|  |  |
| --- | --- |
| 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 |
| 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 | Interval |
| Years of Education | Ordinal |

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

Events = 8

Probability of two heads and one tail = 3 = 3/8 = 37.5%

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

1. Equal to 1

Events = 36

Probability of sum equal to 1 = 0 = 0/36=0

1. Less than or equal to 4

Events = 36

Probability of sum less than or equal to 4 = 5 = 5/36 = 13.8%

1. Sum is divisible by 2 and 3

Events = 36

Probability of sum is divisible by 2 and 3 =24=24/36= 66.7%

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?

Events = 21

Probability that none of the balls drawn is blue = 10 = 10/21 = 47.6%

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

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\*0.120)= 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.

**Use Q7.csv file**

**#mean median mode sd variance and range for Points**

**> mean(q\_data$Points)**

**[1] 3.596563**

**> median(q\_data$Points)**

**[1] 3.695**

**> q\_mode(q\_data$Points)**

**[1] 3.92**

**> sd(q\_data$Points)**

**[1] 0.5346787**

**> var(q\_data$Points)**

**[1] 0.2858814**

**> rangevalue(q\_data$Points)**

**[1] 2.17**

**> #mean median mode sd variance and range for score**

**> mean(q\_data$Score)**

**[1] 3.21725**

**> median(q\_data$Score)**

**[1] 3.325**

**> q\_mode(q\_data$Score)**

**[1] 3.44**

**> sd(q\_data$Score)**

**[1] 0.9784574**

**> var(q\_data$Score)**

**[1] 0.957379**

**> rangevalue(q\_data$Score)**

**[1] 3.911**

**> #mean median mode sd variance and range for score**

**> mean(q\_data$Weigh)**

**[1] 17.84875**

**> median(q\_data$Weigh)**

**[1] 17.71**

**> q\_mode(q\_data$Weigh)**

**[1] 17.02**

**> sd(q\_data$Weigh)**

**[1] 1.786943**

**> var(q\_data$Weigh)**

**[1] 3.193166**

**> rangevalue(q\_data$Weigh)**

**[1] 8.4**

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?

Expected value is 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Kurtosisi and skewness for cars speed**

**q9.speed.kurtosis()**

**-0.5089944204057617**

**q9.speed.skew()**

**-0.11750986144663393**

**Skewness and kurtosis for cars dist**

**q9.dist.kurtosis()**

**0.4050525816795765**

**q9.dist.skew()**

**0.8068949601674215**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**q9b.SP.kurtosis()**

**2.9773289437871764**

**q9b.WT.kurtosis()**

**0.9502914910300326**

**q9b.WT.skew()**

**-0.6147533255357768**

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



Histogram is a plot which show the frequency distribution for set of continuous data. We can use histograms when you have continuous measurements and want to understand the distribution of value and outliers. Histogram take continuous measurements and place them into range of values as bins. Each bin represents the count of observations that fall within that bin. We can use the histogram for measure of central tendency and measures of variability



Box plot are used to show overall patterns for a group, It provides a useful way to visualize the range and other characteristics of responses for a large group.

Median-Middle quartile marks the mid-point of the data, line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less

Inter-quartile range-middle box represents the middle 50% of the data of group. The range of scores from lower to upper quartile as inter quartile range. The middle 50% of scores fall within the inter quartile range.

Upper quartile-75% of the scores fall below the upper quartile range.

Lower quartile-15% of scores fall below the lower quartile.

Whiskers-The upper and low whiskers represent score outside the middle 50%. It stretches wide range of scores than the middle quartile groups.

Probability Density Function-It is used to specify the probability of random variable falling within a particular range of values.

Outliers-It is a data point that differs significantly from other observation. An outlier may be due to variability in the measurement.

**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=255.335

98% confidence interval=273.35

96% confidence interval=264.8

**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.

Mean=41

Median=41

Standard deviation=4.58

1. What can we say about the student marks?

From the data we can tell that the distribution of the student is approximately symmetrical.

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

The nature of the skewness is symmetrical and unimodal when mean, median and mode are equal, distribution has zero skewness

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

If the mean is greater than mode, then the distribution is positive skewness

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

If the median is greater than mode, then the distribution is negative skewness

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

Positive kurtosis value indicates that the distribution has heavier tails than the normal distribution, data that follow a t distribution have a positive kurtosis value, and positive values for the skewness indicate data that are skewed right

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

A distribution with a negative kurtosis value indicates that the distribution has lighter tails and a flatter 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?

Distribution of the data lies right side of the plot which is 50% the data. And the median is around 15.8 apprx.

What is nature of skewness of the data?

Form this plot we can tell that the data is negatively skewed

What will be the IQR of the data (approximately)?   
IQR = 8 apprx.

Q19) Comment on the below Boxplot visualizations?



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

From the boxplot visualization, we can tell that both the boxplot 1 and boxplot 2 are normally distributed with no skewness, both the boxplot have same median but the distribution of the data from q3 to q1 is large in boxplot2 which covers more data and boxplot1 covers less data.

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)

stats.norm.cdf(-35,

loc=cars\_bf.MPG.mean(),

scale=cars\_bf.MPG.std() )

1.451991487329956e-14

* 1. P(MPG<40)

stats.norm.cdf(45,

loc=cars\_bf.MPG.mean(),

scale=cars\_bf.MPG.std() )

0.8766508913473714

c. P (20<MPG<50)

p1 = stats.norm.cdf(50,loc=cars\_bf.MPG.mean(),

scale=cars\_bf.MPG.std() )

p1

0.955992693289364

p2 = stats.norm.cdf(20,loc=cars\_bf.MPG.mean(),

scale=cars\_bf.MPG.std())

p2

0.05712377632115936

p1-p2

0.8988689169682046

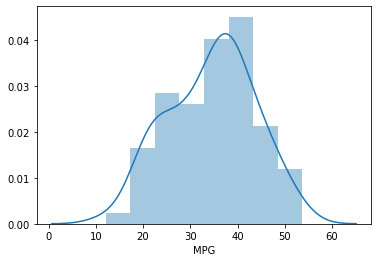
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

By implementing the plot, I came to know that the MPG of cars

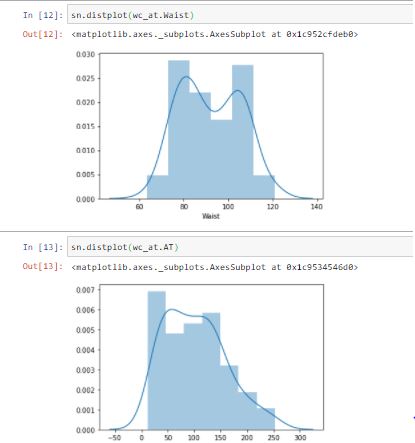
Follows normal distrubution



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

Dataset: wc-at.csv

the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set does not follows Normal Distribution



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

90 % confidence interval-1.645

94% confidence interval-1.88

60% confidence interval-0.841

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

95% of confidence interval-2.0638985616280205

94% of confidence interval-2.1715446760080677

99% of confidence interval-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

Mean of sample=260

Population mean=270

standard deviation of the sample = 90

n = number of items in the sample = 18

t = - 0.471

bulbs lasting less than 260 days on average of 0.3218