q7$Points)** | **3.695** |
| **Mode** | **Mode(Q7$Points)** | **3.92** |
| **Variance** | **Var(Q7$Points)** | **0.28588135** |
| **Standard Deviation** | **Std(Q7$Points)** | **0.53467874** |
| **Range** | **Max(Q7$Points)** | **4.93** |
|  | **Min(Q7$Points)** | **2.76** |

|  |  |  |
| --- | --- | --- |
| **Column1** | **Formula** | **Score** |
| **Mean** | **Mean(Q7$Score)** | **3.21725** |
| **Median** | **Median(Q7$Score)** | **3.325** |
| **Mode** | **Mode(Q7$Score)** | **3.44** |
| **Variance** | **Var(Q7$Score)** | **0.95737897** |
| **Standard Deviation** | **Std(Q7$Score)** | **0.97845744** |
| **Range** | **Max(Q7$Score)** | **5.424** |
|  | **Min(Q7$Score)** | **1.513** |

|  |  |  |
| --- | --- | --- |
| **Column1** | **Formula** | **Weigh** |
| **Mean** | **Mean(Q7$Weigh)** | **17.84875** |
| **Median** | **Median(Q7$Weigh)** | **17.71** |
| **Mode** | **Mode(Q7$Weigh)** | **17.02** |
| **Variance** | **Var(Q7$Weigh)** | **3.19316613** |
| **Standard Deviation** | **Std(Q7$Weigh)** | **1.78694324** |
| **Range** | **Max(Q7$Weigh)** | **22.9** |
|  | **Min(Q7$Weigh)** | **14.5** |

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?

**Ans.**

**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**

**The expected value of the weight of that patient is 145.33.**

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

Cars speed and distance

Use Q9\_a.csv

**Ans)**

1. **skewness(Q9\_a$speed)**

**-0.1105533**

**kurtosis(Q9\_a$speed)**

**2.422853**

**The Skewness is negative i.e The median will lie after the mean and shape of the graph will be tall also there might be outliers.**

1. **skewness(Q9\_a$dist)**

**0.7591268**

**kurtosis(Q9\_a$dist)**

1. **3.248019**

**The Skewness is positive i.e The mean will lie after the median and the shape of the graph will be tall also there might be outliers.**

SP and Weight(WT)

**Use Q9\_b.csv**

1. **skewness(Q9\_b$SP)**

**1.552258**

**kurtosis(Q9\_b$SP)**

**5.723521**

**The Skewness is positive i.e The mean will lie after the median and the shape of the graph will be tall also there might be outliers.**

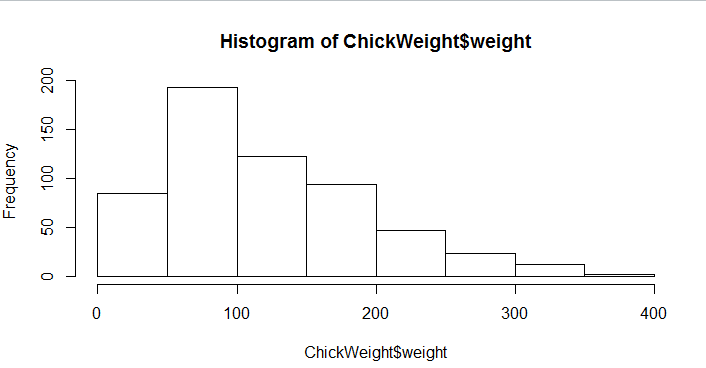
1. **skewness(Q9\_b$WT)**

**-0.5921721**

**kurtosis(Q9\_b$WT)**

**3.819466**

**The Skewness is negative i.e The median will lie after the mean and shape of the graph will be tall also there might be outliers.**

Q10) Draw inferences about the following boxplot & histogram.

**Ans. The Histogram is Positively Skewed i.e the Mean will lie after the median in this case. Also we can see that there are outliers at the very end of the graph .**



**Ans. The Box Plot has most values in the initial part I.e the median lies before the mean for this plot. Also there are outliers and we can assume it to be positively skewed .**

**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?

**Ans.**

|  |  |  |
| --- | --- | --- |
| **94% Confidence Interval** |  |  |
| stats.t.ppf(0.970,df=1999) | t value | 1.8818 |
| 200+1.8818\*(30/2000\*\*0.5) | Upper Limit | 201.26 |
| 200-1.8818\*(30/2000\*\*0.5) | Lower Limit | 198.73 |
| **98% Confidence Interval** |  |  |
| stats.t.ppf(0.99,df=1999) | t value | 2.3282 |
| 200+2.3282\*(30/2000\*\*0.5) | Upper Limit | 201.56 |
| 200-2.3282\*(30/2000\*\*0.5) | Lower Limit | 198.43 |
| **96% Confidence Interval** |  |  |
| stats.t.ppf(0.98,df=1999) | t value | 2.055 |
| 200+2.0550\*(30/2000\*\*0.5) | Upper Limit | 201.37 |
| 200-2.0550\*(30/2000\*\*0.5) | Lower Limit | 198.62 |

**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 | x.mean() | 41 |
| Median | **x.median()** | **41** |
| Variance | **x.var()** | **25.52941176** |
| Standand Deviation | **x.std()** | **5.052663829** |

**The average marks obtained by the students are 41 and the variation from the mean is around 25 marks.**

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

**Ans. The Skewness is Zero in this case.**

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

**Ans. The Skewness is Positive in this case.**

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

**Ans. The Skewness is Negative in this case.**

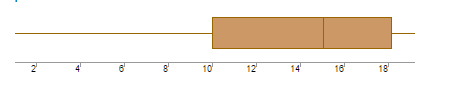
Q16) What does positive kurtosis value indicates for a data ?

**Ans. The Positive kurtosis value indicates the grape will be taller.**

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

**Ans. The Positive kurtosis value indicates the grape will be shorter in height.**

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



What can we say about the distribution of the data?

**Ans. The distribution is Asymmetrical and loaded on the upper or right side of the graph**

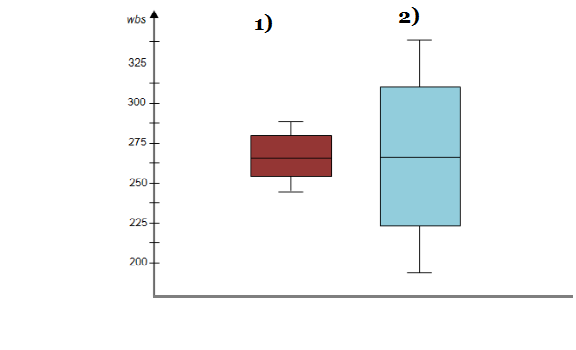
What is nature of skewness of the data?

**Ans. The data is Negatively Skewed.**

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

**Ans. The IQR will be between 10-18.**

Q19) Comment on the below Boxplot visualizations?



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

**Ans. The boxplot 1 has values in between 235-280 and is less dispersed than the values of boxplot 2. By observing both the plots whisker’s level is high in boxplot 2.Mean and median are equal hence distribution is 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)

**1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())**

**=** **0.35**

* 1. P(MPG<40)

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

**=** **0.72**

* 1. P (20<MPG<50)

**stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())- stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())**

= **0.897**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans. **Car MPG mean 34.422075728024666**

**Car MPG median 35.15272697**

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

Dataset: wc-at.csv

Ans.  **Waist Circumference(Waist) = (91.90183486238533, 90.8)**

**Adipose Tissue= (101.89403669724771, 96.54)**

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

**Ans. Confidence Intervals Z scores**

**90% 1.644**

**94% 1.880**

**60% 0.841**

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

**Ans. Confidence Intervals t scores**

**95% 2.063**

**96% 2.171**

**99% 2.796**

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 à pt(tscore,df)

df à degrees of freedom

**Ans.**

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

**= -0.4714045207910317**

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

**= 0.3216**