|  |  |
| --- | --- |
| 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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

Ans. When three coins are tossed, sample space S = {HHH, HHT, HTT, TTT, TTH, THT, HTH, THH} n(S) = 8

Let A is the event that two heads and one tail is obtained

Therefore, A = {HHT, HTH, THH}, n(A) = 3;

p(A) = =

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. When two dice are rolled, sample space

S = {(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)}

n(S) = 36

1. Let A is the event that sum of the outcome euals to 1

A = 0 Therefore, p(A) = 0

1. Let Bis the event that sum of the outcome is less than or equals to 4

B = {(1,1),(1,2),(1,3),(2,2),(2,1),(3,1)} n(B) = 6

p(B) = = =

1. Let C is the event that sum of the outcome is divisible by 2 & 3

C = {(3,3), (4,2), (2,4), (6,6), (1,5), (5,1)} n(C) = 6

P(C) = = =

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 balls = (2 + 3 + 2) = 7

Let S be the sample space.

Then, n(S) = Number of ways of drawing 2 balls out of 7

=7C2

=(2×1)(7×6)

=21

Let E be the event of drawing 2 balls, none of which is blue.

∴n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.

=5C2

=(2×1)(5×4)

=10

∴P(E) = (n(E))/(n(S)) = 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 number of candies for a randomly selected child

=  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 number 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**

**Ans.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | POINTS | SCORE | WEIGH |
| Mean | 3.5965 | 3.21725 | 17.8487 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.891 | 3.54 | 17.43 |
| Variance | 0.285 | 0.957 | 3.1931 |
| Standard Deviation | 0.5346 | 0.9784 | 1.7869 |
| Range | 4.93 – 2.76 = 2.17 | 5.42 - 1.51= 3.911 | 22.9 – 14.5 = 8.4 |

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 value of weight= = = 145.3333

Expected value for weight of randomly selected patient is 145.3333

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Index** | **Speed** | **Distance** |
| **Skewness** | **0.00000** | **-0.117510** | **0.806895** |
| **Kurtosis** | **-1.20000** | **-0.508994** | **0.405053** |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans.**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| **Skewness** | **1.611450** | **-0.614753** |
| **Kurtosis** | **2.977329** | **0.950291** |

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

**Ans**. Given data is positively skewed.



**Ans**. Outliers present in data 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?

**Ans.** For 94% confidence interval Range is **(198.73 – 201.26)**

For 98% confidence interval range is **(198.43 – 201.56)**

For 96% confidence interval range is **(198.62 – 201.37)**

**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.

**Ans.**

Mean =41

Median =40.5

Variance = 24.11

Standard Deviation = 4.91

1. What can we say about the student marks?

**Ans.** we don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

**Ans.** No skewness is present we have a perfect symmetrical distribution

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

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

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

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

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

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

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

**Ans.** A negative kurtosis value indicates that the distribution has lighter tails thanthe normal distribution.

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 a skewed towards left. The whisker range of minimum value is greater than maximum.

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

**Ans.** The Inter Quantile Range = Q3 Upper quartile – Q1 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.** First there are no outliers. Second both the box plot 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 whisker 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)

Ans: 0.3475939251582705

* 1. P(MPG<40)

Ans: 0.7293498762151616

* 1. P (20<MPG<50)

Ans: 1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Mean of MPG = 34.422075728024666

Median of MPG = 35.15272697

Therefore mean and median of MPG column nearly quals to each other hence they are normally

Distributed.

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

Dataset: wc-at.csv

**Answer-** Adipose Tissue (AT) and Waist does not follow Normal Distribution

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

# Z-score of 90% confidence interval

**answer= 1.6448536**

# Z-score of 94% confidence interval

**answer= 1.8807936**

# Z-score of 60% confidence interval

**answer= 0.841621**

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

# t scores of 95% confidence interval for sample size of 25

**Answer=2.0638985616280205**

# t scores of 96% confidence interval for sample size of 25

**Answer=2.1715446760080677**

# t scores of 99% confidence interval for sample size of 25

**Answer=2.796939504772804**

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

from scipy import stats

from scipy.stats import norm

# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

t**=**(260**-**270)**/**(90**/**18**\*\***0.5)

t

t score= -**0.4714045207910317**

# Find P(X>=260) for null hypothesis

# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

p\_value

P value**= 0.32167411684460556**