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

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

**Ans. 2^3=8**

Total probability combinations are –

(HHH, HHT, HTH, HTT, THH, TTH, THT, TTT)

The probability that two heads and one tail are **HHT, HTH, THH**

That is 3 so **probability=Number of favorable outcome/Total number of outcome**

=**3/8 = 0.375**

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.** Two dice are rolled then Sample space is 36

1. Equal to 1 = as there are two dice the result will always start by at least two only
2. Less than or equal to 4 = possible outcomes are {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)} so Probability is = **6/36 =1/6**
3. Sum is divisible by 2 and 3 = possible outcomes are {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)} so probability is **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?

**Ans.** Total balls = (2 + 3 + 2) = 7

through combinations 7\*6/2\*1=21

The probability that none of the balls drawn is blue means only red and green should obtained Total number of balls = (2 +3) = 5

Through combinations 5\*4/2\*1=10

Event of drawing two balls, none of which is blue = **10/21 = 0.476**

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

= 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

= **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?

**Ans.** As there are 9 patients

Probability of selecting each patient = **1/9**

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**

**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**



**Ans.** The above histogram is **right skewed** i.e positively skewed . There is one **outlier** on right side

The boxplot has some **outliers above upper extreme**

**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.**  Using t-distribution, it is found that :

* The **94%** confidence interval is **(198.73, 201.27).**
* The **96%** confidence interval is **(198.61, 201.39)**
* The **98%** confidence interval is **(198.43, 201.57)**

**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*.* 1)** Mean = 41 , Median = 40.5 , Variance = 25.52 and SD = 5.05

**2)** We don’t have any outliers and the data is slightly skewed towards because mean is greater than median means positively skewed

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

**Ans.** The nature of is depends on the mean, median and mode, it may be positive, negative or zero. **if the mean, median of data are equal, hence the kurtosis is zero**.

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

**Ans.**  If the mean is greater than median, then the distribution is **positively skewed.**

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

**Ans.**  If the mean is less than median, then the distribution is **negatively skewed**.

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

**Ans.**  Positive value of kurtosis indicate that distribution is **peaked and possesses thick tails**

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

**Ans.** Negative value of kurtosis indicate that distribution is **flat and has thin tails.**

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



What can we say about the distribution of the data?

**Ans. The above boxplot is not normally distributed the median is towards the higher value**

What is nature of skewness of the data?

**Ans. The data is skewed towards left. The whisker range of minimum value is greater than maximum means negatively skewed.**

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

**Ans.**  **The Inter Quartile Range = Upper Quartile – Lower Quartile = 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.

**Ans. 1) There are no outliers.**

**2) Both the box plots shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum range.**

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

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans.** a) MPG of cars follows 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

**Ans.** Adipose Tissues (AT) and Waist does not follows Normal Distribution

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

**Ans. Given:**

Given confidence intervals are

90%

94%

60%

We need to find the z-scores at these intervals

**Solution:**

For 90% confidence interval:

We have the significance level at 5% (it is two tailed test)

That is:

α = 5% = 0.05

z at α = 0.05 from the z table will be:

z = 1.645.

For 65% confidence interval, we get:

We have the significance level at 3% (as it is two tailed test)

That is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

z = 1.555.

For 60 % confidence interval, we get:

We have the significance level at 20 % ( as it is a two tailed test)

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

z = 0.253

**Therefore, we get that the z score at 90 % confidence interval is 1.645, at 94 % confidence interval is 1.555 and at 60 % confidence interval is 0.253.**

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

**Ans.**  To compute the 95% confidence interval, start by computing the mean and standard error: M = (2 + 3 + 5 + 6 + 9)/5 = 5. σM = = 1.118. Z.95 can be found using the normal distribution calculator and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff points

Confidence Level z

0.90 1.645

0.92 1.75

0.95 1.96

0.96 2.05

With a 90 percent confidence interval, you have a 10 percent chance of being wrong. A 99 percent confidence interval would be wider than a 95 percent confidence interval (for example, plus or minus 4.5 percent instead of 3.5 percent).

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:** For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, 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.