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

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

Three coins are tossed so

Possible outcomes = (HHH,HHT,HTH,THH,TTH,THT,HTT,TTT)

Favorable outcomes = (HHT,THH,HTH)

Probability = Favorable / Possible = **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**
4. **Zero**

When we roll two dice we will always get sum of number as atleast 2, so the probability of getting sum equal to 1 Is ZERO

1. The set of possible outcomes when we roll a die are {1, 2, 3, 4, 5, 6}

So, when we roll two dice there are 6 × 6 = 36 possibilities.

When we roll two dice, the possibility of getting number 4 is (1, 3), (2, 2), and (3,1).

So,

The number of favourable outcomes = 3

Total number of possibilities = 36

Probability = The number of favourable outcomes / Total number of possibilities = 3 / 36 = 1/12.

1. Total number of possible outcomes = 36

Sum of number divisible by both 2 and 3 i.e 6 and 12

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

Probability = Favorable / Possible outcomes

= 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?**

Total number of balls = (2 + 3 + 2) = 7  
n(S) = Number of ways of drawing 2 balls out of 7  
=7C2​  
=(7×6) ​/ (2×1)  
=21  
E = Event of drawing 2 balls, none of which is blue.  
n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.  
=5C2​  
= (5×4) / (2×1)​  
=10  
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)**

|  |  |  |
| --- | --- | --- |
| 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 | Candies count | Probability |  | | A | 1 | 0.015 | 0.015 | | B | 4 | 0.2 | 0.8 | | C | 3 | 0.65 | 1.95 | | D | 5 | 0.005 | 0.025 | | E | 6 | 0.01 | 0.06 | | F | 2 | 0.12 | 0.24 | |  |  |  |  | |  |  | Expected Value E(x) | 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**

**Points**

Mean =3.59656

Median = 3.695,

Mode = 3.92

Min = 2.76

Max = 4.93

Range = 2.17

Variance = 0.27

Std Deviation = 0.526

**Scores**

Mean = **3.2173**

Median = **3.325**

Mode = **3.44**

Min = 1.513

Max = 5.424

Range = 3.911

Variance = **0.927,**

Std Deviation =  **0.963**

**Weigh =**

Mean = **17.848,**

Median = **17.71,**

Mode = **17.02**

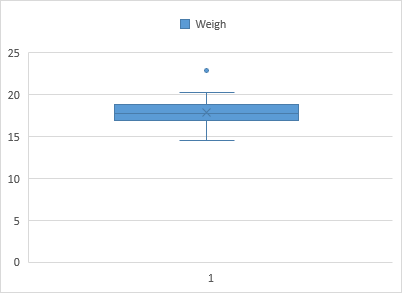
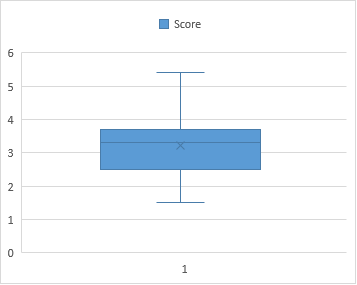
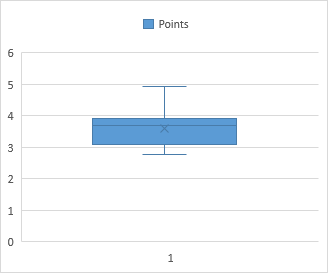
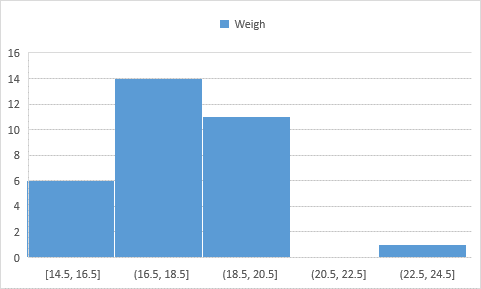
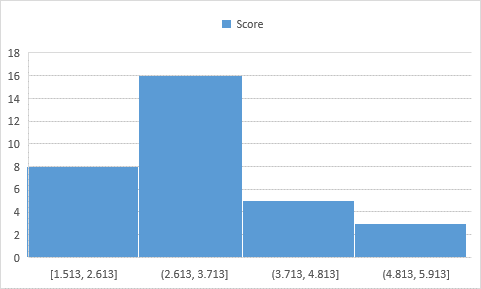
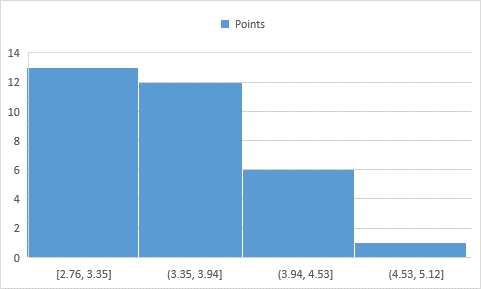
Min = 14.5

Max = 22.9

Range = 8.4

Variance = **3.093,**

Std Deviation =  **1.758**



**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?**

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**

Speed

|  |  |
| --- | --- |
| Mean | 15.4 |
| Median | 15 |
| Mode | 20 |
| Range | 21 |
| Std dev | 5.234501 |
| skew | -0.11751 |
| Kurtosis | -0.50899 |

Slightly left skewed

Distance

|  |  |
| --- | --- |
| Mean | 42.98 |
| Median | 36 |
| Mode | 26 |
| Range | 118 |
| Std dev | 25.51038 |
| skew | 0.8068 |
| Kurtosis | 0.405053 |

Right skewed data

SP and Weight(WT)

Use Q9\_b.csv

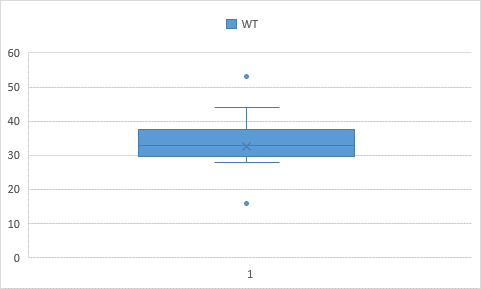
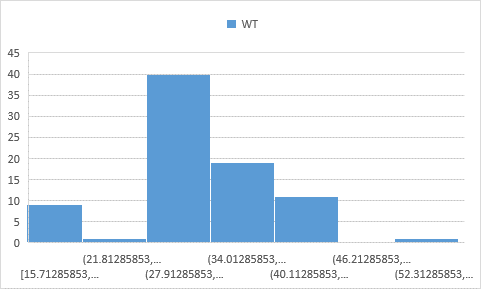
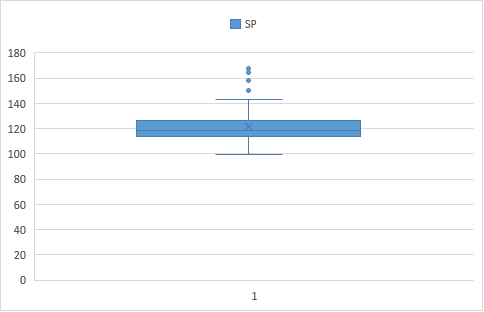
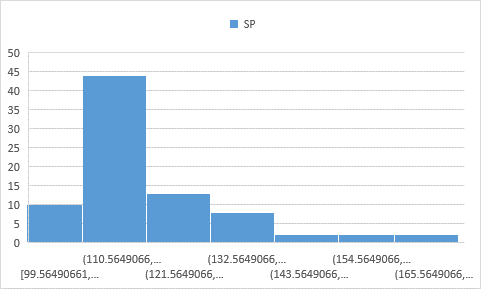
|  |  |  |
| --- | --- | --- |
|  | SP | Weight |
| Mean | 121.5403 | 32.41258 |
| Median | 118.2087 | 32.73452 |
| Max | 169.5985 | 52.99775 |
| Min | 99.56491 | 15.71286 |
| Range | 70.03361 | 37.28489 |
| Std dev | 14.09362 | 7.446417 |
| Skewness | 1.61145 | -0.61475 |
| Kurtosis | 2.977329 | 0.950291 |

SP data set is right skewed

SP data set contains outliers

SP data set also contains a sudden peaked Ness of data between 110 and 121

Weight data set is slightly left skewed



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



**HIstogram**

Right Skewed Data

The most common weight range is 50 to 100

The mean of the data is around 120

The median of data around 100

The mode of the data is around 90



The boxplot appears to be skewed to the left since q3 – q2 > q2 -q1

The box plot contains many outliers

The boxplot also contains a long whisker on the left indicating a left skew or negative skew in data

**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?**

Population = 3,000,000 Sample(n) = 2000

Average Sample (X---) =200

Std deviation Sample = 30

**For Confidence interval 94%**

1-α = 94% , α = 0.06/2 =0.03, Df = n-1 = 1999

T0.03 or 0.97 = 1.88

CI = 200 + t \* 30/√2000

Or 200 - t \* 30/√2000

CI = 200 – 1.2611 & 200 + 1.2611

= **( 198.73 , 201.261 )**

**For Confidence interval 96%**

1-α = 96% , df = n-1 = 1999

α = 0.04/2 =0.02

t0.02 or 0.98  = 2.171

CI = 200 +2.171 \* 30/√2000

Or 200 – 2.171\* 30/√2000

CI = 200 – 1.456 & 200 + 1.456

**= (198.54, 201.45)**

**For Confidence interval 98%**

1-α = 98% , df = n-1 = 1999

α = 0.02/2 =0.01

t0.01 or 0.99  = 2.32

CI = 200 +2.32 \* 30/√2000

(&) 200 – 2.32 \* 30/√2000

CI = 200 – 1.556 & 200 + 1.556

=**= (198.44, 201.55)**

**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?**

Q12 Soln

|  |  |
| --- | --- |
| **Mean** | 41 |
| **Median** | 40.5 |
| **Mode** | 41 |
| **Variance** | 24.11111 |
| **Std Deviation** | 4.910307 |

We can say that majority of the students scores lies around 41 , since we see mean=median=mode , we can say that the data is normally distributed

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

The data is normally distributed

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

Right skew

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

Left skew

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

There is a peaked Ness in data

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

The data is flattened across a range

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



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

Since the whisker is long on the left side we can say that data is left skewed and

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

Left skew

**What will be the IQR of the data (approximately)?**   
  
Q3 =18

Q1 = 10

So IQR = Q3- Q1 = 8

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



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

SOln 19 :-

In the plot 1

All the data are in a range of 250 to 290

The median of the data lies approximately at 262

In the Plot 2

All the data lie in the range between 225 to 310

The median of data lies approximately at 263

**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,loc = 34.42, scale = 9.074 )

P = 0.3465934767750194

* 1. **P(MPG<40)**

stats.norm.cdf(40,loc = 34.42, scale = 9.074 )

P = 0.730704080621096

* 1. **P (20<MPG<50)**

stats.norm.cdf(50,loc = 34.42, scale = 9.074 ) - stats.norm.cdf(20,loc = 34.42, scale = 9.074 )

P = 0.900

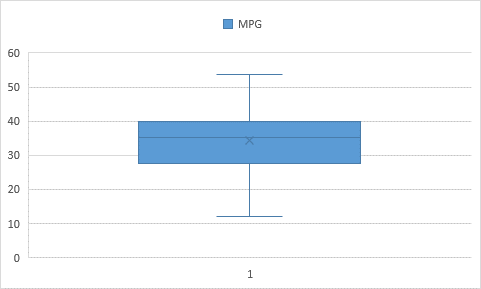
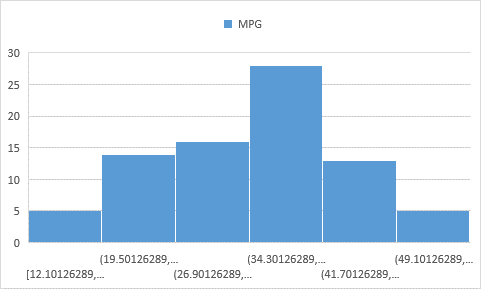
**Q 21) Check whether the data follows normal distribution**

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

**Dataset: Cars.csv**

|  |  |
| --- | --- |
| Mean | 34.42208 |
| Median | 35.15273 |
| Mode | 29.62994 |
| Range | 41.59942 |
| Std Dev | 9.074903 |
| skew | -0.17795 |

Since mean = median = mode is approximately same and skew is almost negligible, we can say that it is normally distributed



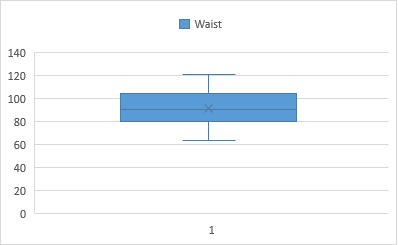
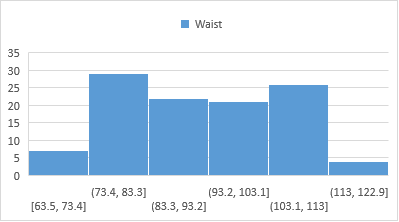
1. **Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution**

**Dataset: wc-at.csv**

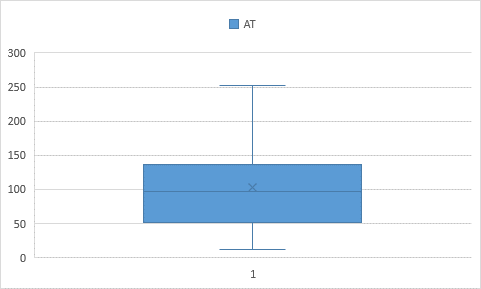
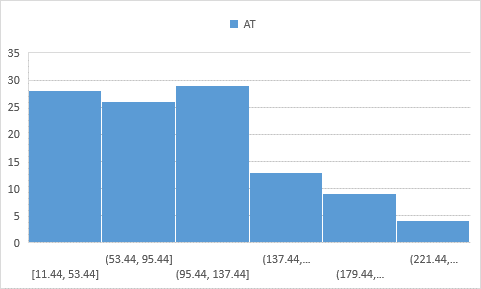
|  |  |  |
| --- | --- | --- |
|  | Waist | AT |
| Mean | 91.90183 | 101.894 |
| Median | 90.8 | 96.54 |
| Mode | 94.5 | 121 |
| Range | 57.5 | 241.56 |
| Std dev | 13.49677 | 57.03134 |
| Skew | 0.134056 | 0.584869 |
| Kurtosis | -1.10267 | -0.28558 |

Waist dataset is normally distributed, approximately mean = median = mode

The dataset is flattened since we have a negative kurtosis



The AT dataset is Right skewed and even this data is slightly flattened



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

For 90% confidence interval

1-α = 90%

α = 0.10/2 = 0.05 or 0.95

Z0.05 = μ + σ / √n

Z0.05 or 0.95 = 1.64

For 94% confidence interval

1-α = 94%

α = 0.06/2 = 0.03

Z0.03 or 0.97 = 1.88

For 60% confidence interval

1-α = 60%

α = 0.4/2 =0.2

Z0.2 or 0.8 = 0.84

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

For 95% confidence interval

1-α = 95% , df = n-1 = 24

α = 0.05/2 =0.025

t0.025 or 0.975  = 2.064

For 96% confidence interval

1-α = 96% , df = n-1 = 24

α = 0.04/2 =0.02

t0.02 or 0.98  = 2.171

For 99% confidence interval

1-α = 99% , df = n-1 = 24

α = 0.01/2 =0.005

t0.02 or 0.995  = 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

Sol

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18









t = - 0.471

df = n-1 = 18-1=17

P from T table for t= -0.471 is **0.3218**

P = 0.3218