|  |  |
| --- | --- |
| 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 | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ordinal |
| 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?

Total possible outcomes will be 8.

**Probability of two heads and one tail is 3/8=0.37**

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

1. Equal to 1

**Its probability is zero**

1. Less than or equal to 4

Total possible outcomes after rolling two dice=36

Combinations having sum less than or equal to 4 are(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)

**So probability here is 6/36 i.e 1/6=0.167**

1. Sum is divisible by 2 and 3

Total possible outcomes after rolling two dice =36

Sum should be divisible by 2 and 3 means Sum should be multiple of 6

i.e (1,5),(2,4),(3,3),(4,2),(5,1),(6,6) means 6 possible situations

So **probability will be 6/36 i.e 1/6=0.167**

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?

Total possible outcomes 2+3+2=7

**Probability of balls drawn other than Blue is 5/7=0.714**

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

**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 \* 0.120

= 0.015 + 0.8+ 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.

**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?

Sum=108+ 110+ 123+ 134+135+145+ 167+187+199

Avg=1308/9=145.333

**Expected Value of the Weight of the patient is 145**

**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**

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



There are 200 Chicks whose weight is 50 to 100.There are no chicks whose weight is more than 350.Very less chicks have weight more than 200.



If center represents the median. This boxplot indicates there are more right outliers. Boxplot is left 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?

**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?

Answer:

1. **Mean is 41**

**Median is 40.5**

**Standard deviation is 5.05266382858645**

**Variance is 25.529411764705884**

1. **Standard deviation is more than 3 means data is scattered means not normally distributed. Most of the students have marks less than 50.Variance is high so prediction will be poor.**

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

Answer -**Data will be normally distributed**.

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

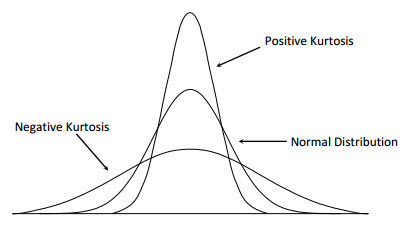
Answer-**Data will be positively skewed means more data will be on right tail.**

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

Answer-**There will be negative skewness as more data will be on the left tail.**

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

**Answer**-Positive values of kurtosis indicate that a distribution is peaked and possess thick tails, indicating the large outliers means less stability in data.



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

**Answer**- If a distribution has negative kurtosis, it means that it has a flatter peak and thinner tails compared to a normal distribution. This simply means that more data values are located near the mean and less data values are located on the tails.

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



What can we say about the distribution of the data?

**Answer**- Here in this Box plot, Median is near about 15.Lower Quartile is longer and left whisker is also long means more left side outliers.

What is nature of skewness of the data?

**Answer**- Box Plot is left skewed means more outliers are on the left side. It is also known as negative skewed. In this mean<median.

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

Q1=10 Q2=15 Q3=18

IQR(Inter Quartile Range)=Q3-Q1=18-10=8

**IQR is 8**  
Q19) Comment on the below Boxplot visualizations?



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

**Answer**-In Boxplot 1,data values is scattered from around 250 to 287 whereas in Boxplot 2, data is scattered from values 200 to 325.Both have same median i.e. about 262.Boxplot is right skewed as Q3-Q2>Q2-Q1 and same is the case with Boxplot 2 as 50>37.Values taken here are not exact but approximately similar.

|  |  |  |
| --- | --- | --- |
| **Column1** | **Boxplot 1(Approx)** | **Boxplot 2(Approx)** |
| Min | 250 | 200 |
| Max | 287 | 325 |
| Q1 | 255 | 225 |
| Q2 | 262 | 262 |
| Q3 | 277 | 312 |
| IQR | 22 | 87 |
| Q3-Q2 | 15 | 50 |
| Q2-Q1 | 7 | 37 |
|  |  |  |

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)

Q 21) Check whether the data follows normal distribution

a)Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Mean of MPG is 34.422075728024666**

**Median of MPG is 35.15272697**

**Mode of MPG is 29.629936**

**It is clear from the histogram that it is near to bell curve and Mean, Median and Mode are almost equal. So Data is near about 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

Done in scripting only……

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

**90% confidence interval**

90% means .9 so degree of freedom is 1-0.9=0.1

We will divide 0.1/2 because value can lie either on left or right side so it comes 0.05

So we use stats.norm.ppf(0.95)

Here standard deviation will be 1.64

**94% confidence interval**

94% means .94 so degree of freedom is 1-0.94=0.06

We will divide 0.06/2 =0.03

1-0.03=0.97

So we use stats.norm.ppf(0.97)

Here standard deviation will be 1.88

**60% confidence interval**

60% means .6 so degree of freedom is 1-0.6=0.4

We will divide 0.4/2 =0.2

1-0.2=0.8

So we use stats.norm.ppf(0.8)

Here standard deviation will be 0.84

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

Done in scripting only……

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

Answer t - statistics for the data is as follows:

t=(x-µ)/(s/sqrt(n))

x = mean of the sample of bulbs =  260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t=(260-270)/(90/sqrt(18)

t=-10/(90/3\*sqrt(2)

t=-10/(30/sqrt(2)

t=(-1\*sqrt(2))/3

**t = - 0.471**

**Degree of freedom is 18-1=17**

The probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.

----------------------THE END-----------------------------------