|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continous |
| Weight of Gold | Continous |
| Distance between two places | Continous |
| Length of a leaf | Continous |
| Dog's weight | Continous |
| 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 | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Answer = 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

Answer:

1. 0
2. 6/36 = 0.166
3. 18/36 = 0.5

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?

Answer:

Total Balls = 2+3+2 = 7

Total probability = (7X6)/(2X1) = 21

Probability to not draw blue ball = (5X4)/(2X1) = 10

= 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

Answer : Expected number of candies = 3.09

Formula = (1\*0.015)+(4\*.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120)

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**

Answer:

For Points : Mean – 3.59, Median – 3.69, Mode – 3.07 and 3.92, Variance – 0.28, STD – 0.53, Range – 2.17

For Score : Mean – 3.217, Median – 3.325, Mode – 3.44, Variance – 0.957, STD – 0.978, Range – 3.910

For Weigh : Mean – 17.848, Median – 17.71, Mode – 17.02 and 18.90, Variance – 3.193, STD – 1.786, Range – 8.399

Points and Weigh have 2 modes, so are bimodel in nature. Weigh have wide range. Points and Score have almost similar dataset considering mean and median values.

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?

Answer: Expected value = 143.88

Probability of each patient = 1/9 = 0.11

Formula = (108\*0.11)+(110\*0.11)+(123\*0.11)+(134\*0.11)+(135\*0.11)+(145+0.11)+(167\*0.11)+(187\*0.11)+(199\*0.11)

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

**Cars speed and distance**

**Use Q9\_a.csv**

Answer: Speed : Skewness = -0.117, Kurtosis = -0.508

Distance : Skewness = 0.806, Kurtosis = 0.405

**SP and Weight(WT)**

**Use Q9\_b.csv**

Answer: SP : Skewness = 1.611, Kurtosis = 2.977

Weight : Skewness = -0.614, Kurtosis = 0.950

Speed and Weight follows negative skewness, which means distribution has tail at starting.

Speed follows negative kurtosis, which means graph of speed has thin or no tails.

SP follows positive skewness, which means distribution has tail at the end and positive kurtosis, which means distribution has sharp peak and wide tails.

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



Answer: For histogram, we have positively skewed data as tail is forming at the end and Mean>Median.

For boxplot, we have outlier at the start, so data is negatively skewed, as tail will form at the start.

**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?

Answer:

94% Confidence interval = 198.73 to 201.26

98% Confidence interval = 198.44 to 201.55

96% Confidence interval = 198.62 to 201.37

Use formula for z values : For 94% - stats.norm.ppf(.970),

for 98% - stats.norm.ppf(.990) and for 96% - stats.norm.ppf(.980)

Confidence interval = mean+z\*(std/sqrt(n) and mean-z\*(std/sqrt(n)

**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?

Answer

Let df=pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

1. Mean = 41 from python code => df.mean()

Median = 40.5 from python code => df.median()

Variance = 25.52 from python code => df.var()

Standard deviation = 5.052 from python code => df.std()

1. We can say that students marks vary with each other with a range of 22 with highest mark being as 56 and lowest as 34.

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

Answer: Data is perfectly symmetric.

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

Answer: Data is positively skewed.

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

Answer: Data is negatively skewed.

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

Answer: Graph have more peakness and more values lie on tail

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

Answer: Graph have less peakness and more values lie in middle.

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



What can we say about the distribution of the data?

Answer: It is not a normal distribution and more values lie between value 10 to 18 (on graph).

What is nature of skewness of the data?

Answer: Data is negatively skewed.

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

Answer: IQR = Q3-Q1 = 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.

Answer: Both are normally distributed

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)

Answer:

a) P(MPG>38) is 0.35

Formula=1 - stats.norm.cdf(38,loc=MPG.mean,scale=MPG.std)

1. P(MPG<40) is 0.73 by

Formula= stats.norm.cdf(40,loc=MPG.mean,scale=MPG.std)

1. P(20<MPG<50) is 0.90

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer: MPG follows normal distribution as per boxplot and histogram

plt.boxplot(cars['MPG']) and plt.hist(cars['MPG'])

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

Dataset: wc-at.csv

Answer: Both AT and Waist follows normal distribution

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

Answer:

90% confidence interval – Z = 0.82 by (stats.norm.cdf(.950))

94% confidence interval – Z = 0.83 by (stats.norm.cdf(.970))

60% confidence interval – Z = 0.78 by (stats.norm.cdf(.800))

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

Answer:

t scores of 95% confidence interval = 2.06 by (stats.t.ppf(0.975,24))

t scores of 95% confidence interval = 2.17 by (stats.t.ppf(0.980,24))

t scores of 95% confidence interval = 2.79 by (stats.t.ppf(0.995,24))

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

Answer: t = (sample mean – population mean)/(std./sqrt(n))

= 260-270/(90/sqrt(18))

= - 0.47

Therefore, probability = stats.t.cdf(-0.47,17) = 0.32