**SURESH LOKU RATHOD**

**DATA SCIENCE BATCH 1ST FEBRUARY 2023**

**ASSIGNMENT NO 1-STATISTIC BASIC 1**

ASSIGNMENT.DOCX

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

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

* {HHH,HHT,HTT,HTH,THT,TTH,THH,TTT}
* n=8
* Probability That Two Head And One Tail Are Obtained Is
* =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

* S=Sample Space

= (1,1)(2,1)(3,1)(4,1)(5,1)(6,1)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(1,3)(2,3)(3,3)(4,3)

(5,3)(6,3)(1,4)(2,4)(3,4)(4,4)(5,4)(6,4)(1,5)(2,5)(3,5)(4,5)(5,5)(6,5)(1,6)(2,6)

(3,6)(4,6)(5,6)(6,6)

* n(S)=36

1. p(Sum is equal to 1)

=0/36

=**0**

1. s={(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

n(s)=6

p(sum is less than or equal to 4)

=6/36

=**0.1666**

1. s={(1,5),(2,4),(3,3),(4,2),(5,1)}

n(s)=5

p (sum is divisible by 2 and 3)

=5/3

=**1.6666**

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?

S =sample space

n(S)= 7C2=21

p(none of the balls drawn id blue)=( 2C1\*3C1+2C2+3C2)/21

=**0.4762**

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

E (No. of candies for children)

=1\*0.015 +4\*0.20 +3\*0.65 +5\*0.005 +6\*0.01 +2\*0.120

= **3.09**

Therefore, expected number of candies for children is 3.09 ~ 3.

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**

**Python code:**

*import pandas as pd*

*data=pd.read\_csv(r"\Users\Downloads\Q7.csv")*

*d=pd.DataFrame(data)*

*d.Points.mode()*

Output:

0 3.07

1 3.92

dtype: float64

d.Score.mode()

0 3.44

dtype: float64

d.Weigh.mode()

0 17.02

1 18.90

dtype: float64

|  |  |  |  |
| --- | --- | --- | --- |
|  | points | score | weigh |
| Count | 32 | 32 | 32 |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Std | 0.534679 | 0.978457 | 1.786943 |
| Min | 2.76 | 1.513 | 14.5 |
| 25% | 3.08 | 2.58125 | 16.8925 |
| 50% | 3.695 | 3.325 | 17.71 |
| 75% | 3.92 | 3.61 | 18.9 |
| Max | 4.93 | 5.424 | 22.9 |

**Conclusion**: Here we can see that the average of data is 3.59, 3.21, 17.84 respectively, Weigh std = 1.78 and var = 3.19 are high as compared to others, Points and Weigh have two modes, Points has a low range = 2.17 as compared to others.

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?

P(choosing one of the patients at random)=1/9=0.1111

E(x)=x\*p(X=x)

= (108\*0.1111)+( 110\*0.1111)+ …….+( 199\*0.1111)

**= 145.3188 ~ 145**

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

**Cars speed and distance**

**Use Q9\_a.csv**

*data=pd.read\_csv(r"\Users\Downloads\Q9\_a.csv")*

*data.skew()*

speed -0.117510

dist 0.806895

data.kurt()

speed -0.508994

dist 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

*data1=pd.read\_csv(r"\Users\Downloads\Q9\_b.csv")*

*data1.skew()*

SP 1.611450

WT -0.614753

*data1.kurt()*

SP 2.977329

WT 0.950291

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



From the above histogram we can conclude that the data is positively skewed.



**Conclusion:** From the above boxplot we can detect some outliers are present in data and we can observe that the data is positively skewed since most of the observations lie on lower end .

**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?

* 94% confidence interval

*from scipy import stats*

*c1=stats.norm.interval(0.94,loc=200,scale=30)*

(143.57619175546247, 256.42380824453755)

* 96% confidence interval

*from scipy import stats*

*c2=stats.norm.interval(0.96,loc=200,scale=30)*

(138.38753268104531, 261.61246731895466)

* 98% confidence interval

*from scipy import stats*

*c3=stats.norm.interval(0.98,loc=200,scale=30)*

(130.2095637787748, 269.7904362212252)

**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?

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

*a.describe()*

|  |
| --- |
| count 18.000000 |
| mean 41.000000 |
| std 5.052664 |
| min 34.000000 |
| 25% 38.250000 |
| 50% 40.500000 |
| 75% 41.750000 |
| max 56.000000 |

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

* : Symmetric

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

* : Positively Skewed

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

* : Negatively Skewed

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

* : Positive value for kurtosis indicates that the distribution is leptokurtic i.e. the distribution is having more peak than the normal distribution.

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

* : Negative value for kurtosis indicates that the distribution is platykurtic i.e. the distribution is having less peak than the normal distribution.

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



What can we say about the distribution of the data?

* : Density of the data is more on the right side.

What is nature of skewness of the data?

* : The data is negatively skewed.

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

* : here , Q1=10, Q3=18

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.

* From both box plot we can observe that there is no outliers in the data.
* Median of both the data is same.
* Compare to the first box plot, second box plot has more variation in the data.
* IQR of second box plot is higher than first box plot.

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)

*import pandas as pd*

*import numpy as np*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

*from scipy import stats*

*cars=pd.read\_csv(r"C:\Users\rajes\Downloads\Cars.csv")*

*plt.boxplot(cars['MPG'])*

*plt.ylabel("MPG")*

*plt.title("Boxplot")*

*plt.show()*

Chart, box and whisker chart

Description automatically generated

*plt.hist(cars['MPG'])*

*plt.xlabel("MPG")*

*plt.ylabel("Frequency")*

*plt.title("Histogram")*

*plt.show()*

Chart, histogram

Description automatically generated

a)

*#P(MPG>38)*

*1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())*

0.3475939251582705

b)

*#P(MPG<40)*

*stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())*

0.7293498762151616

c)

*#P(20<MPG<50)*

*stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())*

0.8988689169682046

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



**:- ( Kolmogorov test for normality (**N < 5000**)**

H0 : The data is normal. v/s

H1 : The data is not normal

*import pandas as pd*

*import numpy as np*

*import matplotlib.pyplot as plt*

*Import seaborn as sns*

*from scipy import stats*

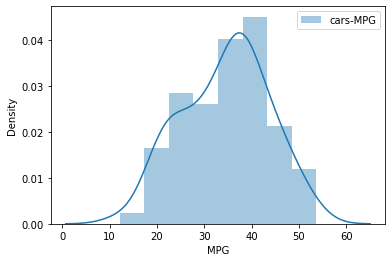
*sns.distplot(cars.MPG,label='cars-MPG')*

*plt.xlabel('MPG')*

*plt.ylabel('Density')*

*plt.legend();*

*plt.show()*



Conclusion: here the p-value is less than 0.05, so we reject the null

Hypothesis and conclude that the data is not normal.

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

Dataset: wc-at.csv



**(Kolmogorov test for normality (**N < 5000**)**

H0: The data is normal.

H1: The data is not normal.

*import pandas as pd*

*import numpy as np*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

*from scipy import stats*

*df=pd.read\_csv(r"C:\Users\Downloads\wc-at.csv")*

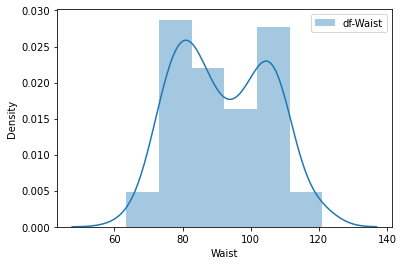
*sns.distplot(df.Waist,label='df-Waist')*

*plt.xlabel('Waist')*

*plt.ylabel('Density')*

*plt.legend();*

*plt.show()*



*sns.distplot(df.AT,label='df-AT')*

*plt.xlabel('AT')*

*plt.ylabel('Density')*

*plt.legend();*

plt.show()

Chart, histogram

Description automatically generated

Conclusion: Here the p-value is less than 0.05, so we reject the null

Hypothesis and conclude that the data is not normal.

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

*import pandas as pd*

*import numpy as np*

*from scipy import stats*

*# z-score of 90% C.I*

*stats.norm.ppf(0.95)*

1.6448536269514722

*# z-score of 94% C.I*

*stats.norm.ppf(0.97)*

1.8807936081512509

*# z-score of 80% C.I*

*stats.norm.ppf(0.8)*

0.8416212335729143

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



*import pandas as pd*

*import numpy as np*

*from scipy import stats*

*# z-score of 90% C.I*

*stats.norm.ppf(0.95)*

1.6448536269514722

*# z-score of 94% C.I*

*stats.norm.ppf(0.97)*

1.8807936081512509

*# z-score of 80% C.I*

*stats.norm.ppf(0.8)*

0.8416212335729143

2.1665866344527562

*# z-score of 99% C.I*

*stats.t.ppf(0.995,25)*

2.787435813675851

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



Solution:

p(x<260)=?

n=18

*> xbar=260*

*> s=90*

*> mu=270*

*> tscore=(xbar-mu)/(s/sqrt(n));tscore*

**-0.4714045**

*> pt(tscore,17)*

**0.3216725**