ALLAMA IQBAL OPEN UNIVERSITY

AIOU LIBRARY
aioulibrary.com

Name:

Registration No.:

Course Code: 794

Program: MSc Forestry Extension

Semester: Autumn 2026

WhatsApp Orders: +92 324 5836626

ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

Course: Statistics (794)

Level: MSc Forestry Extension

Semester: Autumn 2026

Assignment No. 1 (Units 1–5)

Total Marks: 100 | Pass Marks: 40

SOLVED ASSIGNMENT GUIDE

(For Study & Reference)

AIOU LIBRARY aioulibrary.com

Note: This guide is prepared for study and reference purposes only. Students are advised to read the relevant units of the course book carefully and to write the assignment in their own words, with their own examples and presentation, rather than copying this guide word for word.

Question 1: Measurement Scales; Frequency Distribution and a Frequency Table (5 + 5 + 10 Marks)

Question as printed: (a) What do you mean by measurement scales? Explain its types with examples. (b) Define Frequency Distribution. Explain its objectives and importance in statistical analysis. (c) Prepare a frequency table for the data given below: 100, 96, 92, 88, 86, 84, 82, 80, 78, 91, 87, 83, 79, 77, 75, 73, 71, 69, 58, 56, 73, 50, 57, 55, 53, 51, 48, 46, 63, 59, 55, 51, 49, 47, 45, 43, 41, 58, 54, 50, 56, 44, 42, 40, 38, 36, 46, 53, 50, 43 (50 observations).

Introduction

Statistics begins with data, and data begin with measurement. Before any number is recorded, analysed or compared, the investigator must decide how a characteristic is to be observed and expressed in figures. That decision determines what the resulting numbers actually mean and, even more importantly, which arithmetical and statistical operations may legitimately be performed on them. A mark of 80 in an examination, a temperature of 30 degrees, a household income of Rs 60,000 and a jersey number 10 on a player's shirt are all 'numbers', yet they carry very different kinds of information. Part (a) of this question explains the scales on which such measurements are made. Part (b) then shows how raw measurements, once collected, are organised into a frequency distribution, and part (c) applies that technique to a set of fifty observations. Together the three parts cover the first two steps of every statistical investigation: measuring correctly, and organising what has been measured.

Part (a): Meaning of Measurement Scales

Measurement is the process of assigning numbers (or symbols) to objects, persons or events according to a definite set of rules, so that the properties of the numbers reflect the properties of the things measured. A measurement scale is the complete system of rules by which those assignments are made: it specifies what the numbers stand for, whether equal differences between numbers mean equal differences in the characteristic, and whether the numbers have a genuine zero. The scale is therefore not a property of the instrument alone; it is a property of the relationship between the numbers and the characteristic. S. S. Stevens (1946) classified measurement scales into four types, arranged in a hierarchy from the weakest to the strongest: nominal, ordinal, interval and ratio. Each higher scale possesses all the properties of the scales below it plus something extra, and each permits a wider range of statistical techniques. Choosing

the correct scale matters in forestry extension work as much as anywhere else: the species of a tree can only be named, the quality grade of timber can be ranked, temperature can be differenced, but height, girth and yield can be divided — and the scale decides which of these statements is meaningful.

1. Nominal Scale

The nominal scale is the lowest and simplest level of measurement. It merely classifies observations into distinct categories and attaches a label — a name or a number used purely as a name — to each category. The numbers carry no order, no magnitude and no arithmetic meaning; they only say 'same' or 'different'. The only permissible statistical operations are counting the cases in each category, finding the modal (most frequent) category, and tests based on frequencies such as the chi-square test. Examples: the species of trees in a sample plot (deodar, pine, oak, fir), where the labels 1, 2, 3, 4 could be permuted without losing any information; sex of respondents (male/female); blood groups (A, B, AB, O); marital status; district of residence; and the numbers printed on players' jerseys. Averaging jersey numbers or blood-group codes would be meaningless, which is exactly the limitation the nominal scale imposes.

2. Ordinal Scale

The ordinal scale also classifies, but its categories have a definite order or rank: one category is more, better, higher or stronger than another, although the size of the gap between successive categories is unknown and need not be equal. Numbers here express relative position only. Permissible statistics include the median, quartiles and percentiles, and rank correlations such as Spearman's coefficient, but not the arithmetic mean in a strict sense, because equal numerical differences do not represent equal real differences. Examples: the quality grade of timber or fruit

(Grade I, II, III); socio-economic status (low, middle, high); educational level (primary, middle, matric, graduate); the hardness of wood species ranked from softest to hardest; farmers' level of satisfaction with an extension programme (very dissatisfied to very satisfied on a five-point Likert scale); and positions in a class (1st, 2nd, 3rd). The difference between the 1st and 2nd position is not necessarily the same as between the 2nd and 3rd, and that is why the ordinal scale stops short of full arithmetic.

3. Interval Scale

The interval scale has ordered categories with equal intervals between successive values, so differences between numbers are meaningful and can be added and subtracted. Its limitation is the absence of a true zero: zero is an arbitrary point fixed by convention, not the complete absence of the characteristic, and therefore ratios are not meaningful. Because differences are genuine, the mean, standard deviation, and product-moment correlation can be used. The classic examples are temperature on the Celsius or Fahrenheit scales — 20°C is not 'twice as hot' as 10°C, and 0°C does not mean no heat — and calendar time, where the year 0 is conventional. Other examples are intelligence and aptitude test scores, and standardised (z) scores. In agricultural and forestry research, temperature, humidity indices and many test or index scores are interval measurements, and the analyst must remember that only their differences, never their ratios, may be interpreted.

4. Ratio Scale

The ratio scale is the highest level of measurement. It has all the properties of the interval scale — order and equal intervals — together with a true, non-arbitrary zero that indicates the complete absence of the characteristic. Consequently ratios are meaningful: a tree 20 metres tall really is twice as tall as one of 10 metres. All arithmetic operations and all statistical techniques,

including the geometric mean and the coefficient of variation, are permissible. Examples: height, weight, age, income, area, volume, yield per hectare, number of trees per plot, distance, and time duration. Most physical and economic measurements in forestry — girth at breast height, timber volume, biomass, seedling counts — are ratio measurements, which is why forestry data can usually be subjected to the full range of statistical analysis, including the means, regressions and analyses of variance used later in this course.

Scale	Properties	Permissible statistics	Example
Nominal	Classification only; labels	Counts, mode, chi-square	Tree species; blood group
Ordinal	Classification + order; unequal gaps	Median, percentiles, rank correlation	Timber grade; satisfaction level
Interval	Order + equal intervals; arbitrary zero	Mean, SD, correlation (differences only)	Temperature in °C; test scores
Ratio	Order + equal intervals + true zero	All statistics, incl. ratios, CV	Height, weight, yield, income

AIOU LIBRARY aioulibrary.com

Part (b): Frequency Distribution — Definition, Objectives and Importance

A frequency distribution is a systematic arrangement of statistical data in which the values of a variable are grouped into a number of classes (for a continuous or large data set) or listed individually (for a discrete variable), and the number of observations falling in each class — called the frequency of that class — is recorded against it. In short, it is a table that shows how the total frequency is distributed over the different values or classes of the variable. For example, if the marks of 50 students are grouped into the classes 30–40, 40–50 and so on, with 2, 12 … students in them, the resulting table is a frequency distribution. It converts a long, unorganised list of raw figures — from which nothing can be seen at a glance — into a compact summary in which the pattern of the data becomes visible.

Objectives. The main objectives of constructing a frequency distribution are as follows.

- **To condense and simplify raw data:** a mass of scattered figures is reduced to a small table that the mind can grasp at once.
- **To reveal the pattern of the data:** the distribution shows where observations concentrate, how they spread out, and whether the shape is symmetric, skewed, or has gaps and outliers.
- **To facilitate comparison:** two or more sets of data placed in similar class intervals can be compared easily, which is impossible with raw lists.
- **To facilitate further statistical analysis:** almost every later technique — mean, median, mode, standard deviation, correlation, graphs — is computed from, or drawn from, the frequency table rather than from the raw data.
- **To present data attractively:** a well-made table, and the histogram or frequency polygon drawn from it, communicates results to non-statisticians, including farmers and field officers in extension work.

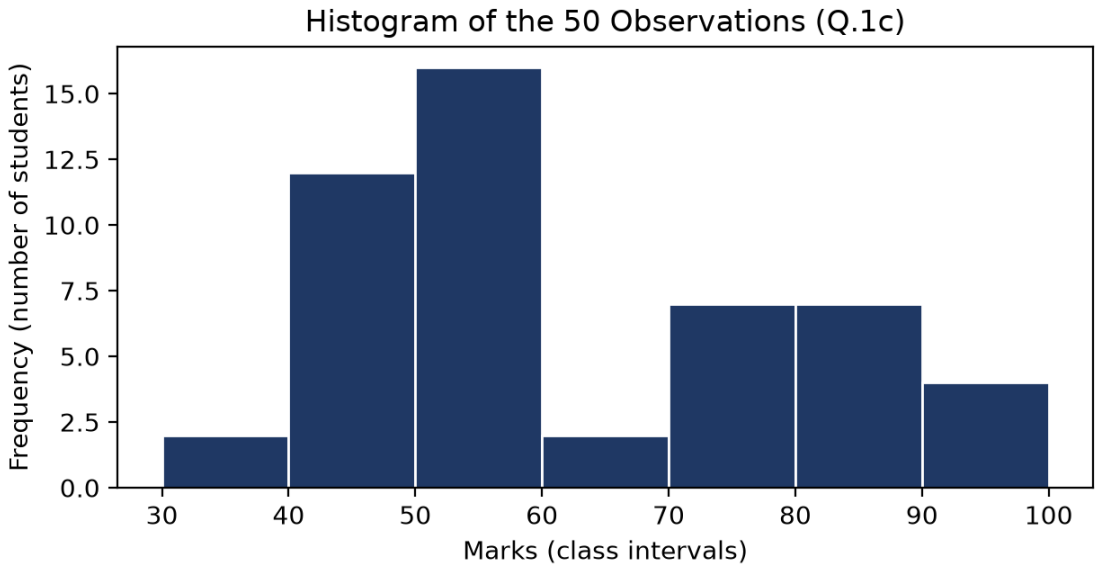
AIOU LIBRARY aioulibrary.com

The importance of the frequency distribution in statistical analysis follows from these objectives. It is the essential first step between data collection and data analysis: no average, measure of dispersion or significance test can be sensibly chosen until the analyst has seen how the data are distributed. It exposes errors and abnormal values at once — an impossible entry or an isolated extreme observation stands out in a table that it would hide in a raw list. It also provides the basis for estimating the probabilities of values falling in given ranges, which underlies all inference. For a forestry extension officer, for example, a frequency distribution of sapling survival rates, of farmers' landholdings, or of tree heights in a plantation immediately shows the typical value, the variability between farmers or plots, and the proportion falling below a critical standard — the three facts on which practical advice depends.

Part (c): Frequency Table for the Given Data

The fifty observations range from a minimum of 36 to a maximum of 100, so the range is 100 − 36 = 64. Using Sturges' rule, the suggested number of classes is $1 + 3.322 \log_{10}(50) \approx 1 + 3.322(1.699) \approx 6.6$, i.e. about 7 classes. A convenient class width is therefore 64/7 ≈ 9.1, rounded to 10. Taking classes 30–40, 40–50, …, 90–100 (the last class including its upper limit 100, where the maximum value falls), tallying each observation, and adding relative and cumulative frequencies gives the following table. The tally was checked observation by observation against the original list, and the frequencies sum to 50, the total number of observations, as they must.

Class interval	Tally (counting aid)	Frequency (f)	Relative frequency	Cumulative frequency
30–40	||	2	0.04	2
40–50	||||||||||||	12	0.24	14
50–60	||||||||||||||||	16	0.32	30
60–70	||	2	0.04	32
70–80	||||||	7	0.14	39
80–90	||||||	7	0.14	46
90–100	||||	4	0.08	50
Total	—	50	1.00	—



Reading the table: the observations concentrate heavily in the middle-lower classes — 30 of the 50 values (60 per cent) lie between 40 and 60 — while there is a striking dip in the 60–70 class, which contains only 2 observations, followed by a secondary cluster of 18 values between 70 and 100. The distribution is therefore irregular, with a gap in the sixties separating a large lower group from a smaller upper group; the histogram above shows this shape at a glance. (Grouped totals: the sum of the 50 observations is 3,128 and their overall mean is 62.56, a figure that sits in the nearly empty 60–70 class — a useful warning that a single average can misrepresent data of this shape.)

Conclusion

Measurement scales decide what numbers mean: nominal scales only name, ordinal scales rank, interval scales give meaningful differences, and ratio scales give meaningful ratios as well, and the analyst must match the technique to the scale. The frequency distribution is the first act of analysis performed on measured data: it condenses raw figures, displays their pattern, supports comparison, and prepares the ground for every later computation, as the table constructed in part (c) demonstrates — it turns fifty loose numbers into a clear picture of concentration, spread and a gap in the sixties.

Question 2: Measures of Central Tendency; Mean, Median and Mode of a Frequency Distribution (10 + 10 Marks)

Question as printed: (a) Explain different measures of location/central tendency. (b) Calculate the mean, median and mode from the following data — Number of persons per house: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; Number of houses: 26, 113, 120, 95, 60, 42, 21, 14, 5, 4.

Introduction

A frequency distribution, as constructed in Question 1, describes a whole set of data, but for most purposes we also need a single figure that stands for the set as a whole — a typical or representative value around which the observations cluster. Such a figure is called a measure of location, or measure of central tendency, or simply an average. Part (a) explains the principal measures — the arithmetic mean, the median, the mode, and the supplementary geometric and harmonic means — together with their merits and limitations. Part (b) then computes the mean, median and mode for a distribution of 500 houses classified by the number of persons per house, a typical piece of demographic data on which housing and forestry-extension planning (for example, fuelwood demand per household) actually depends.

Part (a): The Different Measures of Central Tendency

1. Arithmetic Mean. The arithmetic mean is the most familiar and most widely used average. It is obtained by dividing the sum of all observations by the number of observations: $\bar{x} = \Sigma x/N$ for raw data and $\bar{x} = \Sigma fx/\Sigma f$ for a frequency distribution.

Its merits are that it is simple to understand and compute, it is based on every observation, it is rigidly defined, and it lends itself to further algebraic treatment — the combined mean of two groups, for instance, can be found from their separate means. Its great defect is that it is pulled towards extreme values: a few very large incomes or a few giant trees can drag the mean far away from the bulk of the data, so for strongly skewed distributions it may represent nobody. It also cannot be computed for open-ended distributions or qualitative data.

2. Median. The median is the middle value of the data when they are arranged in ascending or descending order: half the observations lie below it and half above it. For grouped data it is

WhatsApp Orders: +92 324 5836626

located in the median class by the (N/2)th item. The median is unaffected by extreme values, can be found for open-ended classes, and can even be located graphically from an ogive. It is therefore the preferred average for income, landholding and other skewed distributions. Its limitations are that it ignores the actual magnitudes of the observations (only their order matters), it is not based on every value, and it does not admit much algebraic treatment.

3. Mode. The mode is the value that occurs most frequently in the data — the most typical or most fashionable value. In the data of part (i), for example, 3 persons per house is the mode because it occurs in the largest number of houses. The mode is simple, requires no calculation beyond inspection (or a short formula for grouped data), is unaffected by extremes, and is often the most realistic 'typical' value for practical decisions such as the commonest family size or the most demanded seedling type. Its weaknesses are that it may not exist, may not be unique (a distribution can be bimodal), and is ill-defined when frequencies are irregular.

4. Geometric Mean. The geometric mean is the nth root of the product of n observations, GM = $(x_1x_2\ldots x_n)^{(1/n)}$. It is the correct average for ratios, percentages and rates of growth — for example the average annual growth rate of a plantation or of prices over several years — because growth compounds multiplicatively. It gives less weight to large values than the arithmetic mean, but it cannot be computed if any observation is zero or negative.

5. Harmonic Mean. The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals, HM = $N/\Sigma(1/x)$. It is used for averaging rates and speeds when the distance or quantity is fixed — for example the average speed of a vehicle covering equal distances at different speeds, or average price when equal money amounts are spent. It gives the greatest weight to the smallest values and is rarely appropriate for ordinary level data.

For a perfectly symmetrical distribution the three main averages coincide: Mean = Median = Mode. For a moderately skewed distribution they are connected by the empirical relation Mode = 3(Median) − 2(Mean), which is often used as a check. The choice among them is not a matter of taste: the mean suits symmetric data and further statistical work, the median suits skewed data and open-ended classes, and the mode suits the 'most typical case' question. A good average, in Yule's classic statement of requisites, should be clearly defined, based on all observations, simple to calculate and understand, capable of further algebraic treatment, and not unduly affected by extremes or by sampling fluctuations — no single measure satisfies all of these at once, which is why several are reported together.

Part (b): Calculation of Mean, Median and Mode

Here x = number of persons per house and f = number of houses. The products fx and the cumulative frequencies are computed first; the total number of houses is Σf = 500 and the total number of persons is Σfx = 1,888.

x (persons per house)	f (houses)	fx	Cumulative frequency
1	26	26	26
2	113	226	139
3	120	360	259
4	95	380	354
5	60	300	414
6	42	252	456
7	21	147	477
8	14	112	491
9	5	45	496
10	4	40	500
Total	500	1,888	—

Mean. Mean = Σfx / Σf = 1,888 / 500 = 3.776, that is, about 3.78 persons per house.

Median. With N = 500 observations arranged in order, the median is the value of the (N/2)th = 250th item (equivalently the average of the 250th and 251st items). From the cumulative frequencies, the 259 items up to x = 3 include the 250th and 251st items, and both of these items have the value 3. Hence Median = 3 persons per house.

Mode. The largest frequency in the table is 120, corresponding to x = 3. Hence Mode = 3 persons per house — the commonest household size is three persons. (Check by the empirical relation: 3 × Median − 2 × Mean = 9 − 7.552 = 1.448, which does not equal 3; the relation is only approximate and is meant for continuous grouped distributions, so the directly observed mode, 3, stands. The mean exceeding the median and mode shows the distribution is positively skewed, pulled upward by the minority of large households of 7–10 persons.)

Conclusion

The mean (3.78 persons per house) summarises the total population in terms of houses and is the figure needed for planning totals — 500 houses × 3.78 ≈ 1,888 persons — while the median and mode (both 3) describe the typical household, unaffected by the few very large families. Reported together, the three measures show both the centre and the positive skewness of household size, which is exactly the kind of complete picture a good statistical summary should give.

Question 3: Null and Alternative Hypotheses, Steps in Hypothesis Testing; Test of a Claimed Mean (10 + 10 Marks)

Question as printed: (a) Define Null Hypothesis and Alternative Hypothesis. Explain the steps involved in hypothesis testing. (b) A manufacturer claims that the average weight of rice bags produced by a machine is 50 kg. A sample of 49 bags gave a mean weight of 48.5 kg. The

population standard deviation is known to be 4 kg. Test at 5% level of significance whether the machine is producing bags with the claimed average weight.

Introduction

Statistical inference has two great branches: estimation, which guesses the value of a population parameter, and hypothesis testing, which judges a claim about the parameter against sample evidence. Almost every practical question in research and industry is a claim of this kind — a machine fills bags to 50 kg, a new fertiliser out-yields the old one, a teaching method is superior — and the sample data never speak with complete certainty, because another sample would give a slightly different result. Hypothesis testing provides a disciplined procedure for deciding whether the difference between what was claimed and what was observed is small enough to be ordinary sampling fluctuation, or so large that the claim itself must be rejected. Part (a) defines the two competing hypotheses and sets out the steps of the procedure; part (b) applies the procedure to a manufacturer's claim about rice bags.

Part (a): Null Hypothesis, Alternative Hypothesis and the Steps of Testing

Null hypothesis. The null hypothesis, denoted H_0, is the hypothesis of 'no difference' or 'no effect': it states that any apparent difference between the sample result and the population value (or between two samples) is due to chance sampling variation alone, and that the true parameter equals the claimed value. In part (b), H_0: $\mu = 50$ kg. The null hypothesis is the statement actually subjected to the test, and the test can only reject it or fail to reject it — a test never 'proves' H_0 true; it merely finds the evidence insufficient against it, just as a court that acquits has not proved innocence. H_0 is always stated as an equality, because the sampling distribution of the test statistic must be computed for one definite parameter value.

Alternative hypothesis. The alternative hypothesis, denoted H_1 (or H_a), is the hypothesis accepted if the null hypothesis is rejected; it states that a genuine difference exists. It may be two-sided (H_1: $\mu \neq 50$ — the bags differ from 50 kg in either direction), or one-sided (H_1: $\mu < 50$, the bags are underweight, or H_1: $\mu > 50$). The form of H_1 determines the critical region: a two-sided alternative puts half the significance level in each tail of the distribution, while a one-sided alternative puts it all in one tail. The alternative must be fixed from the purpose of the enquiry before the data are examined, never chosen afterwards to suit the figures.

Errors and steps. Two errors are possible in any test. A Type I error rejects a true H_0; its probability is the level of significance α, fixed in advance (commonly 5% or 1%). A Type II error accepts (fails to reject) a false H_0; its probability is β, and $1 - \beta$ is the power of the test. The steps of the standard testing procedure are the following.

- **Step 1 — State H_0 and H_1** clearly, in terms of the population parameter, choosing a one- or two-sided alternative as the problem requires.
- **Step 2 — Fix the level of significance α** (usually 5% or 1%), the risk of a Type I error the investigator is willing to run.
- **Step 3 — Choose the test statistic** appropriate to the problem and its assumptions (z when the population standard deviation is known, t when it is estimated from a small sample, chi-square for variances and independence, F for comparing several means), and note its sampling distribution under H_0.
- **Step 4 — Compute the test statistic** from the sample data.
- **Step 5 — Find the critical value(s)** from the tables at the chosen α (and degrees of freedom where relevant), defining the rejection (critical) region; equivalently, compute the p-value.

- **Step 6 — Decide and interpret:** if the computed statistic falls in the critical region (or $p < \alpha$), reject H_0 and accept H_1; otherwise do not reject H_0. State the conclusion in the language of the original problem, not merely in symbols.

Part (b): Testing the Manufacturer's Claim (z-test)

Given: claimed mean $\mu_0 = 50$ kg; sample size $n = 49$; sample mean $\bar{x} = 48.5$ kg; population standard deviation $\sigma = 4$ kg (known, so the z-test is the correct statistic); level of significance $\alpha = 5\%$. Step 1: H_0: $\mu = 50$ (the machine produces bags of the claimed average weight). H_1: $\mu \neq 50$ (two-sided, because the claim 'is producing bags with the claimed weight' fails if the bags are either underweight or overweight). Step 2: $\alpha = 0.05$. Step 3: the test statistic is $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$, which under H_0 follows the standard normal distribution. Step 4: the standard error is $\sigma/\sqrt{n} = 4/\sqrt{49} = 4/7 = 0.5714$ kg, and $z = (48.5 - 50)/0.5714 = -1.5/0.5714 = -2.625$. Step 5: the two-sided critical value at 5% is $z = \pm 1.96$, so H_0 is rejected if $|z| > 1.96$ (the corresponding p-value is about 0.0087, well below 0.05). Step 6: since $|z| = 2.625 > 1.96$, the result falls in the critical region and H_0 is rejected at the 5% level.

Interpretation: the shortfall of 1.5 kg is 2.625 standard errors away from the claimed mean — a deviation this large would arise by pure chance in fewer than 9 samples out of 1,000 if the machine were truly averaging 50 kg. The evidence is therefore strong that the machine is not producing bags of the claimed average weight; on average the bags are underweight (48.5 kg against a 50 kg claim, a deficiency of 3 per cent), and the machine should be checked and recalibrated. Note carefully what the test does and does not say: it does not prove every bag is underweight, nor even that μ is exactly 48.5; it says the claim $\mu = 50$ is inconsistent with the sample evidence at the stated level of risk (5%) of a Type I error.

Conclusion

Hypothesis testing converts a vague question — 'does the evidence contradict the claim?' — into a six-step procedure with a controlled risk of error: state H_0 and H_1, fix α, choose and compute the statistic, compare with the critical value, and interpret. Applied to the rice bags, $z = -2.625$ exceeds the 5% critical value 1.96, so the manufacturer's claim of a 50 kg average is rejected: the machine is underfilling.

Question 4: One-Sample versus Two-Sample Tests; Comparing Two Teaching Methods (8 + 12 Marks)

Question as printed: (a) Differentiate between one-sample and two-sample hypothesis testing. Discuss their applications. (b) The following data represent the marks obtained by two groups of students taught by different teaching methods: A: 65, 70, 68, 72, 75, 78, 74, 69; B: 60, 62, 64, 66, 68, 65, 63, 61. Test whether there is a significant difference between the mean marks of the two groups. Interpret the result.

Introduction

Question 3 tested one sample against a standard fixed in advance. Very often, however, there is no fixed standard: the investigator has two samples — two teaching methods, two fertilisers, two varieties, two machines — and wants to know whether the populations behind them differ. The machinery of hypothesis testing is the same, but the statistic, the standard error and the applications differ, and confusing the two designs is a common examination and practical error. Part (a) sets out the differences systematically and discusses where each design is used; part (b) carries out a two-sample (independent-samples) t-test on the marks of two groups of eight students.

Part (a): One-Sample and Two-Sample Testing Compared

Point of difference	One-sample test	Two-sample test
Question asked	Does this population's parameter equal a specified/claimed value?	Do the parameters of two populations differ from each other?
Data	One random sample	Two samples (independent, or paired observations)
Hypotheses	H_0: $\mu = \mu_0$ (a fixed constant)	H_0: $\mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0$
Standard error	Based on the single sample: $\sigma/\sqrt{n}$	Based on the variability of both samples combined (pooled where variances are equal)
Typical statistics	z or t with $n - 1$ d.f.	Two-sample t with $n_1 + n_2 - 2$ d.f.; paired t for matched pairs; F in ANOVA extensions
Role of the standard	External: a claim, specification, or past standard	Internal: the other sample itself is the standard

- **One-sample applications:** quality control against a specification (the rice bags of Question 3, filled to a claimed 50 kg); testing whether this year's mean yield equals the long-run average; whether a coin or die is fair; whether the mean germination of a seed lot meets the certified standard; whether students' mean score reaches a required pass standard.
- **Two-sample applications:** comparing two teaching methods, drugs, fertilisers, crop varieties, machines, or training programmes; comparing the mean incomes of two regions or the mean heights of two species; before-and-after studies using the paired form (same plots measured before and after treatment); clinical trials comparing a new drug with a standard drug.
- **Paired versus independent:** a two-sample problem in which each observation in one sample is naturally matched with one in the other (same student before and after, twin plots) must use the paired t-test on the differences; treating paired data as independent wastes information and can reverse the conclusion.

- **Assumptions:** both designs assume random sampling and (for the small-sample t forms) approximately normal populations; the independent two-sample t additionally assumes equal population variances when the pooled form is used, and independence between the samples.

In essence, the one-sample test measures a sample against a yardstick supplied from outside, while the two-sample test measures two samples against each other, the variability of both entering the standard error. The two-sample question is the more common one in experimental science — nearly every experiment compares a treatment with a control or with a rival treatment — and it is the gateway to the analysis of variance used for several groups at once in Assignment No. 2.

Part (b): Two-Sample t-Test on the Marks

Let group A (method A) be sample 1 and group B (method B) sample 2, with $n_1 = n_2 = 8$. Hypotheses: H_0: $\mu_1 = \mu_2$ (the two methods give the same mean marks); H_1: $\mu_1 \neq \mu_2$ (two-sided), at the 5% level. The samples are small and the population variances are unknown, so the two-sample t-test with pooled variance is used.

Quantity	Group A	Group B
Observations	65, 70, 68, 72, 75, 78, 74, 69	60, 62, 64, 66, 68, 65, 63, 61
Sum Σx	571	509
Mean $\bar{x}$	571/8 = 71.375	509/8 = 63.625
Sum of squares Σx^2	40,879	32,439
Sample variance $s^2 = [\Sigma x^2 - (\Sigma x)^2/n]/(n-1)$	$(40,879 - 40,755.125)/7 = 17.696$	$(32,439 - 32,385.125)/7 = 7.125$

Pooled variance: $s_p^2 = [(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2) = [7(17.696) + 7(7.125)]/14 = (123.875 + 49.875)/14 = 173.75/14 = 12.411$, so $s_p = 3.523$. Standard error of the difference of means $= s_p\sqrt{(1/n_1 + 1/n_2)} = 3.523 \times \sqrt{(0.25)} = 3.523 \times 0.5 = 1.761$. Test statistic: $t = (\bar{x}_1 - \bar{x}_2)/SE = (71.375 - 63.625)/1.761 = 7.75/1.761 = 4.40$, with $n_1 + n_2 - 2 = 14$ degrees of freedom. The two-sided

critical value $t_{0.05}(14) = 2.145$. Since the calculated $t = 4.40 > 2.145$ ($p \approx 0.0006$), H_0 is rejected at the 5% level — indeed the difference is significant even at the 1% level, whose critical value is 2.977.

Interpretation: the observed advantage of method A is 7.75 marks on average (71.38 against 63.63), which is more than four times its standard error, a gap this large would occur by chance in only about 6 samples in 10,000 if the methods were equally effective. There is therefore a real, statistically significant difference between the two teaching methods, and on this evidence method A is the more effective one. Two cautions complete an honest interpretation. First, significance is not the same as size or importance in every context, though a 7.75-mark gap on this scale is plainly substantial as well as significant. Second, the conclusion rests on the assumptions of the test — independent random samples of students and roughly normal mark distributions with a common variance; if the groups differed systematically beforehand (for example, if the abler students had been assigned to method A), the difference could reflect selection rather than teaching. Subject to proper random allocation, method A should be preferred.

Conclusion

One-sample tests compare a sample with an external standard; two-sample tests compare two populations through their samples, pooling the information about variability. For the two teaching methods the pooled t statistic is 4.40 on 14 degrees of freedom, far beyond the 5% critical value 2.145, so the difference in mean marks is highly significant and method A is judged superior.

Question 5: The Linear Regression Model and Its Assumptions; Least-Squares Trend and Forecast (8 + 12 Marks)

Question as printed: (a) What is a linear regression model? Explain the assumptions underlying the linear regression model. (b) The following data represent the annual profit (in million rupees) of a company from 2015 to 2024: Year 2015–2024, Profit 30, 34, 36, 40, 44, 47, 50, 54, 57, 61. Fit a straight-line trend equation using the least squares method and estimate the trend values for each year, and forecast the profit for the year 2025.

Introduction

Many statistical problems concern not a single variable but the relationship between variables: how yield responds to fertiliser, how profit grows over time, how the girth of a tree depends on its age. Regression analysis is the technique that expresses such a relationship as an equation, so that the value of one variable can be estimated from another. Part (a) defines the linear regression model and states, one by one, the assumptions on which its validity rests — assumptions that are frequently recited but rarely understood, although every conclusion drawn from a fitted line depends on them. Part (b) fits the simplest regression of all, a straight-line trend of profit on time, by the method of least squares, computes the trend value for each of the ten years 2015–2024, and forecasts the profit for 2025.

Part (a): The Linear Regression Model and Its Assumptions

A linear regression model is a statistical model that expresses the dependent (response) variable Y as a linear function of one or more independent (explanatory) variables plus a random error. For one explanatory variable the model is $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 is the intercept (the value of Y when X = 0), β_1 is the slope or regression coefficient (the average change in Y for a unit

change in X), and ε is the random error term capturing all influences on Y other than X. 'Linear' refers to linearity in the parameters β. The fitted line $\hat{Y} = a + bX$ is obtained by least squares, choosing a and b to minimise the sum of squared vertical deviations $\Sigma(Y - \hat{Y})^2$; with several explanatory variables the model extends to multiple regression, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$. The model serves two purposes: explanation (measuring how strongly and in what direction X affects Y) and prediction (estimating Y for a given X). Its classical assumptions are as follows.

- **Linearity:** the true relationship between Y and X is linear in the parameters; the conditional mean of Y, E(Y|X), lies on the straight line $\beta_0 + \beta_1 X$. A curved relationship forced into a straight line gives biased estimates and misleading predictions.
- **Fixed (non-stochastic) X:** the values of the explanatory variable are fixed by the design or measured without error, and are not themselves random outcomes correlated with the error.
- **Zero mean of errors:** $E(\varepsilon) = 0$ for every X; the error term does not systematically push Y up or down.
- **Homoscedasticity (constant variance):** $\text{Var}(\varepsilon) = \sigma^2$, the same for all observations. If the scatter of errors widens with X (heteroscedasticity), the estimates remain unbiased but their standard errors, and hence the tests, become unreliable.
- **No autocorrelation:** the errors of different observations are independent, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. This is especially important for time-series data such as the profits of part (b), where one year's disturbance may carry into the next.
- **Normality of errors:** ε is normally distributed, $\varepsilon \sim N(0, \sigma^2)$. This is not needed for the least-squares estimates themselves but is required for exact t- and F-tests and confidence intervals in small samples.

- **No perfect multicollinearity (multiple regression):** no explanatory variable is an exact linear combination of the others; otherwise individual coefficients cannot be separated or estimated.
- **Correct specification and adequate data:** no important variable is omitted and no irrelevant one distorts the model, and the number of observations exceeds the number of parameters estimated.

Under assumptions 1–5 (with 6 added for inference) the least-squares estimators are BLUE — the Best Linear Unbiased Estimators (the Gauss–Markov theorem): among all linear unbiased estimators they have the smallest variance. This is why the assumptions matter practically and not merely formally: violate homoscedasticity or independence and the line may still be drawn, but the standard errors attached to it, and every significance test and forecast interval built on them, cease to be trustworthy.

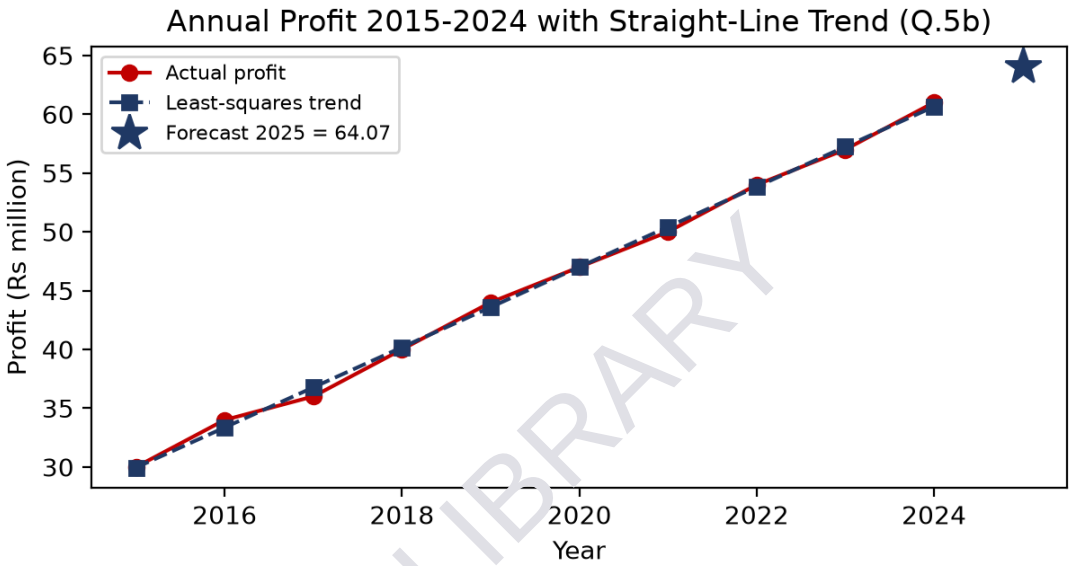
AIOU LIBRARY aioulibrary.com

Part (b): Straight-Line Trend by Least Squares, 2015–2024

Let X denote time and Y the profit (Rs million). Coding time as u = −9, −7, −5, −3, −1, +1, +3, +5, +7, +9 for 2015–2024 makes $\Sigma u = 0$, so the least-squares equations simplify to $a = \Sigma Y/n = \bar{Y}$ and $b = \Sigma uY/\Sigma u^2$. From the table below, $\Sigma Y = 453$, so $a = 453/10 = 45.3$; $\Sigma uY = 563$ and $\Sigma u^2 = 330$, so $b = 563/330 = 1.706$ per unit of u. Since one unit of u is half a year, the annual increase in profit is $2 \times 1.706 = 3.412$ million rupees per year. The trend equation, with origin at the middle of the period (mid-2019/2020, i.e. the point between 2019 and 2020), is therefore $\hat{Y} = 45.3 + 1.706u$, or equivalently, with X = calendar year, $\hat{Y} = -6{,}843.04 + 3.4121 \times \text{Year}$ — the two forms give identical values; the coded form is far easier to use. (Equally, with t = 0 for 2015 … 9 for 2024, the equation is $\hat{Y} = 29.95 + 3.412t$.)

Year	u	Profit Y	uY	u^2	Trend value $\hat{Y} = 45.3 + 1.706u$
2015	−9	30	−270	81	29.95
2016	−7	34	−238	49	33.36
2017	−5	36	−180	25	36.77
2018	−3	40	−120	9	40.18
2019	−1	44	−44	1	43.59
2020	+1	47	47	1	47.01
2021	+3	50	150	9	50.42
2022	+5	54	270	25	53.83
2023	+7	57	399	49	57.24
2024	+9	61	549	81	60.65
Total	0	453	563	330	453.00

The trend values are listed in the last column; as a check, they sum to 453.00, exactly the sum of the actual profits, which must always hold for a least-squares line with an intercept. The actual profits oscillate only slightly around the line (residuals are all within about ±1 million), showing that growth has been remarkably steady at about Rs 3.41 million per year. For the forecast: the year 2025 corresponds to u = +11, giving $\hat{Y}_{2025} = 45.3 + 1.706(11) = 45.3 + 18.77 = 64.07$. The forecast profit for 2025 is therefore approximately Rs 64.07 million (about Rs 64 million). This forecast assumes the same linear growth pattern continues — steady market conditions and no structural break — and, like all extrapolations beyond the observed range, it should be used with corresponding caution: the further beyond 2024 the forecast is pushed, the less reliable it becomes.



Conclusion

The linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$, supported by its assumptions of linearity, fixed X, zero-mean, constant-variance, independent and (for inference) normally distributed errors, is the workhorse of relationship analysis. Fitted by least squares to the company's profits, the trend $\hat{Y} = 45.3 + 1.706u$ shows steady growth of about Rs 3.41 million a year, reproduces the ten observed years closely (trend values summing exactly to the actual total of 453), and forecasts a 2025 profit of about Rs 64.07 million.