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

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

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

There are 23=8 possible outcomes when three coins are tossed: HHH, HHT, HTH, THH, HTT, THT, TTH, TTT (where H represents heads and T represents tails).

Out of these outcomes, there are three ways to get two heads and one tail: HHT, HTH, THH.

Probability=Total number of possible outcomes / Number of desired outcomes​=3/8​

Therefore, the probability of getting two heads and one tail is 3 out of 8, which can be expressed as a fraction:

So, the probability of obtaining two heads and one tail when three coins are tossed is 3/8​ or 0.375 as a decimal.

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

1. **Probability that the sum is equal to 1:**

When rolling two dice, the minimum sum possible is 2 (1 on each die). There is no possible way to get a sum of 1 with two dice since the lowest value on each die is 1.

So, the probability of getting a sum of 1 is 0.

1. **Probability that the sum is less than or equal to 4:**

* The combinations that satisfy this condition are (1, 1), (1, 2), (2, 1), and (1, 3), (2, 2), and (3, 1) as they give sums of 2, 3, and 4 respectively.

There are a total of 36 possible outcomes when two dice are rolled (6 faces on die 1 multiplied by 6 faces on die 2).

So, the probability of getting a sum less than or equal to 4 is 6/36 = 1/6​.

c) **Probability that the sum is divisible by 2 and 3:**

The sums that are divisible by both 2 and 3 are 6 and 12.

For a sum to be divisible by both 2 and 3, the possible combinations are:

* (3, 3) for a sum of 6
* (6, 6) for a sum of 12

So, there are 2 favorable outcomes out of 36 total possible outcomes.

Therefore, the probability of getting a sum divisible by both 2 and 3 is 2/36=1/18

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. Total number of events {(R,R),(R,G),(R,B),(G,R),(G,G),(G,B),(B,R),(B,G),(B,B)}=9

Interested number of events =4

Probability =4/9

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 =(1×0.015)+(4×0.20)+(3×0.65)+(5×0.005)+(6×0.01)+(2×0.120)

=0.015+0.80+1.95+0.025+0.06+0.24

=3.085

Therefore, the expected number of candies for a randomly selected child is 3.085.

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

Points

**:** Mean =3.596563, Median= 3.695, Mode= “numeric”,

Variance= 0.2858814, Standard deviation= 0.5346787.

Score:

Mean= 3.21725, Median= 3.325, Mode= “numeric”,

Variance= 0.957379, Standard deviation= 0.9784574

Note: Mean value are closer for both ‘Point’ and ‘Score’.

Weight:

Mean= 17.84875, Median= 17.71, Mode= “numeric”,

Variance= 3.193166, Standard deviation= 1.786943

**Use Q7.csv 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 E*(*X*)=∑ *n i*=1​ *xi*​⋅*P*(*X*=*xi*​)

Where:

* *E*(*X*) is the expected value of the random variable *X* (in this case, the weight of a patient).
* *xi​ represents each individual weight in the set.*

*P(X=xi​) is the probability of selecting a patient with weight xi​.*

*Given the weights of the patients:={108,110,123,134,135,145,167,187,199}*

*There are 9 patients in total, and assuming random selection without bias, each patient has an equal chance of being chosen, so the probability of selecting any particular patient is p(X=xi​)=1/9*

*E(X)=91​×(108+110+123+134+135+145+167+187+199)*

*=1/9×1208*

*=134.22*

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

**Cars speed and distance**

**Use Q9\_a.csv**

Data[‘speed’].skew()-

0.117509

Data[‘speed’].kurtosis()

-0.5089944

Data[‘dist’].skew()

0.806894

Data[ ‘dist. ’].kurtosis()

0.405052

**SP and Weight (WT**)

**Use Q9\_b.csv**

Data[‘ SP ’].skew()=1.61154

Data[‘ SP’].kurtosis()=2.977328

Data[‘WT’].skew()= -0.61

Data[‘WT’].kurtosis()=0.9502914

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



Here we can see that the major Chick weights fall in the catogory of 50-100g(measures in x) as the maximum which is 200.The minimum weights have afrequency if less than or equal to 5.

The plot is Right sqewed which show that there is lesser concentration ofchick weights in the 300-400gram category .

The expected value should be above 46.45



Solution

Median is less than mean right skewed and we have outlier on the upper side of box plot and there is less data points between Q1 and bottom point.

**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?

The formula for the confidence interval is:

Confidence Interval=Sample Mean±Z×(Sample Standard Deviation/sample Size1/2)

Where:

* Sample Mean (*x*ˉ) = 200 pounds
* Sample Standard Deviation (*s*) = 30 pounds
* Sample Size (*n*) = 2000 men
* *Z* values for different confidence levels:
  + For 94% confidence level: *Z*=1.88
  + For 98% confidence level: *Z*=2.33
  + For 96% confidence level: *Z*=2.05

For a 94% confidence interval:

CI94%  = 200±1.88 \* (30/ ) =200±1.88×0.6708=200±1.2635

For a 96% confidence interval:

CI96%  = 200±2.33 \* (30/ ) =200±2.33×0.6708=200±1.5615

For a 98% confidence interval:

CI98%  = 200±2.05 \* (30/ ) =200±2.05×0.6708=200±1.379

Therefore the 94% confidence interval is approximately (198.736,201.2635) pounds

Therefore the 96% confidence interval is approximately (198.4385,201.5615) pounds

Therefore the 98% confidence interval is approximately(198.6261, 201.3739) pounds

**12)** 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.

Mean: The mean (average) is calculated by summing up all the scores and dividing by the total number of scores.

Mean=∑Scores/Number of Scores

Mean=34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56 Mean=685/18

Mean≈38.055

Median: The median is the middle value of the dataset when arranged in ascending order. If the number of values is odd, it's the middle number; if the number of values is even, it's the average of the two middle numbers.

The data arranged in ascending order: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

There are 18 values, so the median is the average of the 9th and 10th values.

Median=40+41/

​ Median=81/2

Median=40.5

Variance: Variance measures the average squared deviation from the mean.

Variance=∑(xi−Mean)2/N

​ Where *Xi*​ represents each score, *N* is the total number of scores.

Variance=(34−38.0556)2+(36−38.0556)2+⋯+(56−38.0556)2

Variance≈52.798

Standard Deviation: The standard deviation is the square root of the variance.

Standard Deviation=

Standard Deviation≈7.267

1. What can we say about the student marks?

* The mean score of the student is approximately 38.06, indicating the average performance.
* The median score is 40.5, suggesting that half of the scores are above this value and half are below, indicating a slightly higher middle value.
* The variance and standard deviation are measures of the spread or dispersion of the scores around the mean. A higher standard deviation (7.2676) implies a wider spread of scores from the mean, indicating variability or inconsistency in the student's performance across the tests.
* The scores range from 34 to 56, indicating some diversity in test performance, with some scores being notably higher or lower than the average.

Overall, the student's marks show variability with some scores significantly deviating from the mean, suggesting inconsistencies in performance across the tests.

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

Ans When the mean and median of a dataset are equal, it indicates that the data is symmetrically distributed. This equality between the mean and median often suggests that the distribution of the data is approximately symmetric.

In a symmetric distribution:

* The mean, median, and mode (if it exists) are all at the same central point within the data.
* The data tends to have a balanced distribution on both sides of the central point.
* The shape of the distribution looks similar on either side of the center when plotted on a graph.

Regarding skewness:

* Skewness is a measure of the asymmetry of a distribution.
* In a perfectly symmetric distribution where the mean and median are equal, the skewness is zero.
* A skewness of zero indicates that the dataset is perfectly symmetrical with no skew.

Therefore, when the mean and median of a dataset are equal, the nature of skewness in the data is considered to be zero or that there is no skewness present in the distribution.

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

Ans When the mean of a dataset is greater than the median, it suggests that the distribution of the data is positively skewed.

Skewness is a measure of the asymmetry of the distribution of a dataset:

* Positive skewness occurs when the right tail (the larger values) of the distribution is longer or stretched out compared to the left tail. This indicates that there are a few large values that are pulling the mean to the right, making it greater than the median.
* In a positively skewed distribution, the mean is greater than the median because the presence of relatively few unusually high values (outliers) on the right side of the distribution increases the mean but has less impact on the median.
* The majority of the data tends to cluster towards the lower values, and the larger values (outliers) have a disproportionate effect on the mean.

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

When the median of a dataset is greater than the mean, it suggests that the distribution of the data is negatively skewed.

Skewness is a measure of the asymmetry of the distribution of a dataset:

* Negative skewness occurs when the left tail (the smaller values) of the distribution is longer or stretched out compared to the right tail. This indicates that there are a few smaller values that are pulling the mean to the left, making it smaller than the median.
* In a negatively skewed distribution, the median is greater than the mean because the presence of relatively few unusually low values (outliers) on the left side of the distribution reduces the mean but has less impact on the median.
* The majority of the data tends to cluster towards the higher values, and the smaller values (outliers) have a disproportionate effect on the mean.

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

Ans. A positive kurtosis value in a dataset indicates that the distribution has heavier tails and a more peaked central peak compared to a normal distribution. Positive kurtosis is often referred to as leptokurtic.

Kurtosis is a statistical measure that describes the shape of the probability distribution of a dataset in relation to the normal distribution. It measures the "tailedness" or the degree of outliers (extreme values) present in a distribution.

Here's what a positive kurtosis value indicates:

1. **Heavy Tails**: Positive kurtosis suggests that the tails of the distribution are heavier or have more extreme values than would be expected in a normal distribution. It means there are more data points farther from the mean compared to a normal distribution.
2. **Peakedness**: A positive kurtosis value indicates a relatively sharper peak (more peaked) in the center of the distribution compared to a normal distribution. This signifies that the data has more values concentrated around the mean, creating a taller and narrower peak.
3. **Risk of Outliers**: Higher positive kurtosis suggests an increased probability of extreme values or outliers in the dataset. There might be an increased frequency of values far from the mean.

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

Ans. A negative kurtosis value in a dataset indicates that the distribution has lighter tails and a flatter central peak compared to a normal distribution. Negative kurtosis is often referred to as platykurtic.

Kurtosis is a statistical measure that describes the shape of the probability distribution of a dataset in relation to the normal distribution. It measures the "tailedness" or the degree of outliers (extreme values) present in a distribution.

Here's what a negative kurtosis value indicates:

1. **Lighter Tails**: Negative kurtosis suggests that the tails of the distribution are lighter or have fewer extreme values compared to a normal distribution. It means there are fewer data points farther from the mean compared to a normal distribution.
2. **Flattened Peak**: A negative kurtosis value indicates a relatively flatter peak in the center of the distribution compared to a normal distribution. This signifies that the data is more spread out, resulting in a broader and flatter peak around the mean.
3. **Fewer Outliers**: Lower negative kurtosis suggests a decreased probability of extreme values or outliers in the dataset. There might be fewer values far from the mean.

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



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

Answer: Let’s assume above box plot is about ages of the students in a school.

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

And students whose age is above 15 are approx. 40%.

What is nature of skewness of the data?

The data will be left skewed since whisker length on the upper quadrant ishigher than the data on the lower quadrant. Median will be greater than the mean since data is left skewed

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

IQR is the inter quartile range.Here Q1 = 10

Q2 = 14.7

Q3 = 18

IQR = Q3

**–**

 Q1 = 8(approx)

Q19) Comment on the below Boxplot visualizations?



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

Ans. By observing both the plots whisker’s level is high in boxplot 2, mean and

median are equal

hence distribution is symmetrical.

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)

= mean(MPG)=34.42208

= sd(MPG)

=9.131445

= 1 –pnorm(38,mean(MPG),sd(MPG))

= 0.330

= 33%

* 1. P(MPG<40)

=pnorm(40,mean(MPG),sd(MPG))

=0.7293499

=72.3%

c. P (20<MPG<50)

=pnorm(50,mean(MPG),sd(MPG)) - pnorm(20,mean(MPG),sd(MPG))

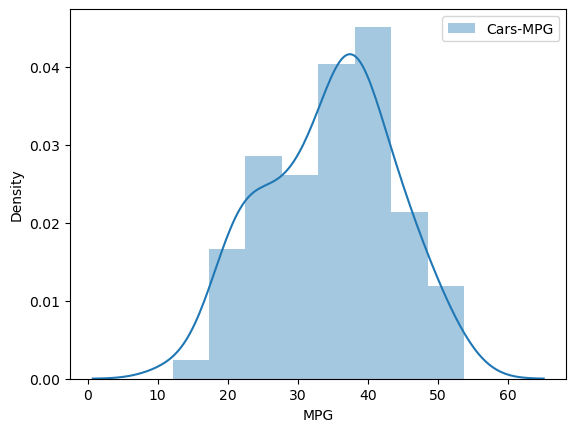
=0.955 -0.057

=0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



cars.MPG.mean()

34.42207572802469

cars.MPG.median()

35.15272697

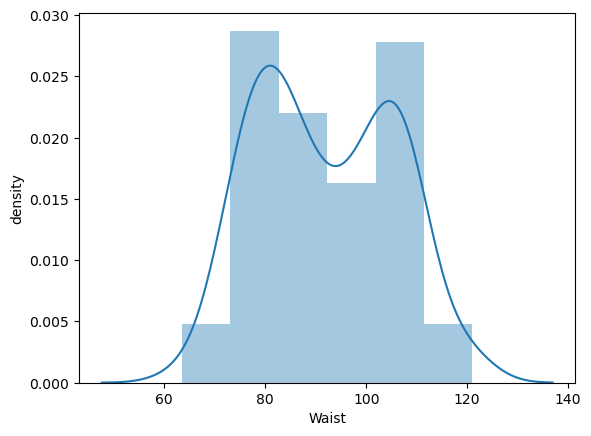
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

# plotting distribution for Waist Circumference (Waist)

sns.distplot(wcat.Waist)

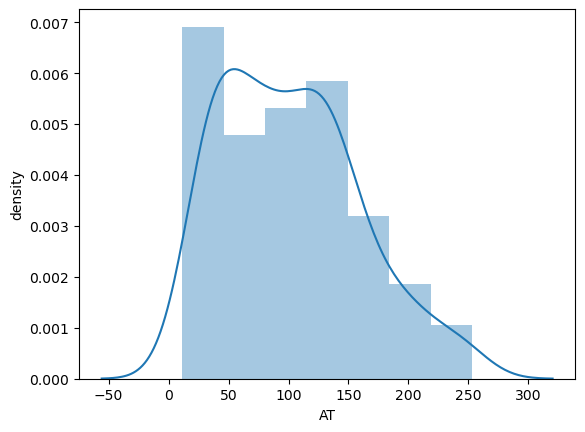
plt.ylabel('density');



# plotting distribution for Adipose Tissue (AT)

sns.distplot(wcat.AT)

plt.ylabel('density');



# WC

wcat.Waist.mean() , wcat.Waist.median()

(91.90183486238531, 90.8)

# AT

wcat.AT.mean() , wcat.AT.median()

(101.89403669724771, 96.54)

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

Z scores

=90%

= 95+2.5

=97.5

=qnorm(0.975)

=1.96

94%

= 94+4=97

=qnorm(0.97)

=1.88

60%

= 60 + 20= 80

= qnorm(0.80)

= 0.841

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

TSCORE CALCULATION

T((1,alpha),(n-1))

Here n = 25

n-1 = 24

Hence t score values will be:

95%

= qt(0.975,24)

= 2.06389996%

 96%

= qt(0.98,24)

=2.171545

99%

= qt(0.995,24)

= 2.79694

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

Answer :

mu=270

n=18

xbar=260

sigma=90

z=x-mu/sigma=260-270/90=-0.11

pnorm(-0.11)=0.4562

p=45%

T=x-mu/s/sqrt(n)=260-270/90/sqrt(18)=-0.4714

Pt-(0.4714,17) =0.3216

P=32%