|  |  |
| --- | --- |
| 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 | Character |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Character |

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 | Ordinal |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| 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?

Ans: Total outcome: 8

Expected outcome:3

Probability=3/8

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

Ans:Total out come=36

1. Equal to 1 =0 (because least value will be 2)
2. Less than or equal to 4 = 3/36 => 1/12 ((1,3),(2,2),(3,1))
3. Sum is divisible by 2 and 3 =((2,4),(4,2),(3,3),(1,5),(5,1),(6,6))

= 6/36 => 1/12

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?

Ans: sample space =7c2 = 21. Let event E be none of the balls is blue = all balls are either red or green or both. n(E) = 5c2 =10 p(E) =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: 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

=   3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weight

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**ANS:**

df.describe().T

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** | range | variance |
| **Points** | 32 | 3.596563 | 0.534679 | 2.76 | 3.08 | 3.695 | 3.92 | 4.93 | 2.17 | 0.2858 |
| **Score** | 32 | 3.21725 | 0.978457 | 1.513 | 2.58125 | 3.325 | 3.61 | 5.424 | 3.911 | 0.9573 |
| **Weigh** | 32 | 17.84875 | 1.786943 | 14.5 | 16.8925 | 17.71 | 18.9 | 22.9 | 8.4 | 3.1931 |

Point: it is left skewed data, much of the data elements lied to left side of mean point.

Score: it is left skewed data, much of the data elements lied to left side of mean point

Weigh: it is right skewed data , much of its data are lying to words right side of mean

data points are seeming to less scattered and most of them lying near to mean data value in all three cases.

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:

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**

**df\_nineA.skew(): speed -0.117510 ,dist 0.806895**

**df\_nineB.kurt(): speed -0.508994 dist 0.405053**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**df\_nineB.skew():SP 1.611450, WT -0.614753**

**df\_nineB.kurt():SP 2.977329 WT 0.950291**

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



**This is a right skewed data.**

**More than 50% chick weight is between 50 to 150**

**Most of the chick weight is between 50 to 100**



Above box plot shows the data is right skewed

Most of its data are lying in Q3

Some of outliers are there at upper 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?

n=2000

N=300000

Xbar=200

S=30

# Avg. weight of Adult in Mexico with 94% CI

stats.norm.interval(0.94,200,30/(2000\*\*0.5))

(198.738325292158, 201.261674707842)

# Avg. weight of Adult in Mexico with 96% CI

stats.norm.interval(0.96,200,30/(2000\*\*0.5))

(198.62230334813333, 201.37769665186667)

# Avg. weight of Adult in Mexico with 98% CI

stats.norm.interval(0.98,200,30/(2000\*\*0.5))

(198.43943840429978, 201.56056159570022)

**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**

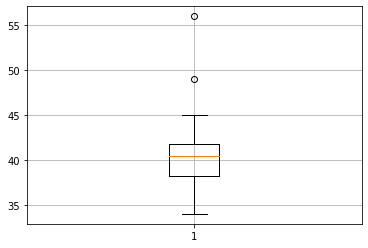
1. Find mean, median, variance, standard deviation.

Mean=41,median=40.5,variance=25.53,standard deviation=5.05

1. What can we say about the student marks?

From the above calculation mean > median which indicates data is slightly right skewed, most of the data lies between 39 to 42

If we go for box plot we can see we have 2 outliers 49,56.



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

When mean = median assuming the data is unimodal the nature of graph will be symmetric and skew=0

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

When mean > median the nature of skewness will be right skewed and have tail towards right.

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

When mean < median the nature of skewness will be left skewed and have tail towards left.

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

Positive kurtosis value indicates that the curve is more picked and thicker tail.

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

A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the data?

From the above boxplot median is lying towards higher value, most of the data are between 10 to 18 and the data is not normally distributed.

What is nature of skewness of the data?

The data is left skewed. The whisker range of minimum value is greater than maximum

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

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.

1.From the above plot it seems as they both share same median in range (250,275).

2.the plot is showing no outliers and normally distributed with skewness ~0 at minimum and maximum whisker range.

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)

round((1-stats.norm.cdf(38,loc=cars\_mpg.mean(),scale=cars\_mpg.std())),4)

=0.3476

* 1. P(MPG<40)

round((1-stats.norm.cdf(40,loc=cars\_mpg.mean(),scale=cars\_mpg.std())),4)

= 0.2707

c. P (20<MPG<50) = p(MPG<50)-p(MPG<20)

p(MPG<50)= round(stats.norm.cdf(50,loc=cars\_mpg.mean(),scale=cars\_mpg.std()),4)

= 0.956

p(MPG<20)

= round(stats.norm.cdf(20,loc=cars\_mpg.mean(),scale=cars\_mpg.std()),4)

= 0.0571

p(MPG<50)-p(MPG<20)=0.9560-0.0571= 0.8989

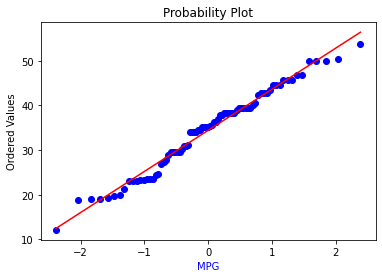
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

stats.probplot(cars\_mpg,dist='norm',plot=plt)

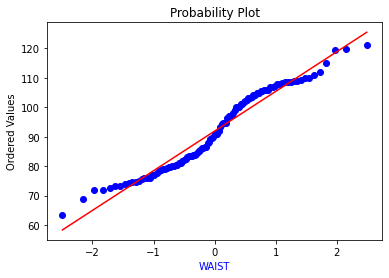
MPG of cars follows normal distribution.

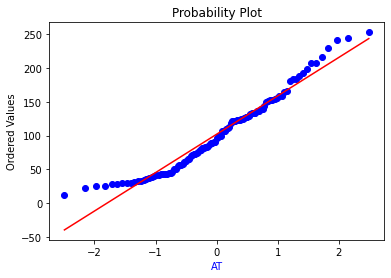


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Neither AT nor Waist from wc-at data follows normal distribution.





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

Z scores of 90% confidence interval: round(stats.norm.ppf(1-0.1/2),4)

1.6449

Z scores of 94% confidence interval: round(stats.norm.ppf(1-0.06/2),4)

1.8808

Z scores of 60% confidence interval: round(stats.norm.ppf(1-0.4/2),4)

0.8416

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

t scores of 95% confidence interval: round(stats.t.ppf(1-0.05,df=24),3)

1.711

t scores of 96% confidence interval: round(stats.t.ppf(1-0.04,df=24),3)

1.828

t scores of 99% confidence interval: round(stats.t.ppf(1-0.01,df=24),3)

2.492

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

n=18

x\_bar=260

std=90

mu=270

t\_score=(x\_bar-mu)/(std/sqrt(n))

(260-270)/(90/sqrt(18))= -0.4714

stats.t.cdf(t\_score,df=17)

= 0.3216

32%