ALLAMA IQBAL OPEN UNIVERSITY

AIOU LIBRARY

Name:

Registration No.:

Course Code: 1350

Program: I.Com

Semester: Autumn 2026

ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

(Department of Statistics)

Introduction to Business Statistics (1350)

Level: I.Com

Semester: Autumn 2026

Assignment No. 2

Total Marks: 100 | Pass Marks: 40

SOLVED ASSIGNMENT GUIDE — For Study & Reference

Note: This guide is for study and reference only. Students are advised to read the relevant units of the course book, understand each answer, and then write the assignment in their own words, in their own handwriting where required. Copying word for word is discouraged; your own analysis and synthesis will be appreciated, as instructed in the assignment booklet itself.

Question 1: a) Define various measures of dispersion and give their formulae. b) The following table gives the marks of students: Marks 30–39 (f=8), 40–49 (f=87), 50–59 (f=190), 60–69 (f=86), 70–79 (f=20). Calculate quartile deviation and mean deviation from the mean. (8+12 = 20 marks)

Introduction

An average by itself never describes a distribution completely. Two classes may both average 55 marks, yet in one every student lies between 50 and 60 while in the other marks stretch from 10 to 95; the teaching problem, the risk and the meaning of the average are entirely different. Dispersion — also called variation or spread — is the property that measures this difference,

and part (a) of this question defines the standard measures of dispersion with their formulae. Part (b) then computes two of them, the quartile deviation and the mean deviation from the mean, for a substantial grouped distribution of the marks of 391 students. The two measures chosen by the examiner are instructive companions: the quartile deviation is a positional measure that ignores the tails altogether, while the mean deviation uses every observation but treats all deviations alike in size. Comparing their values, 5.41 and 6.13 respectively, and their coefficients, is part of the answer, not an ornament added after it.

Part (a): Measures of Dispersion — Definitions and Formulae

Measures of dispersion are of two families. Absolute measures express spread in the original units of the data — marks, rupees, kilograms — and can be compared only between distributions measured in the same unit on a similar scale. Relative measures, usually called coefficients, are pure numbers obtained by dividing an absolute measure by a suitable average; they free the comparison from units, so that the variability of marks can be compared with the variability of wages. Every absolute measure below therefore has a relative companion, and a complete answer names both. The standard measures are the following.

(1) Range. The range is the difference between the largest and the smallest observation: Range = L − S, where L is the largest and S the smallest value; for grouped data it is approximated by the upper boundary of the last class minus the lower boundary of the first. Its coefficient is (L − S) ÷ (L + S). The range is the simplest measure — computable at a glance — and the crudest, since it uses only two observations and ignores everything between them; a single freak value can double it.

(2) Quartile deviation (semi-interquartile range). The quartile deviation is half the difference between the third and the first quartiles: Q.D. = $(Q_3 - Q_1) \div 2$, where Q_1 and Q_3 enclose the middle 50 percent of the observations. Its coefficient is $(Q_3 - Q_1) \div (Q_3 + Q_1)$. Because it depends only on the middle half of the data, it is unaffected by extreme values and computable even for open-ended distributions, at the price of ignoring the outer half of the observations entirely.

(3) Mean deviation (average deviation). The mean deviation is the arithmetic mean of the absolute deviations of the observations from a suitable average, usually the mean or the median: for ungrouped data, M.D. = $\Sigma|X - A| \div n$, and for grouped data, M.D. = $\Sigma f|X - A| \div \Sigma f$, where A is the chosen average and the vertical bars denote that signs are ignored. Its coefficient is M.D. ÷ A. The mean deviation uses every observation — an advance on the range and quartile deviation — but it achieves mathematical simplicity only by discarding signs, which limits its further algebraic development.

(4) Standard deviation and variance. The standard deviation, denoted by σ, is the root-mean-square deviation from the arithmetic mean: $\sigma = \sqrt{[\Sigma(X - \text{Mean})^2 \div N]}$ for ungrouped data and $\sigma = \sqrt{[\Sigma f(X - \text{Mean})^2 \div \Sigma f]}$ for grouped data; its square, σ^2, is the variance. Its coefficient is σ ÷ Mean, and multiplied by 100 it becomes the coefficient of variation, C.V. = (σ ÷ Mean) × 100, the most widely used relative measure in business comparisons. The standard deviation is regarded as the best measure of dispersion because it uses all observations, respects signs through squaring, is rigidly defined, and supports further theory, including the empirical rule for bell-shaped data and the regression and correlation theory of this very assignment.

(5) Lorenz curve and Gini coefficient may be mentioned briefly as a graphical and derived measure of inequality used for incomes and wealth, completing the picture. A sound

comparison of measures closes this part: the range is quick but fragile; the quartile deviation is robust but wasteful of data; the mean deviation is balanced and intelligible but algebraically limited; and the standard deviation is the most complete but the most sensitive to extremes through its squaring. Choosing among them follows the same logic as choosing an average in Assignment No. 1 — purpose, shape of data and further analysis decide.

Part (b): Quartile Deviation and Mean Deviation – Calculation

The printed classes are inclusive (30–39, 40–49, …), the standard first step is therefore the continuity correction. The gap between successive classes is 1 (40 − 39), half of it is 0.5, and subtracting 0.5 from each lower limit and adding 0.5 to each upper limit gives the true classes 29.5–39.5, 39.5–49.5, 49.5–59.5, 59.5–69.5 and 69.5–79.5, each of width 10, with mid-values 34.5, 44.5, 54.5, 64.5 and 74.5. Frequencies are 8, 87, 190, 86 and 20, giving N = Σf = 391. The working table carries every column the two measures need: fX for the mean (required both as the centre for the mean deviation and as a check), cumulative frequency for the quartiles, and f|X − Mean| for the mean deviation.

Table 1: Working table (Mean = 21,539.5 ÷ 391 = 55.09; deviations taken from 55.0882 before rounding).

True class	f	Mid-value X	fX	c.f.	|X – Mean|	f|X – Mean|
29.5–39.5	8	34.5	276.0	8	20.59	164.71
39.5–49.5	87	44.5	3,871.5	95	10.59	921.18
49.5–59.5	190	54.5	10,355.0	285	0.59	111.76
59.5–69.5	86	64.5	5,547.0	371	9.41	809.41
69.5–79.5	20	74.5	1,490.0	391	19.41	388.24

Total	391	—	21,539.5	—	—	2,395.29

Step 1 — The mean. Mean = ΣfX ÷ Σf = 21,539.5 ÷ 391 = 55.0882…, so Mean ≈ 55.09 marks. This value is required for the mean deviation, and it is also a useful landmark: a mean of about 55 in classes centred at 54.5 already suggests a distribution concentrated in the middle class, which the quartiles will now confirm precisely.

Step 2 — Quartiles and quartile deviation. The first quartile corresponds to the (N ÷ 4)th = 97.75th observation. The cumulative frequencies show 95 observations below 49.5 and 285 below 59.5, so the Q_1 class is 49.5–59.5. Interpolating with Q = L + [(position − C) ÷ f] × h, where L = 49.5, C = 95, f = 190 and h = 10: Q_1 = 49.5 + [(97.75 − 95) ÷ 190] × 10 = 49.5 + 0.14 = 49.64. The third quartile corresponds to the (3N ÷ 4)th = 293.25th observation. Since the cumulative frequency reaches 285 at 59.5 and 371 at 69.5, the 293.25th observation lies in the class 59.5–69.5, so with L = 59.5, C = 285, f = 86 and h = 10: Q_3 = 59.5 + [(293.25 − 285) ÷ 86] × 10 = 59.5 + 0.96 = 60.46. Locating the correct quartile class against the cumulative frequency column is the step where errors most commonly enter, and it has been checked twice here. The interquartile range is $Q_3 - Q_1$ = 60.46 − 49.64 = 10.81, and the Quartile Deviation = ($Q_3 - Q_1$) ÷ 2 = 10.81 ÷ 2 = 5.41 marks, with coefficient ($Q_3 - Q_1$) ÷ ($Q_3 + Q_1$) = 10.81 ÷ 110.10 = 0.098. For reference, the median by the same method is 49.5 + [(195.5 − 95) ÷ 190] × 10 = 54.79, which lies between Q_1 and Q_3 as it must.

Step 3 — Mean deviation from the mean. From the last column of the working table, Σf|X − Mean| = 2,395.29 (computed with the unrounded mean 55.0882; using 55.09 throughout changes the total negligibly). Hence Mean Deviation = Σf|X − Mean| ÷ N = 2,395.29 ÷ 391 = 6.13

marks, and its coefficient is 6.13 ÷ 55.09 = 0.111. The interpretation completes the calculation. On average, a student's marks deviate from the mean of 55.09 by about 6.13 marks — roughly 11 percent of the mean — while the middle half of the students is enclosed between about 49.6 and 60.5 marks, a band whose half-width is the quartile deviation of 5.41. The two measures agree in spirit: this is a moderately concentrated distribution, dominated by the great central class of 190 students (nearly half of all 391), with comparatively thin tails of 8 and 20 students at the extremes. Because both tails are thin, the mean deviation (which notices the tails) exceeds the quartile deviation (which ignores them) only modestly — exactly the relationship theory predicts — and either figure, honestly reported with its coefficient, gives a fair picture of spread to accompany the mean.

Conclusion

Dispersion completes what the average begins. The range, quartile deviation, mean deviation and standard deviation form a family of increasing refinement, each with its formula, its absolute value in the units of the data and its relative coefficient for unit-free comparison. For the marks of 391 students (Mean = 55.09), the quartile deviation is 5.41 marks (coefficient 0.098) and the mean deviation from the mean is 6.13 marks (coefficient 0.111): a concentrated, middle-heavy distribution whose spread is about one-tenth of its level by either measure. Averages locate a distribution; measures of dispersion give it width — and no business decision based on an average alone is fully informed.

Question 2: a) What is an index number? Give the uses of an index number. b) Compute the weighted index number for 1964 from the following data with 1960 as the base: Milk — price

1960: 3.95, 1964: 4.25, quantity 1960: 97.75, 1964: 104.36; Cheese — 34.80, 38.90, 78, 83; Butter — 61.56, 59.70, 118, 116. (10+10 = 20 marks)

Introduction

Prices do not stand still, and they do not move together: in the data of part (b), milk and cheese became dearer between 1960 and 1964 while butter became cheaper. No single price can summarise such a movement, and a simple average of the prices themselves would be meaningless because the commodities are sold in different units and quantities. The index number is the statistical device invented for exactly this problem: it compresses the combined movement of a group of related variables into one comparable figure. Part (a) defines the device and gives its uses. Part (b) computes the weighted index number for 1964 on base 1960. Because the booklet asks for 'the weighted index number' without naming a formula, this guide computes the standard base-year weighted aggregative index — Laspeyres' index — as the principal answer, and also shows Paasche's current-year weighted index and Fisher's ideal index, so that the student can see the whole family and state, transparently, which formula any quoted figure comes from. The three answers, 101.25, 101.49 and 101.37, fortunately tell one consistent story.

Part (a): Index Number — Definition and Uses

An index number is a statistical measure that expresses the relative change in a variable, or in a group of related variables, over time (or between places) with respect to a chosen base period, whose value is conventionally set equal to 100. A price index of 101.25 for 1964 with 1960 = 100 means that the price level of the group, weighted as specified, stood about 1.25 percent higher in 1964 than in 1960. The essential features of the concept follow from the definition. An

index number is always relative — a ratio multiplied by 100 — never an absolute amount. It is unit-free, which is what allows milk, cheese and butter to be combined. It refers to a group, not to one commodity alone, except in the elementary case of a price relative. It depends on a base, and the base should be a normal period, free of shocks, and not too remote. And in any serious application it is weighted: commodities matter in proportion to their importance, measured by the quantities bought or sold, because a small rise in the price of a staple consumed daily affects a budget more than a large rise in the price of a rarity.

The construction of an index number involves deliberate choices that the student should be able to name: the purpose of the index; the selection of commodities representative of that purpose; the collection of their prices from reliable sources; the choice of base period; the choice of weights; and the choice of formula — simple aggregative, weighted aggregative (Laspeyres with base-year quantities, Paasche with current-year quantities, Fisher as their geometric mean), or a weighted average of price relatives. Each choice affects the result, which is why two honest indices of the same phenomenon can differ slightly, and why an index number quoted without its base, coverage and formula is incomplete information. The limitations deserve equal honesty: an index measures change in an average sense for a group and conceals opposite movements inside it (butter fell while the index rose); its accuracy is bounded by the price and quantity data; and comparisons over very long periods strain under changing tastes, quality and the appearance of new goods.

The uses of index numbers in business, commerce and public life are extensive. (1) Cost of living (consumer price) indices measure changes in the cost of maintaining a standard of living and directly govern dearness allowances, wage negotiations, rent adjustments and the revision of

minimum wages. (2) Wholesale price indices give early warning of inflation and guide monetary and fiscal policy. (3) Industrial production indices measure the volume of output and the health of industry month by month. (4) Agricultural production and price indices guide food policy, support prices and import decisions. (5) Import and export price and volume indices reveal a country's terms of trade. (6) Share price indices summarise stock-market movement for investors and analysts. (7) Business managers use indices as deflators: dividing a money value such as sales by a price index converts it into a real, volume measure, separating genuine growth from mere price rise — one of the most valuable routine uses in commerce. (8) Governments and courts use indices to adjust pensions, contracts, tax brackets and long-term payments fairly. (9) Regional and international comparisons of price levels rest on spatial index numbers. In short, wherever change over time in a complex aggregate must be watched, summarised and acted upon, the index number is the standard instrument.

Part (b): Weighted Index Number for 1964 (Base 1960)

Notation: p_o, q_o are price and quantity in the base year 1960; p_1, q_1 are price and quantity in the current year 1964. The four product columns needed by the weighted formulae are computed commodity by commodity in the working table. Milk: $p_oq_o = 3.95 \times 97.75 = 386.11$; $p_1q_o = 4.25 \times 97.75 = 415.44$; $p_oq_1 = 3.95 \times 104.36 = 412.22$; $p_1q_1 = 4.25 \times 104.36 = 443.53$. Cheese: $p_oq_o = 34.80 \times 78 = 2{,}714.40$; $p_1q_o = 38.90 \times 78 = 3{,}034.20$; $p_oq_1 = 34.80 \times 83 = 2{,}888.40$; $p_1q_1 = 38.90 \times 83 = 3{,}228.70$. Butter: $p_oq_o = 61.56 \times 118 = 7{,}264.08$; $p_1q_o = 59.70 \times 118 = 7{,}044.60$; $p_oq_1 = 61.56 \times 116 = 7{,}140.96$; $p_1q_1 = 59.70 \times 116 = 6{,}925.20$. The column totals are $\Sigma\, p_oq_o = 10{,}364.59$; $\Sigma\, p_1q_o = 10{,}494.24$; $\Sigma\, p_oq_1 = 10{,}441.58$; $\Sigma\, p_1q_1 = 10{,}597.43$, and every total should be re-added by the student as a first check.

Table 2: Working table for the weighted index numbers (figures rounded to 2 decimals; ratios computed before rounding).

Commodity	p_0q_0	p_1q_0	p_0q_1	p_1q_1
Milk	386.11	415.44	412.22	443.53
Cheese	2,714.40	3,034.20	2,888.40	3,228.70
Butter	7,264.08	7,044.60	7,140.96	6,925.20
Total Σ	10,364.59	10,494.24	10,441.58	10,597.43

Laspeyres' index (base-year weights) — the principal answer. Laspeyres' price index = ($\Sigma p_1q_0 \div \Sigma p_0q_0$) × 100 = (10,494.24 ÷ 10,364.59) × 100 = 101.25. It answers the question: what would the 1960 basket of quantities cost at 1964 prices, as a percentage of what it cost at 1960 prices? The same figure is obtained, as a check, by the weighted average of price relatives with base-year value weights p_0q_0: the relatives are 107.59 for milk, 111.78 for cheese and 96.98 for butter, and weighting them by 386.11, 2,714.40 and 7,264.08 returns exactly 101.25 — agreement between the aggregative and relative routes, as theory requires. Note how the weighting governs the story: butter dominates the basket by value (about 70 percent of base expenditure), so its price fall restrains the index powerfully, even though cheese rose nearly 12 percent.

Paasche's index (current-year weights) = ($\Sigma p_1q_1 \div \Sigma p_0q_1$) × 100 = (10,597.43 ÷ 10,441.58) × 100 = 101.49. It answers the mirror question for the 1964 basket. Fisher's ideal index, the geometric mean of the two, = √(101.25 × 101.49) = 101.37; it is called ideal because it satisfies the time-reversal and factor-reversal tests that the other two fail separately. For completeness, the simple (unweighted) aggregative index would be ($\Sigma p_1 \div \Sigma p_0$) × 100 = (102.85 ÷ 100.31) × 100 =

102.53, noticeably higher because it treats a unit of milk as equal in importance to a unit of butter — a clear demonstration, on the question's own numbers, of why weighting matters and why the booklet rightly demands the weighted index. Interpretation: taking 1960 as 100, the weighted price level of this three-commodity basket in 1964 was approximately 101.3 — a modest net rise of about 1.3 percent over four years, composed of rises of about 7.6 percent (milk) and 11.8 percent (cheese) partly offset by a fall of about 3.0 percent (butter), with the exact figure depending slightly on the weighting formula chosen, as stated openly above.

Conclusion

An index number is a unit-free relative, set on a base of 100, that summarises the combined change of a group of variables; weighted by quantities, it becomes the indispensable instrument of cost-of-living adjustment, inflation watching, production measurement, contract escalation and the deflation of money values into real terms. On the given data the weighted index number for 1964 on base 1960 is 101.25 by Laspeyres' base-weighted formula (Paasche 101.49; Fisher 101.37): the basket's price level rose only slightly overall, because butter — the dominant item by value — became cheaper while milk and cheese became dearer. Naming the formula beside the figure is part of the answer, and this guide has done so throughout.

Question 3: a) What is a scatter diagram? Describe its role in the theory of regression. b) The following data shows the son's height and the father's height. Father's height (X): 59, 61, 63, 65, 67, 69, 71, 73, 75. Son's height (Y): 64, 66, 67, 67, 68, 69, 70, 72, 72. Estimate the regression line y on x. (10+10 = 20 marks)

Introduction

When two variables are observed in pairs — a father's height with his son's, advertising outlay with resulting sales, study hours with marks — the first scientific question is whether they move together at all, and the first instrument for answering it is a picture: the scatter diagram. Part (a) defines the scatter diagram and explains its foundational role in the theory of regression: it is at once the explorer that reveals association, the diagnostic that reveals its form, and the canvas on which the regression line is eventually drawn. Part (b) applies the full least-squares method to nine father–son height pairs, the classic material on which Galton originally built the idea of regression. The fitted line, $\hat{Y} = 35.95 + 0.4833\ X$, is derived with every sum shown, checked through the mean point, and interpreted — including its most famous lesson, regression towards the mean, which the slope of less than one expresses in numbers.

Part (a): Scatter Diagram and Its Role in Regression Theory

A scatter diagram (also called a scatter plot or dot diagram) is a diagram in which paired observations on two variables are plotted as points on a rectangular coordinate system, the independent or explanatory variable X along the horizontal axis and the dependent variable Y along the vertical axis, one point for each pair. No line is drawn and nothing is computed; the cloud of points itself is the information. Its reading follows standard patterns. If the points drift upward from lower-left to upper-right, association is positive; if they drift downward, it is negative. If the drift follows a straight band, the relationship is approximately linear; if the band curves, it is curvilinear. The tighter the band around an imagined central line, the stronger the association; a formless circular cloud indicates little or no association; and a point standing far from the cloud — an outlier — warns of an unusual observation or a recording error that deserves investigation before any calculation proceeds.

The role of the scatter diagram in the theory of regression is foundational in at least five respects. First, it is the preliminary test of whether regression analysis is worth attempting at all: fitting a line to a formless cloud produces a slope near zero and predictions no better than the mean, and the diagram reveals this before effort is wasted. Second, it dictates the form of the model: regression theory offers the straight line $Y = a + bX$ only when the cloud is roughly straight; a visibly curved cloud calls for a transformed variable or a curve, and forcing a line through it yields systematically wrong predictions — too high at the ends, too low in the middle, or the reverse. Third, it motivates the line itself: the regression line is defined as the line that best summarises the cloud, and its least-squares criterion — minimising the sum of squared vertical distances from points to line — is a response to the geometry the diagram displays. The mean point $(\bar{X}, \bar{Y})$, through which the fitted line must pass, is the centre of gravity of the cloud, visible on the diagram before it is proved in algebra. Fourth, the diagram underwrites interpretation: the spread of points around the line is the visual counterpart of the correlation coefficient and of the standard error of estimate, and the danger of extrapolation — predicting beyond the range of X actually observed, where the cloud gives no evidence — is obvious on the picture even to a beginner. Fifth, historically and pedagogically, it is where regression began: Galton's plots of sons' heights against fathers' heights showed the cloud sloping upward but less steeply than the line of equality, the discovery he named regression towards mediocrity. The scatter diagram is thus not a preliminary decoration but the evidentiary basis of the whole regression method, and Figure 5 in part (b) should be studied in that spirit.

Part (b): Regression Line of Y on X — Calculation

The regression line of Y on X by the method of least squares is $\hat{Y} = a + bX$, where $b = \Sigma(X - \bar{X})(Y - \bar{Y}) \div \Sigma(X - \bar{X})^2$ and $a = \bar{Y} - b\bar{X}$. The data contain n = 9 pairs. The sums are: $\Sigma X = 59 + 61 + 63 + 65 + 67 + 69 + 71 + 73 + 75 = 603$, so $\bar{X} = 603 \div 9 = 67.00$ exactly. $\Sigma Y = 64 + 66 + 67 + 67 + 68 + 69 + 70 + 72 + 72 = 615$, so $\bar{Y} = 615 \div 9 = 68.333$. Because $\bar{X}$ is a whole number and the X values are symmetric about 67, the deviations $(X - \bar{X})$ are the tidy sequence −8, −6, −4, −2, 0, 2, 4, 6, 8, whose squares sum to $\Sigma(X - \bar{X})^2 = 64 + 36 + 16 + 4 + 0 + 4 + 16 + 36 + 64 = 240$. The cross-products $(X - \bar{X})(Y - \bar{Y})$, computed pair by pair from the deviations of Y about 68.333 (namely −4.333, −2.333, −1.333, −1.333, −0.333, 0.667, 1.667, 3.667, 3.667), sum to $\Sigma(X - \bar{X})(Y - \bar{Y}) = 116.00$ exactly — a value the student may verify by the equivalent raw-sum identity $\Sigma XY - n\bar{X}\bar{Y}$, with $\Sigma XY = \Sigma(X \cdot Y) = 41{,}321$ and $n\bar{X}\bar{Y} = 9 \times 67 \times 68.333 = 41{,}205$, giving $41{,}321 - 41{,}205 = 116$, in exact agreement.

Table 3: Working table for the regression of Y on X ($\bar{X}$ = 67.00; $\bar{Y}$ = 68.333).

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
59	64	−8	−4.333	34.667	64
61	66	−6	−2.333	14.000	36
63	67	−4	−1.333	5.333	16
65	67	−2	−1.333	2.667	4
67	68	0	−0.333	0.000	0
69	69	2	0.667	1.333	4
71	70	4	1.667	6.667	16
73	72	6	3.667	22.000	36
75	72	8	3.667	29.333	64

ΣX = 603	ΣY = 615	0	0.000	Σ = 116.00	Σ = 240

Therefore b = 116 ÷ 240 = 0.4833 (exactly 29/60), and a = $\bar{Y}$ − b$\bar{X}$ = 68.333 − (0.4833 × 67) = 68.333 − 32.383 = 35.95 (exactly 35.95 = 2,157/60). The estimated regression line of son's height on father's height is $\hat{Y}$ = 35.95 + 0.4833 X. Two checks confirm the arithmetic: the line passes through the mean point, since at X = $\bar{X}$ = 67 it gives $\hat{Y}$ = 35.95 + 0.4833 × 67 = 35.95 + 32.38 = 68.33 = $\bar{Y}$, as least-squares theory guarantees; and fitted values at the ends, $\hat{Y}$(59) = 64.47 against the actual 64 and $\hat{Y}$(75) = 72.20 against the actual 72, track the data closely. The supporting correlation, computed in the manner of Question 4, is r = 116 ÷ √(240 × 58) = 0.983, with Σ(Y − $\bar{Y}$)² = 58, so r^2 = 0.967: about 96.7 percent of the variation in sons' heights in this small set is associated with fathers' heights — a very strong relationship, as the tight upward cloud of Figure 5 shows.

Figure 5: Scatter diagram with regression line of Y on X

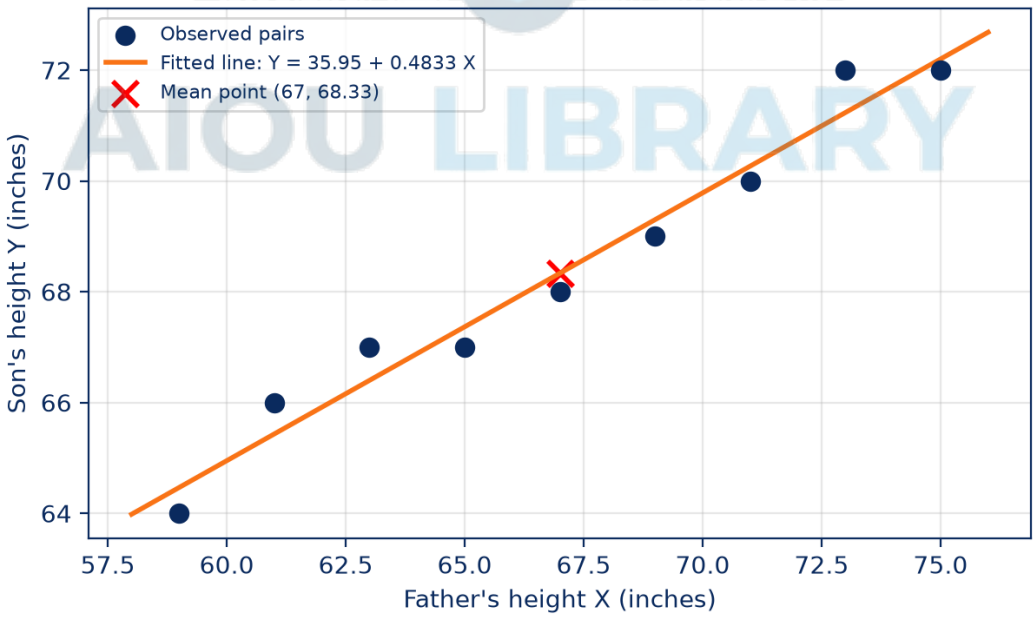

Figure 5: Scatter diagram of the nine father–son pairs with the fitted regression line $\hat{Y}$ = 35.95 + 0.4833X.

Interpretation is the point of the exercise. The slope b = 0.4833 says that, in these data, each additional inch of a father's height is associated on average with about 0.48 inch of additional height in the son — positive, but markedly less than inch-for-inch. That is Galton's regression towards the mean in arithmetic form: sons of unusually tall fathers tend to be tall but somewhat less extreme than their fathers, and sons of unusually short fathers tend to be short but less extreme, because height depends on many factors besides the father's height, and the other factors do not repeat the father's extremity. The intercept a = 35.95 is merely the mathematical anchor of the line at X = 0; it has no practical meaning here, since a father of zero inches does not exist, and quoting it as a prediction would be the extrapolation error against which part (a) warned — predictions are trustworthy only within, or near, the observed range X = 59 to 75, for example $\hat{Y}(71) = 35.95 + 34.32 = 70.27$ inches for the son of a 71-inch father. Finally, association is not sole causation: the line summarises a hereditary-and-environmental pattern in nine pairs; it estimates an average son's height for a given father's height, nothing more, and that disciplined reading is correct statistical practice.

Conclusion

The scatter diagram displays paired data as a cloud of points whose direction, form, tightness and outliers reveal association before any number is computed, and it thereby justifies, shapes and disciplines the regression line that theory then fits. For the nine father–son pairs, least squares gives $\hat{Y} = 35.95 + 0.4833X$, passing exactly through the mean point (67, 68.33), supported by a correlation of 0.983. Its slope of about one-half is the quantitative signature of regression towards the mean — the phenomenon that gave the whole method its name.

Question 4: a) Define the term correlation and product moment co-efficient of correlation. b) Describe the properties of a correlation coefficient. c) Explain the statistical and mathematical models by giving examples. (7+6+7 = 20 marks)

Introduction

Questions 2 and 3 of Assignment No. 1 organised single variables; Question 3 of this assignment fitted a line between two variables. The present question studies the companion measure that states, in one number, how strong and in what direction the linear association between two variables is: the correlation coefficient. Part (a) defines correlation and Karl Pearson's product-moment coefficient with its formula. Part (b) sets out the coefficient's properties — the facts that determine what a computed value does and does not entitle us to say. Part (c) widens the view to the two great kinds of model used throughout quantitative work, the mathematical (deterministic) model and the statistical (probabilistic) model, explaining each with examples and showing why business statistics, dealing with human and economic behaviour, must live mainly in the statistical family.

Part (a): Correlation and the Product-Moment Coefficient

Correlation is the statistical technique — and, by extension, the fact — of measuring the degree and direction of the relationship between two or more variables: the extent to which they vary together. Two variables are positively correlated when they tend to move in the same direction (height and weight; advertising and sales), negatively correlated when they move in opposite directions (price and quantity demanded, ordinarily), and uncorrelated when movement in one carries no tendency for movement in the other. Correlation asserts association only; the warning that correlation does not by itself prove causation is part of the definition's responsible

use, since both variables may be driven by a third, or the association may be coincidence in a small sample — the so-called spurious or nonsense correlation. The scatter diagram of Question 3 is correlation's picture; the coefficient is its number.

The product-moment coefficient of correlation, due to Karl Pearson and universally denoted by r, is defined as the arithmetic mean of the products of the corresponding deviations of the two variables from their means, each deviation expressed in units of its own standard deviation. In working form: $r = \Sigma(X - \bar{X})(Y - \bar{Y}) \div \sqrt{[\Sigma(X - \bar{X})^2 \cdot \Sigma(Y - \bar{Y})^2]}$, equivalently $r = [n\Sigma XY - (\Sigma X)(\Sigma Y)] \div \sqrt{\{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]\}}$ from raw sums. The name records its construction — it is built from the moments (deviation products) of the paired data — and each element of the formula earns its place: the numerator accumulates agreement in sign between the two deviations, so it is large and positive when large X accompanies large Y; the denominator, the geometric mean of the two sums of squared deviations, scales the result into a pure number independent of the units and sizes of X and Y. As a worked anchor from this assignment itself, the father–son data of Question 3 give $r = 116 \div \sqrt{(240 \times 58)} = 116 \div 117.98 = 0.983$, an almost perfect positive linear association. Degrees of correlation are conventionally read as: $|r|$ near 1, very high; about 0.75 to 0.9, high; 0.5 to 0.75, moderate; 0.25 to 0.5, low; near 0, negligible linear association — always with the rider that r measures linear association specifically, a point the properties now formalise.

Part (b): Properties of a Correlation Coefficient

The properties of r may be stated as a connected account. (1) Limits: r always lies between −1 and +1 inclusive. The value +1 denotes perfect positive linear relationship (all points on an upward line), −1 perfect negative linear relationship, and 0 the absence of linear relationship.

No computation can legitimately produce a value outside this range; one that appears to do so announces an arithmetical error. (2) Symmetry: r between X and Y equals r between Y and X; correlation has no direction of causation built into it, unlike regression, whose two lines (Y on X and X on Y) differ. (3) Independence of units and origin: r is a pure number, unaffected by change of scale or origin in either variable — heights in inches or centimetres give the same r — because deviations are standardised by the standard deviations in the denominator. (4) Geometric-mean link with regression: r is the geometric mean of the two regression coefficients, r = √(bYX · bXY), and it carries their common sign; correspondingly, r^2 — the coefficient of determination — gives the proportion of the variation in Y linearly associated with X, the most informative single reading of r (0.967 in the heights example). (5) Sensitivity to the data included: r is computed from every pair and is therefore affected by extreme pairs and by the range of X sampled; restricting the range can shrink r markedly even when the underlying relationship is unchanged. (6) Linearity limitation: r = 0 does not prove independence; a perfect curved relationship, such as Y symmetric in X about its mean, can yield r = 0 while the variables are completely dependent — the scatter diagram must always accompany the coefficient for exactly this reason. (7) No causation: however high, r alone does not establish that X causes Y, nor its direction if causation exists. (8) Sampling variability: an r computed from a small sample can differ appreciably from the population value, and significance testing (beyond this course's present scope) exists to judge whether an observed r could plausibly arise from an uncorrelated population. (9) Relationship to covariance: r is covariance standardised, r = Cov(X, Y) ÷ (σX σY), which is why it inherits covariance's sign while discarding its units. Together these properties

define a disciplined instrument: bounded, symmetric, unit-free, linear in scope, and evidential about association — never, by itself, about cause.

Part (c): Statistical and Mathematical Models

A model is a simplified representation of a real system, built to explain its behaviour or predict its outcomes. Mathematical models — also called deterministic models — express the relationship between variables by exact equations in which, given the inputs, the output is determined completely, with no room for randomness. The simple interest formula I = PRT ÷ 100 is a mathematical model: principal, rate and time fix the interest to the paisa. So are the area of a circle $A = \pi r^2$, the distance law $s = ut + \frac{1}{2}at^2$ of mechanics, the break-even equation of cost accounting in its textbook form (fixed cost divided by contribution per unit gives the exact break-even quantity), and an inventory rule that computes a definite reorder quantity from fixed demand and cost parameters. The strength of such models is precision and certainty within their assumptions; their weakness is that the assumptions — constant demand, no spoilage, exact measurement — are idealisations, and where the real world refuses to hold still, the model's exactness becomes fictitious.

Statistical models — also called probabilistic or stochastic models — represent a relationship that holds on average or in tendency, with an explicit random component absorbing everything the model does not explain. The regression equation of Question 3 is the archetype: Son's height = 35.95 + 0.4833 × Father's height + error, where the error term acknowledges heredity's other inputs, environment, nutrition and chance; the model predicts a mean tendency (r^2 = 0.967 of the variation accounted for), not any individual son's exact height. The demand model of econometrics, the time-series model that splits sales into trend, seasonal, cyclical and

irregular components, the probability model of insurance claims, and the sampling model behind every estimate in inferential statistics are all statistical models. Their strength is honesty about uncertainty: predictions come with margins, and the model can be tested and improved against data. Their weakness is the mirror image: they never yield certainty, and a careless user who treats a statistical prediction as a deterministic promise repeats, in modern dress, the oldest error in forecasting.

The contrast can be fixed in a few lines. A mathematical model contains no random term; the same inputs always give the same output; it is verified by logic and exact measurement; and it fits the inanimate regularities of physics, geometry and accounting identities. A statistical model contains a probability distribution or error term; the same inputs give a distribution of possible outputs centred on a tendency; it is validated by data, goodness of fit and out-of-sample prediction; and it fits biological, social and economic behaviour, where many small unmeasured causes act at once. Modern practice frequently joins the two: a mathematical skeleton — say, the accounting identity that profit equals revenue minus cost — populated by statistical estimates of revenue and cost behaviour; or a deterministic growth curve fitted by statistical least squares. For the student of business statistics the moral is direct: commerce deals with people, markets and chance, so its working models are statistical, its predictions are tendencies with margins, and the deterministic formula should be trusted exactly only where its controlled assumptions have genuinely been secured.

Conclusion

Correlation measures the degree and direction of linear association between two variables, and Pearson's product-moment coefficient r — the standardised mean product of paired deviations

— expresses it as a pure number between −1 and +1, symmetric, unit-free, tied to the regression coefficients through r = √(bYX · bXY), and properly read through r² as a proportion of variation explained, with association never confused for causation. Behind all such measures stands the choice of model: mathematical models determine outcomes exactly under ideal assumptions, while statistical models — regression foremost among them — carry an explicit error term and speak in tendencies. Business statistics is built, correctly, on the second kind, and uses the first only where its assumptions truly hold.

Question 5: a) Explain the classical, a posteriori and mathematical definition of probability. What are the axioms of probability? b) Ten vegetable cans, all of the same size, have lost their labels. It is known that 5 contain tomatoes and 5 contain corn. If 5 are selected at random, what is the probability that all contain tomatoes? What is the probability that 3 or more contain tomatoes? (10+10 = 20 marks)

Introduction

Probability is the language in which inferential statistics speaks: every estimate, test and forecast of business life is qualified by it. Part (a) of this question traces the concept through its three historical definitions. The classical definition, born in the games of chance of the seventeenth century, counts equally likely cases. The a posteriori or empirical definition, forced on the subject by events that are not equally likely by symmetry, reads probability from long-run relative frequency. The mathematical or axiomatic definition, completed by Kolmogorov in 1933, rebuilds the subject on three axioms from which all theorems follow, whatever interpretation of probability the user prefers. Part (b) then applies exact counting — the classical machinery in its sampling-without-replacement form, the hypergeometric setting — to

ten unlabelled cans, five of tomatoes and five of corn. The two answers, 1/252 (about 0.004) for all five cans containing tomatoes and exactly 1/2 for three or more containing tomatoes, are derived in full, with the complete distribution displayed so that each figure can be checked by addition.

Part (a): Three Definitions of Probability and the Axioms

Classical (a priori) definition. If a random experiment can result in N mutually exclusive, exhaustive and equally likely cases, of which M are favourable to an event A, then the probability of A is P(A) = M ÷ N — favourable cases divided by total equally likely cases. A fair coin gives P(head) = 1/2; a fair die gives P(six) = 1/6; a card drawn from a well-shuffled pack gives P(ace) = 4/52 = 1/13. The definition's virtues are simplicity and exactness: the probability is known before any trial, from the symmetry of the apparatus. Its limits are equally clear. It assumes equally likely cases, and defining probability through 'equally likely' flirts with circularity, since likelihood is itself probability in disguise. It applies only to finite case-counts. And most real events — that a debtor will default, that a monsoon will fail, that a new product will succeed — present no symmetric apparatus to count, so the classical definition, historically first and pedagogically indispensable, cannot carry the whole subject.

A posteriori (empirical or relative-frequency) definition. Under this definition, the probability of an event is the limit of its relative frequency in a long sequence of trials performed under identical conditions: P(A) = limit, as n grows, of (number of times A occurs in n trials) ÷ n. If a factory finds 47 defective items in the last 10,000 produced under stable conditions, the empirical probability that the next item is defective is estimated as 0.0047 — not by symmetry, for no symmetry exists, but by the record. Insurance mortality tables, machine failure rates and

batting averages are all empirical probabilities of this kind, and their authority grows with the length and stability of the record, expressing the law of large numbers: relative frequencies settle towards the underlying probability as trials accumulate. The limitations are practical and logical together: the exact limit is never observed, only approximated; conditions must remain stable for the past to speak about the future; and some events — the success of a once-only venture — cannot be repeated at all. The empirical definition therefore supplies the working probabilities of business and science while leaving the theoretical structure to the third definition.

Mathematical (axiomatic) definition. In the axiomatic approach, probability is not defined by cases or frequencies but characterised as a numerical function P assigned to events of a sample space S, satisfying three axioms (stated below) from which the entire calculus is deduced as pure mathematics. The classical and empirical values both satisfy the axioms, so both are embraced as legitimate ways of assigning the numbers; interpretations may differ, the theorems do not. This settlement is what allows one subject to serve dice, defectives and default risk at once.

The axioms of probability (Kolmogorov) are three. Axiom 1 — Non-negativity: for every event A, $P(A) \geq 0$; no event has a negative probability. Axiom 2 — Certainty of the sample space: $P(S) = 1$; something that exhausts all possibilities occurs with probability one, and consequently the impossible event has probability $P(\emptyset) = 1 - P(S) = 0$. Axiom 3 — Additivity: for mutually exclusive events, the probability of their union is the sum of their probabilities: if A and B cannot both occur, $P(A \cup B) = P(A) + P(B)$, extending to any finite (and, in the full theory, countably infinite) collection of pairwise mutually exclusive events. From these three alone flow the working rules

used daily: the complement rule P(not A) = 1 − P(A); the general addition rule P(A ∪B) = P(A) + P(B) − P(A ∩ B); the bound 0 ≤ P(A) ≤ 1 for every event; the multiplication rules for independent and dependent events; and the counting applications of part (b), where probabilities are ratios of combination counts precisely because random selection makes the counted selections equally likely, as the classical definition requires and the axioms permit.

Part (b): The Ten Cans — Hypergeometric Calculation

The selection is made at random, without replacement, from 10 cans of which 5 contain tomatoes (call them successes) and 5 contain corn. The number of equally likely ways to choose any 5 cans from 10 is C(10, 5) = 10! ÷ (5! × 5!) = 252; these 252 selections are the denominator of every probability in this part. The number of selections containing exactly k tomato cans is C(5, k) × C(5, 5 − k): choose which k of the 5 tomato cans, and which 5 − k of the 5 corn cans, complete the handful. This is the hypergeometric pattern — sampling without replacement from a finite divided population — and the complete distribution over k = 0, 1, 2, 3, 4, 5 is 1, 25, 100, 100, 25, 1 favourable selections respectively, summing to 252 as it must. The symmetry of the table (1, 25, 100, 100, 25, 1) is no accident: with equal numbers of tomato and corn cans, getting k tomatoes mirrors getting 5 − k tomatoes — that is, k corn — exactly.

Table 4: Complete distribution of the number of tomato cans in a random selection of 5 from 5 tomato and 5 corn cans.

Tomato cans k	Favourable selections C(5,k)·C(5,5−k)	Probability
0	1 × 1 = 1	1/252 ≈ 0.0040
1	5 × 5 = 25	25/252 ≈ 0.0992

2	10 × 10 = 100	100/252 ≈ 0.3968
3	10 × 10 = 100	100/252 ≈ 0.3968
4	5 × 5 = 25	25/252 ≈ 0.0992
5	1 × 1 = 1	1/252 ≈ 0.0040
Total	252	1.0000

Required probability 1 — all five contain tomatoes. Favourable selections: all 5 tomato cans and none of the corn, C(5, 5) × C(5, 0) = 1 × 1 = 1. Hence P(all tomatoes) = 1 ÷ 252 ≈ 0.0040 — about one chance in 252, or 0.4 percent. Intuition should be trained by the figure: although half the cans are tomatoes, demanding a clean sweep of five is demanding, because each tomato can drawn depletes the stock for the next draw; the sequential check 5/10 × 4/9 × 3/8 × 2/7 × 1/6 = 120 ÷ 30,240 = 1/252 confirms the combination count exactly, and a student is advised to verify any hypergeometric result by this second route whenever time permits.

Required probability 2 — three or more contain tomatoes. 'Three or more' comprises k = 3, 4 and 5, with favourable selections 100 + 25 + 1 = 126. Hence P(3 or more tomatoes) = 126 ÷ 252 = 1/2 = 0.5 exactly. The elegant exactness has a reason the student should notice, since examiners choose such numbers deliberately: by the symmetry displayed in Table 4, selections with 3 or more tomatoes are in perfect balance with selections with 2 or fewer (each totalling 126), and '3 or more' and '2 or fewer' exhaust all possibilities, so each must carry probability one-half. A result that can be both computed and explained is doubly secure. Interpretation for practice: a buyer selecting five unlabelled cans at random faces a negligible 0.4 percent chance of a pure tomato handful, a 50 percent chance of a tomato-majority handful, and — by the

same table, using the complement — a 50 percent chance of a corn-majority or balanced-against handful; the most likely single outcomes, at about 39.7 percent each, are exactly 2 and exactly 3 tomatoes, which is the distribution's centre announcing itself.

Conclusion

Probability began as the counting of equally likely cases (classical), matured into the reading of long-run relative frequencies where no symmetry exists (a posteriori), and was finally secured as mathematics by Kolmogorov's three axioms — non-negativity, P(S) = 1, and additivity for mutually exclusive events — from which all working rules follow. The cans problem shows the classical counting method at its strongest, extended by combinations to sampling without replacement: of the 252 equally likely selections of 5 cans from 10, exactly 1 is all tomatoes, giving a probability of 1/252 ≈ 0.004, while 126 selections contain three or more tomato cans, giving a probability of exactly 1/2 — both results verified by the full distribution and, for the first, by the sequential product rule. With probability in place, the inferential statistics promised in Assignment No. 1 — estimation and testing in the later courses of the programme — has its foundation.