|  |  |
| --- | --- |
| 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 | Interval |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| Religious Preference | Ordinal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

**ANS :** three coins are tossed

Total number of events = 2 ^ 3 = 8

Possible outcomes = {HHH,HHT,HTH,HTT,THH,TTH,THT,TTT}

Probability of getting two heads and one tail is :

P = interested outcome/ total outcomes

Therefore P = **3/8**

Thus the probability of getting two heads and one tail on tossing three coins is 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 :** Two dice are rolled

Total number of events = 6 ^ 2

Possible 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. Sum equal to 1 : **0/36** = **0**
2. Sum less than or equal to 4 : **6/36** = **1/6**
3. Sum is divisible by 2 and 3 : **22/36**

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 number of outcomes = 2+3+2 = 7

Two balls are drawn at random = nC2 = 7C2

= 7 x 6/2 x 1

**Total events = 21**

Balls remaining in bag = 5

Probability of getting none of the balls as blue = 5C2 (since no. of ways to draw the balls out are 2)

= 5 x 4/2 x1

**Interested events = 20/2 = 10**

**So** probability of getting none of the balls as blue is,

= **Interested events / Total events**

= **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.

Child B – probability of having 4 candies = 0.20

**ANS :** Expected value in probability is calculated as summation(x.p(x))

Therefore,

the Expected number of candies for a randomly selected child =

(1\*0.05) + (4\*0.20) + (3\*0.65) + (5\*0.05) + (6\*0.01) + (2\*0.120)

= 0.05 + 0.80 + 1.95 +0.25 + 0.06 + 0.240

= **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.

**ANS:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **Std. Dev.** | **Range** |
| **Points** | 3.596563  Average value for points | 3.695  Middle value | 3.92  Most repeated value | 0.27694  Variability | 0.5262  Std. dev from mean | 2.76-4.93  2.76-min  4.93-max |
| **Score** | 3.21725 | 3.325 | 3.44 | 0.92461 | 0.9630 | 1.513-5.424 |
| **Weigh** | 17.84875 | 17.71 | 17.71 | 3.09338 | 1.7588 | 14.5-22.9 |

Inference :

1) For Points dataset:

- The data is concentrated around median, there are no outliers

- Since the mean > median distribution is Right/+ve skewed

- Since it has 2 peaks it is -ve Kurtosis

2) For Score dataset:

- There are two outliers 5.424, 5.345

- Since the mean < median distribution is Left/ -ve skewed

- Since it has 2 peaks it is -ve Kurtosis

3) For Weight dataset: - There is one outlier 22.9

- Since the mean < median distribution is Left/ -ve skewed

- Since it has only 1 peak it is +ve Kurtosis

**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 :** Assume one of the patients is chosen at random :

Probability of choosing a patient at random = 1/9

Expected value = (108+110+123+134+135+145+167+187+199)/9

= 1308/9

= **145.33**

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

**Cars speed and distance**

|  |  |  |  |
| --- | --- | --- | --- |
|  | skewness | | kurtosis |
|  |  |  |  |
| speed | -0.11751 |  | 0.405053 |
|  |  |  |  |
| distance | 0.806895 |  | 0.405053 |

Inference:

1) Speed is left skewed

It has negative kurtosis and since it is flatter than the normal distribution it is platykurtic in nature.

It has no outlier the data is distributed around the median.

2) Distance is right skewed and has positive kurtosis.

Since it is peaked than normal distribution it is leptokurtic in nature.

It has one outlier 120

**Use Q9\_a.csv**

**SP and Weight(WT)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | skewness | | kurtosis |
|  |  |  |  |
| SP | 1.61145 |  | 2.977329 |
|  |  |  |  |
| WT | -0.61475 |  | 0.950291 |

1) SP is Right skewed and has negative kurtosis

Since it is peaked than normal distribution it is leptokurtic.

It has multiple outliers and the data is not just distributed around the median.

2) WT is left skewed and has positive kurtosis.Since it is peaked than normal distribution it is leptokurtic in nature.It has multiple outliers and the data is not just distributed around the median

**Use Q9\_b.csv**

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



**ANS :** The given graph is positive/ Right skewed since the mean> median>mode.Since it is peaked than normal distribution the distribution seems leptokurtic in nature and is heavy tailed.

There are multiple outliers on the box plot. Considering labels for the given box plots quartiles as Q1, Q2, Q3.

Q1- Lower Quartile

Q2- Median

Q3- Upper Quartile

In the given scenario (Q3-Q2) > (Q2-Q1)

The length of upper whisker is more than lower, and the position of median is more towards the lower quartile. In this case data is skewed to the right.

**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 :** for the random sample the given data that we have is :

Sample population(n) = 2000

Sample mean(x bar) = 200

Standard deviation(s) = 30

Population (N) = 3,000,000

Confidence Interval Estimate= Z => 200 Z

94% Confidence: qnorm(0.97)

[1] 1.880794=Z

200 1.88\* =**198.74 – 201.26**

98% Confidence: > qnorm(0.99)

[1] 2.326348=Z

200 2.33\* =**198.44-201.56**

96% Confidence: > qnorm(0.98)

[1] 2.053749

200 2.05\* = **198.62-201.38**

**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 = 41.176 , Median = 41 , Mode = 41

Std. deviation = 4.9968 , Variance = 24.9688

2)  Mean > Median, This implies that the distribution is slightly skewed towards right. No outliers are present.

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

**ANS :**  normalized skewness

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

**ANS :** Positive / right skewed

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

**ANS :** Negative / left skewed

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

**ANS :** This means the distribution is leptokurtic in nature or it is more peaked than the normal distribution.

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

**ANS :** This means the distribution is platykurtic or flatter as compared with normal distribution with the same Mean and Standard Deviation.

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



What can we say about the distribution of the data?

**ANS :** The data is concentrated more towards the left of the median and the whisker at the left is longer than right.

The data distribution does not seem symmetric in nature.

What is nature of skewness of the data?

**ANS :** The given box plot is left skewed or negative skewed.

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

**ANS :** Considering value for lower quartile as 10 and Upper quartile as 18

IQR = Upper Quartile- Lower Quartile

= 18-10

= 7 approximately

Q19) Comment on the below Boxplot visualizations?



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

**ANS :** Considering the equal concentration of mass at the both sides of the median the given box plots seem symmetric in nature.

The whiskers of both the box plots are equal in length.

There are no outliers for both the box plots and the data is concentrated around the median.

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)

**ANS :** > mean(Cars$MPG)

[1] 34.42208

P(MPG>38):

> sd(Cars$MPG)

[1] 9.131445

> pnorm(38,34.42,9.13)

[1] 0.652513

P(MPG>38)=1-P(MPG<38)

>1 - 0.65

[1] 0.35

P(MPG<40):

pnorm(40,34.42,9.13)

[1] 0.7294571

P (20<MPG<50):

> pnorm(50,34.42,9.13)-pnorm(20,34.42,9.13)

[1] 0.8989178

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS :** To check Normal distribution we can plot a Q-Q Plot

MPG of cars followa 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 :** the Adipose Tissue (AT) and Waist Circumference(Waist) follows Normal Distribution



Adipose tissue follows Normal Distribution



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

**ANS :** 90% 🡪> qnorm(0.95)

[1] 1.644854

94% 🡪> qnorm(0.97)

[1] 1.880794

60% 🡪> qnorm(0.8)

[1] 0.8416212

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

**ANS :** 95% 🡪> qt(0.975,24)

[1] 2.063899

96% 🡪> qt(0.98,24)

[1] 2.171545

99% 🡪 qt(0.995,24)

[1] 2.79694

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 :** Solution: n = 18  
sigma = 90  
mu(pop mean) = 270  
xbar(sample mean) =260

n = 18

sigma = 90

mu = 270

xbar=260

t = (xbar - mu)/(sigma/np.sqrt(n))

t = -0.4714

p\_val = stats.t.cdf(t,17)

p\_val: 0.3216

For the significance value (α) = 0.05  
Since p\_val(0.32) > α(0.05) we conclude that we can accept Ho