|  |  |
| --- | --- |
| **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** | Ordinal |
| **Fahrenheit Temperature** | Interval |
| **Height** | Ratio |
| **Type of living accommodation** | Ordinal |
| **Level of Agreement** | Ordinal |
| **IQ(Intelligence Scale)** | Interval |
| **Sales Figures** | Ratio |
| **Blood Group** | Ordinal |
| **Time Of Day** | Ordinal |
| **Time on a Clock with Hands** | Interval |
| **Number of Children** | Nominal |
| **Religious Preference** | Ordinal |
| **Barometer Pressure** | Ratio |
| **SAT Scores** | Ordinal |
| **Years of Education** | Nominal |

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

All possible outcomes = { (HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT) }

P (2 heads and 1 tail) = 3/8

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

All possible outcomes = 36

1. **Equal to 1** = 0
2. **Less than or equal to 4**

P (Less than or equal to 4) = { (1,1), (1,2), (1,3), (2,1), (2,2), (3,1) }

= 6/36

= 1/6

1. **Sum is divisible by 2 and** 3

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

= 5/36

**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?**

Total number of balls = 7

Let S = Total number of ways drawing 2 balls out of 7

= 7C2 = = 21

E = Event of drawing 2 balls at random, none of which is blue

= 5C2 = = 10

P(E) = =

**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**

Child A: P(1 candy) = 1\*0.015 = 0.015

Child B: P(4 candies) = 4\*0.20 = 0.8

Child C: P(3 candies) = 3\*0.65 = 1.95

Child D: P(5 candies) = 5\*0.005 = 0.025

Child E: P(6 candies) = 6\*0.01 = 0.06

Child E: P(2 candies) = 2\*0.120 = 0.24

Therefore, the 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>**
* **Use Q7.csv file**

Mean:

Points 3.596563

Score 3.217250

Weigh 17.848750

Median:

Points 3.695

Score 3.325

Weigh 17.710

Mode:

Car Points: 0 3.07

1. 3.92

Car Score: 0 3.44

Car Weigh: 0 17.02

1 18.90

Variance:

Points 0.285881

Score 0.957379

Weigh 3.193166

Standard Deviation:

Points 0.534679

Score 0.978457

Weigh 1.786943

Range:

Points: 2.17

Score: 3.911

Weigh: 8.399

Comments:

**Chart, box and whisker chart

Description automatically generated**

Points Score Weigh

count 32.000000 32.000000 32.000000

mean 3.596563 3.217250 17.848750

std 0.534679 0.978457 1.786943

min 2.760000 1.513000 14.500000

25% 3.080000 2.581250 16.892500

50% 3.695000 3.325000 17.710000

75% 3.920000 3.610000 18.900000

max 4.930000 5.424000 22.900000

IQR 0.84 1.029 2.008

U.P.L 5.18 5.153 21.91

L.P.L 4.34 4.124 19.9

* Where,

IQR – Inter Quartile Range

U.P.L – Upper Limit

L.P.L – Lower Limit

* The values above Upper Limit (Score - 5.153 and weigh - 21.91) are called as Outliers
* There are three outliers (5.25, 5.424 and 5.345) in Score dataset.
* There is one outlier (22.9) in Weigh dataset.

**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?**

Total number of patients = 9

One of the patients is chosen at random = 1/9

Expected value of the weight of the patients = Sum [ P(patient chosen) \* weight ]

Expected value = [ ( *\** 108 ) + ( *\** 110 ) + ( *\** 123 ) + ( *\** 134 ) + ( *\** 135 ) + ( *\** 145 ) + ( *\** 167) + ( *\** 187 ) + ( *\** 199 ) ]

= 12 + 12.22 + 13.66 + 14.88 + 15 + 16.11 + 18.55 + 20.78 +22.11

= 145.31

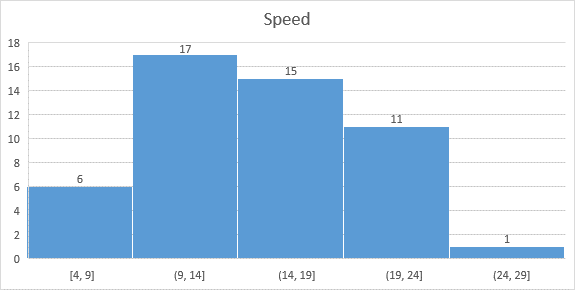
**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

speed

* Mean: 15.4
* Median 15
* Mode 20
* St.Dev 5.287644435
* Variance 27.95918367
* Kurtosis -0.50899442
* Skewness -0.117509861
* Range 21
* Minimum 4
* Maximum 25

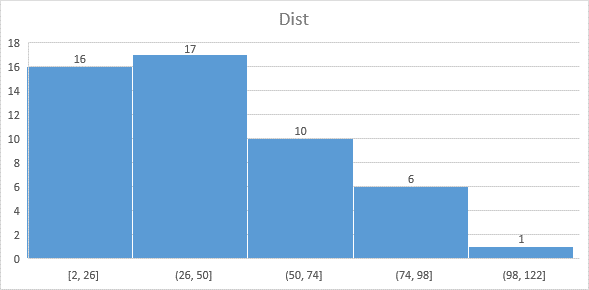


Inferences of Car Speed:

* The value of skewness is -0.117510 which is lying between -0.5 and +0.5, indicating approximately symmetric distribution (Normal Distribution).
* The value of kurtosis is -0.508994 which is less than ZERO (Value < 0), indicating lower degree of peakedness, resulting in Platykurtic distribution.

Distance:

* Mean 42.98
* Median 36
* Mode 26
* St.Dev 25.76938
* Variance 664.0608
* Kurtosis 0.405053
* Skewness 0.806895
* Range 118
* Minimum 2
* Maximum 120



Inferences of Car Distance:

* The value of skewness is 0.806895 which is lying between 0.5 and +1, indicating moderately skewed.
* The value of kurtosis is 0.405053 which is greater than ZERO (Value > 0), indicating high degree of peakedness, resulting in Laptokurtic distribution.

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



Histogram inferences:

* X axis shows the chick weight
* Y axis shows the number of chicks in terms of frequency
* X axis interval = 100
* Number of classes varying the chick weight = 08
* Number of chicks surveyed (APPROXIMATELY) = 603
* Number of chicks having the weight more than 100 = 313



Box plot inferences:

* The distribution is positively skewed
* The distribution is positively skewed because box plot is showing the median closer to the lower or bottom quartile.
* Data constitutes higher frequency of high valued scores (Mean > Median)
* Outliers are present above the maximum value

**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?**

E =

* To find the z score, area to the curve left was determined

AL =

Then value of Z score was found from the Z-table for corresponding confidence interval

* Sample standard deviation = 30 pounds
* n = Sample size = 2000
* E = Error between sample mean and population mean

Then Confidence Interval was calculated by

C.I. = (Average weight – E, Average weight + E)

94% : z score 1.89

E 1.267850543

**C.I. 198.7321495 201.2678505**

96%: z score 2.06

E 1.38189001

**C.I. 198.61811 201.38189**

98%: z score 2.33

E 1.563011516

**C.I. 198.4369885 201.5630115**

**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** |
| Std.Deviation | | **5.05** |

**2)** **What can we say about the student marks?**

Chart, box and whisker chart

Description automatically generated

* Minimum marks scored is 34
* Maximum marks scored is 45 (as per the box plot)
* Average of the marks scored by the student is 41
* Quantile 1 (25%) is 38.25
* Quantile 2 (50%) is median given by the value 40.5
* Quantile 3 (75%) is 41.75
* Inter Quantile Range is 3.5
* Lower limit of the marks scored is 43.5
* Upper limit of the marks scored is 47
* Two student scores (49 and 56) above 47 are considered as outliers

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

Normal distribution or symmetric distribution

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

Positively Skewed distribution

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

Negatively Skewed distribution

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

* It indicates that the data distribution is peaked and possesses thick tails.
* An extreme positive kurtosis indicates a distribution where more of the numbers are in the tails of the distribution instead of around the mean.
* Leptokurtic distributions have positive kurtosis values

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

* It indicates that the data distribution is flat and has thin tails.
* Platykurtic distributions have negative kurtosis values

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



**What can we say about the distribution of the data?**

The distribution is negatively skewed

**What is nature of skewness of the data?**

The distribution is negatively skewed because the whisker and the half box are longer on the left side of the median than on the right side.

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

* The distribution is negatively skewed means data constitutes higher frequency of low valued scores (Mean < Median)
* Q2 – Q1 > Q3 – Q2

15 – 10 > 18 – 15 = 5 > 3

**Q19) Comment on the below Boxplot visualizations?**



Comments:

* Whiskers are same on both the sides of the plot, resulting in symmetric distribution or normal distribution
* The median values for both the boxplots are same i.e., 262.5 wbs
* Variation in the dispersion of data: Less in boxplot1 and more in boxplot2

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

Inferences:

boxplot1:

* Whiskers are same (250 – 237.5 = 12.5 wbs) OR (287.5 – 275 = 12.5 wbs) on both the sides. Therefore, distribution of the data is symmetric
* Whisker starts from 237.5 wbs and increases upto 287.5 wbs
* 25% Quartile (Q1) is 250 wbs
* 50% Quartile (Q2 or MEDIAN) is 262.5 wbs
* 75% Quartile (Q3) is 275 wbs
* IQR = Q3 – Q1 = 275 – 250 = 25

Boxplot 2:

* Whiskers are same (350 – 312.5 = 37.5 wbs) OR (225 – 187.5 = 37.5 wbs) on both the sides. Therefore, distribution of the data is symmetric
* Whisker starts from 187.5 wbs and increases upto 350 wbs
* 25% Quartile (Q1) is 225 wbs
* 50% Quartile (Q2 or MEDIAN) is 262.5 wbs
* 75% Quartile (Q3) is 312.5 wbs
* IQR = Q3 – Q1 = 312.5 – 225 = 87.5

**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)**
* 0.347593925
  1. **P(MPG<40)**
* 0.729349876
  1. **P (20<MPG<50)**
* 0.898868917

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

count 81.000000 mean 34.422076

std 9.131445

25% 27.856252 50% 35.152727 (Median)

75% 39.531633

min 12.101263 max 53.700681

Comments:

* Mean < Median
* 34.422076 < 35.152727 (But almost approximately equal)
* Distribution of data is more along the left side (Histogram) and whisker at the bottom is longer than the top (boxplot) (NOT MUCH VARIATION)
* Therefore, MPG of Cars is APPROXIMATELY SYMMETRICAL DISTRIBUTION

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

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

**Dataset: wc-at.csv**

Waist

count 109.000000 mean 91.901835

std 13.559116

25% 80.000000 50% 90.800000 (Median)

75% 104.000000

max 121.000000 min 63.500000

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Comments:

* Mean > Median
* 91.901835 > 90.800000 (But almost approximately equal)
* Both histogram and box plot of ‘Waist’ suggest that the distribution of the data is approximately symmetrical.
* Therefore, Waist of dataset is NORMAL DISTRIBUTION

Adipose Tissue (AT)

count 109.000000 mean 101.894037

std 57.294763

25% 50.880000 50% 96.540000

75% 137.000000

max 253.000000 min 11.440000

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Comments:

* Mean > Median
* 101.894037 > 96.540000
* Distribution of data is more along the right side (Histogram) and whisker line at the right side is longer than the left side (boxplot)
* Therefore, the Adipose Tissue (AT) of the dataset is POSTIVELY SKEWED DISTRIBUTION

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

90% z score 1.6448536

94% z score 1.880793

60% z score 0.841621

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

Sample size (n) = 25 Degree of freedom = n – 1 = 24

t scores of 95% C.I. = t0.025,24 = 2.064 (From t table)

t scores of 96% C.I. = t0.02,24 = 2.492

t scores of 99% C.I. = t0.005,24 = 2.797

**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**

t - statistics for the data is given as follows:



x = mean of the sample of bulbs =  260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

D.O.F = n – 1 = 18 – 1 = 17

t =

t = -0.471

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.