|  |  |
| --- | --- |
| 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 | Nominal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Nominal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| 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 :

S ={HHH,HHT,HTH,THH,HTT,THT,TTH,TTT}

n(S) = 8

Number of probability that two head and one tail are:

n(A) = 3

P(A)=n(A)/n(S)

P (Getting 2 head &1Tail) = 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

Answer :

S ={(1,1), (1,2), (1,3), (1,4), (1,5),(1,6), (2,1), (2,2), (2,3), (2,4), (2,5),(2,6), (3,1), (3,2), (3,3), (3,4), (3,5),(3,6), (4,1), (4,2), (4,3), (4,4), (4,5),(4,6), (5,1), (5,2), (5,3), (5,4), (5,5),(5,6), (6,1), (6,2), (6,3), (6,4), (6,5),(6,6)}

1. P(Getting equal to 1) = 0/36 = 0
2. P(Getting less than or equal to 4) = 6/36 =1/6 = 0.17
3. P(Getting sum is divisible by 2 and 3) = 24/36 = 8/12 = 2/3 = 0.67

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?

Answer:

Total numbers 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

n(S) = 7C2

= (7x6)/ (2x1)

= 21

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

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

n(E) = 5C2

= (5x4)/(2x1)

= 10

Therefore 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) Expected = Sum [X\*P(x)]

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

Answer:

The Expected number of candies for a randomly selected child

=1 x 0.015 + 4 x 0.20 + 3 x 0.65 + 5 x 0.005 + 6 x 0.01 + 2 x 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**

Answer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Points | Score | Weigh |
| 0 | Mazda RX4 | 3.90 | 2.620 | 16.46 |
| 1 | Mazda RX4 Wag | 3.90 | 2.875 | 17.02 |
| 2 | Datsun 710 | 3.85 | 2.320 | 18.61 |
| 3 | Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| 4 | Hornet Sportabout | 3.15 | 3.440 | 17.02 |
| 5 | Valiant | 2.76 | 3.460 | 20.22 |
| 6 | Duster 360 | 3.21 | 3.570 | 15.84 |
| 7 | Merc 240D | 3.69 | 3.190 | 20.00 |
| 8 | Merc 230 | 3.92 | 3.150 | 22.90 |
| 9 | Merc 280 | 3.92 | 3.440 | 18.30 |
| 10 | Merc 280C | 3.92 | 3.440 | 18.90 |
| 11 | Merc 450SE | 3.07 | 4.070 | 17.40 |
| 12 | Merc 450SL | 3.07 | 3.730 | 17.60 |
| 13 | Merc 450SLC | 3.07 | 3.780 | 18.00 |
| 14 | Cadillac Fleetwood | 2.93 | 5.250 | 17.98 |
| 15 | Lincoln Continental | 3.00 | 5.424 | 17.82 |
| 16 | Chrysler Imperial | 3.23 | 5.345 | 17.42 |
| 17 | Fiat 128 | 4.08 | 2.200 | 19.47 |
| 18 | Honda Civic | 4.93 | 1.615 | 18.52 |
| 19 | Toyota Corolla | 4.22 | 1.835 | 19.90 |
| 20 | Toyota Corona | 3.70 | 2.465 | 20.01 |
| 21 | Dodge Challenger | 2.76 | 3.520 | 16.87 |
| 22 | AMC Javelin | 3.15 | 3.435 | 17.30 |
| 23 | Camaro Z28 | 3.73 | 3.840 | 15.41 |
| 24 | Pontiac Firebird | 3.08 | 3.845 | 17.05 |
| 25 | Fiat X1-9 | 4.08 | 1.935 | 18.90 |
| 26 | Porsche 914-2 | 4.43 | 2.140 | 16.70 |
| 27 | Lotus Europa | 3.77 | 1.513 | 16.90 |
| 28 | Ford Pantera L | 4.22 | 3.170 | 14.50 |
| 29 | Ferrari Dino | 3.62 | 2.770 | 15.50 |
| 30 | Maserati Bora | 3.54 | 3.570 | 14.60 |
| 31 | Volvo 142E | 4.11 | 2.780 | 18.60 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Variance | Standard Deviation | Range |
| Points | 3.596563 | 3.695 | 3.07 , 3.92 | 0.285881 | 0.534679 | 2.17 |
| Score | 3.217250 | 3.325 | 3.44 | 0.957379 | 0.978457 | 3.9110 |
| Weigh | 17.848750 | 17.710 | 17.02, 18.90 | 3.193166 | 1.786943 | 8.3999 |

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?

Answer:

Expected Value of X

X ={108, 110, 123, 134, 135, 145, 167, 187, 199}

n(S) = 9

P(X) = n(X)/n(S)

P(X) = 1308/9

P(X) = 145.3333

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

Answer :

**1) Car Speed and Distance**

|  |  |  |
| --- | --- | --- |
| Index | speed | dist |
| 1 | 4 | 2 |
| 2 | 4 | 10 |
| 3 | 7 | 4 |
| 4 | 7 | 22 |
| 5 | 8 | 16 |
| 6 | 9 | 10 |
| 7 | 10 | 18 |
| 8 | 10 | 26 |
| 9 | 10 | 34 |
| 10 | 11 | 17 |
| 11 | 11 | 28 |
| 12 | 12 | 14 |
| 13 | 12 | 20 |
| 14 | 12 | 24 |
| 15 | 12 | 28 |
| 16 | 13 | 26 |
| 17 | 13 | 34 |
| 18 | 13 | 34 |
| 19 | 13 | 46 |
| 20 | 14 | 26 |
| 21 | 14 | 36 |
| 22 | 14 | 60 |
| 23 | 14 | 80 |
| 24 | 15 | 20 |
| 25 | 15 | 26 |
| 26 | 15 | 54 |
| 27 | 16 | 32 |
| 28 | 16 | 40 |
| 29 | 17 | 32 |
| 30 | 17 | 40 |
| 31 | 17 | 50 |
| 32 | 18 | 42 |
| 33 | 18 | 56 |
| 34 | 18 | 76 |
| 35 | 18 | 84 |
| 36 | 19 | 36 |
| 37 | 19 | 46 |
| 38 | 19 | 68 |
| 39 | 20 | 32 |
| 40 | 20 | 48 |
| 41 | 20 | 52 |
| 42 | 20 | 56 |
| 43 | 20 | 64 |
| 44 | 22 | 66 |
| 45 | 23 | 54 |
| 46 | 24 | 70 |
| 47 | 24 | 92 |
| 48 | 24 | 93 |
| 49 | 24 | 120 |
| 50 | 25 | 85 |

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| Speed | -0.117510 | -0.508994 |
| Distance | 0.806895 | 0.405053 |

**2) SP and Weight**

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| 1 | 104.1854 | 28.76206 |
| 2 | 105.4613 | 30.46683 |
| 3 | 105.4613 | 30.1936 |
| 4 | 113.4613 | 30.63211 |
| 5 | 104.4613 | 29.88915 |
| 6 | 113.1854 | 29.59177 |
| 7 | 105.4613 | 30.30848 |
| 8 | 102.5985 | 15.84776 |
| 9 | 102.5985 | 16.35948 |
| 10 | 115.6452 | 30.92015 |
| 11 | 111.1854 | 29.36334 |
| 12 | 117.5985 | 15.75353 |
| 13 | 122.1051 | 32.81359 |
| 14 | 111.1854 | 29.37844 |
| 15 | 108.1854 | 29.34728 |
| 16 | 111.1854 | 29.60453 |
| 17 | 114.3693 | 29.53578 |
| 18 | 117.5985 | 16.19412 |
| 19 | 114.3693 | 29.92939 |
| 20 | 118.4729 | 33.51697 |
| 21 | 119.1051 | 32.32465 |
| 22 | 110.8408 | 34.90821 |
| 23 | 120.289 | 32.67583 |
| 24 | 113.8291 | 31.83712 |
| 25 | 119.1854 | 28.78173 |
| 26 | 114.5985 | 16.04317 |
| 27 | 120.7605 | 38.06282 |
| 28 | 119.1051 | 32.83507 |
| 29 | 99.56491 | 34.48321 |
| 30 | 121.8408 | 35.54936 |
| 31 | 113.4846 | 37.04235 |
| 32 | 112.289 | 33.23436 |
| 33 | 119.9211 | 31.38004 |
| 34 | 121.3926 | 37.57329 |
| 35 | 111.289 | 32.70164 |
| 36 | 115.0131 | 31.91122 |
| 37 | 114.0934 | 28.754 |
| 38 | 116.9094 | 27.87992 |
| 39 | 116.9094 | 28.6305 |
| 40 | 128.4613 | 30.11543 |
| 41 | 116.3926 | 37.39252 |
| 42 | 115.7488 | 35.02718 |
| 43 | 117.4613 | 30.52743 |
| 44 | 114.0934 | 28.34398 |
| 45 | 114.381 | 33.07863 |
| 46 | 117.1051 | 32.62192 |
| 47 | 118.2087 | 36.49862 |
| 48 | 116.4729 | 33.91006 |
| 49 | 127.9094 | 28.0706 |
| 50 | 118.289 | 33.45847 |
| 51 | 118.289 | 33.21395 |
| 52 | 118.289 | 33.43671 |
| 53 | 120.4043 | 40.39816 |
| 54 | 143.3926 | 37.62069 |
| 55 | 135.3926 | 37.25439 |
| 56 | 126.4043 | 40.58907 |
| 57 | 110.4613 | 30.14754 |
| 58 | 118.289 | 32.73452 |
| 59 | 112.6452 | 30.61528 |
| 60 | 115.5766 | 37.66287 |
| 61 | 130.2087 | 36.88815 |
| 62 | 117.6685 | 37.86041 |
| 63 | 126.0481 | 43.39099 |
| 64 | 125.3123 | 40.72283 |
| 65 | 128.1284 | 40.15948 |
| 66 | 126.5985 | 15.71286 |
| 67 | 132.4846 | 37.97996 |
| 68 | 133.6802 | 41.57397 |
| 69 | 133.3123 | 40.47204 |
| 70 | 158.3007 | 37.14173 |
| 71 | 164.5985 | 15.82306 |
| 72 | 133.416 | 44.01314 |
| 73 | 133.1401 | 43.35312 |
| 74 | 124.7152 | 52.99775 |
| 75 | 121.8642 | 42.6187 |
| 76 | 132.8642 | 42.77822 |
| 77 | 169.5985 | 16.13295 |
| 78 | 150.5766 | 37.92311 |
| 79 | 151.5985 | 15.76963 |
| 80 | 167.9445 | 39.4231 |
| 81 | 139.8408 | 34.94861 |

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.611450 | 2.977329 |
| WT | -0.614753 | 0.950291 |

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



Answer:

1) The most of the data points are connected in the range 50-125 with frequency 200.

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

So the expected value the above distribution is 75.

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

2) The above boxplot we can see that are outliers beyond 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?

Answer:

X+/-(Z1- α. σ/sqrt(n)

Degree of freedom = 2000-1 = 1999

Confidence interval = 94%

1. σ / 2) = 1-0.025 = 0.975

For Confidence interval for 94% is (198.74, 201.26)

Margin of error (E) of 94% is 1.26167.

Confidence interval for 98% is (198.44, 201.56)

Margin of error (E) of 98% is 1.56056.

Confidence interval for 96% is (198.622, 201.378)

Margin of error (E) of 96% is 1.3777.

**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?

Answer:

(1)

|  |  |
| --- | --- |
| Mean | 41.0 |
| Median | 40.5 |
| Variance | 25.529411764705884 |
| Standard Deviation | 5.05266382858645 |

(2) Marks are normally distributed. There are no visible outliers.

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

Answer:

Symmetrical distribution

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

Answer:

Right Skewed distribution

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

Answer:

Left Skewed

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

Answer:

A distribution with a position kurtosis value indicates that the distribution has heavier tails than the normal distribution.

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

Answer:

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?

What is nature of skewness of the data?

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

1. Let’s as assume above Boxplot is about the age of the student in a school.

50% of the people are above 10 year old and the remaining are below 10 year.

And the student’s above 15 year old students is approximately 40%

1. It is left Skewed, median > mean.
2. IOR = Q3-Q1 = 18-10

Approximately = 8.

Q19) Comment on the below Boxplot visualizations?



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

Answer:

By observing both the plots whisker’s level is high in Boxplot 2, mean and median are equal hence distribution is Symmetrical.

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)
  3. P (20<MPG<50)

Answer:

1. The P(MPG>38) is : 0.3475939251582705
2. The P(MPG<40) is : 0.7293498762151616
3. The P(20<MPG<50) is : 0.8988689169682046

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

Answer:

1. MPG of Cars does follow normal distribution approximately (as mean and median are approx. same)

MPG of Cars Mean: 34.422075728024666

MPG of Cars Median: 35.15272697

1. Both the Adipose Tissue (AT) and Waist Circumference(Waist) data set do follow the normal distribution approximately (as mean and median of both the data are approximately same)

Waist Circumference Mean: 91.90183486238533

Waist Circumference Median: 90.8

Adipose Tissue Mean: 101.89403669724771

Adipose Tissue Median: 96.54

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

Answer:

Z score of 90% confidence interval is 1.6448536269514722

Z score of 94% confidence interval is 1.8807936081512509

Z score of 60% confidence interval is 0.8416212335729143

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

Answer:

T scores of 95% confidence interval is 2.0638985616280205

T scores of 96% confidence interval is 2.1715446760080677

T scores of 99% confidence interval is 2.796939504772804

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

Answer:

Mean = 270 days = u

Sample size = 18 = n

Sample mean = 260 = x

Sample Standard Deviation = 90 = s

Solutions:

t = x-u/s/sqrt n

t = 260-270/90/sqrt18

t = -10/90/4.2426406871

t = -10/21.213203435

t = -0.471404520

P\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... (Using cdf function)

P\_value= 0.32167411684460556

P\_value=stats.t.sf(abs(t\_score),df=n-1)... (Using sf function)

P\_value= 0.32167411684460556

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.17%Assuming significance value α = 0.05 (Standard Value)(If p\_value < α ; Reject Ho and accept Ha or vice-versa)

Thus, as p-value > α ; Accept Ho i.e. The CEO claims are false and the avg life of bulb > 260 days.