|  |  |
| --- | --- |
| 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 | Nominal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Ratio |
| 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 | Ratio |

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

**Answer**. Total Outcomes = HHH , HHT , HTH , THH , HTT , THT , TTH , TTT P(2 head 1 tail) = 3/8

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

**Answers.**

1. 0/36
2. 6/36
3. 6/36

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.** Total number of balls =(2+3+2)=7

Let S be the sample space

Then,n(S)= Number of ways of drawings 2 balls out of 7= 7C2

n(S)= 7C2=7\*6/2\*1=21

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

n(E) = Number of ways of drawing 2 balls out of (2+3) balls=5C2

n(S) = 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

**Solution.** 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

= 3.09

Expected number of candies for a randomly selected child = 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**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weight** |
| **Mean** | **3.596563** | **3.217250** | **17.848750** |
| **Median** | **3.695** | **3.325** | **17.71** |
| **Mode** | **3.92** | **3.44** | **17.02** |
| **Variance** | **0.2858813508064** | **0.9573896774193** | **3.19316612903255** |
| **STD** | **0.5346479** | **0.9784857** | **1.786943** |
| **Range** | **2.17** | **3.911** | **8.399** |

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 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.

To Find: Expected Value

Expected Value = (probability \* Value)

[ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex: 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**

**Solution.**

|  |  |  |
| --- | --- | --- |
|  | **Skewness** | **Kurtosis** |
| **Speed** | -0.11751 | -0.50899 |
| **Distance** | 0.806895 | 0.405053 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Solution.**

|  |  |  |
| --- | --- | --- |
|  | **Skewness** | **Kurtosis** |
| **SP** | 1.61145 | -0.61475 |
| **Weight(WT)** | 2.977329 | 0.950291 |

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



**Inference**: The above histogram is positively skewed, where most of the

values are around the left tail of the distribution.

> Right tail of the distribution is longer



**Answer**. Histogram is right or positively skewed and outlier are upper side of boxplot.

**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?

**Answers.**

1. Confidence interval for 94% ) stats.norm.interval(0.94,200,30/(np.sqrt(2000)))

(198.738325292158, 201.261674707842

1. Confidence interval for 98% ) stats.norm.interval(0.98,200,30/(np.sqrt(2000)))

(198.43943840429978, 201.5605615957002)

1. Confidence interval for 96% ) stats.norm.interval(0.96,200,30/(np.sqrt(2000)))

(198.62230334813333, 201.3776966518666)

**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 = 41.0 Median = 40.5 Variance = 24.11 Standard deviation= 4.91

2) Mean is higher than the Median, which indicates that the data is not normal, it is right skewed. Most of the data points are present above mean

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

**Answer.** Data is Normalized and there is zero skewness**.**

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

**Answer.** When mean > median, distribution is positively skewed

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

**Answer.** when median > mean, distribution is negatively skewed

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

**Answer.** Positive skewness value indicates that thinner peak**.**

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

**Answer.** Negative skewness value indicates that thicker peak and wider tails**.**

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



What can we say about the distribution of the data?

**Ans.** Not normal distribution

What is nature of skewness of the data?

**Ans.** Negative skewness

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

**Ans.** 10-18=8

Q19) Comment on the below Boxplot visualizations?



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

**Ans.**  Boxplot 1: Q1=250 Boxplot 2: Q1=225

Q3=275 Q2=300

Q2=265 (median) Q2=265 (median)

1) In both the boxplots, mean and median are almost same, hence it follows normal distribution.

2) Both are non-skewed

3) No outliers are present in both.

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)

**Ans**. a) P(MPG>38)

# P(MPG>38)

1-stats.norm.cdf(38,Cars.MPG.mean(),Cars.MPG.std())

= 0.34669238536888103

b) P(MPG<40)

# P(MPG<40)

stats.norm.cdf(40,Cars.MPG.mean(),Cars.MPG.std())

= 0.7306083416219199

c) P(20<MPG<50)

#P(20<MPG<50)

stats.norm.cdf(50,Cars.MPG.mean(),Cars.MPG.std())- stats.norm.cdf(20,Cars.MPG.mean(),Cars.MPG.std())

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: Cars dataset has 5 numerical columns. By using using Shapiro Wilcoxson test, these columns are checked to see whether the data follows normal distribution .

**\***Null: Data is not normally distributed

**\***Alt: Data is normally distributed

**\*** MPG column

Stat,p= shapiro(MPG)

print('Statistics= %.3f, p= %.3f %(stat,p)')

alpha = 0.05

if alpha < p:

print('reject null:sample looks Gaussian')

else:

print('fail to reject null:sample does not look Gaussian')

Statistics= %.3f, p= %.3f %(stat,p)

reject null:sample looks Gaussian

**\***similarling tried with remaining four columns

As the probability value(p-value) of the columns HP, VOL, SP, WT is less than Significance value(0.05), these are not normally distributed

=>But ‘MPG’ column is normally distributed.

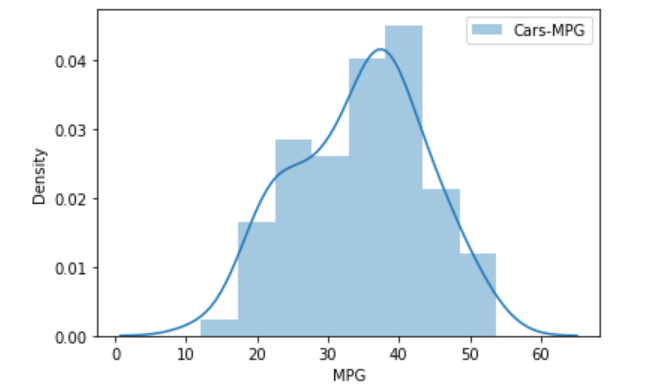
Another method:

sns.distplot(Cars.MPG, label='Cars-MPG')

plt.xlabel('MPG')

plt.ylabel('Density')

plt.legend()



* In this plot it is clearly shows that plot is similar to the bell shape, so‘MPG’ column is normally distributed.

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

Dataset: wc-at.csv

**Ans**. Wc-at dataset has 2 numerical columns- Waist and AT.

**\***Null: Data is not normally distributed

**\***Alt: Data is normally distributed

**\*** Waist column

Stat,p= shapiro(Waist)

print('Statistics= %.3f, p= %.3f %(stat,p)')

alpha = 0.05

if alpha < p:

print('reject null:sample looks Gaussian')

else:

print('fail to reject null:sample does not look Gaussian')

Statistics= %.3f, p= %.3f %(stat,p)

fail to reject null:sample does not look Gaussian

**\***Null: Data is not normally distributed

**\***Alt: Data is normally distributed

**\*** AT column

Stat,p= shapiro(AT)

print('Statistics= %.3f, p= %.3f %(stat,p)')

alpha = 0.05

if alpha < p:

print('reject null:sample looks Gaussian')

else:

print('fail to reject null:sample does not look Gaussian')

Statistics= %.3f, p= %.3f %(stat,p)

fail to reject null:sample does not look Gaussian

As the probability value(p-value) of both the columns Waist and AT is less than Significance value (0.05), these are not normally distributed.

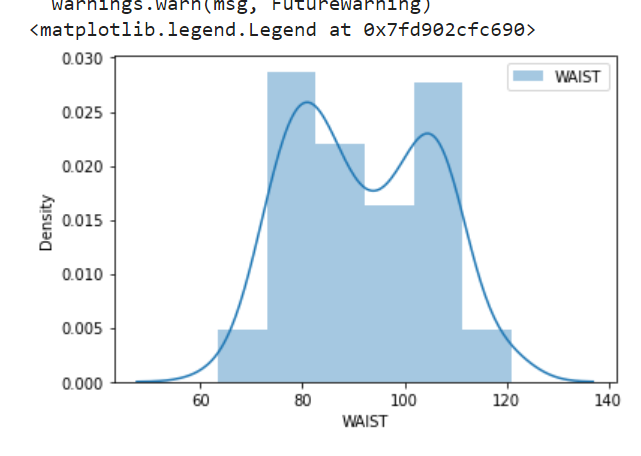
**Another Method:**

sns.distplot(df.Waist, label='WAIST')

plt.xlabel('WAIST')

plt.ylabel('Density')

plt.legend()



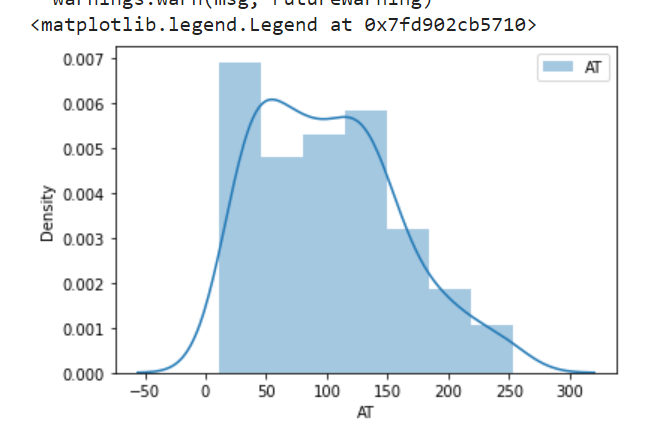
**\***It does not looks like bell shape, , so column ‘WAIST’ is not normally distributed.

sns.distplot(df.AT, label='AT')

plt.xlabel('AT')

plt.ylabel('Density')

plt.legend()



**\*** It does not looks like bell shape, , so column ‘AT’ is not normally distributed.

**\*** By using QQ Plot also we can also find normality

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

**Solution.** 1) stats.norm.ppf(0.90,loc=0,scale=1)

1.2815515655446004

2) stats.norm.ppf(0.94,loc=0,scale=1)

1.5547735945968535

3) stats.norm.ppf(0.60,loc=0,scale=1)

0.253347103135799

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

**Solution.** stats.t.ppf(0.95,25-1)= 1.7108820799094275

stats.t.ppf(0.96,25-1)= 1.8280511719596342

stats.t.ppf(0.99,25-1)= 2.4921594731575762

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.** Mu: 270 N=18 X-bar: 260 S=90 P(X<=260)=?

To calculate T score T = (X – μ) / [ σ/√(n) ].

Tscore = (260 -270)/(90/(np.sqrt(18)) Tscore = -0.4714045207910317

Probability = stats.t.cdf((Tscore,df) =stats.t.cdf(0.471,17)= 0.32167253567098364