|  |  |
| --- | --- |
| 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 | Qualitative |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Qualitative |

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 | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ (Intelligence Scale) | Ratio |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| 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 the total number of possible combinations are 23 = 8.

These combinations are HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

The number of combinations which have two heads, and one tail are:

HHT, HTH, THH which makes them 3 in number.

Therefore, the probability of getting two heads and one tails in the toss of three coins simultaneously is defined as:

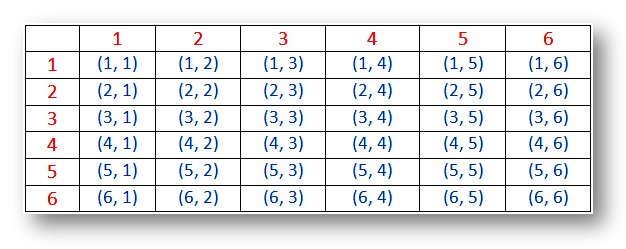
P (Two heads and One Tail) = Number of desired outcomes

= 3 = 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

Ans) Probability for rolling two dice with the six-sided dots such as 1, 2, 3, 4, 5 and 6 dots in each die. When two dice are thrown simultaneously, thus number of events can be 62 = 36 because each die has 1 to 6 number on its faces. Then the possible outcomes are shown in the below table.



A) P (Equal to 1) = **0**

B) P (Less than or equal to 4) = (1, 3) (2,2) (3,1) (1,1) (1,2) (2,1)

=6/36=**1/6.**

C) Sum is divisible by 2 and 3= (1,5) (2,4) (3,3) (4,2) (5,1) (6,6)

= 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 Number of Balls=2+3+2=7

Let S be the sample space.

Then, n(S) = No. of ways of drawing 2 balls out of 7= 7C2

= 7\*6/2\*1 = 21

Let E= Event of drawing 2 balls, none of which is blue.

Therefore n(E) = No. of ways of drawing 2 balls out of (2+3) balls =5C2

= 5\*4/2\*1 = 10

P (E)=N (E)/N(S)=**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)

Expected number of candies for a randomly selected child = 3.09

**Step-by-step explanation:**

Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

= 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**

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) The weights (X) of patients at a clinic (in pounds), are 108, 110, 123, 134, 135, 145, 167, 187, 199

one of the patients is chosen at random.

expected Value = ∑ (probability \* Value)

 ∑ P(x). E (x)

there are 9 patients

Probability of selecting each patient = 1/9

E (x) =108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) = 1/9 1/9   1/9 1/9   1/9   1/9   1/9   1/9 1/9

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

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

= (1/9) (1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

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

**Car’s speed and distance**

**Use Q9\_a.csv**

#Skewness Inference: 1. Speed distribution is left skewed (negative skewness) 2. Distance distribution is right skewed (positive skewness)

**Ans.** speed -0.117510

dist 0.806895

#Kurtosis Inference: 1. Speed distribution is platykurtic (negative kurtosis i.e., flatter than normal distribution) 2. Distance distribution is leptokurtic (positive kurtosis i.e., peaked than normal distribution)

speed -0.508994

dist 0.405053

SP and Weight (WT)

Use Q9\_b.csv

#Skewness Inference: 1. SP distribution is Right skewed (Positive skewness) 2. WT distribution is Left skewed (Negative skewness)

Ans. SP 1.611450

WT -0.614753

#Kurtosis Inference: Both the SP and WT distributions are leptokurtic (have positive kurtosis i.e., Peaked than normal distribution)

SP 2.977329

WT 0.950291

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



Ans) In Histogram we can see that it is Right skewed.

Most of the point lies 50 and 100



Ans) This means that there must be at least two outlier .

Upper quartile is more than lower quartile.

Mean is more than median.

**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) Sample mean = 200 Sample SD = 30 Population size (N) = 2000

**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: 41.0 VAR: 25.529411764705884 Median: 40.5

standard deviation: 5.05266382858645

1. What can we say about the student marks?

Ans) There are 2 Outliers in Student's marks: 49 and 56

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

Ans) Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.

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

Ans) The mean, mode and median can be used to figure out if you have a positively or negatively skewed distribution. ... If the mean is greater than the median, the distribution is positively skewed.

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

Ans) If the mean is less than the median, the distribution is negatively skewed.

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

Ans) Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. ... A leptokurtic distribution has a higher peak and taller (i.e., fatter and heavy) tails than a normal distribution.

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

Ans) Negative values of kurtosis indicate that a distribution is flat and has thin tails. ... A platykurtic distribution is flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter (i.e., lighter, and thinner) tails).

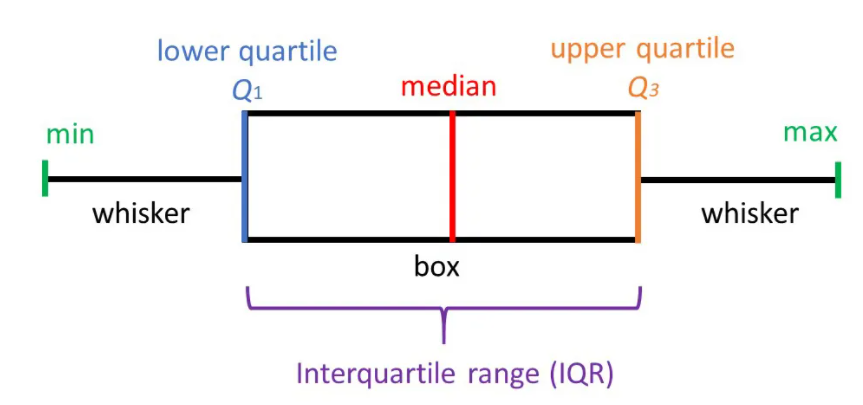
Q18) Answer the below questions using the below boxplot visualization.



A) What can we say about the distribution of the data?

Ans) Brief explanation.   
  
 In descriptive statistics, a box plot or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.



**Minimum Score**

The lowest score, excluding outliers (shown at the end of the left whisker)= Q1 -1.5\*IQR

**Lower Quartile**

Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile) =Q1=10

**Median**

The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value and half are less=15.5

**Upper Quartile**

Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value.

Q3=18

**Maximum Score**

The highest score, excluding outliers (shown at the end of the right whisker) = Q3 + 1.5\*IQR

**Whiskers**

The upper and lower whiskers represent scores outside the middle 50% (i.e., the lower 25% of scores and the upper 25% of scores).

**The Interquartile Range (or IQR)**

This is the box plot showing the middle 50% of scores (i.e., the range between the 25th and 75th percentile) =8

B) What is nature of skewness of the data?

Ans) It is Negatively SKEWED.

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

Ans) The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g., Q3−Q1).

Q3=18, Q1=10

The IQR of the data (approximately)=8

Q19) Comment on the below Boxplot visualizations?



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

Ans) Median of both the boxes are same = 262.5(approx.)

IQR of Box 1 is = 25(approx.), IQR of Box 2 is= 75(approx.)

**Upper Quartile for Box 1(Q3) = 275 Upper Quartile for Box 2(Q3) =300**

**Lower Quartile for Box 1 (Q1) = 250 Lower Quartile for Box 2(Q1) =225**

**Maximum Score for Box 1=275+1.5\*25 Maximum Score for Box 2=300+1.5\*75**

**=275+37.5 =300+112.5**

=312.5 =412.5

**Minimum Score for Box 1=250-1.5\*25= 212.5 Minimum Score for Box 2=225-1.5\*75=112.5**

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)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans) MPG of Cars does follow normal distribution approximately (as mean and median are approx. same).It is a symmetric distribution where most of observations cluster around the center peak and the probabilities further away from the Mean tapped off easily in both directions Extreme values of the distribution are most unlikely.

Hence the data mostly normally distributed

It has 2 parameters

1 Mean

2 STANDARD DEVIATION width of normal probability curve on X axis

Properties

1 Bell shaped

2 Mean Median and Mode values are equal

3 symmetrical

4 Skewness is o

5 Kurtosis is o

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

Dataset: wc-at.csv

Ans) Both the Adipose Tissue (AT) and Waist Circumference(Waist) data set do follow the normal distribution approximately (as mean and median of both the data are approximately same)

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

Ans)

stats.norm.ppf (0.95) #z-score of 90% confidence interval is AREA= 1+CI/2= 1+0.90/2=O.95= 1.6448

stats.norm.ppf (0.97) #z-score of 94% confidence interval is AREA= 1+CI/2= 1+0.94/2=O.97= 1.8807

stats.norm.ppf (0.8) #z-score of 60% confidence interval is AREA= 1+CI/2= 1+0.60/2=O.8= 0.8416

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

Ans)

# t scores of 95% confidence interval for sample size of 25(A=1+0.95/2=0.975)

stats.t.ppf (0.975,24) # df = n-1 = 24=2.0638985616280205

# t scores of 96% confidence interval for sample size of 25(A=1+0.96/2=0.98)

stats.t.ppf (0.98,24) = 2.1715446760080677

# t scores of 99% confidence interval for sample size of 25(A=1+0.99/2=0.995)

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

Ans) Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.17%. Assuming significance value α = 0.05 (Standard Value) (If p\_value < α; Reject Ho and accept Ha or vice-versa).Thus, as p-value > α ; Accept Ho i.e. The CEO claims are false and the avg life of bulb > 260 days.