|  |  |
| --- | --- |
| 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 | Continuous |
| 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) | Ordinal |
| Sales Figures | Nominal |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Nominal |
| SAT Scores | Ordinal |
| Years of Education | Nominal |

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

**Solution:**

Two coins are tossed. The possible events are 23=8. These events are HHH, HHT,HTH,THH,TTH,THT,HTT,TTT.

Out of these, three events are having two heads and one tail i.e. HHT,HTH,THH.

So, this gives Probability P(Two heads and one tail) = 3/8 = 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

**Solution:**

When two dices are rolled, total number of occurred events are 6\*6=36.

{1,1},{1,2},{1,3},{1,4},{1,5},{1,6}

{2,1},{2,2},{2,3},{2,4},{2,5},{2,6}

{3,1},{3,2},{3,3},{3,4},{3,5},{3,6}

{4,1},{4,2},{4,3},{4,4},{4,5},{4,6}

{5,1},{5,2},{5,3},{5,4},{5,5},{5,6}

{6,1},{6,2},{6,3},{6,4},{6,5},{6,6}

1. **Equal to 1:** The probability of getting sum equals to 1 is **0**, as total possible occurrences are as below:

And minimum achieved sum is 2.

1. **Less than or equal to 4:**

Total occurred outcomes that gives sum less than or equal to 4 is 6. These combinations are {1,1},{1,2},{1,3},{2,1},{2,2}, {3,1}

6/36=1/6=**0.1667**.

1. **Sum is divisible by 2 and 3**

**Sum is divisible by 2**

Total occurred outcomes that gives sum divisible by 2 and 3

{1,5},{2,4},{3,3},{4,2},{5,1},{6,6}

is 6/36= **0.1667**.

**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?**

Solution:

The probability that none of the drawn ball is blue and given by

5C2/7C2=(5\*4/1\*2)/(7\*6/1\*2)=10/21=**0.4762.**

**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

Solution:

The expected number of candies

µ= ∑x. p(x)

µ=x1.p(x1)+x2.p(x2) +x3.p(x3)+x4.p(x4)+x5.p(x5)+x6.p(x­6)

µ =1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

µ =0.015+0.8+1.95+0.025+0.06+0.24

**µ =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**

**Solution:**

Here, all asked statistical parameters are calculated in the excel sheet. This excel sheet is attached along with mail.

1. **Points**

* **Mean- 3.5965625**
* **Median- 3.695**
* **Mode- 3.92**
* **Variance- 0.285881351**
* **Standard deviation- 0.534678736**
* **Range- 2.17**

1. **Score**

* **Mean-3.21725**
* **Median-3.325**
* **Mode-3.44**
* **Variance-0.957378968**
* **Standard deviation-**
* **Range-0.978457443**

1. **Weigh**
   * **Mean- 17.82451613**
   * **Median-17.6**
   * **Mode-17.02**
   * **Variance-3.280185591**
   * **Standard deviation- 1.811128265**
   * **Range-8.4**

**Comment:**

For all parameters i.e. points and score, the relation between mean, median, and mode is given by Mean < Median<Mode. But for weigh, mean> median. But mode<median<mode.

**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?

**Solution:**

The expected value of the weight of that patient is given by

E(X)=108+110+123+134+135+145+167+187+199/9= 1308/9=145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Solution:**

**Kindly note that calculations are attached in excel file**

**Skewness for speed and distance**

For speed, skewness= -0.117509861, that shows mean<median<mode. The skewness is negative that shows data are skewed left. The data is not symmetrical.

For distance, skewness= 0.80689496, that shows mean>median>mode. The skewness is positive that shows data are skewed right and not symmetrical.

**Kurtosis for speed and distance**

The kurtosis is a statistical parameter hat measures whether the data are heavy tailed or light tailed relative to normal distribution. High kurtosis means more outliers and low kurtosis means less outliers.

For speed, kurtosis= -0.50899442, that shows negative kurtosis means light tailed distribution.

For distance, kurtosis= 0.405052582, that shows positive kurtosis means high tailed distribution and data points are more centered around its mean.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Kindly note that calculations are attached in excel file**

**Skewness for speed and weight**

**For speed, skewness=** 1.611450196. This shows positive skewness means data points are skewed right and not symmetrical

**For weight, skewness=** -0.614753326. This shows negative skewness means data points are skewed left and not symmetrical.

**Kurtosis for speed and weight**

**For speed, kurtosis=** 2.977328944. This shows positive kurtosis means high tailed distribution and data points are more centered around its mean.

**For weight, kurtosis=** 0.950291491. This shows negative kurtosis means light tailed distribution.

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



**Explanation:**

* The uniformity in histogram shows consistency in data. As it can be seen in above figure, it is not normally distributed. It is positively skewed.
* The maximum data are concentrated in the 50-150. These data points are having 200 frequencies.
* Skewness: By observing above histogram plot, we can say that the plot is positively skewed as long tails is appeared on right side. In this, mean>median. (This is also called a third Moment of Business Decision)
* Kurtosis: By observing above histogram plot, the plot is sharper in peak. That is positive kurtosis. (This is also called a fourth moment business decision)



**Explanation:**

* The summary of a set of data that is measured using and interval scale is known as “Box Plot” and “Whisker Plot”. It represents five number summary of set of data. These are divided as minimum, first quartile, median, third quartile, and maximum.
* **Minimum(Q0 or 0th percentile):** the lowest data point in data
* **Maximum(Q4 or 100th percentile):** the highest data point
* **First Quartile(Q1 or 25th percentile):** it covers the 25% of the data.
* **Median (Q2 or 50th percentile):** it covers the 50% of the data.
* **Third Quartile (Q3 or 75th percentile):** it covers the 75% of the data.
* In the above plot, outliers are present on upper extreme side of box plot.
* There is less data points between Q1 and lower limit.
* The box plot is showing positive skewness means most data points are concentrated on left 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?**

**Solution:**

In this problem, population standard deviation is not given. So, to calculate confidence interval, t-table will be used.

Given data:

N(population) =3,000,000

n(sample size) =2,000

(sample mean) =200

s (standard deviation of sample) =30

degree of freedom=n-1=2000-1=1999

Let us calculate first, standard error which is given by

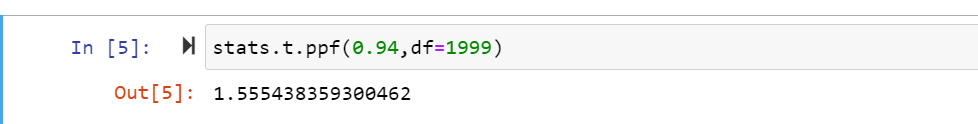
Standard error =

To calculate t-value, let us use below formula,

As, population mean is not given. So, stats library from scipy package in python is used to calculate t-score.

CI based on t-score is calculated using below formula:

1. **For confidence interval of 94%**

Here, 1-α=0.94

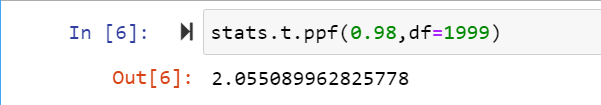
t0.94,1999=1.56

Upper limit of CI=200+1.56(0.6708) = 201.05

Lower limit of CI=200-1.56(0.6708)= 198.95

CI range for 94% confidence interval = [198.95,201.05]

1. **For confidence interval of 98%**

Here, 1-α=0.98

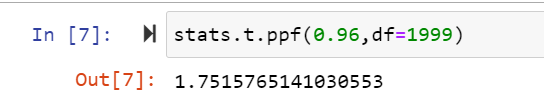
t0.98,1999=2.055

Upper limit of CI=200+2.055(0.6708) = 201.38

Lower limit of CI=200-2.055(0.6708)= 198.62

CI range for 98% confidence interval = [198.62,201.38]

1. **For confidence interval of 96%**

Here, 1-α=0.96

t0.96,1999=1.7516

Upper limit of CI=200+1.7516 (0.6708) = 201.174

Lower limit of CI=200-1.7516 (0.6708) = 198.825

CI range for 96% confidence interval = [198.825,201.174]

**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?

**Solution:**

1. **Mean**

Mean = (34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

= **41**

1. **Median**

For odd n,

Median = (X[n/2]+X[(n/2)+1])/2=(40+41)/2 = **40.5**

1. **Variance**

Variance = = 24.11

1. **Standard Deviation**

Standard deviation = sqrt(variance)=4.910

1. **Comments on students’ marks**

Maximum obtained mark = 56

Minimum obtained mark = 34

Obtained marks are increasing in order.

Also, mode of given marks is 41.

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

**Solution:**

No skewness is present as data distribution is symmetrical because mean and median of

data are equal.

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

**Solution:**

Positive skewness is present.

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

**Solution:**

Negative skewness is present.

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

**Solution:**

Positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.

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

**Solution:**

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?

**Answer:**

Let’s assume above box plot is about point scored in a football game.

* The range of above data is 19-1=17.age’s of the students in a school.
* By observing above box plot, middle 50% of data shows scores value in between 10 to 18.
* The upper quartile for score data is Q3= 18
* The lower quartile for score data is Q1=10
* The upper whisker contains 25% of data of the scores above 18.
* The lower whisker contains 25% of data of the scores below 10.

What is nature of skewness of the data?

**Answer:**

The nature of data is Left Skewed as most of the data concentrate on right side. Median is greater than Mean.

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

**Answer:**

Upper Quartile(Q3)=18, Lower Quartile(Q1)= 10

(Inter Quartile Range) IQR = Q3-Q1

IQR = 18-10

IQR =8

Q19) Comment on the below Boxplot visualizations?



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

**Answer:**

Let us assume that above box plot is showing product sold by two different salesmen.

General Observations:

* Medians of both box plots are same.
* The distribution of data is normal/symmetrical.
* No outlier in upper quartile and lower quartile.

**Box Plot 1 observations:**

* Upper quartile Q3 = 280.5
* Lower quartile Q1 = 256.5
* IQR (Inter Quartile Range) = Q3-Q1=280.5-256.5=24
* 1.5 IQR above the third quartile is Q3+1.5IQR=280.5+1.5(24) =316.5
* 1.5 IQR below the third quartile is Q1-1.5IQR=256.5-1.5(24) =184
* Outlier in upper extreme= Q3+1.5(IQR)=280.5+1.5(24)=316.5
* Outlier in lower extreme= Q1-3(IQR)=256.5-3(24)=184

**Box Plot 2 observations:**

* Upper quartile Q3 = 315.5
* Lower quartile Q1 = 223.6
* IQR (Inter Quartile Range) = Q3-Q1=315.5-223.6=91.9
* 1.5 IQR above the third quartile is Q3+1.5IQR=315.5+1.5(91.9) =453.35
* 1.5 IQR below the third quartile is Q1-1.5IQR=223.6-1.5(91.9) =85.75
* Outlier in upper extreme= Q3+3(IQR)=315.5+3(91.9)=591.2
* Outlier in lower extreme= Q1-3(IQR) = 223.6-3(91.9)=-52.1

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

**Solution:**

For calculations, formula in excel sheet is used. And related excel sheets are attached with this assignment.

* 1. P(MPG>38)

P(MPG>38)= interested events/total events=33/81=0.407407

* 1. P(MPG<40)

P(MPG<40)= interested events/total events=61/81=0.753086

c. P (20<MPG<50)

P(20<MPG<50)= interested events/total events=69/81=0.85185

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Here, distribution of MPG data is plotted using Q-Q plot. The Q-Q plot is a scatterplot of two quantities plotted against each other. Here, first z-score is calculated and then sorted MPG data is plotted with respect to z-score. The plotted distribution said to be normal when both sets of quantiles follows the straight line.

Here, I have used excel to calculate z-score and to plot Q-Q plot. Below figure is showing Q-Q plot. And as it can be seen quantiles follows straight line, so the MPG data are **normally distributed.**

It 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

As discussed in previous answer, in Q-Q plot if both quantiles follow the straight line results into normal distribution. The Q-Q plot for Waist circumference and adipose tissue data is plotted in below figures. As observed in both figures, it follows the straight line resulting into normal distribution.

**Waist circumference data are normally distributed.**

**Adipsoe tissue data are also normally distributed.**

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

**Solution:**

To calculate z-scores, python is used.

1. For 90% confidence interval

Area to left =(1+0.9)/2=0.95 from z-table, z-score is

z-value= 1.65

1. For 94% confidence interval

Area to left =(1+0.94)/2=0.97 from z-table, z-score is

z-value= 1.89

1. For 60% confidence interval

Area to left =(1+0.60)/2=0.8 from z-table, z-score is

z-value= 0.85

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

**Solution:**

As, population mean is not given. So, stats library from scipy package in python is used to calculate t-score.

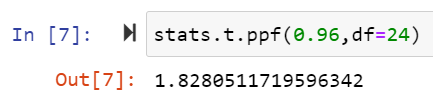
Here, degree of freedom=n-1=25-1=24

1. For 95% confidence interval

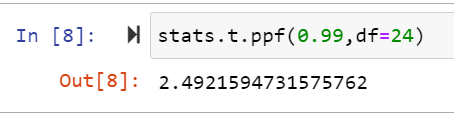


t1-0.95,24 = 1.7108

1. For 96% confidence interval



t1-0.96,24 = 1.8281

1. For 99% confidence interval

t1-0.99,24 = 2.4921

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

**Solution:**

Given data:

Sample mean = = 260

Population mean = μ = 270

Sample size = n = 18

Sample standard deviation = s = 90

Degree of freedom = df = n-1=18-1=17

Let us assume hypothesis

Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

1. First, t-value is calculated using below formula

= -0.4714

The p-value for t value -0.4714 and df of 17 is 0.32167

****

The decision rule for the p-value method:

if p-value (p) > level of significance (α), we fail to reject Null Hypothesis  
if p-value (p) ≤ level of significance (α), we reject Null Hypothesis

so, here p=0.32167 > 0.05, we accept the null hypothesis.

**So the claim of government company is true and an average life of light bulb last 270 days .**