Answer Key – Statistics Test

1) **Identify the variables that are continuous or discrete?**
    a. Time & weight are continuous. Country & color are discrete

Explanation:
A **continuous variable** is a variable whose value is obtained by measuring.

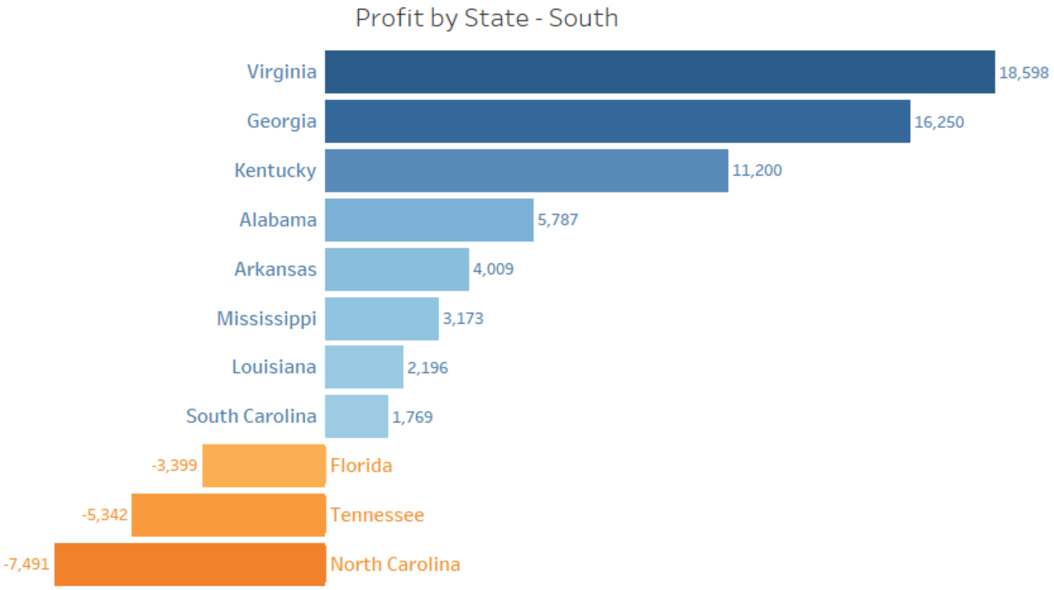
A **discrete random variable** X has a countable number of possible values.

2) **A histogram is what representation of the continuous variable ?**
    a. Probability distirbution

Explanation:
A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable

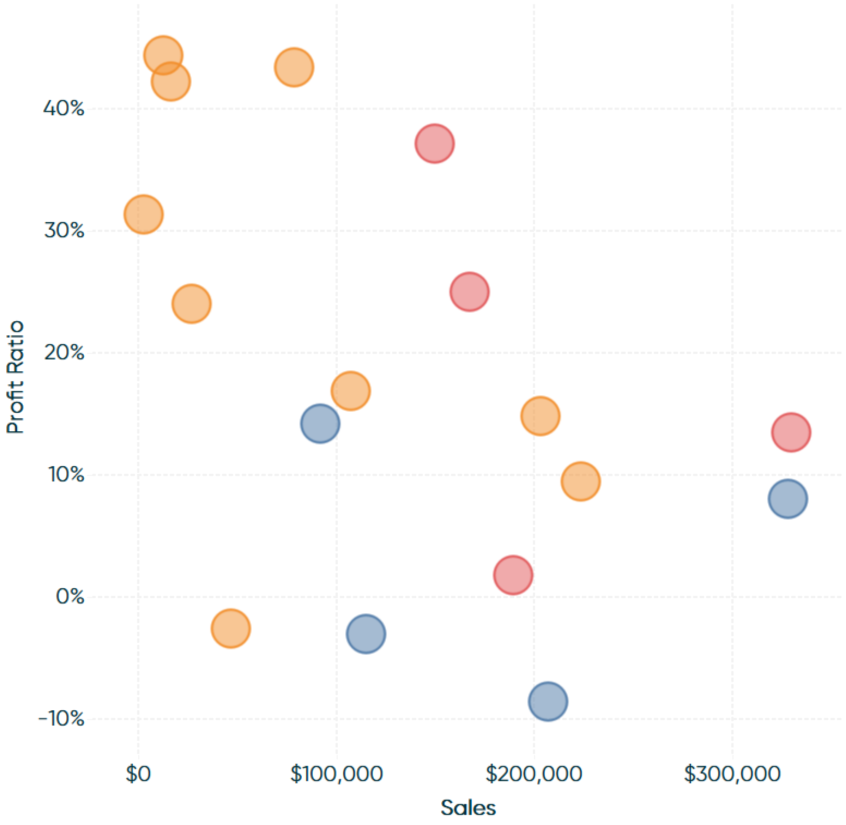
3) **What method of data representation is best suited to the demonstration of data results if that data is of differing nominal values and needs to represent quantitative data on X axes ?**
    a. Bar chart

Explanation:
A bar chart uses bars to show comparisons between categories of data. These bars can be displayed horizontally or vertically. A bar graph will always have two axis. One axis will generally have numerical values, and the other will describe the types of categories being compared.
Example:



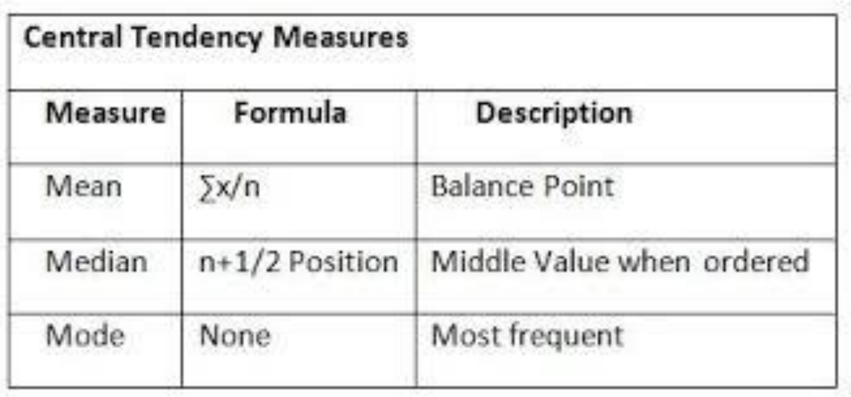
4) **In a project the business wants to see the relationship between revenue generated and YoY Sales. Which plot is the best ?**
   a. Scatter plot

Example:



5) **Millions of Americans work from home during office hours and following is a sample data of individuals who work at home 18,54,20,46,25,48,53,27,26,37,40,36,42,25,27,33,28,40,45,25 Find the Mean and mode?**
   a. 34.75,25

Explanation:



Central Tendency Measures		
Measure	**Formula**	**Description**
Mean	∑x/n	Balance Point
Median	n+1/2 Position	Middle Value when ordered
Mode	None	Most frequent

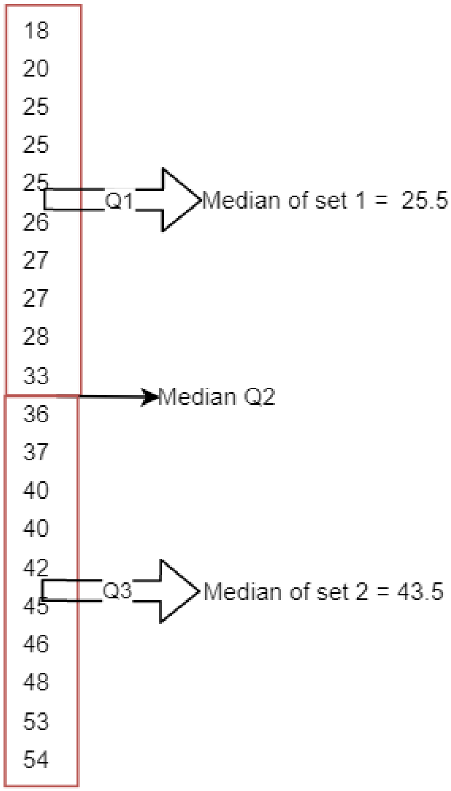
6) **The median age of population of all adults is 36 years. Using the median age obtained in Q5 (above) comment whether the at-home workers tend to be younger or older than the population of adults?**
   a. At home workers are slightly younger

Explanation:
The median age achieved in question 5: 34.5

Sorting the numbers in Question 5 in ascending order
We get Median = (33+36)/2 = 34.5
The median age of Americans work from home during office hours < 36
Therefore, the at-home workers tend to be younger than the older population of adults.

7) **Using the sample data from Q5 compute the first and third quartile?**
   a. 25.5,43.5

Explanation:
Arrange the data in ascending order



18
20
25
25
25
26
27
27
28
33
36
37
40
40
42
45
46
48
53
54

8) **If the value 18 in Q5 is replaced with 38, the standard deviation?**
   a. Decreases

Explanation: By definition Standard Devaition is deviation from the mean.
Here, when we have the number 18 – The difference between mean (34.75) and this number is more and thereby it contributes more to the SD.
But when it is replaced with 38 closers to the mean than 18 and thereby the SD decreases.

9) **What is the relation between Standard Deviation and variance when the sample is less than 30?**

   a. The SD = SQRT(variance)

   SQRT = Square Root

10) **If a population consists of distinct 5000 records sorted between 50 and 5050 and if the last 10 samples are taken for analysis. What is the range of the sample?**

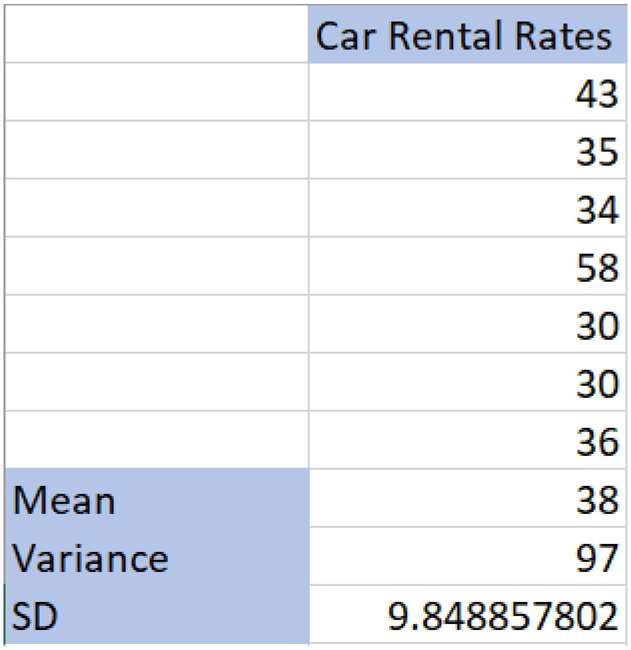
   a. 10

Explanation:
The question says there are 5000 distinct records sorted between 50 – 5050:
Which means the set is 50,51

11) **Car rental rates per day for a sample of seven Eastern US cities are as follows CityBoston Dallas Atlanta Ohio New York Miami Pittsburgh Rate ($)43 35 34 58 30 30 36. Compute the Mean, variance, standard deviation for the car rental rates**

   a. 38,97,9.85

Explanation:



	Car Rental Rates
	43
	35
	34
	58
	30
	30
	36
Mean	38
Variance	97
SD	9.848857802

12) **A similar sample of seven Western US cities showed a sample mean of $ 38 per day and variance and SD as 93, 9.64. What can you infer from this?**

   a. Eastern shows more variation

13) **The sales report about the pharmaceutical company in million $ for the 21 states in US has been provided in the spread sheet 8408,1374,1872,8879,2459,11413,608,14138,6452,1850,2818,1356,10498,7478,4019,4341,739,2127,3653,5794,8305.**

1. Provide the five number summary of the box plot(min,Q1,Q2,Q3,max)

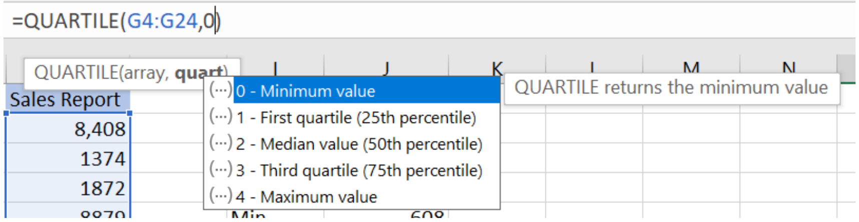
   a. None of the above

Explanation:

Sales Report
8,408
1374
1872
8879
2459
11413
608
14138
6452
1850
2818
1356
10498
7478
4019
4341
739
2127
3653
5794
8305

	Sales report
Min	608
Q1	1872
Median	4019
Q3	8305
Max	14138

I did this with the formula: As discussed in class.



```
=QUARTILE(G4:G24,0)
```

Sales Report
8,408
1374
1872

QUARTILE(array, quart)
I
J
K
L
M
N
0 - Minimum value
1 - First quartile (25th percentile)
2 - Median value (50th percentile)
3 - Third quartile (75th percentile)
4 - Maximum value
QUARTILE returns the minimum value

14) Compute the IQR, lower and upper limits from the above

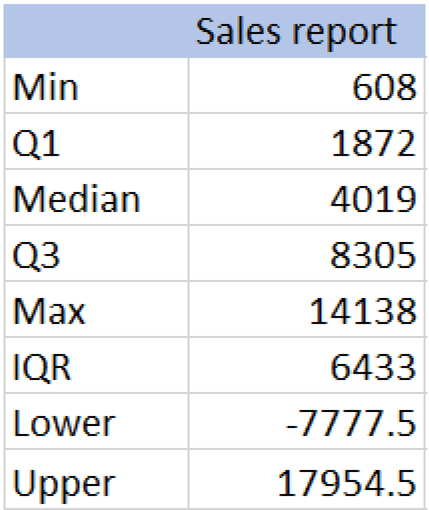
a. 6433,-7777.5,17955

Explanation:

IQR = Q3-Q1

Lower = Q1-1.5*(IQR)

Upper = Q3+1.5*(IQR)



	Sales report
Min	608
Q1	1872
Median	4019
Q3	8305
Max	14138
IQR	6433
Lower	-7777.5
Upper	17954.5

15) **From the above does the data contain any outlier.**

   a. No

Explanation: Anything less than the Lower whisker and greater than the upper whisker are called outliers which is not here in this case.

16) **Ohio state has the highest sales at $14,138 million. Suppose a data entry error has been made as $ 41,138 million. Would this been identified as an outlier and corrected?**

   a. Yes, 41,138 would be an outlier

Explanation Anything less than the Lower whisker and greater than the upper whisker are called outliers which is greater in this case.

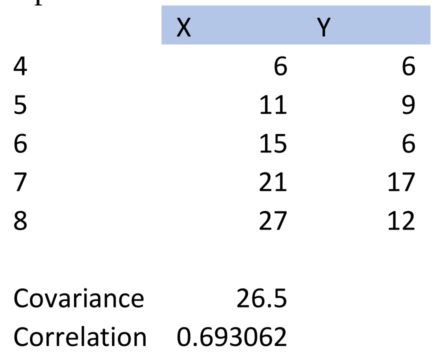
17) **What is the range of correlation coefficient?**

   a. -1 to +1

18) **Sample observations were taken between x and y as follows. X = 6,11,15,21,27 and Y=6,9,6,17,12. Compute the covariance and correlation coefficient**

   a. 26.5,0.693

Explanation: A B



	X	Y
4	6	6
5	11	9
6	15	6
7	21	17
8	27	12
Covariance	26.5	
Correlation	0.693062	

Formulas used for both are as follows (as seen in class):

Covariance: =COVARIANCE.S(A4:A8,B4:B8) - .S as we have sample observations stated in the question

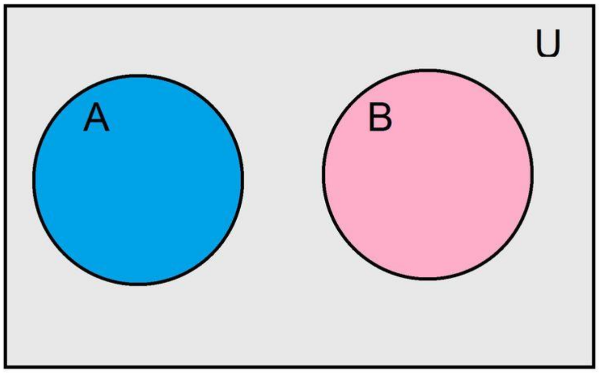
Correlation = =CORREL(A4:A8,A4:B8)

19) 1.The Sum of probabilities of all events is 1
2. The probability lies between -1 to +1
3. In a mutually exclusive event P(AnB) = 1
4. In a mutually exclusive event P(AUB) = 1

From the above please check the appropriate option:

a. 1 & 4 are correct.

Explanation:
Probability lies between 0 & 1. And so, the sum of probabilities of all events can be 1.
For a Mutually exclusive event:



P(AnB) = 0 as there will be no area common to both.
But P(AUB) = 1, Which gives us 1 & 4 true among all the above.

20) When 2 coins are tossed and the probability of getting 2 heads is 0.25 what is the size of sample space?

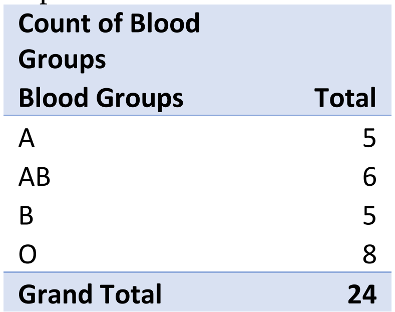
a. 4

Explanation:
When two coins are tossed the Sample Space S = {(h,h),(h,t),(t,h),(t,t)}
P(two heads) = ¼ = 0.25
Size of Sample space = 4 (as we have 4 events on the whole)

21) Twenty-four people had a blood test and the results are shown below. A, B , B , AB , AB , B , O , O , AB , O , B , A, AB , A , O , O , AB , O , O , A , AB , O , B , A. If a person is selected randomly from the group of twenty-four people, what is the probability that his/her blood type is not O?

a. 0.667

Explanation:



Count of Blood Groups	
Blood Groups	**Total**
A	5
AB	6
B	5
O	8
Grand Total	**24**

On the whole we have 24 people tested,

Out if which we will need to find the probability that a random person selected is not from O group.
1-(8/24) = 2/3 = 0.667
(I used Pivot table to sum up the data as per blood group and counts.)

22) CSK winning IPL 2019 (0.8 probability)
MI winning IPL 2019 (0.2 probability)
CSK winning Champions trophy 2019 (0.6 probability)
MI winning Champions trophy 2019 (0.4 probability)
In the above events what are the mutually exclusive events?

a) 3 & 4

Explanation:
IPL 2019 has already been done and both the team went to the finals (CSK and MI). So, the winning of CSK or MI was dependent on each other's performance (Mutually inclusive)

But Champions trophy has not started yet and performance of each team depends only on themselves and not on other in this stage… And hence mutually exclusive.

23) Using the above data what is the probability of CSK winning IPL 2019 or CSK winning Champions trophy 2019?

a. 0.92

Explanation:
P(AorB) = P(AUB) = P(A)*P(B) = 0.8*0.6 = 0.48
P(AandB) = P(A)+P(B)-P(AUB) = 0.8+0.6-0.48 = 0.92

24) Using the above data what is the probability of CSK winning IPL 2019 and not winning Champions trophy 2019

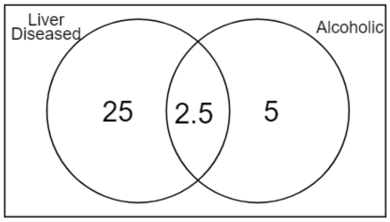
a. 0.32

Explanation:
The question asks for (P(A) and 1-P(B)) = P(A) * 1-P(B) = 0.8*(1-0.6) = 0.8*0.4 = 0.32

25) A case study to find out a patient's probability of having liver disease if they are an alcoholic. Patient has liver disease - Past data tells you that 25% of patients entering your clinic have liver disease. Patient is an alcoholic - Five percent of the clinic's patients are alcoholics. Among those patients diagnosed with liver disease, 10% are alcoholics.

a. 0.5

Explanation:
Assume there are a total of 100 patients coming to the clinic. Out of which 25% = 25 have liver disease and 5% = 5 are alcoholic



And the question says 10% of patients diagnosed with liver disease are alcoholics = 10%(25) = 2.5.

We need to find what is the probability that patient will have liver disease if they are alcoholic. 2.5 out of 5 Alcoholics have liver disease which is 50%.

50% = probability of 0.5.

26) **In a Normal distribution**
    a. mean = median = mode

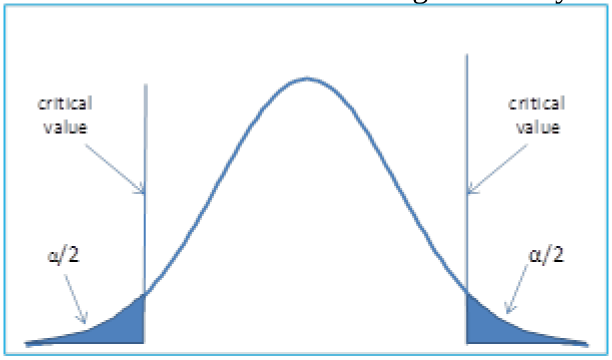
27) **Which test to be performed when we have only the mean of population and sample is less than 30?**
    a. T- TEST

28) **A sample of size 50 is drawn from a population of mean 100 and Standard deviation 25. What is the Standard deviation of the sample?**
    a. 3.53

Explanation:
n = 50, S.D = 25, mean = 100.
To find: Standard Error(s) = 25/sqrt(50) = 25/7.071 = 3.53

29) **A pharma company manufactures thousands of tablets every day. The manufacturing team gets a complaint stating that the weight of a tablet named zingx has changed from its actual claimed weight of 100mg with population standard deviation 20. The company wants to test this and submit a report to the concerned authority stating the proof of this complaint. The company takes 50 samples with mean 105 to test this. What is the Null and alternate hypothesis?**
    a. H0 = 100 and Ha ≠ 100

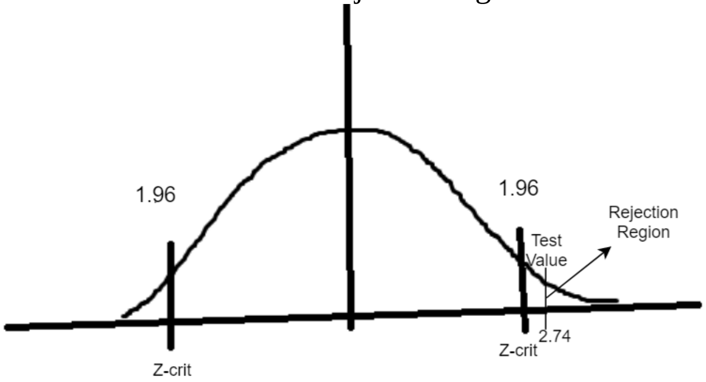
Explanation:
The claim of the complaint states that the weight of the tablet has “changed” and does not say whether is has decreased or increased. In that case it can go both ways so double tail test.



30) **At Ohio University the mean score of scholarship exam for fresh applications is 900 and the population standard deviation is 180. Every year the HOD uses sample applications to determine the change in the examination score. A sample of 200 applications with a sample mean of 935 is used to perform hypothesis test. What is the result?**
    a. Calculated value 2.74, Reject the Null hypothesis

Explanation:
Mean = 900
S.D = 180
Sample n = 200, sample mean = 935.

Sample size > 30 so Z – Test.
Null Hypothesis H0: mean = 900
Alternate Hypothesis Ha ≠ 900
Level of Significance = 1-95% = 1-0.95 = 0.05
Critical Value for 95% confidence interval = 1.96
Z = (X-mean) / S.E , S.E = 180/sqrt(200) = 12.73
Z = (935-900)/12.73 = 2.74 which falls in the rejection region.



31) The California university performs the hypothesis test for the same scenario as Ohio university on 6 samples with the population mean as 900 with samples as 935,925,850,875,945,915. What is the result?

a. Calculated value 2.57, Accept the Null hypothesis

Explanation:
Here the number of samples are less than 30. So, we go with T-test. And I take two variable as the question states same scenario as previous question. Critical

California Sample
935
925
850
875
945
915

t-Test

	California Sample
Mean	907.5
Variance	1377.5
Observations	6
Hypothesized Mean	900
df	5
t Stat	0.494983913
P(T<=t) one-tail	0.320796656
t Critical one-tail	2.015048373
P(T<=t) two-tail	0.641593312
t Critical two-tail	2.570581836

T-stat is less than the critical value so fails to reject the null hypothesis.

32) What is the purpose of a goodness-of-fit test?

a. To assesses whether the central tendency, variability and distribution of sample is different from that of the population

Explanation: The trap here is not about Chi-Square test but about goodness-of-fit. Goodness of fit is not only in Chi-Square there are other methods that is used to test if sample data fits a distribution from a certain population.