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

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) | 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/Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans: H: Heads and T: Tails

Total number of Outcome=8(2\*2\*2)

Observation of 2H and 1T = 3

Probability of 2H and 1T = Observation of 2H and 1T/ Total number of Outcome

=3/8

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

Total number of Outcome=36(6\*6)

1. Equal to 1

Ans: P(1) = 0/36 = 0

1. Less than or equal to 4

Ans : P(<= 4) = 6/36 = 0.166

1. Sum is divisible by 2 and 3  
   Ans: (1,1),(1,2),(1,3),(1,5),(2,2),(2,4),(2,6),(3,1),(3,3),(3,5),(3,6),(4,1),(4,2),(4,4),(4,5),(4,6),(5,1),(5,3),(5,4),(5,5),(6,2),(6,3),(6,4),(6,6)

No. of outcomes divisible by 2 and 3 = 24

P(2,3) = 24/36 = 0.666

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: Total number of Balls: 7

Number of ways getting 2 balls from total balls = nCr=n!r!(n−r)!

= 7!/(2!\*(7-2)!)

=7\*6\*5!/(2\*1(5)!)

=(7\*6)/2

=21

Number of non-blue balls = 5

Number of ways getting non blue balls = 10

P(non blue) = 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 value = P(x)\*x

= 1\*.015+4\*.20+3\*.65+5\*.005+6\*.01+2\*.120

=2.37

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

* For Points,Score,Weigh>

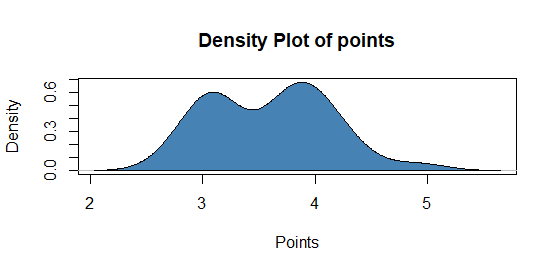
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

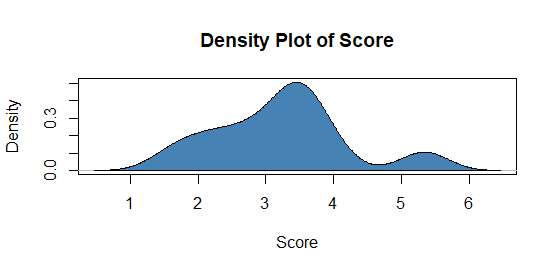
**Use Q7.csv file**

Ans:

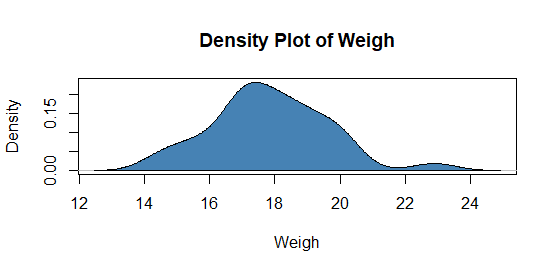
|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.5965 | 3.2175 | 17.8487 |
| Median | 3.69 | 3.215 | 17.6 |
| Mode | Multimode 2.76,3.07,3.08,3.15,3.9, 3.92, 4.08,4.22 | Bimode 3.44, 3.57 | Bimode 17.02 , 18.9 |
| Variance | 0.2858 | 0.9574 | 3.1932 |
| Standard Deviation | 0.5346 | 0.9784 | 1.7870 |
| Range | 2.76- 4.93 | 1.513- 5.424 | 14.5- 22.9 |

**Points**

**Score**



**Weigh**



Among all the features the std deviation of Weigh is high which means the data points of the given dataset is far generally from the mean i.e the datapoints are widely spread for the weigh feature

Score feature has high Skewness compared to other feature and the skew is towards positive side indicating positive skewness.

Weigh feature has high Kurtosis compared to other feature as it has peakedness as well as heavy tails compared to other features whereas Points feature has wider peak and light tails

> skewness(`Q7`$Weigh)

[1] 0.3690453

> kurtosis(`Q7`$Weigh)

[1] 0.3351142

> skewness(`Q7`$Score)

[1] 0.4231465

> kurtosis(`Q7`$Score)

[1] -0.02271075

> skewness(`Q7`$Points)

[1] 0.2659039

> kurtosis(`Q7`$Points)

[1] -0.7147006

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 is mean of the dataset

Or Probability\*X

=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

= 145.33(mean)

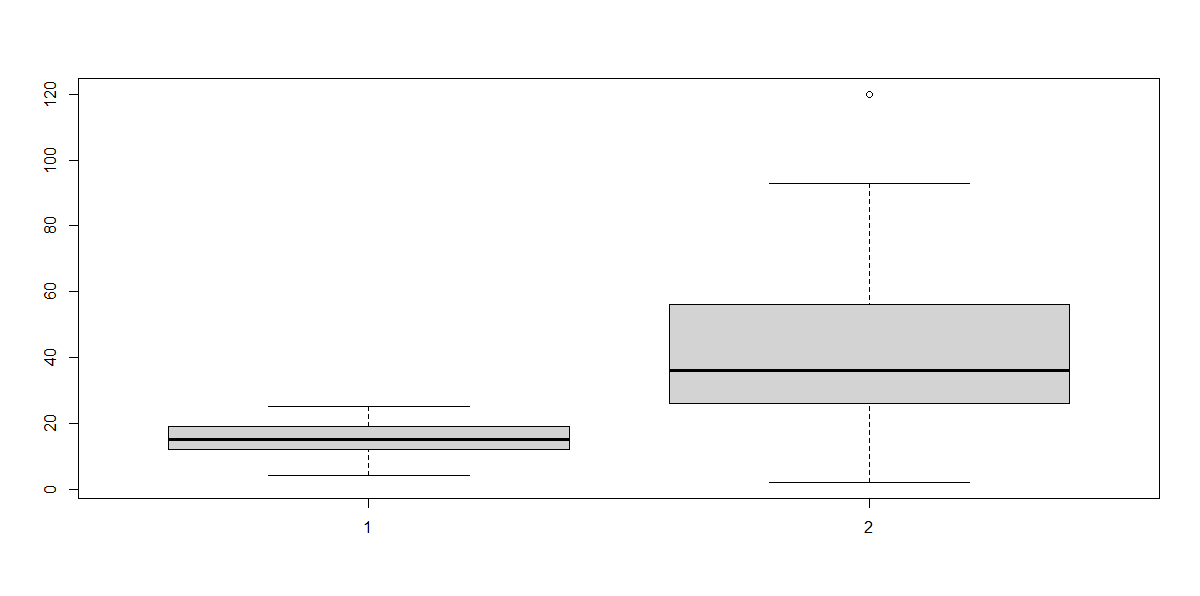
**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

Ans:

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| speed | -0.1105533 | -0.6730924 |
| dist | 0.7591268 | 0.1193971 |



For Speed feature the distribution is skewed negatively indicating negative skewness with wider peak and light tail (-ve kurtosis) and for dist feature the distribution is skewed positively indicating positive skewness with high peak and heavy tail(+ve Kurtosis) compared to speed.

As both the skewness value is btwn -0.5 to +0.5 the distribution of both the plot is nearly to symmetric

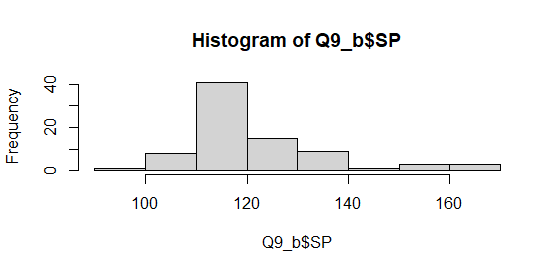
**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans:

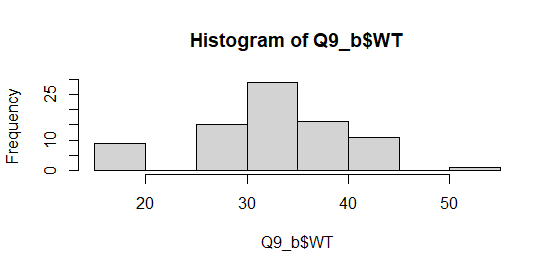
|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.552258 | 2.583072 |
| WT | -0.5921721 | 0.7257402 |

**SP**



Most of the data points falls btwn 110-130 in the above distribution

**WT**



Most of the data points falls btwn 30-40 in the above distribution

For SP feature the distribution is skewed towards positive indicating positive skewness with high peak and heavy tail (+ve kurtosis) and for WT feature the distribution is skewed towards negative side indicating negative skewness with wider peak and light tail(-ve kurtosis) compared to speed.

The skewness value of SP is more than 1 so the distribution has highly skewed and for the WT the skewness value is greater than -0.5 indicating moderately skewed.

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



Ans: In the given histogram data is not normally distributed and the skew is towards right i.e tail of the distribution is towards positve side and mass of the distribution is clustered towards left side indicating **positive skewness distribution** where mode is greater than mean and median.



Ans: In the given box plot data is not normally distributed and the above/top whisker length is longer and median is close towards bottom indicating that the given distribution of data is **positively skewed** and also it has outliers which is pointed above**.** For the positive skewed the mean>median

**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: Theoretical way : Interval estimate = x̄ ± t\* s / (√n)

To calculate t value

df =n-1

stats.t.ppf(x, df) where x is the % value of confidence

Sample mean of a weigh (x̄)= 200 pounds

Sample std deviation of weigh (s) = 30 pounds

Sample size (n)= 2000

Using python: as n>1 so std deviation of sample/sqrt(n)

94%confidence Interval:

np.round(stats.t.interval(.94,loc=200,scale=(30/np.sqrt(2000)),df=1999),4)

([198.7376, 201.2624])

96%confidence Interval:

np.round(stats.t.interval(.96,loc=200,scale=(30/np.sqrt(2000)),df=1999),4)

([198.6214, 201.3786])

98%confidence Interval:

np.round(stats.t.interval(.98,loc=200,scale=(30/np.sqrt(2000)),df=1999),4)

([198.4382, 201.5618])

**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.

Ans: Mean: 738/13 = 41

Median: 40+41 = 40.5

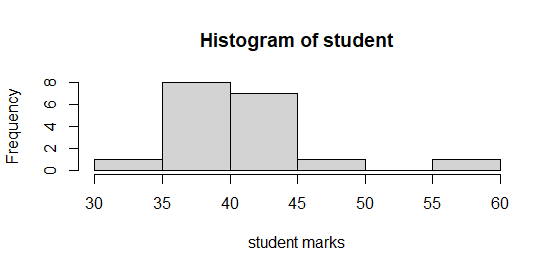
Mode: 41

Variance: 25.5294

Standard Deviation: 5.0526

1. What can we say about the student marks?

Ans:



Based on the student marks distribution mass of the students fall bwtn the 35 to 45 and std deviation is high indicating that each value has spread away from mean.

Based on the distribution it is positively skewed with high peakedness (+ve kurtosis) as it is positive most of the marks scored by students falls towards positive side

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

Ans: When mean=median=mode then there is no skewness and the curve is symmetric which is normal distribution i.e data is normally distributed

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

Ans: Positive skewness as there is skew towards +ve side or right side in distribution

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

Ans: Negative skewness as there is skew towards -ve side or left side in distribution

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

Ans: Positive Kurtosis means the distribution has heavier tails/wider tails and high peak/narrow peak and width of peak is less when compared to normal distribution (where value is 0)

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

Ans: Negative Kurtosis means the distribution has Lighter tails/thinner tails and wider peak/flatten and width of peak is more when compared to normal distribution (where value is 0)

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



What can we say about the distribution of the data?

Ans: The given distribution of data is visually represented as box plot and it is not normally distributed and it has skew towards left i.e left side of whisker has more length compared to right and median > mean indicating that the distribution is **negatively skewed**

What is nature of skewness of the data?

Ans : The distribution is **negatively skewed**

What will be the IQR of the data (approximately)?   
Ans: IQR = Upper quartile – Lower Quratile

= 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.

Ans:

* Both box plot is plotted against same wbs scale
* The median of both the boxplot is same i.e around ~262.5
* As the distribution of both side of the box is equal  
  (i.e. median divides the box equally on both side and both side whisker length is also same), we can say that both the box plots are **normally** **distributed**.
* Box plot2 has more variety data range i.e the upper extreme-lower extreme value is more in box plot 2 than Box plot1
* Box plot2 has the highest value of wbs when compared to Boxplot1 which is ~350

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): 34.76%

np.round(1-stats.norm.cdf(38,cars\_df.MPG.mean(),cars\_df.MPG.std()),4)

0.3476

* 1. P(MPG<40): 72.93%

np.round(stats.norm.cdf(40,cars\_df.MPG.mean(),cars\_df.MPG.std()),4)

0.7293

* 1. P (20<MPG<50): 1.31%

mpg1 = stats.norm.cdf(50,cars\_df.MPG.mean(),cars\_df.MPG.std())

mpg2 = 1-stats.norm.cdf(20,cars\_df.MPG.mean(),cars\_df.MPG.std())

np.round(mpg1-mpg2,4)

0.0131

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

*import seaborn as sn*

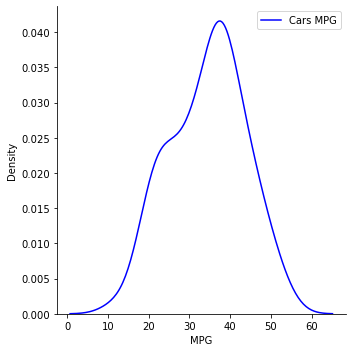
*import matplotlib.pyplot as plt*

*sn.displot(cars\_df.MPG,label="Cars MPG",kind='kde',color = 'blue', legend=True)*

*plt.xlabel("MPG")*

*plt.legend()*

*plt.show()*



*print('Mean of Cars MPG ---',cars\_df.MPG.mean(),'\n Median of Cars MPG --- ',cars\_df.MPG.median())*

Mean of Cars MPG --- 34.422075728024666

Median of Cars MPG --- 35.15272697

The distribution is approximately symmetric as both mean and median are almost equal and from the distribution graph also it is almost symmetric and it is dense at the center and less at the tails

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

Dataset: wc-at.csv

Ans:

**Waist**

*sn.displot(df.AT,label='AT',kind='kde')*

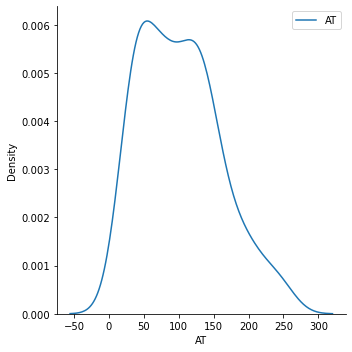
*plt.legend()*

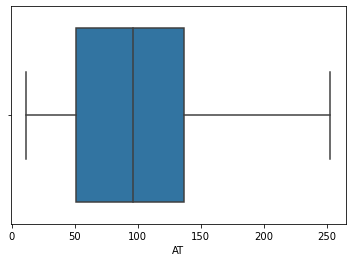
*plt.show()*

*sn.boxplot(df.AT)*

*plt.show()*

*print('Mean of AT ---',df.AT.mean(),'\n Median of AT --- ',df.AT.median())*





Mean of AT --- 101.89403669724771

Median of AT --- 96.54

As the mean is not equal to median and also from the distribution we can see the more data is spread towards +ve side i.e right whiskers is large compared to left .That’s why waist data is not normally distribution

**AT**

*df=pd.read\_csv('/Users/Akshay/Documents/DataScience/ExcelR/Assignments/DataScience/Basic Statistics\_Level-1/wc-at.csv')*

*df.head(5)*

*sn.displot(df.Waist,label='Waist',kind='kde')*

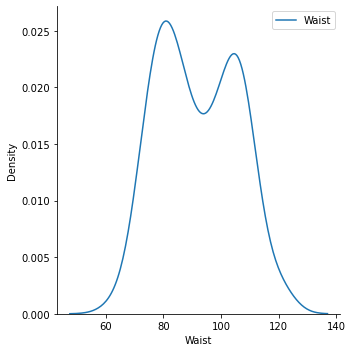
*plt.legend()*

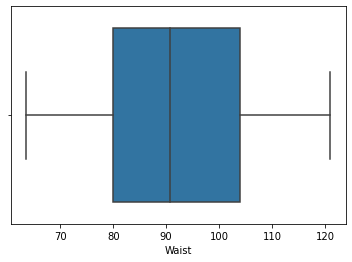
*plt.show()*

*sn.boxplot(df.Waist)*

*plt.show()*

*print('Mean of Waist ---',df.Waist.mean(),'\n Median of Waist --- ',df.Waist.median())*





Mean of Waist --- 91.90183486238533

Median of Waist --- 90.8

As the mean is approximately equal to median and also from the distribution (box plot) we can see that both the whiskers are almost equal. That’s why waist data is approximately normally distribution

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

Ans:

1-alpha = confidence

alpha = 1- confidence

restArea = alpha/2

Z-score = stats.norm.ppf(confidence+ restArea)

|  |  |
| --- | --- |
| Confidence Interval | Z-score |
| 90% | np.round(stats.norm.ppf(.95),4)  1.6449 |
| 94% | np.round(stats.norm.ppf(.97),4)  1.8808 |
| 60% | np.round(stats.norm.ppf(.80),4)  0.8416 |

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

Ans: n=25

df=n-1=24

1-alpha = confidence

alpha = 1- confidence

restArea = alpha/2

Z-score = stats.t.ppf(confidence+ restArea,df=n-1)

|  |  |
| --- | --- |
| Confidence Interval | t-score |
| 95% | np.round(stats.t.ppf(.975,df=24),4)  2.0639 |
| 96% | np.round(stats.t.ppf(.98,df=24),4)  2.1715 |
| 99% | np.round(stats.t.ppf(.995,df=24),4)  2.7969 |

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

Ans:

x̄ = 260

s = 90

n = 18

P(x< 260)=?

To calculate t value

df =n-1

stats.t.cdf(x, df)

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

-0.4714045207910317

np.round(stats.t.cdf(t,df=17),4)

0.3217

So there is a chance of 32.17% probability that the 18 randomly selected bulbs would have an average life of no more than 260 days