Name :- Rajshri Narsing Ekbote

8th feb 2022 Data Science Weekday Batch by Prasad Oruganti.

STATISTICS

Assignment -1

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 | 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 | Interval |
| SAT Scores | Ordinal |
| Years of Education | Ratio |

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

Ans :- sample s={HHH,HHT,HTH,THH,HTT,TTH,TTT}

Probability of getting two heads and one tail{HHT,HTH,THH}

3/8 or 0.375

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

1. Equal to 1

Ans : - o

1. Less than or equal to 4

Ans : - 1/6

1. Sum is divisible by 2 and 3

Ans : - 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 :- 10/21

Total no o balls = 2+3+2= 7

P (E)= n(E)/n(s) (5\*42\*1)/(7\*62\*1)=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 :- Probability o expected no. o candies for randomly selected child

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 :- .csv file**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean** | 3.596563 | 3.21725 | 17.84875 |
| **Median** | 3.695 | 3.325 | 17.71 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Varience** | 0.285881 | 0.957379 | 3.193166 |
| **Standard Deviation** | 0.534679 | 0.978457 | 1.786943 |
| **Range** | 2.17 | 3.911 | 8.4 |

* Mean of a point and score are approximately same but we can say that the mean of weigh is greater than the mean of point and score.
* Median of a points and score are approximately same but median of weigh is greater than median of a point and score.
* Mode is approximately same of points and score and mode of weigh is greater than that of mode of points and score .
* If we observe the variance of points and score there is less variance in points as compared to score but there is high variance of weigh as compared to points and score here we can say that low variance is good for our whatever may be its outcome.
* Here std. dev. of weigh is more than the std. dev. of score and points and score has more std. dev. than points but it is observed that points have a less std. dev. than score and weigh.
* Range is different for points, score and weigh.

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:- Consider, where is weight of patient.

X=(108,110,123,134,145,167,187,199,) P(X)=(1/9+1/9+1/9+1/9+1/9+1/9+1/9+1/9+1/9)

Sum of P(X)=9/9=1 We know that

E(X) = mean = X.P(X)

=1/9(108+110+123+134+135+145+167+187+199)

=145.33

Expected value of the weight of patient chosen at random is 145.33

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

**Cars speed and distance**

**Ans :- a.csv**

|  |  |  |
| --- | --- | --- |
|  | **speed** | **distance** |
| **skewness** | **-0.1139548** | **0.7824835** |
| **kurtosis** | **-0.6730924** | **0.1193971** |

**Car speed is negatively high skewed and distance is positively skewed.**

**SP and Weight(WT)**

**Ans :- b.csv**

|  |  |  |
| --- | --- | --- |
|  | **sp** | **wt** |
| **skewness** | **1.552258** | **-0.5921721** |
| **kurtosis** | **2.583072** | **0.7257402** |

* **Sp is positively high skewness am WT is negatively moderated skewed.**

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



Ans:-

- we can observed from the histogram that the histogram positively high skewed

And it has some flatness so we can say that it is lpatykurtic.

* The boxplot have some outlier on the upper extreme o the boxplot the deviation from the upper quartile to upper extream is high where as deviation is low from quartile to lower extream.

**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:- n=2000

std dev(σ)=30

sample mean=300000 std error= 30/√200=0.67

α=1-(confidence interval/100)=0.06 Degree of freedom(DF)=2000-1=1999 Critical Probability=1-(α/2)=0.97

We have std dev of sample so we apply t Distribution Critical value= 1.881, error=critical value \*std error=1.26

# Confidence interval94%

=(200+1.26, 200-1.26) = (201.26, 198.74)

FOR 98% confidence interval,

α= 1-(confidence level/100)=0.02, Critical probability= 1-(0.02/2)=0.99, DF=1999,

critical value at df= 2.33,

error= critical value\*std error= 1.56

Confidence interval of 94%= (200+1.56,200-1.56) = (201.56, 198.44)

**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 . Mean =41**

**Median = 40.5**

**Variance = 25.52**

**Standard deviation =5.05664**

2. we can say that the mass of students marks between 38-42.

Skewness (1.52)is positive because mass of marks in leftside of plot.

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

Ans :- Data is normalized and there is no skewness.

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

Ans :- Negative skewness it implies mass of the distribution consentrated on

right side

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

Ans :- Positive skewness it implies mass of the distribution consentrated on

left side

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

Ans :-Positive kurtosis value indicates that thinner peak and wide tails.

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

Ans :- Negative kurtosis value indicates wide peak and thinner tails.

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



What can we say about the distribution of the data?

Ans :- Not normally distributed

What is nature of skewness of the data?

Ans :- :- Negative skewness it implies mass of the distribution consentrated on

right side and median is greater than mean.

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

Ans :- The interquartile range will be 10-18.

Q19) Comment on the below Boxplot visualizations?



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

Ans:-

* + First boxplot is symmetrical that is whisker on both the sides of first boxplot have same length.
  + In the second boxplot it is also a symmetrical because the whisker on the both side are equal.
  + By calculating IQR of the both boxplot then the IQR will be different because they have the different data sets that is we can say that they have different values of quartiles.
  + So conclusion is both boxplot are symmetrical but in nature they are different.

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)

Ans.

1) P(MPG>38)=0.3475

2) P(MPG<40)= 0.7293

3) P (20<MPG<50)=0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans :- yes, the data set follows normal distribution

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

Dataset: wc-at.csv

Ans :- yes, the data set follows normal distribution

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

Ans :-

Z scores of 90% confidence interval qnorm(0.95)

[1] 1.644854

94%confidence interval

qnorm(0.97) [1] 1.880794

60%confidene interval qnorm(0.8)

[1] 0.841621

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

Ans.

## T score of 95% confidence interval

1+0.95/2=0.975

Degree of freedom n-1=24 (0.975,24)

R code >qt(0.97,df=24) = 2.063899

## T score of 96% confidence interval

1+0.96/2=0.98 (0.98,24)

R code > qt (0.98,df=24) = 2.171545

## T score of 99% confidence interval

1+0.99/2=0.995 (0.995,24)

R code > qt (0.995,df=24) = 2.79694

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 :- mean (μ)=260 , (n)=18, x=270, std. dev.(s)=90

t stat formula:- (x-µ)/(s/√n)

=(270-260)/(90/√18)

=0.47

Degree of freedom =n-1

Then df=18-1

=17

P score = (0.47, 17) By using log table,

= 0.32 as 32%

R code pt(0.47, df=17) = 0.3221639