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

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, ordinal, interval, ratio |
| High School Class Ranking | Ordinal, nominal, interval |
| Celsius Temperature | Interval, ordinal, ratio |
| Weight | Ratio, interval, ratio |
| Hair Color | Nominal, ordinal |
| Socioeconomic Status | Ordinal, interval, ratio |
| Fahrenheit Temperature | Interval, ordinal, ratio |
| Height | Ratio, ordinal , interval |
| Type of living accommodation | Ratio, ordinal , interval |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | Ordinal, interval, ratio |
| Sales Figures | Ratio, interval |
| Blood Group | Nominal, ordinal, interval, ratio |
| Time Of Day | Nominal, ordinal, interval |
| Time on a Clock with Hands | Nominal, ordinal, interval, ratio |
| Number of Children | Ratio, ordinal, interval |
| Religious Preference | ordinal |
| Barometer Pressure | Interval & ratio |
| SAT Scores | Interval & ratio |
| Years of Education | Ordinal, interval, ratio |

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

A3) Probability= 3/8 = 0.375

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

1. Equal to 1 = 0
2. Less than or equal to 4 = 1/6
3. Sum is divisible by 2 and 3 = 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?

Answer- 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- 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.

**Use Q7.csv file**

**Answer**-

|  |  |  |  |
| --- | --- | --- | --- |
|  | points | score | weigh |
| Mean | **3.596563** | **3.217250** | **17.848750** |
| Median | **3.695** | **3.325** | **17.710** |
| Mode | **3.07** | **3.44** | **17.02** |
| Variance | **0**.**285881** | **0.957379** | **3.193166** |
| Standard Dev. | **0.534679** | **0.978457** | **1.786943** |
| Range | **2.17** | **3.9110000000000005** | **8.399999999999999** |

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- 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answer: skew:** Index 0.000000

speed -0.117510

dist 0.806895

**kurtosis:** Index -1.200000

speed -0.508994

dist 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Answer: skew:** SP 1.611450

WT -0.614753

**Kurtosis:** SP 2.977329

WT 0.950291

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



Answer

Histogram: -the data is skewed on the right side i.e. data is positively skewed

-there are no outliers in the data

Box plot: data is distributed on the right and positively skewed and there are outliers present

**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 - *Avg. weight of Adult in Mexico with 94% CI:* (198.738325292158, 201.261674707842)

*Avg. weight of Adult in Mexico with 98% CI:* (198.43943840429978, 201.56056159570022)

*Avg. weight of Adult in Mexico with 96% CI:*

(198.62230334813333, 201.37769665186667)

**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: 1. *# Mean*

41.0

*# Median*

40.5

*# Variance*

25.52941176470588

*# Standard Deviation*

5.05266382858645

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?

Answer: no skewness, symmetric

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

Answer: right skewed(tail on the right)

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

Answer: left skewed(tail on the left side)

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

Answer: peakness(sharp peak) and less variation.

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

Answer: 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?

Answer: it is not a normal distribution

What is nature of skewness of the data?

Answer: it is left skewed

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

Answer: 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.

Answer: 1. Median of both the boxplots is same aprox

2. it is not skewed in +ve and –ve direction

3. no outliers present

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: *# P(MPG>38)*

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

0.3475939251582705

*# P(MPG<40)*

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

0.7293498762151616

*# P (20<MPG<50)*

stats**.**norm**.**cdf(0.50,cars**.**MPG**.**mean(),cars**.**MPG**.**std())**-**stats**.**norm**.**cdf(0.20,cars**.**MPG**.**mean(),cars**.**MPG**.**std())

1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer: cars**.**MPG**.**mean()

34.422075728024666

cars**.**MPG**.**median()

35.15272697

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

Dataset: wc-at.csv

Answer:

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

Answer: *# Z-score of 90% confidence interval*

stats**.**norm**.**ppf(0.95)

1.6448536269514722

*# Z-score of 94% confidence interval*

stats**.**norm**.**ppf(0.97)

1.8807936081512509

*# Z-score of 60% confidence interval*

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

*# t scores of 95% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.975,24) *# df = n-1 = 24*

2.0638985616280205

*# t scores of 96% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.98,24)

2.1715446760080677

*# t scores of 99% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.995,24)

2.796939504772804

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: *# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))*

t**=**(260**-**270)**/**(90**/**18**\*\***0.5)

t

-0.4714045207910317

*# Find P(X>=260) for null hypothesis*

*# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function*

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

p\_value

0.32167411684460556

*# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function*

p\_value**=**stats**.**t**.**sf(abs(**-**0.4714),df**=**17)

p\_value

0.32167411684460556