|  |  |
| --- | --- |
| 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 |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Continuous |

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

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

Ans:- the Total combinations are, HHH,HHT,HTH,HTT,THH,THT,TTH,TTT

These are total 8 combinations. The two head and one tail combinations are 3. So probability, P=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

Ans:- There are total 6\*6=36 pairs.

1. Equal to 1

There are no combinations with sum equals to 1 so probability,

P=0/36

P=0

1. Less than or equal to 4

Combinations: (1,1),(1,2),(1,3),(2,1),(2,2),(3,1)

Total comb.=6

So probability,P=6/36=1/6

1. Sum is divisible by 2 & 3 both

Combinations: (1,5),(2,4),(3,3),(4,2),(5,1),(6,6) =6 combinations

So probability,P=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?

Ans:- Two times ball is drawn, so 1st time it would be total 7 outcomes and 2nd time it would be 6 outcomes so total outcomes=7\*6=42.

And total non-blue balls =2+3=5 for 1st time and 4 for 2nd time

So total non-blue balls outcome=5\*4=20

So probability of none of balls are blue is:

P=20/42=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

Ans:- Expected number of candies for a randomly selected child

=summation(candies count \* probability)

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

Ans:-

For code see the Assignment1.ipynb file.

1. Mean Values:

Points: 3.5965625000000006

Score: 3.2172499999999995

Weigh: 17.848750000000003

Mean values are the average value of that particular column.

1. Median Values:

Points: 3.6950000000000003

Score: 3.325

Weigh: 17.71

Median values are the middle value of that particular column when arranged.

1. Mode Values:

Points: 3.07 & 3.92

Score: 3.44

Weigh: 17.02 & 18.90

Mode values are the most repeated value of that particular column.

1. Variance Values:

Points: 0.28588135080645166

Score: 0.9573789677419356

Weigh: 3.193166129032258

Variance is the average square difference of datapoints (values) from the

mean.

1. Standard deviation:

Points: 0.5346787360709716

Score: 0.9784574429896967

Weigh: 1.7869432360968431

Standard deviation is square root of variance.

1. Range Values:

Points:(2.76, 4.93)

Score:(1.513, 5.424)

Weigh:(14.5, 22.9)

Range gives max and min value of that data set (Column).

So we can say Points contains lowest Variance and Std values.

**For code see the Basic statistics 1.ipynb file.**

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?

Ans:- No. of total patients=9

So probability of randomly chosen 1 patient is=1/9

So expected value of the weight of the patient

=

=145.33

**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**

**Ans:- a)**

Skewness:

speed -0.117510

dist 0.806895

Kurtosis:

speed -0.508994

dist 0.405053

b)

Skewness:

SP 1.611450

WT -0.614753

Kurtosis:

SP 2.977329

WT 0.950291

**For code see the Basic statistics 1.ipynb file.**

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



Ans:- 1) Histogram:-

* The histogram is right skewed (Positively Skewed), meaning the more data lies in the left part of the graph.
* ChickWeight 100 contains the highest number of data points.
* This data contains only one mode value that is 100.

2) Boxplot:-

- The density of the points are downwards so its most data points are in left side.

- The dots shows the outliers which are on the positive side of the data(Right side).

- As most of the data on the left side so it is positively skewed (Right skewed).

**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:- N=3000000

n=2000

sample mean=200

Std= 30 pounds

t94= 1.8818

t98= 2.3282

t96= 2.0550

For sample,

Confidence interval=mean ± t \* (std/√n)

1. For 94% Confidance interval:

=200-1.882\*(30/√2000) to 200+1.882\*(30/√2000)

=198.73 to 201.26

1. For 98% confidence interval:

=200-2.328\*(30/√2000) to 200+2.328\*(30/√2000)

=198.43 to 201.56

1. For 96% Confidence interval:

=200-2.055\*(30/√2000) to 200+2.055\*(30/√2000)

=198.62 to 201.378

**For code please see the Basic statistics 1.ipynb file.**

**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?

Ans:- 1) Mean of marks:41.0

Median of marks:40.5

Variance of marks: 25.529412

Standard deviation of marks: 5.052664

**For code please see the Basic statistics 1.ipynb file.**

2)

a) Highest and lowest marks are 56 & 34.

b) the average marks obtained by students are 41.

c) mean and median are almost equal so data is almost normally distributed.

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

Ans: if mean and median are equal then the data is normally distributed means skewness would be zero.

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

Ans: If mean> median then data would be positively skewed or right skewed.

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

Ans: If median > mean then data would be negatively skewed or left skewed.

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

Ans: Positive kurtosis meaning data contains higher peaks means may contain high amount of outliers because of extreme values.

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

Ans: Negative kurtosis meaning data contains flat peaks or maybe no peaks at all. Means may contain less amount of the outliers than normally distributed.

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)?

Ans: a) Data is on higher side so more data distributed on left side means less data is on the right side.

1. More data is on left side so its Positively skewed or we can say right skewed.
2. IQR means inter quartile range means range contains middle 50% of the data.

There are 4 Quartiles.Q1=25% data

Q3=75% of data so finding 50% data

IQR=Q3-Q1

Here Q1 = 10.2 and Q3= 18.3 approximately

So IQR=18.3-10.2

IQR=8.1

Q19) Comment on the below Boxplot visualizations?



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

Ans:

a) both data contains same median i.e. approx. 262.

b) both the boxplots looks equally distributed means median is almost in center.

c) 1st data is denser than 2nd data means 1st one is less distributes(negative kurtosis) and 2nd one is more distributed (positive kurtosis).

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)

Ans: P(MPG>38) 0.3475939251582705

P(MPG<40) 0.7293498762151616

P(20<MPG<50) 0.8988689169682046

**For code please see the Basic statistics 1.ipynb file.**

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

Ans:

a) as we can see here data skewness is -0.1779, and for normally distributed data skewness should not exceed -0.5 to 0.5, so here data is normally distributed.

b) here for waist skewness is 0.1340 means its normally distributed

and for AT its skewness is 0.5848 means its slightly positively skewed or we can say almost normally distributed.

**For code please see the Basic statistics 1.ipynb file**.

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

Ans:

Z90= 1.2815

Z94= 1.5547

Z60= 0.2533

**For code please see the Basic statistics 1.ipynb file.**

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

Ans:

T95=1.7108

T96= 1.8280

T99=2.4921

**For code please see the Basic statistics 1.ipynb file.**

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

Ans:-

n=18

std=90 days

population mean = 270

sample mean= 260

T value=

So T value=-0.4714

So probability for bulb having life not more than 260 days = 32.167%

**For code see the Basic statistics 1.ipynb file.**