Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| 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 |

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

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

Ans) Total number of events= {hhh, hht, htt, ttt, tth, thh, hth, tht} =8

Interested events=3

Probability=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)** Total number of outcomes when two dice are rolled=6\*6=36.

(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) Equal to 1 = **0% probability**

b) Less than or equal to 4= 6/36 = **1/6**

c) sum is divisible by 2 and 3

{ 2 3 4 5 6 7

3 4 5 6 7 8

4 5 6 7 8 9

5 6 7 8 9 10

6 7 8 9 10 11

7 8 9 10 11 12}

Probability=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?

Ans) total number of events= ===21

Interested events===10

Probability that none of the balls is blue =10/21=0.47

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.

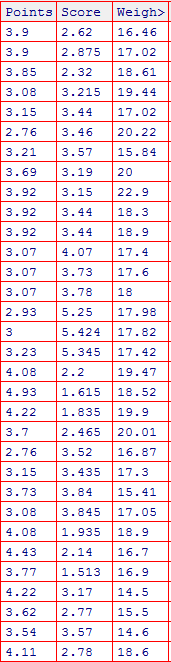
Child B – probability of having 4 candies = 0.20

Ans) Expected number = E(x) ==1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120= **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) df.describe()

(#will describe the data sets by providing values like mean, std, max and min, etc)

df.mode() (#will show the mode fo each column)

df.var() (#will provide the variance for given dataset for columns with numerical 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?

Ans: EV=Σx/n ==145.33

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

**Cars speed and distance**

****

Ans)

print('skewness value for speed and distance is', np.round(df1.speed.skew(), 2),

'and', np.round(df1.dist.skew(), 2), 'respectively')

skewness value for speed and distance is -0.12 and 0.81 respectively

print('Kurtosis value for speed and distance is', np.round(df1.speed.kurt(), 2),

'and', np.round(df1.dist.kurt(), 2), 'respectively')

Kurtosis value for speed and distance is -0.51 and 0.41 respectively

Inferences: as you can see from the above data, there is a huge difference in the kurtosis values when e1071 and moments package are compared with each other. This is due to different equations used by the packages to find kurtosis.

**Q.9b)**

****

**SP and Weight(WT)**

Ans)

print('skewness value for SP and WT(weight) is', np.round(df2.SP.skew(), 2),

'and', np.round(df2.WT.skew(), 2), 'respectively')

skewness value for SP and WT(weight) is 1.61 and -0.61 respectively

print('Kurtosis value for SP and WT(weight) is', np.round(df2.SP.kurt(), 2),

'and', np.round(df2.WT.kurt(), 2), 'respectively')

Kurtosis value for SP and WT(weight) is 2.98 and 0.95 respectively

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



.

Ans: The above boxplot suggests that the distribution has lots of outliers towards upper extreme

**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: AVG\_WGT1 = stats.norm.interval(0.97, loc = 200, scale = 30)

print('Average weight of adult in Mexico at 94% confidence interval', np.round(AVG\_WGT1, 3))

Average weight of adult in Mexico at 94% confidence interval [134.897 265.103]

AVG\_WGT2 = stats.norm.interval(0.99, loc = 200, scale = 30)

print('Average weight of adult in Mexico at 98% confidence interval', np.round(AVG\_WGT2, 3))

Average weight of adult in Mexico at 98% confidence interval [122.725 277.275]

AVG\_WGT3 = stats.norm.interval(0.98, loc = 200, scale = 30)

print('Average weight of adult in Mexico at 96% confidence interval', np.round(AVG\_WGT3, 3))

Average weight of adult in Mexico at 96% confidence interval [130.21 269.79]

**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) df = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

df = pd.DataFrame(df)

df.mean()

0 41.0

dtype: float64

df.median()

0 40.5

dtype: float64

df.std()

0 5.052664

dtype: float64

df.var()

0 25.529412

dtype: float64

**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) no skewness, symmetric

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

Ans) Right skewed(tail on the right side).

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

Ans) Left Skewed(tail on the left side).

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

Ans) peakness (sharp peak) and less variation.

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

Ans) less peakness (Broad peak) and more variation.

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



What can we say about the distribution of the data?

Ans) it is not a Normal Distribution

What is nature of skewness of the data?

Ans) It is left skewed.

What will be the IQR of the data (approximately)?   
Ans) Inter Quartile Range =Upper Quartile- 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) 1) The median of the two boxplots are same approximately 260.

2) The boxplots are not skewed in +ve or –ve direction.

3) Outliers doesn’t exist in both of the boxplots.

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) print("Probabilty that 'MPG' > 38 = ", np.round(1-stats.norm.cdf(38, loc = df4.MPG.mean(), scale = df4.MPG.std()), 3))

Probabilty that 'MPG' > 38 = 0.348

print("Probabilty that 'MPG' < 40 = ", np.round(stats.norm.cdf(40, loc = df4.MPG.mean(), scale = df4.MPG.std()), 3))

Probabilty that 'MPG' < 40 = 0.729

print("Probabilty that 20 <'MPG' < 40 = ", np.round((1-stats.norm.cdf

(20, loc = df4.MPG.mean(), scale = df4.MPG.std())) -

(stats.norm.cdf(40, df4.MPG.mean(), scale = df4.MPG.std())) , 3))

Probabilty that 20 <'MPG' < 40 = 0.214

print("Probabilty that 'MPG' < 40 = ", np.round(stats.norm.cdf(70, loc = 60, scale = 10), 5))

Probabilty that 'MPG' < 40 = 0.84134

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans) Follows Normal distribution as indicated by qq-plot.

import statsmodels.api as smf

import pylab as py

smf.qqplot(df["MPG"],line='45')

py.show()



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

Dataset: wc-at.csv

Ans) waist follows Normal Distribution from the below QQ-plot

> import statsmodels.api as smf

import pylab as py

smf.qqplot(df5["Waist"],line='45')

py.show()



import statsmodels.api as smf

import pylab as py

smf.qqplot(df5["AT"],line='45')

py.show()



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

Ans) print('Z scores at 90% confidence interval is', np.round(stats.norm.ppf(.95), 2))

print('Z scores at 94% confidence interval is', np.round(stats.norm.ppf(.97), 2))

print('Z scores at 60% confidence interval is', np.round(stats.norm.ppf(.80), 2))

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

Ans) print(' t scores at 95% confidence interval is', np.round(stats.t.ppf(0.975, df = 24), 2))

print(' t scores at 96% confidence interval is', np.round(stats.t.ppf(0.98, df = 24), 2))

print(' t scores at 99% confidence interval is', np.round(stats.t.ppf(0.995, df = 24), 2))

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) t\_value = (260 - 270)/(90/np.sqrt(18))

print('critical value = ', np.round(t\_value, 2))

print('probabilty for average life of no more than 260 days is', np.round(stats.t.cdf(t\_value, df=17), 2))

**critical value = -0.47**

**probabilty for average life of no more than 260 days is 0.32**