|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Discrete data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete data |

**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 | Nominal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Ratio |
| Socioeconomic Status | Interval |
| Fahrenheit Temperature | Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Interval |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Ratio |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | nominal |
| Religious Preference | Ratio |
| 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?**

Sol: P(H H T) + P(H T H) + P(T H H)

=1/8 +1/8+1/8

= 3/8

Or (Alternative approach)

Binomial distribution.

**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**

**a) Equal to 1:**

There is no outcomes which corresponds sum is equal to one. i.e.0/36. Probability is 0

**b) Less than or equal to 4:**

N (Event (Sum is less than or equal to 4)) / N (Event (Two dice rolled))

= 6 / 36 = 1/6 = 0.166 = 16.66%

**c) Sum is divisible by 2 and 3:**

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?**

Sol:10/21P (2R, 3G, 2B)

P (5/7, 4/6) = 20/42 I.e. 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**

**E = Σ (candy count \* probability)**

**Given the probabilities and candy counts for each child:**

**Child A: Candy count = 1, Probability = 0.015**

**Child B: Candy count = 4, Probability = 0.20**

**Child C: Candy count = 3, Probability = 0.65**

**Child D: Candy count = 5, Probability = 0.005**

**Child E: Candy count = 6, Probability = 0.01**

**Child F: Candy count = 2, Probability = 0.120**

Sol:Expected number of candies (E) = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 4.125

The expected number of candies for a randomly selected child is 4.125.

**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**

Sol:

**Points:**

**1.Mean:**

Mean Points = (Sum of Points) / (Number of Data Points)

118.83 / 32 = 3.7059375

**2.Median:**

Ascending order:

2.76, 2.76, 2.93, 3.08, 3.08, 3.15, 3.15, 3.21, 3.54, 3.62, 3.69,

3.7, 3.73, 3.77, 3.9, 3.9, 3.92, 3.92, 3.92, 4.08, 4.08, 4.11,

4.22, 4.22, 4.43, 4.93

Median Points = (3.69 + 3.7) / 2 = 3.695

**3.Mode:**

Mode Points = 3.92

**4.Variance:**

Squared differences from the mean:

(2.76 - 3.7059375)^2, (2.76 - 3.7059375)^2, ...

Sum of squared differences = 3.21949359375

Variance Points = Sum of squared differences / (Number of Data Points- 1) =

0.1059843729

**5.Standard Deviation:**

Standard deviation is the square root of the variance.

Standard Deviation Points = √(0.1059843729) = 0.3256120679

**6.Range:**

Minimum Points = 2.76

Maximum Points = 4.93

Range Points = 4.93 - 2.76 = 2.17

**Score:**

**1.Mean:**

Sum of Scores = 103.08

Number of Data Points = 32

Mean Score = 103.08 / 32 = 3.22125

**3.Median :**

Arrange the Scores in ascending order:

1.513, 1.615, 1.935, 2.14, 2.2, 2.32, 2.465, 2.62, 2.77, 2.78, 2.875, 3.15, 3.215,

3.44, 3.435 3.46, 3.52, 3.57, 3.57, 3.78, 3.815, 3.84, 3.845, 4.07, 5.25,5.345,5.424

Median Score = (3.57 + 3.57) / 2 = 3.57

**3.Mode:**

Mode Score = 3.44 (appears twice)

**4.Variance:**

Sum of squared differences = 7.7856975

Variance Score = Sum of squared differences / 31 = 0.250178145

**5.Standard Deviation:**

Standard Deviation Score = √(0.250178145) = 0.500178479

**6.Range:**

Minimum Score = 1.513

Maximum Score = 5.424

Range Score = 5.424 - 1.513 = 3.911

**Weight:**

**1.Mean:**

Sum of Weights = 571.16

Number of Data Points = 32

Mean Weight = 571.16 / 32 = 17.860

**2.Median :**

14.5, 15.41, 15.5, 15.84, 16.46, 16.7, 16.87, 16.9, 17.02, 17.02, 17.05, 17.3, 17.4,

17.42, 17.44, 17.6, 17.82, 17.98, 18.3, 18.6, 18.61, 18.9, 19.44, 19.47, 19.9, 20,

20.01,20.22, 20.22, 22.9

Median Weight = (17.42 + 17.44) / 2 = 17.43

**3.Mode:**

Mode Weight = No mode (no repeated value)

**4.Variance:**

Sum of squared differences = 183.6745

Variance Weight = Sum of squared differences / 31 = 5.9243927419

**5.Standard Deviation:**

Standard Deviation Weight = √(5.9243927419) = 2.4335948659

**6.Range:**

Minimum Weight = 14.5

Maximum Weight = 22.9

Range Weight = 22.9 - 14.5 = 8.4

**Comments and Inferences:**

* In the "Points" column, the mean (3.7059) is roughly equal to the median (3.695), suggesting a relatively symmetric distribution.
* In the "Score" column, the mean (3.22125) and median (3.57) are slightly different, which might indicate some skewness.
* The "Weight" column has a wider spread of values, as indicated by its larger range (8.4) compared to the other columns.
* The standard deviation values show how much the data points deviate from the mean. The "Score" column has the highest standard deviation, indicating more variability.
* .The absence of a mode in the "Weight" column indicates that no weight value appears more than once.
* The variance and standard deviation values for "Weight" (5.92 and 2.43, respectively) are relatively larger, suggesting higher variability compared to the other columns. This could be due to the larger range of weights in the dataset.
* Overall, the "Points" and "Score" columns seem to have similar central tendencies, while

**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?**

Sol:

Weights (X) of patients: 108, 110, 123, 134, 135, 145, 167, 187, 199

Number of patients (n): 9

Expected Value (E[X]) = (Sum of (X \* Probability) for all values of X)

E[X] = (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)

E[X] = 12 + 12.22 + 13.67 + 14.89 + 15 + 16.11 + 18.56 + 20.78 + 22.11

E[X] = 145.34

So, the expected value of the weight of the randomly chosen patient is approximately 145.34 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

Sol:

**• Skewness :**

• Index 0.000000

• speed -0.117510

• dist 0.806895

**• Kurtosis:**

• Index -1.200000

• speed -0.508994

• dist 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

Sol:

**• Skewness:**

• Unnamed: 0 0.000000

• SP 1.611450

• WT -0.614753

**• Kurtosis:**

• Unnamed: 0 -1.200000

• SP 2.977329

• WT 0.950291

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

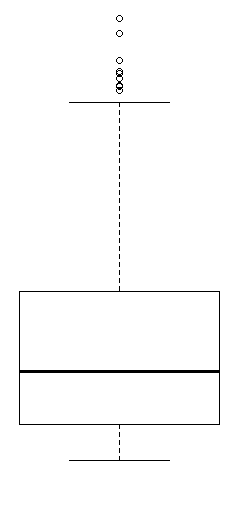


Sol:

• The peak is in between 50-100.

• The least weight are in between 350-400.

• It represents that the data is right skewed



Sol:

• The box plot is right skewed (Positively)with having outliers.

• The median and mean are not equal and median lies towards the Q1.

**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?**

**Sol:** X+/-(Z1-α.σ/sqrt(n)

Degrees of freedom= 2000-1= 1999

Confidence interval= 94%

(1-σ/2)= 1-0.03) =0.97

for confidene interval for 94% is 1.882

Confidence interval for 98%= 2.33

Confidence interval for 96% = 2.05

**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?

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

**Sol:**Symetrical

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

**Sol:**Right skewed

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

**Sol:**Left skewed

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

**Sol:** The data is notmally distributed and kurtosis value is 0.

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

**Sol:** The distribution of the data has lighter tails and a flatter peaks than the normaldistribution.

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



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

**Sol:** Let’s assume above box plot is about age’s of the students in a school.

50% of the people are above 10 yrs old and remainig are less.

And students who’s age is above 15 are approx 40%

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

**Sol:** Left skewed, median is greater than mean.

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

**Sol:** Approximately= -8

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



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

Sol:

**•** Both the box plot median lies on the same place close to 263

• The data is Symmetrically distributed around this central value as the mean and median is 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)**

Sol:

a. (df.MPG > 38).sum()

The no.of favourable outcome are 33

The no.of Total outcome are 81

So. P(MPG>38) is 0.407

b. (df.MPG < 40).sum()

The no.of favourable outcome are 61

The no.of Total outcome are 81

So. P(MPG<40) is 0.753

c. ((20<df.MPG) & (df.MPG<50)).sum()

The no.of favourable outcome are 69

The no.of Total outcome are 81

So. P(MPG>38) is 0.851

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

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

**Dataset: Cars.csv**

Sol:

The data is Normally distributed. As Mean is slightly near to median.

Mean = 34.422

Median = 35.152

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

**Dataset: wc-at.csv**

Sol:

**•** In Adipose Tissue the data is not Normally Distributed. As Mean > Median .

Mean = 101.89

Median = 96.54

• In Waist Circumference the data is Normally Distributed. As Mean is slightly near to median.

Mean = 91.90

Median = 90.8

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

Sol:

•AL = 1+CL/2 = 1+0.90/2 = 0.95

Z-score = 2.1

• AL = 1+CL/2 = 1+0.94/2 = 0.97

Z-score = 2.7

• AL = 1+CL/2 = 1+0.60/2 = 0.80

Z-score = 1.3

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

**Sol:**To compute the 95% confidence interval, start by computing the mean and standard error: M = (2 + 3 + 5 + 6 + 9)/5 = 5. σM = = 1.118. Z.95 can be found using the normal distribution calculator and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff point.

Confidence Level z

0.90 1.645

0.92 1.75

0.95 1.96

0.96 2.05

With a 90 percent confidence interval, you have a 10 percent chance of being wrong. A 99 percent confidence interval would be wider than a 95 percent confidence interval (for example, plus or minus 4.5 percent instead of 3.5 percent).

**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:

sample\_mean = 260

sample\_size = 18

population\_mean = 270

sample\_std\_dev = 90

Source Code:

import scipy.stats as stats

standard\_error = sample\_std\_dev / (sample\_size\*\*0.5)

t\_score = (sample\_mean - population\_mean) / standard\_errordf = sample\_size - 1

probability = stats.t.cdf(t\_score, df)

Probability is 0.321 or 32.1%

Probability that 18 bulbs would have an average life of no more than 260 days is 32.1% only