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

Assignment 1 – Pooja Pandey

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

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

Ans**. 3/8 (HHT,HTH,THH)**

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

1. Equal to 1 - **ZERO**
2. Less than or equal to 4 – **6/36=1/6[(1,1),(1,2),(1,3),(2,2),(2,1),(3,1)]**
3. Sum is divisible by 2and 3–**6/36=1/6 [(1,5),(5,1),(2,4),(4,2),(3,3),(6,6)]**

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?

**Total No. of balls = (2 + 3 + 2) = 7  
Let S be the sample space.  
Then, n(S) = Number 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.  
n(E) = Number 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: **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**

**Ans: MEAN: Points 3.596563**

**Score 3.217250**

**Weigh 17.848750**

**dtype: float64**

**MEDIAN : Points 3.695**

**Score 3.325**

**Weigh 17.710**

**dtype: float64**

**Mode: Points Score Weigh**

**0 3.07 3.44 17.02**

**1 3.92 NaN 18.90**

**Variance: Points 0.285881**

**Score 0.957379**

**Weigh 3.193166**

**dtype: float64**

**Standard Deviation: Points 0.534679**

**Score 0.978457**

**Weigh 1.786943**

**dtype: float64**

**Range:**

**INFERENCE:**

**FROM THIS, WE CONCLUDE THAT ALL THE VALUES ARE CLOSER FOR POINTS AND SCORE AS COMPARE**

**TO WITH WEIGHT,**

**ALSO POINTS AND Weigh HAVE 2 RECURRING NUMBERS WHERE AS SCORE HAS ONLY**

**ONE NUMBER WHICH IS RECURRING**

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: Mean= 145.33

**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**

**Cars speed and distance**

**Use Q9\_a.csv**

Ans = Skewness of speed and Distance

speed -0.117510

dist 0.806895

Kurtosis of speed and Distance

speed -0.508994

dist 0.405053

Inference : As the skewness of speed is negative, which means it is left skewed (as the skewness value is very close to 0 , we can say that it is slightly left skewed) also the kurtosis of the speed is less than 3, which is platykurtic (nearly flat curve) which means that it tends to produce fewer outlier than the Normal distribution.

For distance , the Skewness is positive, but again close to zero which means it is slightly right skewed also the kurtosis is less than 3 which means it is platykurtic (nearly a flat curve) has fewer outlier than the Normal distribution.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans = Skewness of SP and Weight(WT)

SP 1.611450

WT -0.614753

Kurtosis ofSP and Weight(WT)

SP 2.977329

WT 0.950291

Inference : As the skewness of SP is Positive which means it is right skewed and kurtosis **2.9** indicated data is mesokurtic.

For Weight (WT) The skewness is negative, but close to zero which means it is slightly left skewed and kurtosis of WT is 0.95 which is less than 3 which means it is platykurtic (nearly a flat curve) has fewer outlier than the Normal distribution.(the dataset has lighter tails than a normal distribution)

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



Ans = Inferences for Histogram

The most of the data points are concerated in the range 50-100 with frequency 200.

And least range of weight is 400 somewere around 0-10.

Skewness- we can notice a long tail towards right so it is heavily right skewed.

Inferences for Boxplot

From the given Boxplot, we can say that the distribution has lots of outliers towards upper extreme.

**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=** Given Values:

N= 2000

S = 30

Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (200) as the sample statistic.

**For confidence interval of 94%**

Standard Error (SE) of the mean is = s/√n = 30/√2000 = 0.67

Find critical value. The critical value is a factor used to compute the margin of error. To express the critical value as a t score (t\*), follow these steps.

Compute alpha (α): α = 1 - (confidence level / 100) =0.06

Critical probability (p\*): p\* = 1 - α/2 = 1 - 0.06/2 = 0.97

Degrees of freedom (df): df = n -1 = 2000 - 1 = 1999

The critical value is the t score having 1999 degrees of freedom and a probability equal to 0.97. From the t

Chart, we find that the critical value is 1.96.

**Margin of Error = 1.96\*0.94 = 1.84**

Range of Confidence interval for 94% = sample statistics ± margin of error

=200 ± 1.84

= 201.84 , 198.16

Therefore we can be 94% confident that the population mean falls within the range of 200±1.84

**FOR Confidence interval of 96%**

Compute alpha (α): α = 1 - (confidence level / 100) =0.04

Critical probability (p\*): p\* = 1 - α/2 = 1 - 0.04/2 = 0.98

Critical value = 1.96

**Margin of Error = 1.96\*0.96 = 1.88**

Range of Confidence interval for 96% = sample statistics ± margin of error

=200 ± 1.88

= 201.88 , 198.12

Therefore we can be 94% confident that the population mean falls within the range of 200±1.88

**FOR Confidence interval of 98%**

Compute alpha (α): α = 1 - (confidence level / 100) =0.02

Critical probability (p\*): p\* = 1 - α/2 = 1 - 0.02/2 = 0.99

Critical value = 1.96

**Margin of Error = 1.96\*0.98 = 1.92**

Range of Confidence interval for 98% = sample statistics ± margin of error

=200 ± 1.92

= 201.92 , 198.08

Therefore we can be 94% confident that the population mean falls within the range of 200±1.92

**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?

Ans= Mean = 41.55

Median=41

Variance=28.73

Standard Deviation=5.36

We can say that the average marks obtained by the student is 41

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

Ans= It will be a symmetrical distribution

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

Ans= Right Skewed

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

Ans= Left Skewed

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

Ans= It indicates that the distribution has heavier tails then the normal distribution.

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

Ans= It indicates that the distribution has lighter tails then the normal distribution

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



What can we say about the distribution of the data?

Ans= The maximum data or 50% of the data is greater than a value of 10 , the average of the above data is somewhere around 15.

What is nature of skewness of the data?

Ans= Left Skewed (median>mean)

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

Ans = Q3-Q1 = 18-10 = 8

IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

Ans=

The line in the middle of the box plot 1 is close to Q1, which indicates that the distribution for box plot 1 is positively skewed. Conversely, the line in the middle of the box plot 2 is near the center of the box, which means the distribution has little skew at all.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

ANS = stats.norm.cdf(X,loc=mean,scale=std)

For a) P(MPG>38) = 0.3475

**1-stats.norm.cdf(38,cars\_data.MPG.mean(),cars\_data.MPG.std())** ## for data on the ride side we are using 1- stats.norm.cdf(X,loc=mean,scale=std)

**output : 0.3475939251582705**

1. P(MPG<40) = 0.7293

**stats.norm.cdf(40,cars\_data.MPG.mean(),cars\_data.MPG.std()) ##**data is on left side

**Output:** 0.7293498762151616

1. P(20<MPG<50) = 0.0131

**stats.norm.cdf(50,cars\_data.MPG.mean(),cars\_data.MPG.std())-(1-stats.norm.cdf(20,cars\_data.MPG.mean(),cars\_data.MPG.std()))**

**Output:**0.013116469610523374

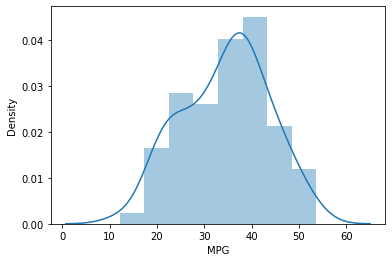
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans= Using distplot from seaborn

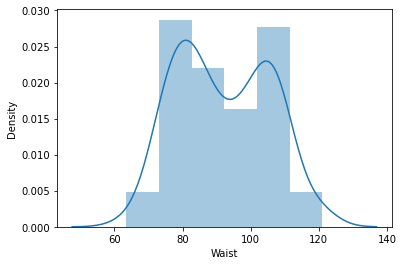
1. Import seaborn as sns
2. sns.distplot(cars\_data.MPG)  
   Output:



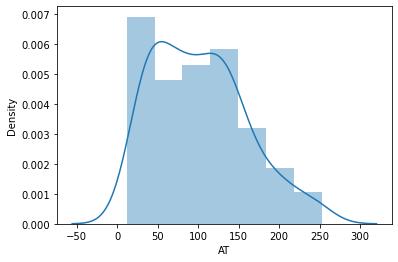
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

sns.distplot(q21b.Waist)



sns.distplot(q21b.AT)



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

Ans= for confidence interval 60%

AL(1+CL)/2 = (1+0.60)/2 = 0.80

We will look out 0.80 in Z table and will add the row and column value

**Zs = 0.8+0.05 = 0.85**

for confidence interval 90%

AL(1+CL)/2 = (1+0.90)/2 = 0.95

We will look out 0.95 in Z table and will add the row and column value

**Zs = 1.6+0.05 = 1.65**

for confidence interval 94%

AL(1+CL)/2 = (1+0.94)/2 = 0.97

We will look out 0.97 in Z table and will add the row and column value

**Zs = 1.8+0.09 = 1.89**

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

Ans = for 95% CI , n=25

t(1-0.95),(25-1) =t0.05,24 with the help of t table = 2.06

or

from scipy import stats

stats.t.ppf(0.975,df=24)

**= 2.06**

for 96% CI , n=25

t(1-0.96),(25-1) =t0.04,24 with the help of t table = 2.17

or

from scipy import stats

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

**= 2.17**

for 99% CI , n=25

t(1-0.99),(25-1) =t0.005,24 with the help of t table = 2.79

or

from scipy import stats

stats.t.ppf(0.995,df=24)

**= 2.79**

Q 24**)**A Government companyclaims 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=

Find T score  
Given:

X= 260

µ= 270

s =90

n=18

**t= (x- µ)/(s/√n)**

t= -0.471

df = n-1 = 17

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

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

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

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

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

p\_value

**= 0.32**

the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming

the mean life of the bulbs is 300 days.