ALLAMA IQBAL OPEN UNIVERSITY



AIOU LIBRARY

Name:

Registration No.:

Course Code: 1430

Program: BA/B. Com/AD

Semester: Autumn 2026

ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

Course: Business Statistics (1430) • Level: BA/B. Com/AD

Semester: Autumn 2026

Assignment No. 1 (Units 1–4)

Total Marks: 100 • Pass Marks BA: 40 • Pass Marks (AD): 50

SOLVED ASSIGNMENT GUIDE

(For Study & Reference)

Note: This guide is prepared for study and reference purposes. Students are advised to read the relevant units of the course book, understand the concepts and methods, and write the assignment in their own words.

Question 1: Why We Draw a Histogram; Constructing the Graph of a Frequency Distribution (20 Marks)

Q. 1 Why we draw a histogram? Explain. Write the steps to construct the graph of frequency distribution for equal and unequal class intervals.

Introduction

Statistics begins with raw data, and raw data in their original form are almost impossible to read. A business may record the daily sales of hundreds of customers, a hospital may record the weights of hundreds of babies, and a college may record the marks of hundreds of students, yet a long unorganised list of numbers conveys no clear message. The first task of the statistician is therefore to organise the data into a frequency distribution — a table that shows the classes into which the values fall and the frequency, that is the number of observations, in each class. The second task is to present that distribution in a picture, because a well-drawn diagram communicates in a few seconds what a table takes several minutes to reveal. The histogram is the most important diagram for a continuous frequency distribution, and this answer explains

why it is drawn and then sets out, step by step, how the graph of a frequency distribution is constructed, first when the class intervals are equal and then when they are unequal.

What Is a Histogram?

A histogram is a graph of a grouped continuous frequency distribution in which the class intervals are marked along the horizontal axis and a rectangle is erected on each interval whose area is proportional to the frequency of that class. When all the intervals are equal, the height of each rectangle is simply the frequency itself, so the eye can compare frequencies directly by comparing heights. Three features distinguish a histogram from an ordinary bar diagram. First, the variable on the horizontal axis is continuous, so the rectangles touch one another with no gap between them; a gap in a histogram has a real meaning, namely that a class contains no observations. Second, it is the area of a rectangle, and not merely its height, that represents frequency — this becomes decisive when intervals are unequal. Third, a histogram can be drawn only for a continuous variable grouped into class intervals, whereas a bar diagram is used for discrete data and for attributes such as regions, products or years.

Why Do We Draw a Histogram? — The Purposes

- To give an immediate visual summary: a histogram converts a table of numbers into a shape. A manager, a doctor or a teacher who has no time to study a table can see at once where most of the observations lie and how widely they are spread.
- To reveal the shape of the distribution: the histogram shows whether the distribution is symmetrical like a bell, skewed to the right with a long tail of high values, skewed to the left, flat, or has two humps (bimodal). Shape is the first clue to the behaviour of the

phenomenon — incomes, for example, are almost always right-skewed, while heights of adults are nearly symmetrical.

- To locate the centre and the spread: the position of the tallest rectangle shows the modal class, the region where observations concentrate, and the total width of the base shows the range. A narrow, tall histogram means consistent, uniform values; a wide, flat one means great variability.
- To detect gaps, clusters and outliers: an isolated rectangle far from the main body, or an empty interval inside the range, warns the investigator of unusual observations, possible recording errors, or a mixture of two different populations in the same data.
- To compare distributions: two histograms drawn on the same scale — for example the output of two factories, or the marks of two sections of a class — allow a fair visual comparison of centre, spread and shape which would be tedious from tables alone.
- To estimate averages graphically and to judge later analysis: the histogram gives a rough check on the calculated mean, median and mode, helps in smoothing the data into a frequency curve, and indicates whether procedures that assume a roughly normal shape are reasonable for the data in hand.

In short, we draw a histogram because the human eye understands a shape faster and remembers it longer than a column of figures. The table remains necessary for exact calculation, but the histogram is what turns the table into knowledge.

Steps to Construct the Graph When Class Intervals Are Equal

Step 1 — Make the distribution continuous. If the classes are given in the inclusive form, such as 1–3, 4–6, 7–9, first convert them into true class limits by applying the correction factor,

which is half the gap between the upper limit of one class and the lower limit of the next. Here the gap is 1, the correction factor is 0.5, and the classes become 0.5–3.5, 3.5–6.5, 6.5–9.5 and so on. Only a continuous series can be drawn correctly as a histogram.

Step 2 — Choose suitable scales. Mark the true class limits along the horizontal (X) axis and the frequencies along the vertical (Y) axis, choosing scales so that the figure is neither too flat nor too tall; a common guide is that the height of the diagram should be about three-quarters of its width.

Step 3 — Erect the rectangles. On each class interval as base, draw a rectangle whose height equals the frequency of that class. Because all bases are equal, heights alone compare the frequencies correctly.

Step 4 — Join the rectangles without gaps. Adjacent rectangles must touch, expressing the continuity of the variable. Each rectangle may be lightly shaded or coloured in the same shade so that the whole figure appears as one continuous area.

Step 5 — Label the diagram completely. Give the figure a clear title, label both axes with the variable and its unit, and mention the scales used. A diagram without a title and axis labels is incomplete, however neatly it is drawn.

Step 6 (optional) — Superimpose the frequency polygon or curve. Joining the mid-points of the tops of the rectangles gives the frequency polygon, and smoothing that polygon freehand gives the frequency curve, both of which rest on the histogram as their foundation. The following simple example illustrates the result: the marks of 60 students grouped in equal classes of

width 10 produce rectangles of heights 3, 7, 14, 18, 11 and 7, and the shape immediately shows that most students cluster in the middle classes 30–40, with fewer students at both extremes.

Figure: A histogram (equal class intervals)

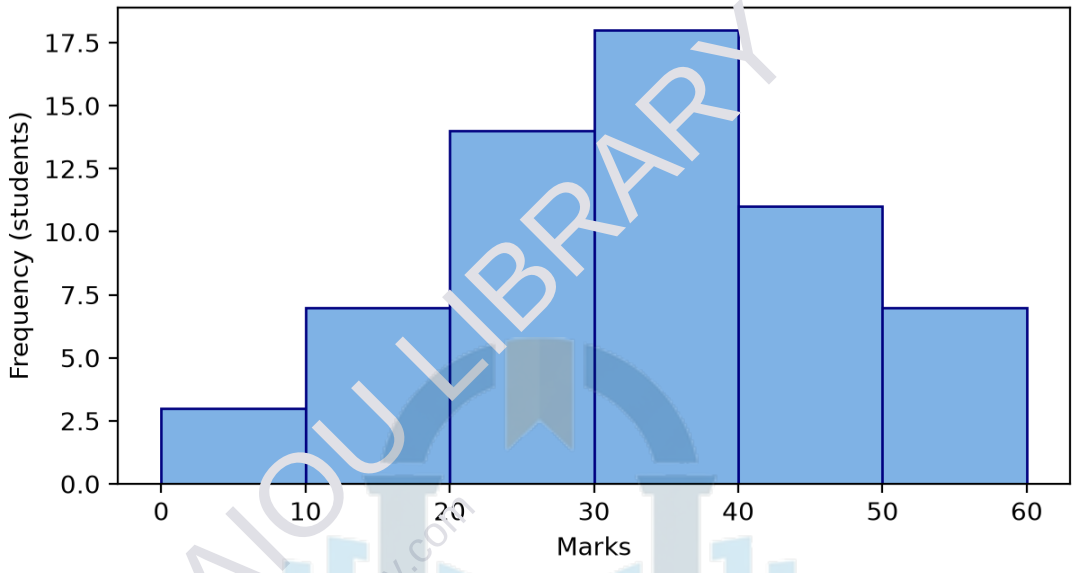


Figure 1: A histogram for an equal interval distribution. The rectangles touch, and because the bases are equal, the heights compare the frequencies directly.

Steps to Construct the Graph When Class Intervals Are Unequal

Unequal intervals arise whenever the data are grouped with narrow classes where observations are dense and wide classes where they are thin — income and wage distributions are the classic example, with classes such as 0–10, 10–20, 20–50 and 50–100 (thousand rupees). If, in such a case, a rectangle were drawn with height equal to the raw frequency, a wide class would automatically look more important than it really is, because its base is larger. The governing principle must therefore be restored: area, not height, must be proportional to frequency. The steps are the same as before up to the drawing of the rectangles, but one extra calculation is inserted.

Step 1 — Convert the classes into continuous (true-limit) form, exactly as in the equal-interval case. Step 2 — Identify the smallest class interval and take it as the standard width. Step 3 — Compute the frequency density, also called the adjusted frequency, for every class by the rule: frequency density = frequency ÷ class width. Equivalently, the adjusted frequency = frequency × (smallest width ÷ width of the class). Step 4 — Draw each rectangle with its true class interval as the base but with the frequency density (adjusted frequency) as its height. The area of every rectangle is then exactly proportional to its frequency, and the comparison between classes is honest. Step 5 — Title, label and scale the diagram as before, and note on the diagram that heights represent frequency densities.

Table 1: Worked example — adjustment for unequal class intervals

Class	Width	Frequency	Frequency density (f ÷ width)	Adjusted height (standard width = 10)
0–10	10	8	8 ÷ 10 = 0.8	8
10–20	10	15	15 ÷ 10 = 1.5	15
20–50	30	30	30 ÷ 30 = 1.0	10
50–100	50	25	25 ÷ 50 = 0.5	5

The example in Table 1 shows why the adjustment matters. The class 20–50 has the largest raw frequency, 30, but spread over a width of 30 its density is only 1.0 per unit, lower than the density 1.5 of the class 10–20. Without adjustment the histogram would falsely suggest that observations concentrate most thickly in 20–50; with adjustment the rectangles tell the truth. The same adjustment is carried over to the frequency polygon and frequency curve for unequal intervals: the points are plotted at class mid-points against frequency densities, never against raw frequencies.

Conclusion

To conclude, the histogram is drawn to make a frequency distribution visible — to show its centre, spread, shape, gaps and unusual values at a glance and to support comparison and further analysis. Its construction follows a disciplined sequence: make the classes continuous, choose scales, and erect touching rectangles whose areas are proportional to frequencies. With equal intervals the heights may be the frequencies themselves; with unequal intervals the heights must be frequency densities obtained by dividing each frequency by its class width. Observing this single rule — area represents frequency — is what separates a correct histogram from a misleading picture.

Question 2: Frequency Curve and Ogive for the Weights of Premature Babies (10 + 10 Marks)

Q. 2 The Complex Hospital in Islamabad has the following data on the weights in pounds of 200 premature babies: Class 0.5–0.9, frequency 10; 1.0–1.4, 19; 1.5–1.9, 24; 2.0–2.4, 27; 2.5–2.9, 29; 3.0–3.4, 44; 3.5–3.9, 47. Construct a frequency curve and ogive and explain what types of information these can provide. (a) Is the program in compliance with the policy? (b) Does your 13-interval relative frequency distribution help you to answer part (a) better than the 7-interval distribution?

Introduction

Birth weight is one of the most important indicators of the health of a newborn baby, and for premature babies it is a matter of survival: the lower the weight, the greater the care the baby needs. The data given in the question record the weights, in pounds, of 200 premature babies born at a hospital in Islamabad, grouped into seven classes of width 0.5 lb (printed in the

inclusive form 0.5–0.9, 1.0–1.4 and so on). A table alone, however, does not show the pattern of the weights. This answer first organises the data with true class boundaries, relative frequencies and cumulative frequencies, then constructs the two graphs asked for — the frequency curve and the (less-than) ogive — explains the information each of them provides, and finally discusses parts (a) and (b) honestly, because those sub-parts refer to a "program", a "policy" and a "13-interval distribution" that the printed question itself does not define; the correct statistical treatment of that situation is explained rather than an invented verdict being offered.

Organising the Data

The printed classes are in inclusive form with a gap of 0.1 between successive classes, so the correction factor is 0.05 and the true (continuous) classes become 0.45–0.95, 0.95–1.45, 1.45–1.95, 1.95–2.45, 2.45–2.95, 2.95–3.45 and 3.45–3.95. Table 2 gives the mid-points needed for the frequency curve, the relative frequencies, and the cumulative frequencies needed for the ogive. The frequencies sum to exactly 200, as stated in the question, which is the first check on the working.

Table 2: Weights of 200 premature babies — working table for the frequency curve and the ogive

Class (lb)	True class	Mid-point	Frequency (f)	Relative frequency	Cumulative frequency (less than)
0.5–0.9	0.45–0.95	0.70	10	0.050 (5.0%)	10
1.0–1.4	0.95–1.45	1.20	19	0.095 (9.5%)	29
1.5–1.9	1.45–1.95	1.70	24	0.120 (12.0%)	53
2.0–2.4	1.95–2.45	2.20	27	0.135 (13.5%)	80
2.5–2.9	2.45–2.95	2.70	29	0.145 (14.5%)	109
3.0–3.4	2.95–3.45	3.20	44	0.220 (22.0%)	153

3.5–3.9	3.45–3.95	3.70	47	0.235 (23.5%)	200

Constructing the Frequency Curve

A frequency curve is the smoothed form of a frequency polygon. The steps are: (i) plot a point for each class at its mid-point (horizontal axis) against its frequency (vertical axis); (ii) add two imaginary points of zero frequency at the mid-points of the empty classes just below the first class (0.20) and just above the last class (4.20), so that the curve begins and ends on the base line; (iii) join the plotted points by a smooth freehand curve — not by straight line segments — taking care that the total area under the curve is approximately equal to the total area of the histogram, since that area represents the total frequency of 200. The resulting curve is shown in panel (a) of Figure 2. It rises steadily from the lowest weights to the highest: the distribution is negatively skewed, that is, it has a long tail towards the small weights on the left, while the bulk of the babies are concentrated at the heavier end, in the classes 3.0–3.4 and 3.5–3.9 which together account for 91 babies, or 45.5 per cent of the total.

Constructing the Ogive

An ogive is a graph of the cumulative frequency distribution. For the “less than” ogive the steps are: (i) form the cumulative frequencies by adding the frequencies class by class, as in the last column of Table 2; (ii) plot each cumulative frequency against the upper true boundary of its class — the points are therefore (0.95, 10), (1.45, 29), (1.95, 53), (2.45, 80), (2.95, 109), (3.45, 153) and (3.95, 200), starting from (0.45, 0); (iii) join the points by a smooth curve, which must rise steadily and can never fall, because cumulative frequencies only increase. The ogive is shown in panel (b) of Figure 2. A “more than” ogive can be drawn in exactly the same way by

accumulating the frequencies from the bottom upwards and plotting against the lower true boundaries; where the less-than and more-than ogives intersect, the vertical line through the intersection gives the median.

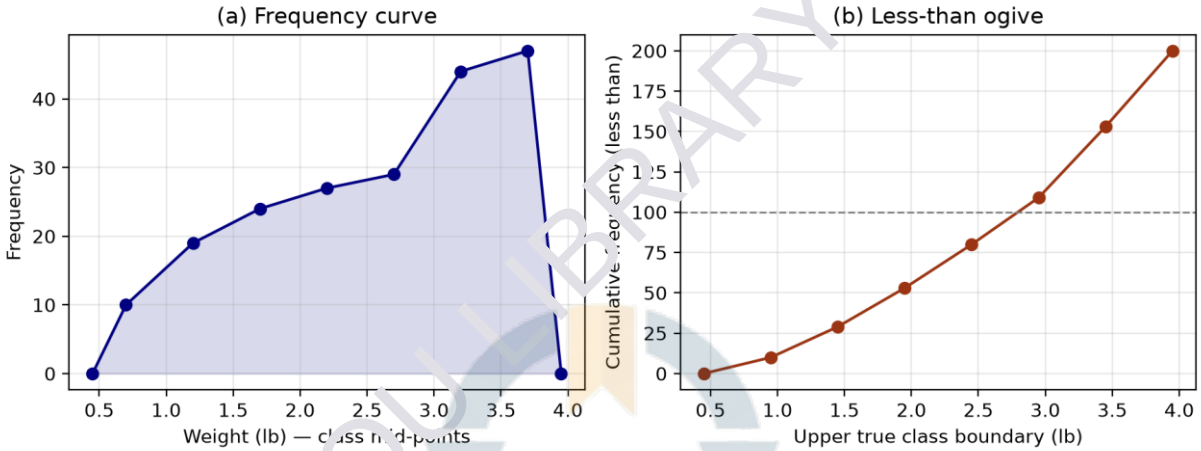


Figure 2: (a) Frequency curve and (b) less-than ogive for the weights of 200 premature babies.

What Information Do These Graphs Provide?

- From the frequency curve — the shape of the distribution: here the steady rise towards the heavier classes with a tail towards low weights tells the hospital that most premature babies are born reasonably heavy, but a minority are born very light and will need intensive care; the curve also shows the modal region (3.5–3.9 lb), the absence of gaps, and it allows two distributions, for example this year and last year, to be compared fairly even when their totals differ, by drawing relative frequency curves.
- From the ogive — numbers and proportions below or above any weight: reading horizontally from a cumulative frequency to the curve and then down gives the corresponding weight, and reading in the reverse direction gives the number of babies below any chosen weight. For example, the ogive shows that 80 babies (40 per cent) weighed less than about 2.45 lb, and 109 (54.5 per cent) weighed less than about 2.95 lb.

- From the ogive — positional averages: drawing a horizontal line at N/2 = 100 meets the ogive at a weight of about 2.9 lb, which is the median weight; lines at N/4 = 50 and 3N/4 = 150 give the lower and upper quartiles (approximately 1.9 lb and 3.4 lb), and any percentile, such as the lightest 10 per cent of babies, can be read in the same way. These readings are estimates, but they are obtained in seconds without any calculation.
- For planning: combined, the two graphs tell the hospital how many incubators and intensive-care places the very light babies require, whether the pattern of birth weights is improving over time, and what proportion of babies falls below any clinical threshold of concern — exactly the kind of information on which nursery staffing and equipment decisions rest.

Parts (a) and (b): An Honest Note and the Correct Method

A careful reading of the question booklet shows that parts (a) and (b) are carried over from a longer textbook exercise: the booklet does not state any hospital “program”, does not quote any “policy” (for example a rule such as “at least a stated percentage of babies must weigh more than a stated number of pounds”), and does not provide the raw data from which a 13-interval distribution could be formed. A student must not invent a policy or a verdict. The correct answer therefore states the position plainly — on the information printed, compliance cannot be judged, because no policy standard is given — and then demonstrates the method by which the judgement would be made the moment the standard is known: convert the distribution to relative frequencies (Table 2), read from the ogive or the cumulative column the proportion of babies on the required side of the policy threshold, and compare that proportion with the proportion the policy demands.

Part (b) can be answered as a matter of method. A 13-interval distribution is obtained by halving the class width (from 0.5 lb to about 0.27 lb), so each printed class is split into two narrower classes. Narrower classes locate a policy threshold much more precisely: with seven wide classes, a threshold that falls inside a class forces the student to guess how the frequency of that class divides on the two sides of the threshold, whereas with thirteen classes the boundary of a narrow class usually lies very close to the threshold and the cumulative proportion can be read almost exactly. Finer grouping also reveals local irregularities that wide classes smooth away. The cost is that each narrow class contains fewer babies, so the relative frequencies become less stable. The answer to (b) is therefore: yes, the 13-interval relative frequency distribution helps to answer a compliance question better than the 7-interval distribution, because compliance is judged by a proportion relative to a fixed threshold, and the finer distribution measures that proportion with much less interpolation error — provided the policy and its threshold are actually stated.

Conclusion

To conclude, the frequency curve for the 200 premature babies rises steadily towards the heavier classes and shows a distribution skewed towards low weights, while the less-than ogive converts the same data into cumulative form from which the median (about 2.9 lb), the quartiles and the number or percentage of babies below any weight can be read directly. Together the two graphs turn a hospital table into planning information. Parts (a) and (b), as printed, lack the policy statement they refer to; the professionally correct response is to say so, to supply the relative frequency distribution and the ogive readings by which compliance would

be tested, and to recognise that a finer 13-interval distribution would test any such threshold more accurately than the 7-interval distribution given.

Question 3: Mean and Weighted Mean; Mean and Standard Deviation by Direct and Coding Methods (10 + 10 Marks)

Q. 3 (a) Differentiate between mean and weighted mean. (b) The following data represent the weights in pounds of a sample of packages carried out last month by a small airfreight company. Classes: 10–10.9 (f = 1), 11–11.9 (4), 12–12.9 (6), 13–13.9 (8), 14–14.9 (12), 15–15.9 (15), 16–16.9 (11), 17–17.9 (8). Calculate mean and standard deviation using direct and coding methods.

Introduction

The arithmetic mean is the most familiar of all averages, but it treats every observation as equally important. In real business data that assumption often fails: a student's final grade depends more on the final examination than on a class test, a consumer price index depends more on food than on recreation, and the average cost of a firm's purchases depends on how much of each item was bought. The weighted mean exists for exactly such situations. Part (a) of this answer draws the distinction carefully, and part (b) then computes the mean and standard deviation of a grouped distribution of package weights twice — once by the direct method and once by the coding (short-cut) method — showing that the two methods are algebraically identical and differ only in the size of the numbers handled.

Part (a): Mean and Weighted Mean — The Difference

The (simple) arithmetic mean of observations is the sum of the values divided by their number, $\bar{x} = \Sigma x / n$; for grouped data it is $\bar{x} = \Sigma fx / \Sigma f$, where the frequency f simply records how many

times each value occurs. The weighted mean assigns to each value an explicit weight w expressing its relative importance, size or exposure, and is computed as $\bar{x}_p = \Sigma wx / \Sigma w$. The essential points of difference are as follows.

- Importance of items: the simple mean assumes all items are equally important; the weighted mean is built on the recognition that items differ in importance, and the weights measure that difference.
- Formula: simple mean divides by the count of items (or the total frequency); weighted mean divides by the sum of the weights, Σw, which need not equal the number of items at all.
- Nature of the multiplier: in the grouped simple mean the multiplier f is a frequency — it counts occurrences. In the weighted mean the multiplier w is an importance measure — credit hours, quantities purchased, population sizes, expenditure shares — which may have nothing to do with counting.
- When each is appropriate: the simple mean suits data whose items are naturally comparable and equally significant, such as the daily temperatures of a month. The weighted mean is compulsory when items are not equally significant — averaging percentages, rates, grades, prices or examination marks without weights gives a positively misleading result.
- Typical uses: the weighted mean is used for grade-point averages (grades weighted by credit hours), index numbers (prices weighted by quantities or expenditure), average cost per unit when batches are bought at different prices and in different quantities, and

standardised rates such as the standardised death rate (rates weighted by standard population).

- Relationship between the two: the simple mean is a special case of the weighted mean in which all weights are equal. If every w is the same constant, it cancels from numerator and denominator and the weighted formula reduces exactly to the simple formula.

A compact example fixes the idea. Suppose a student scores 80 in a 2-credit course and 60 in a 4-credit course. The simple mean is (80 + 60)/2 = 70, but the weighted mean is (2×80 + 4×60)/(2 + 4) = 400/6 = 66.7. The weighted figure is the correct summary of the student's performance, because the weaker subject carries twice the credit; the simple mean overstates the performance by ignoring that fact.

Part (b): Mean and Standard Deviation of the Package Weights

The eight classes are of equal width. Their mid-points are 10.45, 11.45, 12.45, 13.45, 14.45, 15.45, 16.45 and 17.45 lb, rising by exactly 1, which makes the coding method particularly clean. The total frequency is N = Σf = 1 + 4 + 6 + 8 + 12 + 15 + 11 + 8 = 65 packages.

Direct Method

By the direct method, $\bar{x} = \Sigma fx / \Sigma f$ and the (grouped-data) variance is $s^2 = \Sigma f(x - \bar{x})^2 / \Sigma f$, computed in practice through the equivalent computing form $\Sigma fx^2 / \Sigma f - \bar{x}^2$. Table 3 sets out the working: Σfx = 964.25, so the mean is $\bar{x}$ = 964.25 / 65 = 14.83 lb (more precisely 14.835). The sum of $f(x - \bar{x})^2$ over the eight classes is 205.38, giving variance = 205.38 / 65 = 3.16 and standard deviation = √3.16 = 1.78 lb. (If the 65 packages are treated strictly as a sample and the divisor N − 1 = 64 is used, the standard deviation is 1.79 lb; the conclusion is unaffected.)

Table 3: Direct method — mean and standard deviation of package weights

Class (lb)	f	Mid-point x	fx	f(x – x̄)²
10–10.9	1	10.45	10.45	19.22
11–11.9	4	11.45	45.80	45.82
12–12.9	6	12.45	74.70	34.12
13–13.9	8	13.45	107.60	15.34
14–14.9	12	14.45	173.40	1.78
15–15.9	15	15.45	231.75	5.68
16–16.9	11	16.45	180.95	28.70
17–17.9	8	17.45	139.60	54.72
Total	65	—	964.25	205.38

Coding (Short-Cut) Method

The coding method replaces each mid-point x by a coded value u = (x − A) / h, where A is an assumed mean and h is the common class width. Taking A = 14.45 (the mid-point of the middle class) and h = 1, the coded values are simply −4, −3, −2, −1, 0, 1, 2, 3. The working gives Σfu = 25 and Σfu² = 215. Then x̄ = A + (Σfu / Σf) × h = 14.45 + (25/65) × 1 = 14.45 + 0.385 = 14.835 lb, identical to the direct result. For the spread, variance = h² × [Σfu² / Σf − (Σfu / Σf)²] = 1 × [215/65 − (25/65)²] = 3.308 − 0.148 = 3.16, so the standard deviation is again √3.16 = 1.78 lb.

Table 4: Coding method — the same calculation with small numbers

Class (lb)	f	Mid-point x	u = (x – 14.45)/1	fu	fu²
10–10.9	1	10.45	−4	−4	16
11–11.9	4	11.45	−3	−12	36
12–12.9	6	12.45	−2	−12	24
13–13.9	8	13.45	−1	−8	8
14–14.9	12	14.45	0	0	0
15–15.9	15	15.45	1	15	15

16–16.9	11	16.45	2	22	44
17–17.9	8	17.45	3	24	72
Total	65	—	—	25	215

The agreement of the two methods is not a coincidence to be hoped for but a result to be checked: coding is only a change of origin and scale, and a change of origin does not alter the standard deviation at all, while the mean is shifted back by exactly the amount that was subtracted. The practical lesson for business work is that the coding method should be preferred whenever mid-points are large or awkward, because it replaces numbers such as 17.45 and their squares with single digits, drastically reducing both labour and the risk of arithmetical error — the reason the method was standard practice long before electronic calculators and remains the cleanest way to check a computer printout by hand.

Conclusion

To conclude: the weighted mean differs from the simple mean in recognising that items carry different importance, measured by weights, and the simple mean is merely its equal-weights special case. For the airfreight packages, both the direct and the coding methods give the same verified results — a mean weight of 14.83 lb and a standard deviation of about 1.78 lb (1.79 lb on the sample divisor) — so a typical package weighs just under 15 lb, with most packages lying roughly within two pounds of that average, information the company can use directly for load planning and pricing.

Question 4: The Median from Grouped Data; Computing Median and Mode (10 + 10 Marks)

Q. 4 (a) Write a procedure to calculate the median from grouped data. Also explain the advantages and disadvantages of median. (b) From the following data, compute the median and mode. Classes: 1–3 (f = 18), 4–6 (90), 7–9 (44), 10–12 (21), 13–15 (9), 16–18 (9), 19–21 (4), 22–24 (5).

Introduction

The median is the middle value of a distribution — the value that divides the data into two equal halves, with half the observations below it and half above it. Because it depends on position rather than on the size of every observation, it is the natural average to use when data are skewed, when extreme values would distort the mean, or when the variable can only be ranked. Part (a) gives the step-by-step procedure for finding the median from grouped data and weighs the merits and demerits of the measure; part (b) applies the median formula, and the corresponding mode formula, to the distribution given in the question.

Part (a): Procedure for Calculating the Median from Grouped Data

Step 1 — Make the classes continuous. If the series is in inclusive form (1–3, 4–6, …), convert it into true class limits using the correction factor, half the gap between successive classes. Step 2 — Prepare the cumulative frequency column. Add the frequencies from the top downwards to obtain the “less than” cumulative frequencies. Step 3 — Compute the median position, N/2, where N = Σf is the total frequency. Step 4 — Locate the median class: scan the cumulative frequency column and select the first class whose cumulative frequency is equal to or just greater than N/2; the median lies inside that class. Step 5 — Apply the interpolation formula:

Median = L + [(N/2 − cf) / f] × h, where L is the lower true limit of the median class, cf is the cumulative frequency of the class preceding the median class, f is the frequency of the median class, and h is its width. The formula rests on the assumption that the f observations inside the median class are spread evenly across its width, so the median is the point that cuts the class in the required proportion. Step 6 — Check the result: the median must lie inside the median class, and it should be verified against the ogive, where the value read off at cumulative frequency N/2 gives the same answer graphically.

Advantages of the Median

- It is unaffected by extreme values: a single very large income, or one abnormally heavy parcel, does not move the median at all, whereas it can drag the mean far from the typical value. This is why median income and median house price are the figures quoted in official statistics.
- It is easy to understand and, for ungrouped data with few items, easy to find — it is simply the middle observation after arranging the data in order.
- It can be located graphically from the ogive, and even by inspection in a cumulative table, without any elaborate calculation.
- It is meaningful for open-ended distributions, where classes such as “100 and above” make the mean impossible to compute because the mid-point of the open class is unknown; the median needs only the middle class, which is usually closed.
- It is the appropriate average for skewed distributions and for data measured on an ordinal scale, and it always corresponds to a real positional value of the distribution rather than a possibly impossible figure.

Disadvantages of the Median

- It ignores the magnitude of most observations: only the position of the middle item matters, so two very different distributions can share the same median. It is therefore not based on all the observations in the fuller sense in which the mean is.
- It is not capable of further algebraic treatment: medians of subgroups cannot be combined into the median of the whole group, and the median does not enter the standard formulas of sampling theory as conveniently as the mean does.
- For grouped data it is only an estimate, because the interpolation formula assumes an even spread of values inside the median class which may not hold in fact.
- It can be erratic in small samples: adding or removing one or two observations near the centre can shift the median noticeably, and for an even number of ungrouped observations it must be defined as the average of the two middle values, which may be a value no observation actually takes.
- Arranging a large mass of raw data in order, which the median requires, can itself be laborious when the data are not already grouped.

Part (b): Computation of the Median and the Mode

The classes are inclusive with a gap of 1, so the correction factor is 0.5 and the true classes are 0.5–3.5, 3.5–6.5, and so on, each of width h = 3. The total frequency is N = 18 + 90 + 44 + 21 + 9 + 9 + 4 + 5 = 200, and the cumulative frequencies are shown in Table 5.

Table 5: Working table for the median and the mode

Class	True class	Frequency (f)	Cumulative frequency
1–3	0.5–3.5	18	18
4–6	3.5–6.5	90	108

7–9	6.5–9.5	44	152
10–12	9.5–12.5	21	173
13–15	12.5–15.5	9	182
16–18	15.5–18.5	9	191
19–21	18.5–21.5	4	195
22–24	21.5–24.5	5	200

Median: the median position is N/2 = 100. The first cumulative frequency reaching or exceeding 100 is 108, so the median class is 3.5–6.5, with L = 3.5, cf = 18, f = 90 and h = 3. Substituting, Median = 3.5 + [(100 − 18) / 90] × 3 = 3.5 + (82/90) × 3 = 3.5 + 2.73 = 6.23. The median therefore lies near the top of the second class, which is exactly what the cumulative figures lead us to expect: 100 of the 200 observations are exhausted only after 82 of the 90 observations of that class have been passed. Half the observations are below about 6.23 and half are above it.

Mode: the modal class is the class of maximum frequency, namely 3.5–6.5 with $f_1 = 90$, flanked by $f_0 = 18$ (preceding class) and $f_2 = 44$ (following class). The mode formula is Mode = $L + [(f_1 - f_0) / ((f_1 - f_0) + (f_1 - f_2))] \times h = 3.5 + [(90 - 18) / ((90 - 18) + (90 - 44))] \times 3 = 3.5 + [72 / (72 + 46)] \times 3 = 3.5 + (72/118) \times 3 = 3.5 + 1.83 = 5.33$. The most typical value of the distribution is therefore about 5.33. Note that mode (5.33) is less than median (6.23): the mass of the distribution is concentrated in the low classes with a long thin tail stretching to 24, the signature of a positively skewed distribution, in which the usual ordering mode < median < mean is found.

Conclusion

To conclude, the median of grouped data is found by converting the classes to true limits, cumulating the frequencies, locating the class that contains the (N/2)th observation, and

interpolating within that class. Its great strengths are its resistance to extreme values and its availability even for open-ended and skewed data; its weaknesses are that it disregards magnitudes, resists further algebra, and, in grouped form, is an interpolation-based estimate. For the data of this question, the median is 6.23 and the mode is 5.33, and the two together describe a distribution concentrated at low values with a right-hand tail.

Question 5: The Five-Step Process of Hypothesis Testing; A Test for a Population Proportion (10 + 10 Marks)

Q. 5 (a) Discuss the five-step process for hypothesis testing using a standardized scale. (b) From a total of 10,200 loans made by a state employees' credit union in the most recent 5-year period, 350 were sampled to determine what proportion was made to women. This sample showed that 39% of the loans were made to women employees. A complete census of loans 5 years ago showed that 41% of the borrowers then were women. At a significance level of 0.02, can you conclude that the proportion of loans made to women has changed significantly in the past five years?

Introduction

Business decisions are constantly made in the face of claims: that a new training method raises sales, that a machine fills packets to the stated weight, that the proportion of loans going to women employees has changed. A sample can never prove such a claim with certainty, because another sample would give a slightly different result; what statistics offers instead is a disciplined procedure for deciding whether the sample evidence is strong enough to reject an existing belief, while controlling the risk of a wrong decision. That procedure is hypothesis testing. Part (a) sets out its standard five steps in the form in which the test statistic is

expressed on a standardised (z) scale, and part (b) applies all five steps, in order, to the credit union's loans.

Part (a): The Five-Step Process Using a Standardised Scale

Step 1 — State the null and alternative hypotheses. The null hypothesis, H_o, is the statement of no effect, no difference or no change — the existing belief that the sample is invited to challenge, expressed in terms of the population parameter, for example H_o: P = 0.41. The alternative hypothesis, H_1, is the statement that will be accepted if H_o is rejected; it may be two-sided (≠, meaning "has changed in either direction"), or one-sided (>, or <) when the direction of interest is specified in advance. The hypotheses must be fixed before the sample is examined, and H_o always contains the equality, because it is H_o whose sampling distribution the test uses.

Step 2 — Choose the significance level and the test statistic. The significance level, α, is the probability of a Type I error that the investigator is willing to accept — the risk of rejecting a true H_o. Common choices are 0.05, 0.02 and 0.01; a smaller α means stronger evidence is demanded before rejecting. At the same time the test statistic appropriate to the parameter and the sample size is selected: a z-statistic when the population standard deviation is known or the sample is large, and a t-statistic when the standard deviation must be estimated from a small sample.

Step 3 — Establish the decision rule on the standardised scale. "Standardised scale" means that the test statistic is expressed in standard-error units — how many standard errors the sample result lies from the hypothesised parameter value. From α and the form of H_1, the critical value(s) are read from the standard normal (or t) table: for a two-sided test at α = 0.02 the

critical values are ±2.33, at α = 0.05 they are ±1.96, and at α = 0.01 they are ±2.58. The decision rule is then stated in advance: reject H_o if the computed standardised statistic falls in the rejection region beyond the critical value; otherwise do not reject H_o.

Step 4 — Compute the standardised test statistic from the sample. The general form is: standardised statistic = (sample statistic − hypothesised parameter) ÷ standard error of the statistic. For a proportion this is $z = (\hat{p} - P) / \sqrt{[P(1 - P)/n]}$, where the standard error is computed under H_o, that is, using the hypothesised P. This single number summarises the whole sample evidence on a scale whose behaviour, when H_o is true, is fully known.

Step 5 — Decide and conclude in the language of the problem. The computed statistic is compared with the critical value and H_o is either rejected or not rejected — note that we never "accept" or "prove" H_o; a non-rejection means only that the evidence is insufficient to reject it. The statistical decision is then translated into a practical conclusion about the original question, together, where useful, with the p-value, the probability of a sample at least as extreme as the one obtained if H_o were true. Throughout, the two possible errors must be kept in mind: a Type I error (rejecting a true H_o, probability α) and a Type II error (failing to reject a false H_o, probability β, which falls as the sample size grows).

Part (b): Has the Proportion of Loans to Women Changed?

Step 1 — Hypotheses. The census figure of five years ago, P = 0.41, is the established position, and the question asks whether the proportion has changed — in either direction — so the test is two-sided: H_o: P = 0.41; H_1: P ≠ 0.41. Step 2 — Significance level and statistic: α = 0.02 as given, and with n = 350 the normal approximation to the sampling distribution of $\hat{p}$ is fully justified (nP and n(1 − P) are both far above 5), so the z-statistic is used. Step 3 — Decision rule:

for a two-sided test at α = 0.02, 1 per cent of the area lies in each tail and the critical values are z = ±2.33. Reject H_o if the computed z is less than −2.33 or greater than +2.33.

Step 4 — Computation. The sample proportion is $\hat{p}$ = 0.39 from n = 350 loans. The standard error, computed under H_o, is $\sigma_p = \sqrt{[P(1 - P)/n]} = \sqrt{[(0.41 \times 0.59)/350]} = \sqrt{(0.000691)} = 0.0263$. Hence z = (0.39 − 0.41) / 0.0263 = −0.02 / 0.0263 = −0.76. (A finite population correction, √[(10,200 − 350)/(10,200 − 1)] = 0.983, could be applied because the sample is drawn from a stated population of 10,200 loans; it reduces the standard error slightly to 0.0258 and changes z only to −0.77, so the decision is identical and the simpler standard form is reported.) Step 5 — Decision and conclusion: the computed value z = −0.76 lies well inside the non-rejection region between −2.33 and +2.33, and its two-sided p-value is about 0.45, far above 0.02. We therefore do not reject H_o.

In practical language: no, at the 2 per cent level of significance we cannot conclude that the proportion of loans made to women has changed in the past five years. The fall from 41 per cent to 39 per cent in the sample is only three-quarters of one standard error — exactly the size of movement that ordinary sampling variation produces when nothing real has changed. This is not proof that the proportion is still exactly 41 per cent; it is the statement that a sample of 350 provides no significant evidence of change. If the credit union needs to detect a change as small as two percentage points, it must examine a considerably larger sample, because only a larger n shrinks the standard error enough to make such a small difference statistically visible.

Conclusion

To conclude, hypothesis testing on a standardised scale proceeds in five disciplined steps — state H_o and H_1, fix α and choose the statistic, set the critical-value decision rule before

computing, calculate the standardised statistic in standard-error units, and decide and interpret. Applied to the credit union, the procedure gives z = −0.76 against critical values of ±2.33 at the 0.02 level, so the null hypothesis stands: the proportion of loans made to women, 39 per cent in the sample against 41 per cent five years ago, has not changed significantly.

AIOU LIBRARY
aioulibrary.com
AIOU LIBRARY