|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | discrete |
| Results of rolling a dice | discrete |
| Weight of a person | Continues |
| Weight of Gold | Continues |
| Distance between two places | Continues |
| Length of a leaf | Continues |
| Dog's weight | Continues |
| 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 | Interval |
| Hair Color | nominal |
| Socioeconomic Status | ordinal |
| Fahrenheit Temperature | interval |
| Height | interval |
| Type of living accommodation | nominal |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | ordinal |
| Sales Figures | interval |
| 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 | interval |

- Ratio includes zero values

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

Ans:3/8=.375

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1

Ans:0

1. Less than or equal to 4

Ans: 1/6=0.16666

1. Sum is divisible by 2 and 3

Ans:1/6=0.16666

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: 20/42=.47619

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:sum of(p(i)\*x(i))= 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:

Points Score Weigh

mean 3.596563 3.217250 17.848750

Std 0.534679 0.978457 1.786943

Median(50%) 3.695000 3.325000 17.710000

Var 0.285881 0.957379 3.193166

Mode 3.07,3.92 3.44 17.02,18.90

Min 2.76000 1.513000 14.50000

Max 4.930000 5.424000 22.900000

-Min -max will give the range

- weigh and score is having outliers

-weigh is normally distributed

**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:

b=pd.DataFrame(data={'weight':[108, 110, 123, 134, 135, 145, 167, 187, 199]})

b.mean()

Mean=145.333

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Skew:**

Index -1.200000

speed -0.508994

dist 0.405053

**kurtosis:**

Index -1.200000

speed -0.508994

dist 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Skew:**

SP 1.611450

WT -0.614753

**kurtosis:**

SP 2.977329

WT 0.950291

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



Ans:

* 1. Inferences about histogram:

Most common range=50 to 100

Bin size: 50

Mean: 75

Positively skewed graph

b. Inferences about box plot:

Outliers are present, positively skewed, median, quartile values and density of data can be obtained from the box plot

**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:

from scipy import stats

import math

t1=stats.t.ppf(.97,1999)

t2=stats.t.ppf(.99,1999)

t3=stats.t.ppf(.98,1999)

a1=(200+(t1\*(30/math.sqrt(2000))),200-(t1\*(30/math.sqrt(2000))))

a2=(200+(t2\*(30/math.sqrt(2000))),200-(t2\*(30/math.sqrt(2000))))

a3=(200+(t3\*(30/math.sqrt(2000))),200-(t3\*(30/math.sqrt(2000))))

(201.2623910556929, 198.7376089443071)- 94%

(201.5618139516784, 198.4381860483216)- 98%

(201.3785962570268, 198.6214037429732)-96%

**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?

import pandas as pd

import numpy as np

df=pd.DataFrame(data={'a':[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]})

df.mean()

df.median()

df.mode()

df.std()

df.var()

Mean: 41.0, median:40.5, variance: 25.529412, standard deviation: 5.052664

Positively skewed(mean>median)

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

Zero skewness

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

positive skewness

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

negative skewness

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

Distribution is peaked

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

Distribution is more flat compared to normal distribution

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



What can we say about the distribution of the data?

Negatively skewed distribution, and not a normal distribution

What is nature of skewness of the data?

Negatively skewed

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

IQR= 18-12=8

Q19) Comment on the below Boxplot visualizations?



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

Both have the same mean value (262.5)

IQR range of second is higher than first

Kurtosis is more in first box plot than second

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)

car=pd.read\_csv("C:/Users/ashiq/Desktop/csv/assignment\_1/Cars.csv")

from scipy import stats

r=car['MPG']

an1= 1-(stats.norm.cdf(38,loc=car.MPG.mean(),scale=car.MPG.std()))

an2= (stats.norm.cdf(40,loc=car.MPG.mean(),scale=car.MPG.std()))

an3= (stats.norm.cdf(50,loc=car.MPG.mean(),scale=car.MPG.std()))-(stats.norm.cdf(20,loc=car.MPG.mean(),scale=car.MPG.std()))

a. 0.3475939251582705

b. 0.7293498762151616

c. 0.8988689169682046

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Mean!=median,

So not normal disrtibution

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

Dataset: wc-at.csv

Mean!=median,

So not normal distribution in both cases

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

Z1: 1.6448536269514722

Z2: 1.8807936081512509

Z3: 0.8416212335729143

z1=stats.norm.ppf(0.95)

z2=stats.norm.ppf(0.97)

z3=stats.norm.ppf(0.8)

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

T1: 2.0638985616280205

T2: 2.1715446760080677

T3: 2.796939504772804

T3=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

Ans:p=(stats.norm.cdf(260,loc=260,scale=90))

P=0.5

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom