|  |  |
| --- | --- |
| 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 | CONTINUOUS(wavelength),Discrete(number of colors) |
| 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 | NOMINAL |
| Level of Agreement | ORDINAL |
| IQ(Intelligence Scale) | ORDINAL |
| Sales Figures | RATIO |
| Blood Group | NOMINAL |
| Time Of Day | RATIO |
| Time on a Clock with Hands | INTERVAL |
| Number of Children | NOMINAL |
| 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?

Ans:When 3 coins are tossed, the possible outcomes are HHH, TTT, HTT, THT, TTH, THH, HTH, HHT.

The sample space is S = { HHH, TTT, HTT, THT, TTH, THH, HTH, HHT}

Number of elements in sample space, n(S) = 8

Let E1 be the probability of getting two heads and one tail={HHT,HTH,THH}

So P(getting two heads and one tail)=3/8=>0.375

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

* Equal to 1
* Less than or equal to 4
* Sum is divisible by 2 and 3

Ans:When two dice are rolled the possible outcomes are

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (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) |

a)The probability that sum is equal to 1 is zero

b) The probability that sum is less than or equal to 4

outcomes less than or equal to 4

=

total no of outcomes

6/36=>1/6

P (getting sum <=4)=1/6

c) The probability that sum is divisible by 2 and 3 is 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 no of chances to draw ***2*** balls at random from ***7*** coloured balls,**n(s)=7C2=21**  
let **E** be an event to draw ***2*** balls other than **blue** .  
no of chances to draw two balls other than blue are,**n(E)=2C2+3C2+2C1\*3C1=1+3+6=10**  
the probability that none of balls drawn is blue is,**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

Expected Value  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

Expected number of candies for a randomly selected child

=  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

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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Points** | **Score** | **Weigh** |
| Mean | 3.597 | 3.217 | 17.85 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.2858814 | 0.957379 | 3.193166 |
| Standard Deviation | 0.526258 | 0.963048 | 1.758801 |
| Range | [2.76, 4.93] | [1.513, 5.424] | [14.5, 20.22] |

Inferences :

* Points has the Least Variance, Standard Deviation, and Range
* Points & Score data - Negative skewness, Weigh has Positive Skewness

**Use Q7.csv file**

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?

Ans:

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**

**Skewness**

**speed -0.117510 Fairly Symmetrical**

**dist 0.806895 Moderately Skewed**

**Kurtosis**

**speed -0.508994 Platykurtic distribution**

**dist 0.405053 leptokurtic distribution**

**Ans:**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Skewness**

**SP 1.611450 -> Highly Skewed**

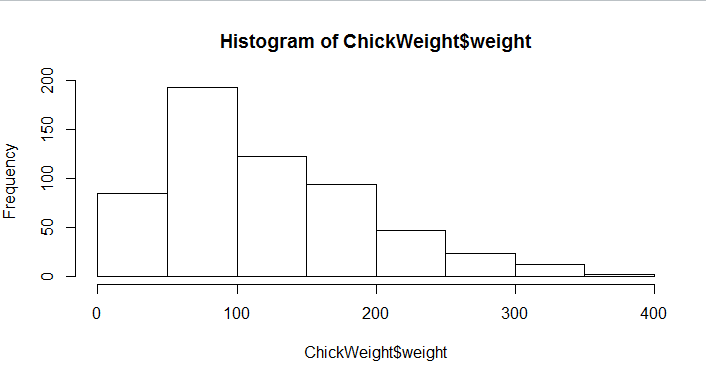
**WT -0.614753 -> Moderately Skewed**

**Kurtosis**

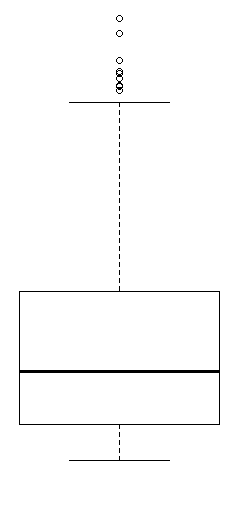
**SP 2.977329 -> leptokurtic distribution**

**WT 0.950291 -> leptokurtic distribution**

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



* Data is Positively Skewed
* Data is present on the Left side and has a long tail on the right side
* Mode of the data is 100 with 200freq
* Data Follows Unimodal and approximately normal distribution



Ans: The above boxplot has a long tail on the upper quartile

It has 7 outliers

Median of the data lies close to the 1st quartile

**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: 94% CI= (143.54415570565965, 256.45584429434035)

96% CI= (138.34730111522666, 261.6526988847733)

98% CI= (130.15355671679083, 269.84644328320917)

**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.

Ans: mean 41.000000

std 5.052664

Median 40.5

Var 25.529411

* What can we say about the student marks?
* Range is 56 to 34
* Most of the students scored 41

Ans: The above data clearly follows normal distribution since it can be represented on the graph and we would see a bell shaped curve.

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

If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.

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

Ans : It is Positively Skewed since β>0

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

Ans : It is Negatively Skewed since β<0

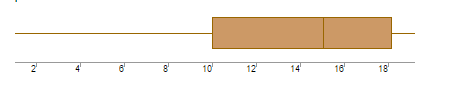
Q16) What does positive kurtosis value indicates for a data ?

Ans : **Positive** values of **kurtosis** indicate that a distribution is peaked and possess thick tails

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

Ans : **Positive** values of **kurtosis** indicate that a distribution is peaked and possess thick tails

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



What can we say about the distribution of the data?

* The above box plot has no outliers
* It has a very long tail on the left side
* Data is present on the Right side

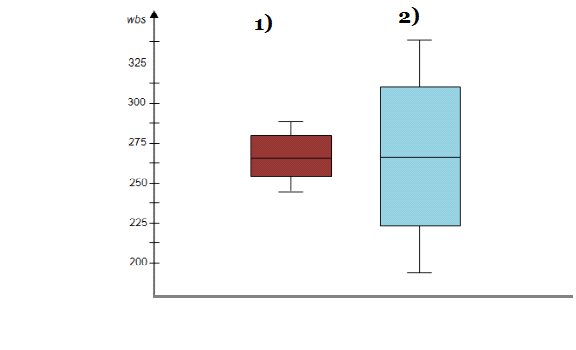
What is nature of skewness of the data?

* We can say that the above box plot is negatively skewed since the median value lies closer to the upper quartile

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

* IQR can be calculated using the formula Q3-Q1
* ->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.

* **Both the box plots have the same median value** **lie at same point (262.5)**
* **Both Plots have equal distribution of data above and below Median which means plots have Normal Distribution, to be specific data is symmetric about mean value.**
* **Kurtosis is negative in 1st case as it has thinner tails**

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

* + P(MPG>38)
  + P(MPG<40)
    - P (20<MPG<50)

Ans: 1.Probability that MPG>38 is 0.3475939251582705

Python code =1-stats.norm.cdf(38,loc=mpg.mean(),scale=mpg.std())

2.Probability that MPG<40 is 0.7293498762151616

Python code = stats.norm.cdf(40,loc=mpg.mean(),scale=mpg.std())

3.Probability that 20 < MPG <50 is

Probability of mpg<50 – Probabilty of mpg <20 =0.898868916968204

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: **MPG of Cars follow Normal Distribution**

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

Dataset: wc-at.csv

Ans:It does not follow normal distribution WC column and AT

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

Ans: 90% = 1.6448536269514722

94 %= 1.8807936081512509

60% = 0.8416212335729143

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

|  |  |
| --- | --- |
| 95% | 2.0638985616280205 |
| 96% | 2.1715446760080677 |
| 99% | 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

Ans: The t statisatic value for the above problem is

populationmean=270

samplemean=260

samplestd=90

sample=18

t=(samplemean-populationmean)/(samplestd/np.sqrt(sample))

t value is

-0.4714045207910317

stats.t.cdf(- 0.471,df=17)

0.3218140331685075

The probability that on an avg bulb lasts less than 260 days is 0.3218140331