|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete(Countable) |
| 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(countable) |
| Number of tickets in Indian railways | Discrete(Countable) |
| Number of times married | Discrete(Countable) |
| Gender (Male or Female) | Discrete(categorical) |

**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) | Ordinal |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Nominal |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Ordinal |
| SAT Scores | Interval |
| Years of Education | Nominal |

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

Sample space of the Event,

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

Number of observations with two heads and one tail = **3**

Total number of observations = **8**

Probability(Two heads and one tail) = Favourable Outcomes / Total Outcomes

= **3/8**

= **0.375**

Therefore, the probability of obtaining two heads and one tail when 3 coins are tossed is **0.375**.

**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**

Sample space, **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.Equal to 1:**

Observations with sum is equal to 1 = 0

Total Observations = 36

Probability ( Sum is equal to 1) = **0/36** = 0

**2.Less than or equal to 4:**

Observations with sum less than or equal to 4 = 6

Total No: of Observations = 36

Probability ( Sum less than or equal to 4 ) **= 6/36**

**= 0.166**

Probability of sum less than or equal to **4** is **0.166**

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

Observations with Sum is divisible by **2** and **3** = **24**

Total Number of Observations = 36

Probability ( Sum is divisible by 2 and 3 ) = 24/36

**= 0.66**

**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?**

Probability of drawing two balls with none of the blue balls =

= Total probability - Probability of drawing two balls that are blue balls

= 1 - [ (2/7) + (1/6) ]

= 1 - [ 0.285 + 0.166 ]

= 1 - 0.451

= **0.549**

**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**

Expected No: of candies for randomly selected child, E(X) = P(x) \* n

= 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

= 3.090

**= 3.09**

Expected number of candies for a randomly selected child is **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 Data**

**Mean** = Sum of all observations / total no: of observations **= 3.5965625**

**Median** = mid-point **= 3.695**

**Mode** = Highest No: of Observation = **3.92**

**Variance** = **s2** = ∑ni=1(xi−μ)/ 2n−1 **= 0.28588**

**Standard Deviation** = 1/n∑ni=1(xi−¯x)2 = **0.5346**

**Range** = min, max = [2.76, 4.93]

Inference--> The mean is less than median which is less than mode hence,t distribution is negatively skewed.

**- Score Data**

**Mean** = Sum of all observations / total no: of observations = **3.2172**

**Median** = mid-point = **3.325**

**Mode** = Highest No: of Observation = **3.44**

**Variance** = **s2** = ∑ni=1(xi−μ)/ 2n−1 = **0.9573**

**Standard Deviation** = 1/n∑ni=1(xi−¯x)2 **= 0.9784**

**Range** = min, max = [5.424, 1.513]

Inference -> Here too, mean<media<mode = negatively skewed.

**- Weight Data**

**Mean** = Sum of all observations / total no: of observations **= 17.84875**

**Median** = mid-point **= 17.71**

**Mode** = Highest No: of Observation =**17.02**

**Variance** = **s2** = ∑ni=1(xi−μ)/ 2n−1 **= 3.1931**

**Standard Deviation** = 1/n∑ni=1(xi−¯x)2 = **1.78694**

**Range** = min, max = [14.5, 22.9]

Inference -> Here mean>median.>mode therefore, the distribution is positively skewed.

**Q8) Calculate Expected Value for the problem below**

* **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?**

The expected values = ∑ P(x).E(x)

Here P(x) is 1/9 for all observations.

Therefore, Expected Values =

(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 randomly chosen patient is **145.33**

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

**Cars speed and distance**

1. **Skewness** = ∑(Xi - x̅ )3 /(N-1) sd3
2. **Speed**= -0.1175
3. **Distance** = 0.8068

**Inference**: Speed is negatively skewed and distance is positively skewed. So, when we draw the graph we will could see the distribution of data in speed is inclined towards right. While in distance the distribution is inclined towards left side.

1. **Kurtosis** = n x Σni(Yi – Ȳ)4 / (Σni(Yi – Ȳ)2)2
2. **Speed** = -0.50899442
3. **Distance** = -0.405052582

**Inference**: In both cases of speed an distance, there is a negative value. Hence we can say that it is platykurti and the distribution is data is even.

**SP and Weight(WT)**

**Use Q9\_b.csv**

1. **Skewness** = ∑(Xi - x̅ )3 /(N-1) sd3
2. **SP** = 1.6114
3. **Weight** = -0.614753326

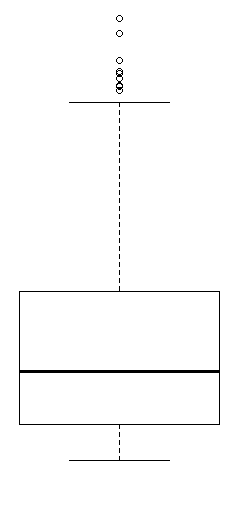
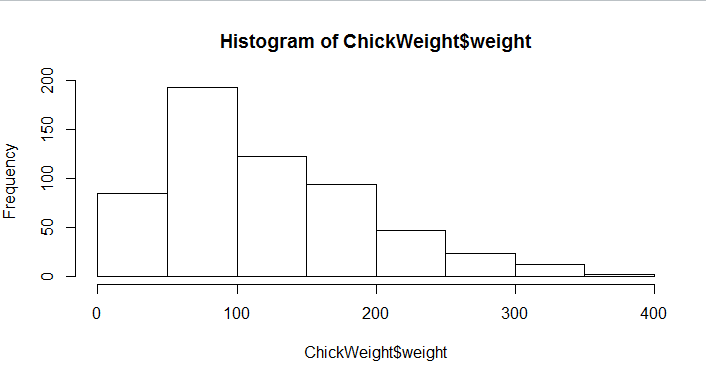
**Inference:** we can see that Skewness of SP is is positive and hence the distribution is inclined towards left side and for the weight it is negatively skewed and hence the distribution of data is inclined towards right side of the graph.

1. **Kurtosis** = n x Σni(Yi – Ȳ)4 / (Σni(Yi – Ȳ)2)2
2. **SP** = 2.9773
3. **Weight** = 0.9502

**Inference:** as for SP the kurtosis is greater that than zero and hence it is leptokurtic therefore, the distribution is concentrated in a region while for weight the kurtosis is near to zero hence, it is mesokurtic an the distribution is normal.

**Inference:**

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



**Inference**: from the box-plot graph we ca understand that our distribution has some outliers and from the histogram graph we understand that 200 chicks are of weight 50-100 which is the most occurring meaning, most of the chicks fall under 50-100 weight category.

**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?**

Here the parameters are as follows:

**N** = 3,000,000

**n** = 2000

**Mean** =200

**SD** = 30

We can calculate confidence intervals with the formula **= x̅ ± Z(σ / √n)**

1. For 94% confidence interval: 200 ± 1.881(30/√2000) = 200 ± 1.2681
2. For 98% confidence interval: 200 ± 2.326(30/√2000) = 200 ± 1.5603
3. For 96% confidence interval: 200 ± 2.05(30/√2000) = 200 ± 1.3751

**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**

* **Find mean, median, variance, standard deviation.**
* **What can we say about the student marks?**

**Mean** = ΣXi / n = 661 / 18 = 36.72

**Median** = n/2 = 18/2 = 9th term = 40

**Mode** = 41

From the above numbers: mean < median < mode --> the distribution is negatively skewed.

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

When the mean, median and mode are equal then the distribution is said to be

symmetric in nature and is normally distributed.

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

When mean > median then it is positively skewed.

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

When median > mean then it is negatively skewed.

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

Positive kurtosis indicates that the distribution is concentrated at a point and we can

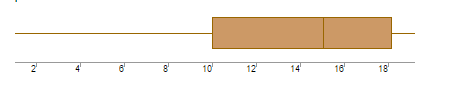
find the peak.

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

Negative kurtosis indicates that the distribution is not concentrated at any point

and is uniformly distributed.

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



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

By drawing conclusions from the above box-plot, we can assume that data is concentrated more on the left side of the scale.

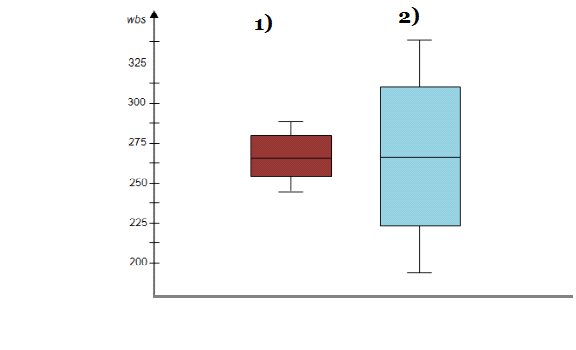
**What is nature of skewness of the data?**

According to the inference drawn from the box-plot, we can assume that data is negatively skewed.

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

**IQR** = Q3 - Q1 = 16 - 12 = 4.

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



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

1. box-plot is more concentrated and while **(2)** box-plot has a more distributed data. When we talk about range of both box-plots **(1)** has the range of **237.5 - 287.5 = 50** and **(2)**  has the range of 200 - 350 = 150 and hence **(2)** has more range. The kurtosis value of **(1)** will be greater than **(2)** since **(1)** is more concentrated.

Another inference is that the median values of both the box-plots are equal.

**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)**

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

a) P(MPG>38) = 55 / 81 = 0.679 **≈ 0.68**

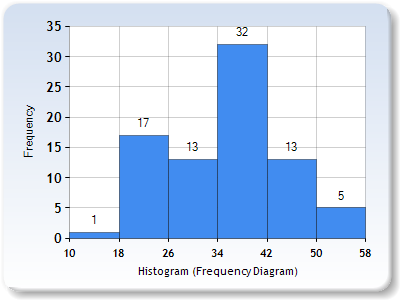
b) P(MPG<40) = 61 / 81 = 0.753 **≈ 0.75**

c) P(20<MPG<50) = 69 / 81 = 0.851 **≈ 0.85**

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

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

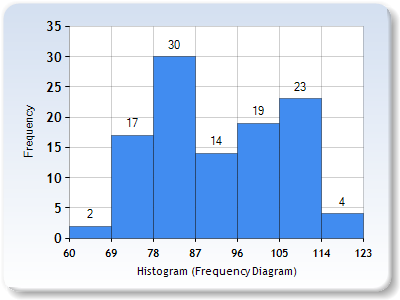
**Dataset: Cars.csv**



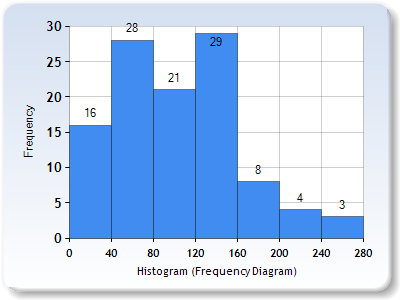
From histogram we get a complete picture that the distribution is normal

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

**Dataset: wc-at.csv**



From the above analysis we understand that Waist data is not normally distributed.



From the above code we can understand that AT data is not normally distributed.

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

from scipy import stats

from scipy.stats import norm

#z-score of 90% confidence interval

stats.norm.ppf(0.95)

#z-score of 94% confidence interval

stats.norm.ppf(.97)

#z-score of 60% confidence interval

stats.norm.ppf(-60)

**90%= 1.645**

**94%= 1.880**

**60%= 0.253**

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

|  |  |
| --- | --- |
| **Confidence Interval** | **t-Scores** |
| 95% | 2.262 |
| 96% | 2.054 |
| 99% |  |

**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

**t** = - 0.471

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