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

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

23 =8

HHH,**HHT**,**HTH**,**THH**,HTT,TTH,THT,TTT

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

2 dice has **36** outcomes

(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)

1. 0/36=**0**
2. 3+2+1=6

6/36=**1/6**

c)1+1+1+1+1+1=6

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?

2+3+2=7(Total balls)

Probability of ball being Blue is 2/7

So the probability that none of the balls drawn is blue is 1-2/7=**5/7**

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 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12

= **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 file** for answer

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?

(108+110+….+199)/9=**145.33 (approx 145)-pounds**

**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 & histogramFrequency of chickweight$weight is highest between 50 and 100 and the data is right skewed/positive skewed.**



**Its has many outliers on the Upper limit that means the values of outliers are high-values that makes the mean inaccurate.**

**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?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CI =** | |  |  | | --- | --- | | **X̄ ± Z×** | **s** | |  | | **√n** | | |
| **94%** = | | | |  |  | | --- | --- | | 200 ± 1.8808× | 30 | |  | | √2000 | | |

**=198.738 – 201.262**

|  |  |
| --- | --- |
| **96%**  = 200 ± 2.0537× | 30 |
|  |
| √2000 |

**=198.622 – 201.378**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **98%** = | |  |  | | --- | --- | | 200 ± 2.3263× | 30 | |  | | √2000 | |

=**198.439 – 201.561**

**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

Mode=41 (as it appears most of the times)

Variance=((34-41) 2+ (36-41)2+……+(56-41)2)/18=25.53

Std=root of 25.53=5.05

2) 18 students have an average of 41

min value is 34

max in 56

most repeated marks was 41(mode)

std is approx 5 i.e the spread of data

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

* When mean and median are equal that means the distribution is symmetric, so nature of skewness becomes zero

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

* This makes the data positively skewed/right skewed

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

* This makes the data negatively skewed/left skewed

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

* Positive values of kurtosis indicate that a distribution is peaked and possess thick tails

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

* Negative values of kurtosis indicate that a 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?

* Std of the data is low as the data is not too much spread
* All the values lie b/w 10-18
* It is Negatively skewed/Left Skewed
* This also indicates that mean< median here
* As the median almost lies to 3rd quartile most of the values lies on left of the median

What is nature of skewness of the data?

* It is Negatively skewed/Left Skewed

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

* IQR=Q3-Q1
* 18-10=**8(approx)**

Q19) Comment on the below Boxplot visualizations?



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

* Data for boxplot 1 has less std as the spread is less
* Data for boxplot 2 has more std as the spread is more
* Median is same for both of them so that means most of the data for both of them lie at same value
* IQR will be different for them- less for 1, more for 2
* As the spread of data for 1 is low it will give more accuracy compared to 2

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)

round(1-stats.norm.cdf(x = 38, loc = 34.42, scale = 9.13),2)

**35%**

* 1. P(MPG<40)

round(stats.norm.cdf(x = 40, loc = 34.42, scale = 9.13),2)

**73%**

* 1. P (20<MPG<50)

round(stats.norm.cdf(x = 50, loc = 34.42, scale = 9.13)-stats.norm.cdf(x = 20, loc = 34.42, scale = 9.13),2)

**90%**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

**Dataset: Cars.csv**

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

**Dataset: wc-at.csv**

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

* **90%= 1.645**
* As 90% =0.90, 1-0.90=0.1(alpha)
* We divide alpha by 2, so 0.1/2=0.05, 1-0.05=**0.95**
* **round(stats.norm.ppf(0.95),3)**
* **94%= 1.881**
* 0.94,1-0.94=0.06
* 0.03,1-0.03=**0.97**
* **round(stats.norm.ppf(0.97),3)**
* **60%= 0.842**
* 0.60,1-0.60=0.4
* 0.2,1-0.2=**0.8**
* **round(stats.norm.ppf(0.80),3)**

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

* **95%= 2.064**
* As 95% =0.95, 1-0.95=0.05(alpha)
* We divide alpha by 2, so 0.05/2=0.025, 1-0.025=**0.975**
* **round(stats.t.ppf(0.975,24),3)**
* **96%= 2.172**
* 96%=0.96,1-0.96=0.04
* 0.02,1-0.02=**0.98**
* **round(stats.t.ppf(0.98,24),3)**
* **99%= 2.797**
* 99%=0.99,1-0.99=0.01
* 0.005,1-0.005=**0.995**
* **round(stats.t.ppf(0.995,24),3)**

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

t - statistics for the data is given as follows:

t=

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=

**t =** **- 0.471**

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with (n-1=18-1)=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 (32%)** assuming the mean life of the bulbs is 300 days.