Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| 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 |

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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Ratio |

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

**Sol: -** Total outcomes = 2\*2\*2 = 8

Interested events = (H,H,T), (H,T,H), (T,H,H) = 3

3/8

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

1. Equal to 1 = 0 (minimum sum obtained by adding outcomes of 2 dice is 2)
2. Less than or equal to 4

**Sol: -** Total Outcomes = 6\*6 = 36

P(X<=4) = [P(X=4)+P(X=3)+P(X=2) ]/Total Outcomes

= {[(1,3),(2,2),(3,1)]+[(1,2),(2,1)]+[(1,1)]}/36

= (3+2+1)/36

= 6/36 = 1/6

1. Sum is divisible by 2 and 3

Sol: - Total Outcomes = 6\*6 = 36

Interested Events = (1,5),(2,4),(3,3),(4,2),(5,1),(6,6) = 6

Probability = 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?

Sol: - Total no of balls = 2+3+2 = 7

Total Events (2 balls randomly drawn from 7 balls) = 7C2 = 21

Excluding 2 blue balls, there are balls

Interested events = 2 balls drawn randomly from 5 balls = 5C2

Probability = 5C2/7C2 = 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

Sol: - E(X) = X.P(X) = 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.



**Sol:-**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weight** |
| **Mean** | 3.57 | 3.22 | 17.85 |
| **Median** | 3.66 | 3.33 | 17.71 |
| **Mode** | 3.08 | 3.44 | 17.02 |
| **Variance** | 0.28 | 0.93 | 3.09 |
| **SD** | 0.53 | 0.96 | 1.76 |
| **Range** | 2.17 | 3.91 | 8.40 |

1. No huge variation between mean, median & mode => data follows Normal distribution
2. Points data has less SD, so data are more close to the mean
3. Range of Points data is less

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?

**Sol:-**Expected Value = Mean = (108+110+123+134+135+145+167+187+199)/9 = 145.33

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

**Cars speed and distance**



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Speed** | **Dist** | **Comments** |
| **Skewness** | 1.179446 | -0.84489 | +veskewnewss for Speed and -ve for Dist |
| **Kurtosis** | 2.407355 | 0.249561 | +ve kurtosis for both Speed and Dist |

**SP and Weight(WT)**



|  |  |  |  |
| --- | --- | --- | --- |
|  | **SP** | **WT** | **Comments** |
| **Skewness** | -0.40769 | -1.28736 | -ve or left skewness for both SP and WT |
| **Kurtosis** | -0.86374 | 1.152952 | -ve kurtosis for SP and +ve for WT |

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



**Solution:-**

1. Positive Skewness (tail is towards right side)
2. Positive Kurtosis (thin peak)
3. Mean > Median (Since skewness is on right side)



1. Right Skewed
2. Has outliers at the 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 ?

|  |  |
| --- | --- |
| N | 30,00,000 |
| n | 2,000 |
| sample mean | 200 |
| sample SD | 30 |

Confidence Interval Formula: ± t 1-α, n-1 (s/

|  |  |  |  |
| --- | --- | --- | --- |
| Confidence % | 94% | 98% | 96% |
| t(1-α),(n-1) | t(0.97,1999) | t(0.99,1999) | t(0.98,1999) |
| t Values | 1.88 | 2.32 | 2.05 |
| s/sqrt(n) | 0.67 | 0.67 | 0.67 |
| Lower CI | 198.74 | 198.44 | 198.62 |
| Upper CI | 201.26 | 201.56 | 201.38 |

**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 | 40.5 |
| Variance | 24.11 |
| SD | 4.91 |

1. What can we say about the student marks?

* The distribution of marks is Right Skewed (Mean>Median)
* 2 Outliers are present

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

No Skewness (Symmetric Distribution)

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

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

Q16) What does positive kurtosis value indicates for adata ?

**Sol: -** Thinner peak of the distribution

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

**Sol: -** Wider peak of the distribution

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



What can we say about the distribution of the data?

**Sol: -** Distribution is not normal

What is nature of skewness of the data?

**Sol: -** Left Skewed

What will be the IQR of the data (approximately)?   
**Sol: -** IQR = Q3 – Q1 = 18 – 10 = 8

Q19) Comment on the below Boxplot visualizations?



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

**Solution:-**

* Both are normally distributed
* No Outliers
* For both, mean = median = mode = 265 (approximately)

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)

1-stats.norm.cdf(38, np.mean(car['MPG']), np.std(car['MPG']))

**0.34669238536888103**

[Same can be achieved in R using 1-pnorm(38,mean(cars$MPG),sd(cars$MPG))]

* 1. P(MPG<40)

stats.norm.cdf (40, np.mean(car['MPG']), np.std(car['MPG']))

**0.7306083416219199**

c. P (20<MPG<50)

stats.norm.cdf(50, np.mean(car['MPG']), np.std(car['MPG']))-stats.norm.cdf(20, np.mean(car['MPG']), np.std(car['MPG']))

**0.9009686820346051**

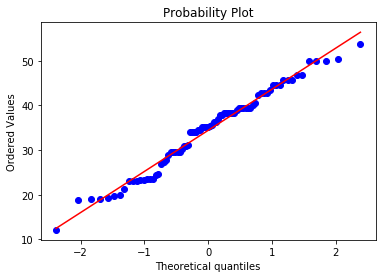
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

car = pd.read\_csv("C:\\Users\\hardi\\OneDrive\\Documents\\Excelr\_1\\Python\\Basic Statistics Level 1\\Cars.csv")

stats.probplot(car['MPG'],dist="norm",plot=pylab)



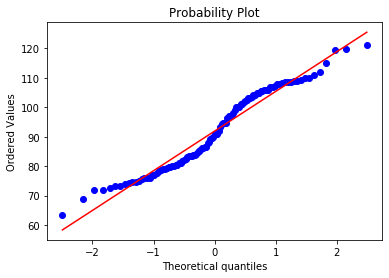
Since most of the data points fall along the QQline, we can say that data is Normal

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

Dataset: wc-at.csv

**Sol: -** wc = pd.read\_csv("C:\\Users\\hardi\\OneDrive\\Documents\\Excelr\_1\\Python\\Basic Statistics Level 1\\wc-at.csv")

stats.probplot(wc['Waist'],dist="norm",plot=pylab)[The same can be achieved in R using qqnorm(wc$Waist) and qqline(wc$Waist)]



Many data points are away from the normal line(QQline). Also the line in not in between Q1 and Q3, hence data is not Normal for “Waist”.

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

* @90% = stats.norm.ppf(0.95)= 1.64[In R we can use qnorm(0.95)]
* @94% = stats.norm.ppf(0.97)= 1.88
* @60% = stats.norm.ppf(0.80)= 0.84

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

* @95% = stats.t.ppf(0.975,24)= 2.06[In R we can use :- qt(0.975,24)]
* @96% = stats.t.ppf(0.98,24) = 2.171545
* @99% = stats.t.ppf(0.995,24) = 2.79694

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 260days

Sol:-Python Code - stats.t.pdf(tscore,df)[Equivalent rcode🡪pt(tscore,df)]

df🡪 degrees of freedom

µ = 270, n = 18, = 260, s = 90

t score = (sample mean – population mean)/ (sample SD/sqrt(n))

= (260-270)/(90/)

* stats.t.pdf(t score, df)
* stats.t.pdf(-0.4714, 17)
* 0.35