**Q1) Identify the Data type for the Following**:

|  |  |
| --- | --- |
| 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 |

**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) | Ratio |
| Sales Figures | Ratio |
| 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 | Ratio |
| Years of Education | Ratio |

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

* Total number of events= {HHH, HHT, HTT, TTT, TTH, THH, HTH, THT} =8

events=3

1. Probability that two heads and one tail are obtained

={ HHT , HTH , THH }

= **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

* Total number of outcomes when two dice are rolled=6\*6=36.

{(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. sum is Equal to 1

events =0

Probability that sum is Equal to 1

P(X) =**0**

1. Less than or equal to 4

Events = {(1,3) (1,2) (1,1) (2,1) (2,2) (3,1)}

= **6**

Probability that Less than or equal to 4

P(X) = 6/36 = **1/6**

c) sum is divisible by 2 and 3

{(1,1) (1,2) (1,3) (1,5) (2,1) (2,2) (2,4) (2,6) (3,1) (3,3) (3,5) (3,6)

(4,2) (4,4) (4,6) (5,1) (5,2) (5,3) (5,4) (5,5) (6,2) (6,3) (6,4) (6,6)}

Events = 24

Probability sum is divisible by 2 and 3

P(X) = 24/36 = **0.67**

**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 events

= =

=**21**

* events==10
* Probability that none of the balls is blue

=10/21

=**0.47**

**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 number

= E(x)

= (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.

**Use Q7.csv file**

* **stat.desc(df) ##pastecs library**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| nbr.val | 32 | 32 | 32 |
| nbr.null | 0 | 0 | 0 |
| nbr.na | 0 | 0 | 0 |
| min | 2.76 | 1.513 | 14.5 |
| max | 4.93 | 5.424 | 22.9 |
| range | 2.17 | 3.911 | 8.4 |
| sum | 115.09 | 102.952 | 571.16 |
| median | 3.695 | 3.325 | 17.71 |
| mean | 3.596563 | 3.21725 | 17.84875 |
| SE.mean | 0.094519 | 0.172969 | 0.31589 |
| CI.mean.0.95 | 0.192772 | 0.352772 | 0.644262 |
| var | 0.285881 | 0.957379 | 3.193166 |
| std.dev | 0.534679 | 0.978457 | 1.786943 |
| coef.var | 0.148664 | 0.304129 | 0.100116 |

**Conclusion :**

* Here, standard deviation is high in Weight than that of in point and score.
* Variance is high in Weight.
* Here, Mean and Median of each variable are mostly similar.so we may conclude that data is normally distributed.
* There is no null value in the variables of dataset.
* There are 32 entries (row) in the dataset.
* Point range is between 2.76 and 4.93, Score range is between 1 .51 and 5.424, weight range is between 14.5 and 22.9.

**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

= Σx / n

=

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

* from scipy.stats import kurtosis

from scipy.stats import skew

print(skew(df, axis=0, bias=True))

[-0.11395477 0.78248352]

print(kurtosis(df, axis=0, bias=True))

[ -0.57714742 0.24801866]

|  |  |  |
| --- | --- | --- |
|  | Speed | Dist |
| Skewness | -0.11395477 | 0.78248352 |
| Kuertosis | -0.57714742 | 0.24801866 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

* from scipy.stats import kurtosis

from scipy.stats import skew

print(skew(df, axis=0, bias=True))

[1.58145368 -0.60330993]

print(kurtosis(df, axis=0, bias=True))

[2.72352149 0.81946588]

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skewness | 1.58145368 | -0.60330993 |
| Kuertosis | 2.72352149 | 0.81946588 |

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



* From the Histogram ,we may conclude that the distribution is right skewed.
* Most of the data points fall between 50 and 100. There is less data-points fall between 350 and 400.



* From the box-plot ,we may conclude that there is outliers in the dataset

And the distribution is 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?**

: N = 3,000,000 n=2000, = 200, s= 30

Confidence Interval Estimate= Z => 200 Z

94% Confidence: qnorm(0.97)

[1] 1.880794=Z

200 1.88\*

**Lower limit =198.74**

**Upper limit = 201.26**

98% Confidence: qnorm(0.99)

[1] 2.326348=Z

200 2.33\*

**Lower limit =198.44**

**Upper limit = 201.56**

96% Confidence: qnorm(0.98)

[1] 2.053749

200 2.05\*

**Lower limit =198.62**

**Upper limit = 201.38**

**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?

**1**) > mean(A$scores)

[1] 41

> median(A$scores)

[1] 40.5

> var(A$scores)

[1] 25.52941

> sd(A$scores)

[1] 5.052664

1. **Conclusion:**

* Mean is grater than Median, This implies that the distribution is slightly skewed towards right. No outliers are present.
* On an average A student obtain 40 scores.
* There is high variation in the scores of students.

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

* There is no skewness, and distribution is symmetric

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

* Right skewed distribution (tail on the right side)

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

* Left Skewed distribution (tail on the left side).

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

* Positive values of kurtosis indicate that distribution is more peaked and possesses thick tails.

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

* negative values of kurtosis indicate that distribution is less peaked and possesses thin tails.

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



What can we say about the distribution of the data?

* Negative skewed distribution

What is nature of skewness of the data?

* Left skewed distribution

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

* Inter Quartile Range

=Upper Quartile- Lower Quartile

=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.

**Conclusion:**

* The median of the two boxplots are same approximately 262.
* Both follows normal distribution.
* Outliers doesn’t exist in both of the boxplots.
* Here, Range will be less in boxplot 1 than in boxplot 2.

**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

df.mean()

HP 117.469136 MPG 34.422076 VOL 98.765432 SP 121.540272 WT 32.412577

df.std()

HP 57.113502 MPG 9.131445 VOL 22.301497 SP 14.181432 WT 7.492813

from scipy import stats

print(1-stats.norm.cdf(38,34.422076,9.131445))

print(stats.norm.cdf(40,34.422076,9.131445))

print(stats.norm.cdf(50,34.422076,9.131445)-stats.norm.cdf(20,34.422076,9.131445))

* 1. P(MPG>38)

**0.34759394041453007**

* 1. P(MPG<40)

**0.7293498604157946**

* 1. P (20<MPG<50)

**0.8988689076273199**

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

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

* From normal Q-Q plot ,it Follows Normal distribution



* From normal Q-Q plot ,MPG variable Follows Normal distribution

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

Dataset: wc-at.csv

Conclusion:

waist follows Normal Distribution from the below QQ-plot

> qqnorm(wc\_at$Waist)

> qqline(wc\_at$Waist)



Adipose Tissue follows normal distribution

qqnorm(wc\_at$AT)

qqline(wc\_at$AT) 

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

* Price - 90% CI

norm.interval(0.90,loc = 2219.576610,scale = 580.803956)

(1264.2391164256364, 3174.9141035743633)

* speed - 94% CI

norm.interval(0.94,loc = 52.011024,scale = 21.157735)

(12.217691249041998, 91.804356750958)

* hd - 60% CI

norm.interval(0.60,loc = 416.601694,scale = 258.548445)

(199.00183278074124, 634.2015552192588)

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

* stats.t.interval(0.975,df=24)

**(-2.3909493151293866, 2.390949315129389)**

* stats.t.interval(0.98,df=24)

**(-2.4921594731575762, 2.4921594731575762)**

* stats.t.interval(0.995,df=24)

**(-3.0905135487060877, 3.090513548706097)**

**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

* µ=270 , =260 , SD=90 , n=18
* df=n-1=18-1= 17

tscore=

=

= -10/21.23

= -0.47

> pt(-0.47,17)

[1] **0.3221639**

* probability that 18 randomly selected bulbs would have an average life of no more than 260 days

= 0.32

= **32%**