|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | int |
| Results of rolling a dice | int |
| Weight of a person | float |
| Weight of Gold | float |
| Distance between two places | float |
| Length of a leaf | float |
| Dog's weight | float |
| Blue Color | char |
| Number of kids | int |
| Number of tickets in Indian railways | int |
| Number of times married | int |
| Gender (Male or Female) | char |

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

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

**Ans:** When three coins are tossed, there are 8 possible outcomes since each coin can either land heads (H) or tails (T). The possible outcomes are: HHH,HHT,HTH,HTT,THH,THT,TTH,TTT

Out of these, the outcomes with two heads and one tail are: HHT,HTH,THH

So probability of getting two heads and one tail=no. favourable outcomes/ no. of total outcomes = 3/8

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

1. Equal to 1

**Ans**: 0

Because will not get sum 1 after rolling two dice at a time.

1. Less than or equal to 4

**Ans:** Total possible outcomes = 36

The possible outcomes with a sum less than or equal to 4 are:

Sum = 2: (1, 1), Sum = 3: (1, 2), (2, 1), Sum = 4: (1, 3), (2, 2), (3, 1)

Probability that sum is less than or equal to 4= 6/36 = 1/6

1. Sum is divisible by 2 and 3

**Ans:** Total possible outcomes = 36

Favorable outcomes are: (1, 5),(2, 4),(3, 3),(4, 2),(5, 1), (6, 6)

Probability that 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?

**Ans:** Total balls = 2+3+2 = 7

Probability of not getting blue ball in 1st draw =5/7

Probability of not getting blue ball in 2nd draw =5/6

the probability that none of the balls drawn is blue = 5/7 \* 5/6 = 25/42

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:** E[X]=∑ P(X )\*x = (1\*0.015)+(2\*0.20)+(3\*0.65)+(4\*0.005)+(5\*0.01)+(6\*0.120) = 3.14

​

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**

**Ans: Point Score Weigh**

**Mean** 3.596563 3.21725 17.84875

**Median** 3.695 3.325 17.71

**Mode** 3.92 3.44 17.02

**Variance** 0.2858814 0.957379 3.193166

**S.D.** 0.5346787 0.9784574 1.786943

**Range** 2.76 4.93 1.513 5.424 14.5 22.9

Mean and median are approx.equal for point, score and weigh , hence we can say that they are normally distributed.

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:** E[X]=(108+110+123+134+135+145+167+187+199)/9 = 1278/9 = 142

Expected Value of the Weight of that patient = 142 pound

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:** Skewness(speed)= -0.1139548 Skewness(distance)= 0.7824835

Kurtosis(speed)= 2.422853 Kurtosis(distance)= 3.248019

Speed and distance are moderately skewed.

Kurtosis value for speed is 2.422 which indicates that distribution is too peaked.

Kurtosis value for distance is 3.248 which can be recognized as mesokurtic .

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans:** Skewness(SP) =1.581454 Skewness(WT) = -0.6033099

Kurtosis(SP) = 5.723521 Kurtosis(WT) = 3.819466

SP is highly skewed and WT is moderately skewed.

SP is highly peaked and WT isdistribution is peaked and possesses thick tails.

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



**Ans:** From above histogram we can conclude that data is positively/ right skewed.



**Ans:** From above boxplot, we get to see that there are outliers present at right side.

**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:** Confidence Interval=Sample Mean± (Critical Value×sqrt(Sample Size​ )/

Sample Standard Deviation​)

*Z*=(1−Confidence Level)​/2

1. **94% Confidence Interval:** Confidence Interval≈(198.87,201.13)
2. **98% Confidence Interval:** Confidence Interval≈(197.54,202.46)
3. **96% Confidence Interval:** Confidence Interval≈(198.14,201.86)

**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.

**Ans:** Mean= 41

Variance= 25.52941

Standard deviation= 5.052664

1. What can we say about the student marks?

**Ans:** On an average students are having 41 marks

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

**Ans:** When mean=median , we can say that data is normally distributed or symmetric in nature.

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

**Ans:** It is said that distribution is right/positively skewed.

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

**Ans:** It is said that distribution is left/negatively skewed.

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

**Ans:** Positive kurtosis indicates that there are more extreme values in the distribution than would be expected in a normal distribution.

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

**Ans:** Negative kurtosis indicates that there are fewer extreme values in the distribution than would be expected in a normal distribution.

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



What can we say about the distribution of the data?

**Ans:** Asymmetric distribution

What is nature of skewness of the data?

**Ans:** Negatively/ left skewed.

**Ans:** It is negatively skewed.

What will be the IQR of the data (approximately)?   
**Ans:** 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.

**Ans:** Data points of boxplot 1 ranges from 230 to 280 and of boxplot ranges from 190 to 335 approx. Median is same for both boxplots. No outliers are present for 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)

**Ans:** a=read.csv('D:\ExcelR\Data science\Assignments\Assignment1\\Cars.csv')

a

View(a)

b=a$highway.mpg

b

probability\_greater\_than\_38 <- sum(b > 38) / length(b)

# Print the result

cat("Probability P(MPG > 38):", probability\_greater\_than\_38)

Probability P(MPG > 38): 0.09268293

* 1. P(MPG<40)

**Ans:** probability\_less\_than\_40 <- sum(b < 40) / length(b)

cat("Probability P(MPG < 40):", probability\_less\_than\_40)

Probability P(MPG < 40): 0.9170732

* 1. P (20<MPG<50)

**Ans:** Probability\_greater\_than\_20\_less\_than\_50 <- sum(b>20 & b<50) / length(b)

cat("Probability P(20 < MPG < 50):", probability\_greater\_than\_20\_less\_than\_50)

Probability P(20 < MPG < 50): 0.9414634

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** hist(highway.mpg)

From histogram got to see that it is positively skewed.

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

Dataset: wc-at.csv

**Ans:** From histogram got to see that AT is positively skewed and Waist is normally distributed.

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

**Ans:** # Confidence levels

confidence\_levels <- c(0.90, 0.94, 0.60)

# Calculate Z-scores

z\_scores <- qnorm((1 + confidence\_levels) / 2)

# Print the results

for (i in seq\_along(confidence\_levels)) {

cat(sprintf("%d%% Confidence Interval Z-score: % .2f\n", confidence\_levels[i] \* 100, z\_scores[i]))

}

90% Confidence Interval Z-score: 1.64

94% Confidence Interval Z-score: 1.88

60% Confidence Interval Z-score: 0.84

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

**Ans:** confidence\_levels <- c(0.95, 0.96, 0.99)

sample\_size <- 25

df <- sample\_size - 1

t\_scores <- qt((1 + confidence\_levels) / 2, df)

for (i in seq\_along(confidence\_levels)) {

cat(sprintf("%d%% Confidence Interval t-score for df=%d: %.2f\n", confidence\_levels[i] \* 100, df, t\_scores[i]))}

95% Confidence Interval t-score for df=24: 2.06

96% Confidence Interval t-score for df=24: 2.17

99% Confidence Interval t-score for df=24: 2.80

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:** sample\_mean <- 260

population\_mean <- 270

sample\_sd <- 90

sample\_size <- 18

# Calculate the t-score

t\_score <- (sample\_mean - population\_mean) / (sample\_sd / sqrt(sample\_size))

# Degrees of freedom

df <- sample\_size - 1

# Calculate the probability using the t-distribution

probability <- pt(t\_score, df)

# Print the result

cat("Probability that 18 randomly selected bulbs would have an average life of no more than 260 days:", probability, "\n")

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days: 0.3216725