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

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

Ans) Total outcomes = 2x2x2 = 8

No. of outcomes for 2 heads and 1 tail = 3, therefore probability = 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

Ans) a) Minimun sum of two dices = 2 (when both are one)

Maximum sum of two dices = 12 (when both are 6)

Probability when sum is 0 = **0**

b) No of outcomes for sum <= 4 is =6 ((1,1),(1,2),(1,3),(2,1),(2,2),(3,1))

Total outcomes = 36, probability = 6/36 = **1/6**

c) No of outcomes divisible by 2 and 3 = 6 ((1,5),(2,4),(3,3),(4,2),(5,1)&(6,6))

Total outcomes = 36, 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?

No of favorable outcomes = 10

Total outcomes = 21, 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 \*1 = 0.015 ~ 0 candy

Child B – probability of having 4 candies = 0.20 \* 4 = 0.8 ~ 1 candy

Child C - probability of having 3 candies = 0.65 \* 3 = 1.95 ~ 2 candies

Child D – probability of having 5 candies = 0.005 \* 5 = 0.025 ~ 0 candy

Child E - probability of having 6 candies = 0.01 \* 6 = 0.06 ~ 0 candy

Child F - probability of having 2 candies = 0.12 \* 2 = 0.24 ~ 0 candy

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:-

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.5965625 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.07 or 3.92 | 3.44 | 17.02 or 18.9 |
| Variance | 0.285881351 | 0.95737897 | 3.193166129 |
| Std deviation | 0.534678736 | 0.97845744 | 1.786943236 |
| Range | 2.17 | 3.911 | 8.4 |

The mean and median of all columns are similar therefore their distributions will be very close to a normal distribution.

There are two modes for Points column and Weigh column, hence they have bimodal data set.

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 weight = average weight = 145.33 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
|  | speed | dist |
| skewness | -0.11751 | 0.806895 |
| kurtosis | -0.03531 | 0.360105 |

The data set of Speed column has a negative skewness, which means it is left tailed data set. As the skewness value is close to zero, there is only very small amount of skewness. Whereas the Distance column is positively skewed and has a tail on the right end of the distribution and it is highly skewed distribution.

The negative kurtosis for Speed dataset indicates there is a flat peak and light tails for it and the positive kurtosis for the Distance dataset indicates it has a sharp peak and heavy tails.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skew | 1.61145 | -0.61475 |
| kurt | 2.977329 | 0.950291 |

The skewness and kurtosis values are positive for the SP column, which indicates it is a right tailed dataset, has a heavy tail and sharp peak.

There is a negative skewness and positive kurtosis value for the WT column, which indicates that it is a left-tailed dataset having a slightly sharp peak.

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



This is a positively skewed dataset having tail on the right side.



The above boxplot indicates there are many outliers in the data set at the upper extreme. Also there is an indication of positive skewness.

**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?

Sample size = 2000

Std dev = 30

Sample mean = 200 pounds

We use z-test as the sample size is greater than 30.

C.I. = 0.94, finding value for 0.94+0.03=0.97

Upper value = scipy.stats.norm.ppf(0.97,200,30) = 256.4238

Lower value = scipy.stats.norm.ppf(0.03,200,30) = 143.576

C.I. = 0.98, finding value for 0.98+0.01=0.99

Upper value = scipy.stats.norm.ppf(0.99,200,30) = 269.79

Lower value = scipy.stats.norm.ppf(0.01,200,30) = 130.21

C.I. = 0.96, finding value for 0.96+0.02=0.98

Upper value = scipy.stats.norm.ppf(0.98,200,30) = 261.612

Lower value = scipy.stats.norm.ppf(0.02,200,30) = 138.387

**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?

Mean = 41, Median = 40.5, variance=24.11, standard deviation = 4.91

Student marks is approximately a normal distribution as the mean and median are too close.

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

Skewness = 0, when mean and median are equal.

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

Skewness>0, positive skewness

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

Skewness<0, negative skewness

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

It indicates that the distribution will have heavy tail and sharp peaks.

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

It indicates that the distribution will have light tails and flat peaks.

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



What can we say about the distribution of the data?

The distribution is left-tailed in nature.

What is nature of skewness of the data?

Negatively-skewed

What will be the IQR of the data (approximately)?   
IQR = 75th percentile = 25th percentile = 18-10 = 8

Q19) Comment on the below Boxplot visualizations?



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

Both the boxplots indicate the data is normally distributed due to the symmetry about the median value. Boxplot 1 has much less data points compared to Boxplot 2, even though their medians are the same.

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) n = 81, mean=34.42, std dev=9.13

P(MPG>38)= 1 - stats.norm.cdf(38,34.42,9.13) = 0.3474 = 34.74%

P(MPG<40) = stats.norm.cdf(40,34.42,9.13) = 0.7294 = 72.94%

P(20<MPG<50)=stats.norm.cdf(50,34.42,9.13)-stats.norm.cdf(20,34.42,9.13) = 0.8989 = 89.89% ~ 90%

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Mean(34.42) and median(35.15) are too close for the MPG column, hence it approximately follows a 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

The Waist data follows a normal distribution approximately as its mean(91.9) and median(90.8) are almost same. Whereas the AT data does not follow a normal distribution as their mean(101.9) and median(96.5) are not same.

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

For 90%, Z = stats.norm.ppf(0.9) = 1.2815

94%, Z = stats.norm.ppf(0.94) = 1.554

60%, Z = stats.norm.ppf(0.6) = 0.2533

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

95%, t = stats.t.ppf(0.95,df=24) = 1.7108

96%, t = stats.t.ppf(0.96,df=24) = 1.828

99%, t = stats.t.ppf(0.99,df=24) = 2.492

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

Claimed average = 270

N = 18

Sample mean = 260

SD = 90

To find P(X<=260), using t-test,

t-statistic = = -0.4714

Code in R,

pt(t-statistic,df) = pt(-0.4714,17) = 0.32167 = 32.16%

The probability that 18 random samples of bulbs will last no more than 260 days is 32.16%.