TE/Sem VI/Quantitative Analysis Solution

May 2022

QP: 93900

Q2.

A.

The questionnaire studies can be classified on the basis of :

(*i*) The degree to which the questionnaire is formalized/structured,

(*ii*) The disguise or lack of disguise of the questionnaire, and

(*iii*) The communication method used.

When no formal questionnaire is used, interviewers adapt their questioning to each inteview as it progresses or perhaps elicit responses by indirect methods such as showing pictures on which the respondent comments. When a prescribed sequence of questions is followed, it is referred to as *structured* study. On the other hand, when no prescribed sequence of questions exists, the study is *non-structured.*

When questionnaires are constructed so that the objective is clear to the respondents they are *non-disguised* ; on the other hand, when the objective is not clear the questionnaire is a *disguised* one. Using these two bases of classification,

1. Non-disguised structured,
2. Non-disguised non-structured,
3. Disguised structured, and
4. Disguised non-structured.

Merits of Questionnaire Method

1. Easy to adopt
2. Less expensive
3. Better when questions are of personal nature

Limitations of Questionnaire Method

1. Informants have to be literate
2. Uncertainty of response
3. Answers may not be accurate

Merits of Schedule Method

1. Easy to adopt even when respondents are illiterate

2. Response can be calibrated
3. Answers are generally accurate

Limitations of Schedule Method

1. Comparatively costly
2. Success depends on the enumerators
3. Experience and training is required
4. The way interview is conducted decides the type of data

To make this method work effectively the following suggestions are made :

1. The questionnaire should be so framed that it does not become an undue burden on the respondents, otherwise they may not return them back.

2. Prepaid postage stamp should be affixed.

3. The sample should be large.

4. It should be adopted in such enquiries where it is expected that the respondents would return the questionnaire because of their own interest in the enquiry.

5. Its use should be preferred in such enquiries where there could be a legal compulsion to supply the information so that the risk of non-response is eliminated.

Guidelines for drafting Questionnaire

1. Covering Letter
2. Number of questions should be small
3. Questions should be arranged logically
4. Questions should be short and simple
5. Ambiguous question should be avoided
6. Personal questions should be avoided
7. Instructions to the informants
8. Question should be capable of objective answer
9. As far as possible 'Yes' or 'No' type of questions should be asked
10. Should be specific
11. Questionnaire should look attractive
12. Questions requiring calculations should be avoided
13. Pre-testing the questionnaire
14. Cross –checks
15. Method of tabulation

B.

	Town A	Town B

	Male	Female	Total	Male	Female	Total
Coffee Drinkers	25%	20%	45%	25%	15%	40%
Non-Coffee Drinkers	35%	20%	55%	30%	30%	60%
Total	60%	40%		55%	45%	

C.

Point estimators are functions that are used to find an approximate value of a population parameter from random samples of the population. They use the sample data of a population to calculate a point estimate or a statistic that serves as the best estimate of an unknown parameter of a population.

i. Consistency:

Consistency tells us how close the point estimator stays to the value of the parameter as it increases in size. The point estimator requires a large sample size for it to be more consistent and accurate.

You can also check if a point estimator is consistent by looking at its corresponding expected value and variance. For the point estimator to be consistent, the expected value should move toward the true value of the parameter.

(Example 9.2) Let $Y_1, \ldots, Y_n$ denote a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$. Show that $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a consistent estimator of μ.

ii. Unbiasedness:

The most efficient point estimator is the one with the smallest variance of all the unbiased and consistent estimators. The variance measures the level of dispersion from the estimate, and the smallest variance should vary the least from one sample to the other.

Generally, the efficiency of the estimator depends on the distribution of the population. For example, in a normal distribution, the mean is considered more efficient than the median, but the same does not apply in asymmetrical distributions.

Q3.

A.

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

i. Z-Test for single mean

The **One-Sample z-test** is used when we want to know whether the difference between the mean of a sample mean and the mean of a population is large enough to be statistically significant, that is, if it is unlikely to have occurred by chance. The test is considered robust for violations of normal distribution and it is usually applied to relatively large samples (N > 30) or when the population variance is known, otherwise you might consider using t-test.

Assumptions
1. Mean and variance of the population are known.
2. The test statistic follows normal distribution.

How To

Run**: Statistics→Basic Statistics→One Sample z-Test for Mean...**

Select the variable. For summarized data please use the **Statistics→Basic Statistics→One Sample z-Test for Mean (use summarized data)...** command.

Enter the **population mean hypothesized value** μ_0 and the **population variance** σ^2 (known).

If the population standard deviation σ is known instead of the variance – square the standard deviation value to calculate the population variance value.

ii. Z-Test for Different of Mean

Requirements: Two normally distributed but independent populations, σ is known

Hypothesis test

Formula: $$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means), σ_1 and σ_2 are the standard deviations of the two populations, and n_1 and n_2 are the sizes of the two samples.

The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. What is the likelihood that the population means of concentrations of the element are the same for men and women?

Null hypothesis: H_0: $\mu_1 = \mu_2$

or H_0: $\mu_1 - \mu_2 = 0$

alternative hypothesis: H_a : $\mu_1 \neq \mu_2$

or: H_a : $\mu_1 - \mu_2 \neq 0$

$$z = \frac{28 - 33 - 0}{\sqrt{\frac{14.1^2}{75} + \frac{9.5^2}{50}}} = \frac{-5}{\sqrt{2.65 + 1.81}} = -2.37$$

The computed z-value is negative because the (larger) mean for females was subtracted from the (smaller) mean for males. But because the hypothesized difference between the populations is 0, the order of the samples in this computation is arbitrary— $\bar{x}_1$ could just as well have been the female sample mean and $\bar{x}_2$ the male sample mean, in which case z would be 2.37 instead of –2.37. An extreme z-score in either tail of the distribution (plus or minus) will lead to rejection of the null hypothesis of no difference.

The area of the standard normal curve corresponding to a z-score of –2.37 is 0.0089. Because this test is two-tailed, that figure is doubled to yield a probability of 0.0178 that the population means are the same. If the test had been conducted at a pre-specified significance level of $\alpha < 0.05$, the null hypothesis of equal means could be rejected. If the specified significance level had been the more conservative (more stringent) $\alpha < 0.01$, however, the null hypothesis could not be rejected.

B. Slope = -1.8
   Intercept = 8.9

C. R^2 = 0.85867 / 0.85894
   R is 0.927
   F cal is 15.3978 / 15.28

Q4.

A. Stratified Sampling

Stratified random sampling or simply stratified sampling is one of the random methods which by using the available information concerning the population, attempts to design a more sufficient sample than obtained by the simple random procedure.

While applying stratified random sampling technique, the procedure followed is given below:

1. The universe to be sampled is subdivided (or stratified) into groups which arc mutually exclusive and include all items in the universe.

2. A simple random sample is then chosen independently from each group.

This sampling procedure differs from simple random sampling in that in the latter the sample items are chosen at random from the entire universe. In stratified random sampling the sampling is designed so that a designated number of items is chosen from each stratum. In simple random sampling the distribution of the sample among strata is left entirely to chance.

Some of the issues involved in setting up a stratified random sample are:

1. Base of Stratification

What characteristic should be used to subdivide the universe into different strata ? As a general rule, strata are created on the basis of a variable known to be correlated with the variable of interest and for which information on each universe element is known. Strata should be constructed in a way which will minimize differences among sampling units within strata, and maximize difference among strata.

For example, if we are interested in studying the consumption pattern of the people of Delhi, the city of Delhi may be divided into various part (such as zones or wards) and from each part a sample may be taken at random. Before deciding on stratification we must have knowledge of the traits of the population. Such knowledge may be based upon expert judgment, past data, preliminary observations from pilot studies, etc.

The purpose of stratification is to increase the efficiency of sampling by dividing a heterogeneous universe in such a way that (*i*) there is as great a homogeneity as possible within each stratum and (*ii*) a marked difference is possible between the strata.

2. Number of strata

How many strata should be constructed ? The practical considerations limit the number of strata that is feasible, costs of adding more strata may soon outrun benefits. As a generalization more than six strata may be undesirable.

3. Sample size within strata

How many observations should be taken from each stratum ? When deciding this question we can use either a proportional or a disproportional allocation. In proportional allocation, one samples each stratum in proportion to its relative weight. In disproportional allocation this is not the case. It may be pointed out that proportional allocation approach is simple and if all one knows about each stratum is the number of items in that stratum, it is generally also the preferred procedure. In disproportional sampling, the different strata are sampled at different rates. As a general rule when variability among observations within a

Merits. 1. *More representative.* Since the population is first divided into various strata and then a sample is drawn from each stratum there is a little possibility of any essential group of the population being completely excluded. A more representative sample is thus secured. C.J. Grohmann has rightly pointed out that this type of sampling balances the uncertainty of random sampling against the bias of deliberate selection.

2. *Greater accuracy.* Stratified sampling ensures greater accuracy. The accuracy is maximum if each stratum is so formed that it consists of unifrom or homogeneous items.

3. *Greater geographical concentration.* As compared with random sample, stratified samples can be more concentrated geographically, *i.e.*, the units from the different strata may be selected in such a way that all of them are localised in one geographical area. This would greatly reduce the time and expenses of interviewing.

Limitations. 1. Utmost care must be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous items as otherwise the results may not be reliable. If proper stratification of the population is not done, the sample may have the effect of bias.

2. The items from each stratum should be selected at random. But this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.

3. Because of the likelihood that a stratified sample will be more widely distributed geographically than a simple random sample cost per observation may be quite high.

B. Performance of Regression Model

i. MAE:

In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where,

$\hat{y}$ – *predicted value of* y

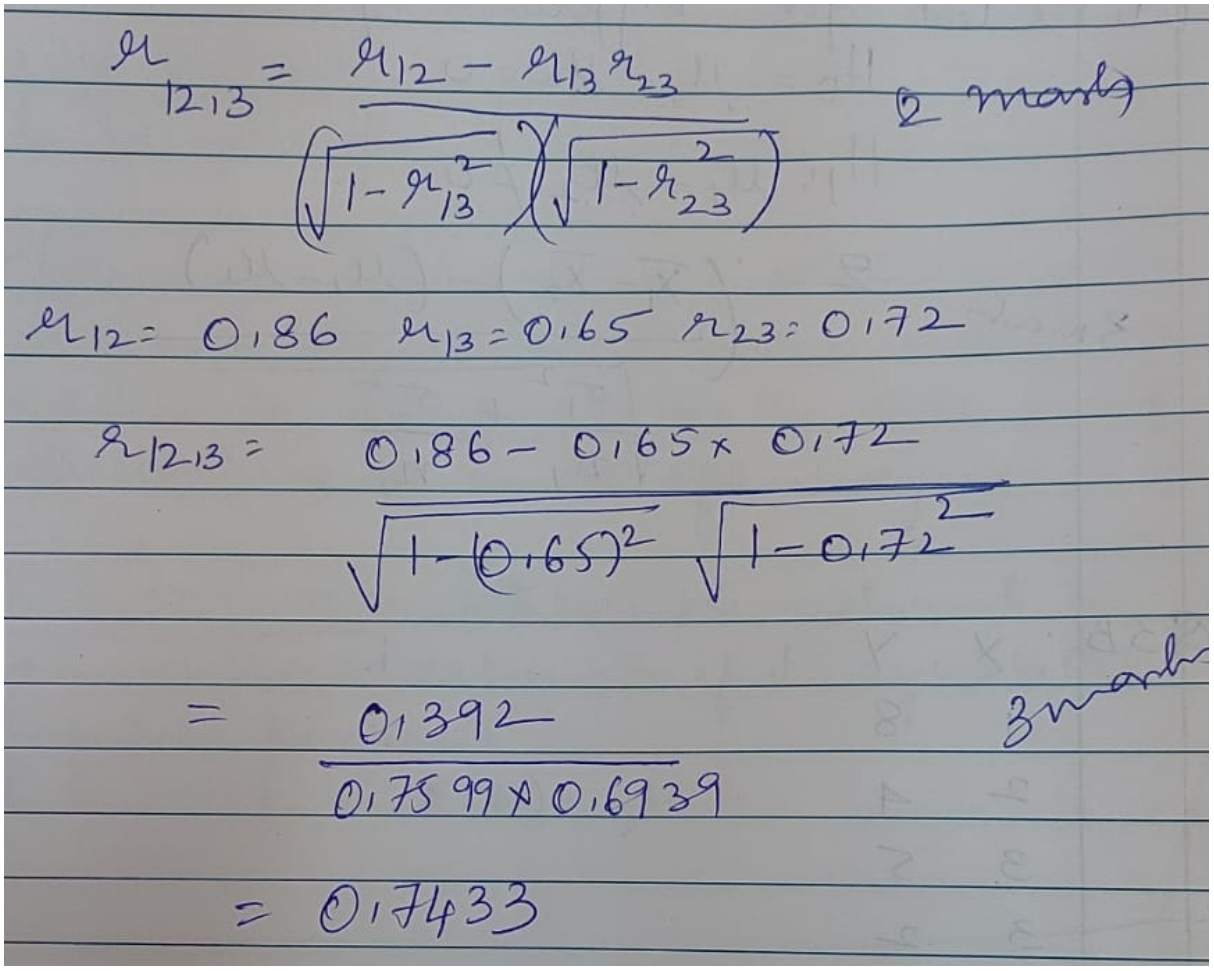
$\bar{y}$ – *mean value of* y

ii. MAPE:

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

C.



D.

Concept of Diagrammatic Presentation

- It is a technique of presenting numeric data through pictograms, cartograms, bar diagrams, and pie diagrams. It is the most attractive and appealing way to represent statistical data. Diagrams help in visual comparison and they have a bird's eye view.

- Under pictograms, we use pictures to present data. For example, if we have to show the production of cars, we can draw cars. Suppose the production of cars is 40,000, we can show it by a picture having four cars, where 1 car represents 10,000 units.
- Under cartograms, we make use of maps to show the geographical allocation of certain things.
- Bar diagrams are rectangular and placed on the same base. Their heights represent the magnitude/value of the variable. The width of all the bars and the gaps between the two bars are kept the same.
- Pie diagram is a circle that is subdivided or partitioned to show the proportion of various components of the data.
- Out of the given diagrams, only one-dimensional bar diagrams and pie diagrams are there in our scope.

Advantages of Diagrammatic Presentation

(1) Diagrams are attractive and impressive: The data presented in the form of diagrams can attract the attention of even a common man.

(2) Easy to remember: (a) Diagrams have a great memorising effect. (b) The picture created in mind by the diagrams last much longer than those created by figures presented through the tabular forms.

(3) Diagrams save time : (a) They present complex mass data in a simplified manner. (b) The data presented in the form of diagrams can be understood by the user very quickly.

(4) Diagrams simplify data: Diagrams are used to represent a huge mass of complex data in a simplified and intelligible form which is easy to understand.

(5) Diagrams are useful in making comparison: It becomes easier to compare two sets of data visually by presenting them through diagrams.

(6) More informative : Diagrams not only depict the characteristics of data but also bring out other hidden facts and relations which are not possible from the classified and tabulated data.

E.

a).

$$H_0 : \mu = 25 \; vs \; H_1 : \mu \neq 25$$

b). Mean, $\bar{x}$

$$\bar{x} = \frac{23+25+30+20+20+12}{6}$$

$$\bar{x} = 21.6667$$

Standard deviation, S

$$S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{181.33333}{5}} = 6.02218$$

tSTAT

$$tSTAT = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{21.66667-25}{\frac{5}{\sqrt{6}}}$$

$$= -1.63299$$

c). Conclusion :

The null hypothesis is not rejected as tSTAT is less than the table value. Hence the manufacturer's claim is valid at 1% level of significance

F.

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate. The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.

If the likelihood function is differentiable, the derivative test for determining maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved explicitly; for instance, the ordinary least squares estimator maximizes the likelihood of the

linear regression model. Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function.

From the perspective of Bayesian inference, MLE is generally equivalent to maximum a posteriori (MAP) estimation under a uniform prior distribution on the parameters. In frequentist inference, MLE is a special case of an extremum estimator, with the objective function being the likelihood.

Advantages.

- Simple to apply
- Lower variance than other methods (i.e. estimation method least affected by sampling error) and unbiased as the sample size increases.
- The method is statistically well understood
- Able to analyze statistical models with different characters on the same basis. Maximum likelihood provides a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for a large variety of estimation situations.
- Once a maximum-likelihood estimator is derived, the general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference.

Disadvantages.

- Computationally intensive and so extremely slow (though this is becoming much less of an issue)
- Frequently requires strong assumptions about the structure of the data
- The estimates that are obtained using this method are often biased. That is, they contain a systematic error of estimation. This is true for small samples. The optimality properties may not apply for small samples.
- MLE is inapplicable for the analysis of non-regular populations (Non-regular distributions are models where a parameter value is constrained by a single observed value)