|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Discrete data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete data |

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 | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ordinal |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Solution: when the three coins are tossed, the sample space will be

Sample space = {(H,H,H) (H,H,T) (H,T,H) (H,T,T) (T,H,H) (T,H,T) (T,T,H) (T,T,T)}

Condition: two heads and one tail A ={(H,H,T) (H,T,H) (T,H,H)}

P(A) = 3/8

**P(A) = 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

Solution: the sample space is = {(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)}

A: sum is equal to 1=

P(A)=0/36

**P(A)=0**

B: sum is less than or equal to 4

P(B)=6/36

**P(B) = 0.166**

C: sum is divisible by 2 and 3

P(C)=6/36

**P(C)=0.166**

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?

Solution: using permutation and combination,

Total no of Chances to draw 2 Balls at Random From 7 Colored Balls, **N(S)=7C2=21**

Let **E** be an event to draw 2 balls other than Blue.

No of Chances to draw two balls other than Blue,

**N(E)=2C2+3C2+2C1.3C1=1+3+6=10**

The Probability that none of the Balls drawn is Blue is,

**P(E)=N(E)/N(S)=10/21**

**P(E)=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

Solution:

= 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**

Expected no 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**

Solution: file attached in the name **“a\_1 Q7 solution vinayak”**

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?

Solution:

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**

Expected Value of the Weight of that patient = **145.33**

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

Cars speed and distance

Use Q9\_a.csv

Solution: file attached in the name of **“a\_1 Q9\_a sol vin”**

SP and Weight(WT)

Use Q9\_b.csv

Solution: file attached in the name of **“a\_1 Q9\_b sol vin”**

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



**Histogram:**

1.ChickWeight data is right skewed or positively skewed.

2.More than 50% ChickWeight is between 50 to 150.

3.Most of the ChickWeight is between 50 to 100.



**Boxplot:**

1. The data is right skewed.

2. There are outliers at upper side

**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?

The information given is:

* Sample **mean**of x̅=200
* Sample **standard deviation** of  s = 30.
* Sample **size**of n = 2000.

The **interval** is:

X̅±t\*s/√n

* In which **t** is the critical value for the two-tailed confidence interval.

Considering a **94%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 1.8916**, hence:

X̅-t\*s/√n = 200-1.8916 \* 30/√2000 = 198.73

X̅+t\*s/√n = 200+1.8916 \* 30/√2000 = 201.27

The **94%** confidence interval is **(198.73, 201.27).**

Considering a **96%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.0673**, hence:

X̅-t\*s/√n = 200-2.0673\* 30/√2000 = 198.61

X̅+t\*s/√n =200+2.0673\*30/√2000 = 201.39

The **96%** confidence interval is **(198.61, 201.39).**

Considering a **98%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.3452**, hence:

X̅-t\*s/√n = 200-2.3452\* 30/√2000 = 198.43

X̅+t\*s/√n =200+2.3452\*30/√2000 = 201.57

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.

Solution: attached the file in the name of **“a\_1 Q12 sol vin”**

1. What can we say about the student marks?

Solution: using boxplot which is plotted for the given data, we can say that most the students marks lies between 35 to 45 .

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

Solution:

Skewness refers to deviation or asymmetry that deviates from the bell curve or normal distribution in a set of data. If the curve shifts to the left or to the right then it is said to be skewed.

Skewness can be said as a representation of the extent to which a given distribution varies from a normal distribution . normal distribution has a skewness of zero.

When mean, median and mode is equal then it is normal distribution and the nature of the skewness is zero.

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

Solution: when mean>median,the nature of the skewness is positive (positively skewed)

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

Solution: when median>mean ,the nature of skewness is negative(negatively skewed)

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

Solution: kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure to observe whether the data is heavy-tailed or light-tailed in a normal distribution.

Normal distribution have a kurtosis of 3.

Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where most of the numbers are located in the tails of the distribution instead of around the mean.

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

Solution: A negative kurtosis having a lighter tails and stretched around centre tails means most of the data points are present near the mean.

A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

a) first 25% of data value is less than 10, next 25% of data value lies between 10 to 15.2, next 25% of data value lies between 15.2 to 18.1 and last 25% of data value is greater than 18.1

b) According to boxplot, median > mean.

So, the nature of skewness is negatively skewness.

c) Here Q1=10, Q2=15.2 and Q3=18.1

So, IQR is 18.1-10=8.1

Q19) Comment on the below Boxplot visualizations?



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

Solution: from 1st boxplot:

The values lies between 245 to 290. The first 25 % of values lies between 245 to 255 next 25% of values lies between 255 to 265 and next 25% values lies between 265 to 280.last 25% values lies between 280 to 290.

From 2nd boxplot:

The values lies between 190 to 340. The first 25 % of values lies between 190 to 225 next 25% of values lies between 225 to 255 and next 25% values lies between 255 to 305.last 25% values lies between 305 to 340.

Therefore, by comparing the both boxplots, we can say that the median value of both distribution is same but in 1st distribution most of value is similar or nearly comparison of 2nd distribution.

The IQR for 1st distribution is 280-255=25

The IQR for 2nd distribution is 305-225=80

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)
  3. P (20<MPG<50)

Solution: file attached in the name of “a\_1 Q20 sol vin”

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Solution: file attached in the of “ a\_1 Q21\_a sol vin”

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

Dataset: wc-at.csv

Solution : file attached in the of “a\_1 Q21\_b sol vin”

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

Solution: file attached in the name of “a\_Q22 sol vin”

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

Solution: file attached in the name of “a\_1 Q23 sol vin”

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

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

solution:

assuming null hypothesis as h0 = avg life of bulb = 260 days

alternative hypothesis as h1 = avg life of bulb ≠ 260 days

for calculation refer file “a\_1 Q24 sol vin”

the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.3216