ICS Part-I Statistics — Punjab Board

CHAPTER 2

Representation of Data

Key Concepts · Definitions · Short & Long Questions · MCQs · Diagrams · Exam Tips

Chapter 2: Representation of Data

Most scientific experiments and surveys result in the collection of data in the form of batches of numbers, usually called a sample. To extract useful information from this raw sample, it must first be organized and summarized. This chapter covers the two broad approaches to doing that: tabular methods (classification, tabulation, frequency distributions, and cumulative frequency distributions) and graphical methods (bar diagrams, pie diagrams, histograms, frequency polygons, and scatter plots).

A frequency distribution condenses raw data into a compact table by grouping similar values into classes and counting how many observations fall in each class. Building one well requires deciding on the number of classes, the class width, and the class boundaries, and the same idea extends naturally to cumulative frequencies (running totals) and, for two variables measured together, bivariate frequency distributions. Graphs then turn these tables into pictures that are often easier to interpret than the numbers alone, and this chapter covers the standard toolkit: simple, multiple, and sub-divided bar diagrams; pie diagrams; histograms for equal and unequal class widths; frequency polygons and cumulative frequency polygons (ogives); and scatter plots for paired data.

Learning Objectives

- Explain classification and tabulation, and identify the parts of a well-constructed table
- Construct a frequency distribution for continuous, discrete, and categorical data
- Calculate the number of classes, class width, and class boundaries for grouped data
- Compute relative frequency, cumulative frequency, and cumulative relative frequency
- Distinguish simple, multiple, and sub-divided bar diagrams, and construct each

- Construct a pie diagram using the correct sector-angle formula
- Construct a histogram for both equal and unequal class intervals
- Construct a frequency polygon and a cumulative frequency polygon (ogive)
- Construct and interpret a scatter plot for bivariate data
- Construct a bivariate frequency distribution for paired measurements

Key Concepts

2.2 Classification

Classification is the process of arranging observations into different classes or categories according to some common characteristic. Data may be classified by one or more characteristics at a time: when data is classified according to one characteristic it is called one-way classification, by two characteristics at a time it is two-way classification, and by three characteristics it is three-way classification.

2.3 Tabulation

Tabulation is the process of making tables, or arranging data into rows and columns. Tabulation may be simple, double, triple, or complex depending on the number of characteristics involved. A well-constructed table has several standard parts: the title (a brief, self-explanatory heading at the top describing the table's contents); column captions, collectively called the box-head (brief, clear headings for each column, arranged in order of importance); row captions, collectively called the stub (brief, clear headings for each row); the body of the table (the actual entries in each cell, which may be arranged qualitatively, quantitatively, chronologically, geographically, or alphabetically); source notes (given at the end of the table, indicating the compiling agency, publication, date, and page); and prefatory notes and footnotes (used to explain the table's contents, with footnotes usually indicated by asterisks).

Spacing and ruling are used to enhance a table's effectiveness and to separate items within it: thick, double, or single lines separate row and column captions, and dots or dashes (never zeroes) are used to indicate no entry in a cell.

2.4 Frequency Distribution

A frequency distribution is a compact tabular form of data that displays categories of observations according to their magnitude and frequency, grouping similar or identical values together. These categories are also called groups, class intervals, or classes, and the number of values falling in a given category is its frequency, denoted f. The relative frequency (r.f) of a category is its frequency divided by the total frequency; the sum of all relative frequencies should equal 1 (except for rounding), and relative frequencies are especially useful for comparing distributions built from samples of different sizes.

Building a frequency distribution for continuous data follows five steps. First, calculate the range: Range = maximum value - minimum value. Second, decide the number of classes c, using c = 1 + 3.3 log(n) or approximately c = square root of n, where n is the total number of observations — too few wide classes loses variation in the data, while too many narrow classes hardly groups the values at all. Third, decide the class width h = Range / c (approximately), rounding to a convenient number. Class limits are the stated end points of each class interval; because it's best if no observation falls exactly on a boundary, class limits are converted to class boundaries by extending them one more decimal place, so that the upper boundary of one class equals the lower boundary of the next, keeping classes mutually exclusive. Fourth, decide the starting point of the first class, usually just below the minimum value so that the class midpoint (average of its lower and upper limits) is well placed. Fifth, tally each observation into its class to obtain the frequency of each class.

For discrete data, since each observation is already a whole number, the possible values are simply listed and tallied directly, without needing to compute a range or class width. For categorical data, the categories themselves are listed and tallied. An open-end class is one where either the lower limit of the first class or the upper limit of the last class is not a fixed number (for example, 'Below 5' or '75 and above') — these arise naturally in some practical situations, such as age distributions.

2.5 Cumulative Frequency Distribution

A cumulative frequency distribution displays class intervals alongside their cumulative frequencies. The cumulative frequency (c.f) of a class is obtained by adding the frequencies of all preceding classes, including that class itself, and it indicates the total number of values less than or equal to the upper limit of that class. The cumulative relative frequency (c.r.f) is the cumulative frequency divided by the total frequency (equivalently, the running sum of relative frequencies); multiplying c.r.f by 100 gives the percentage cumulative frequency. When comparing two or more distributions of different sample sizes, relative or percentage cumulative frequencies should be used instead of raw cumulative frequencies, since raw counts are distorted by differing sample sizes. The same logic applies directly to discrete data.

2.6 Graphic Representation of Data

A single well-made graph can convey more than pages of prose: graphs help check assumptions about the data, provide a subjective check on formal calculations, suggest which statistical analysis is appropriate, communicate results clearly to readers, and help spot variability and outliers (values highly inconsistent with the rest of the data). Good graphing practice includes clearly labelling both axes with variable names and units, keeping the scale uniform along each axis, keeping the diagram simple, choosing a clear and concise title, using identical scales when comparing graphs, and — for scatter plots — never joining the dots, since this can create the illusion of a pattern in random scatter.

2.6.1-2.6.3 Bar Diagrams

A simple bar diagram represents a discrete or categorical data set by drawing a bar for each category or value, with the bar's height equal to its frequency, and the values or categories placed along the x-axis. The gaps between bars emphasize the gaps between the discrete values the variable can take. A multiple bar diagram extends this to compare two or more related sets of data at once, drawing a group of bars (one per set) for each category, side by side. A sub-divided (component) bar diagram is used when a simple bar represents a total that can be broken into segments — for example, a bar for total population can be sub-divided into male and female portions within the same bar.

2.6.4 Pie Diagram

A pie diagram (or pie chart) divides a circular region into sectors representing the components of a whole. Each sector's angle Q is found using Q = (Component Part / Total) x 360 degrees, since a full circle is 360 degrees. Each sector is shaded differently so the parts are visually distinguishable. Pie diagrams are useful for displaying how a whole splits into component parts, and for comparing such divisions across different times or groups.

2.6.5 Histogram

A histogram gives a visual impression of the distribution of grouped data by drawing rectangles: class boundaries are marked along the x-axis and frequencies along the y-axis. For equal class widths, the height of each rectangle is simply proportional to (equal to) its class frequency. For unequal class widths, the area of each rectangle — not its height — must be proportional to frequency, so the height is instead obtained as an adjusted frequency: class frequency divided by class width. The area under a full histogram equals the sum of the areas of all its rectangles (width x frequency for each class), which for equal-width classes equals class width multiplied by the total frequency. A histogram may also be constructed for discrete grouped data, by centering a rectangle on each possible value with equal width to either side (commonly 0.5). The main advantages of a histogram over unprocessed data are that it reveals the range, location, and skewness of the data, and can flag out-of-control or unusual observations.

2.6.6 Frequency Polygon and Cumulative Frequency Polygon

A frequency polygon is a closed geometric figure displaying a frequency distribution: mid-values of the class boundaries are marked on the x-axis, frequencies on the y-axis, the corresponding points are joined, and the resulting line is extended to meet the x-axis at both ends. It can equivalently be built by joining the upper midpoints of a histogram's rectangles and extending both ends to the axis. A smoothed frequency polygon is called a frequency curve, useful for visually assessing symmetry or skewness.

A cumulative frequency polygon (also called an ogive) plots upper class boundaries against cumulative frequencies (or cumulative relative/percentage frequencies): it is an increasing curve that starts at zero height at the lower boundary of the first class and rises to the total frequency at the upper boundary of the last class. It can be used to read off values corresponding to particular cumulative proportions, such as quartiles or percentiles. For discrete data, the cumulative frequency polygon instead makes a vertical jump at each data value (equal to that value's frequency) and stays flat between data values, since there is nothing to accumulate in between.

2.6.7 Scatter Plots

When two variables are measured on each individual — for example, height and weight — the resulting n pairs of observations (xi, yi) form a bivariate data set. A scatter plot displays this by placing one variable on the x-axis and the other on the y-axis, marking each pair as a point (a cross or dot). Bivariate data of this kind broadly falls into two types: paired measurements on the same variable (such as morning versus evening milk yield from the same cows, where the line of equality at 45 degrees is a useful visual reference), and two related but different measurements (such as soil nitrogen versus crop yield, where the interest is simply whether the two are related, since it makes no sense to directly compare values of different variables to each other).

2.7 Bivariate Frequency Distribution

When two variables are considered together, the resulting table is called a bivariate frequency distribution. It is constructed the same way as an ordinary frequency distribution, except that each pair of observations (xi, yi) is allocated to a cell formed by the intersection of a class of the first variable and a class of the second variable — for example, allocating a (height, weight) pair to the cell corresponding to its height class and its weight class simultaneously.

Important Definitions

What is classification in statistics?: The process of arranging observations into different classes or categories according to some common characteristic.

What is tabulation?: The process of making tables, or arranging data into rows and columns; may be simple, double, triple, or complex.

What is a frequency distribution?: A compact table that displays categories (classes) of observations along with the number of observations (frequency) falling in each category.

What is relative frequency?: The frequency of a class divided by the total frequency; the sum of all relative frequencies in a distribution should equal 1, except for rounding.

What is cumulative frequency?: The running total obtained by adding the frequencies of all preceding classes, including the class itself; it shows how many values are less than or equal to that class's upper limit.

What is an open-end class?: A class in a frequency table where either the lower limit of the first class or the upper limit of the last class is not a fixed number, such as 'Below 5' or '75 and above'.

What is a histogram?: A graph of a grouped frequency distribution using adjoining rectangles over class boundaries, where rectangle height (equal widths) or area (unequal widths) is proportional to frequency.

What is a scatter plot used for?: To visually display and study the relationship between two variables measured on the same individuals, by plotting each pair of values as a point.

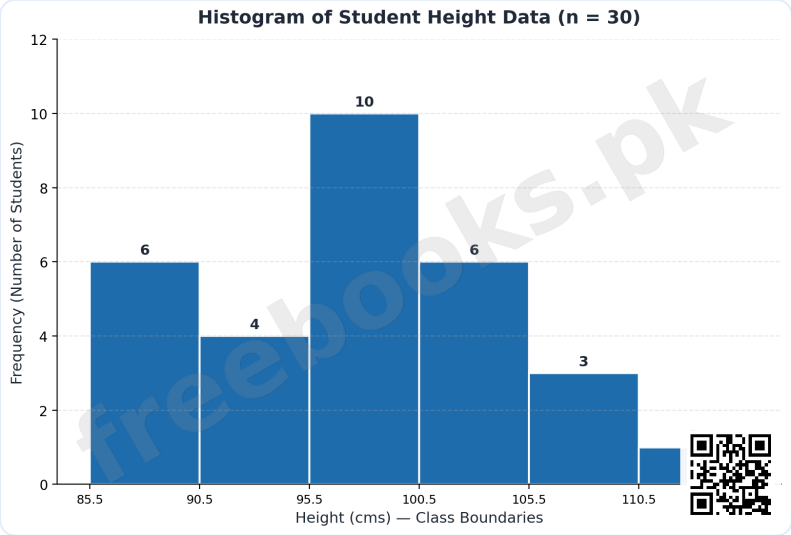
Key Facts and Relations

Topic	Key Fact / Relation
Range	Range = Maximum value - Minimum value
Number of classes (c)	c = 1 + 3.3 log(n), or c = sqrt(n) approximately
Class width (h)	h = Range / number of classes (approximately)
Relative frequency	r.f = frequency of class / total frequency

Topic	Key Fact / Relation
Cumulative relative frequency	c.r.f = cumulative frequency / total frequency
Percentage cumulative frequency	Percentage c.f = c.r.f x 100
Pie chart sector angle	Q = (Component Part / Total) x 360 degrees
Histogram rectangle height (unequal class width)	Adjusted frequency = class frequency / class width
Area of a histogram rectangle	Area = width of class x frequency of class

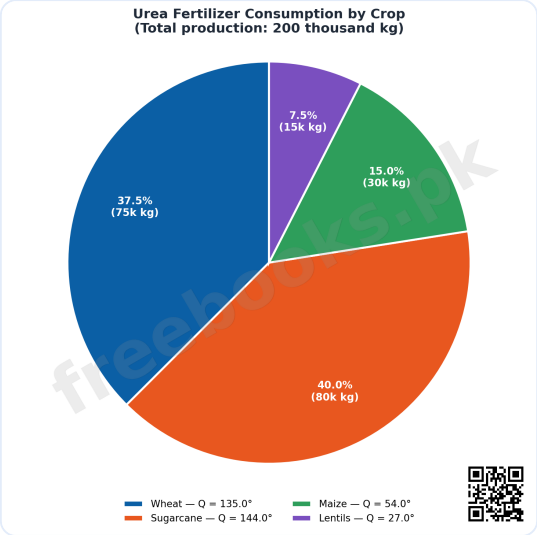
Diagrams

Histogram of a Frequency Distribution: A histogram built from the worked student-height example, showing class boundaries on the x-axis and frequency as bar height on the y-axis, for equal-width classes



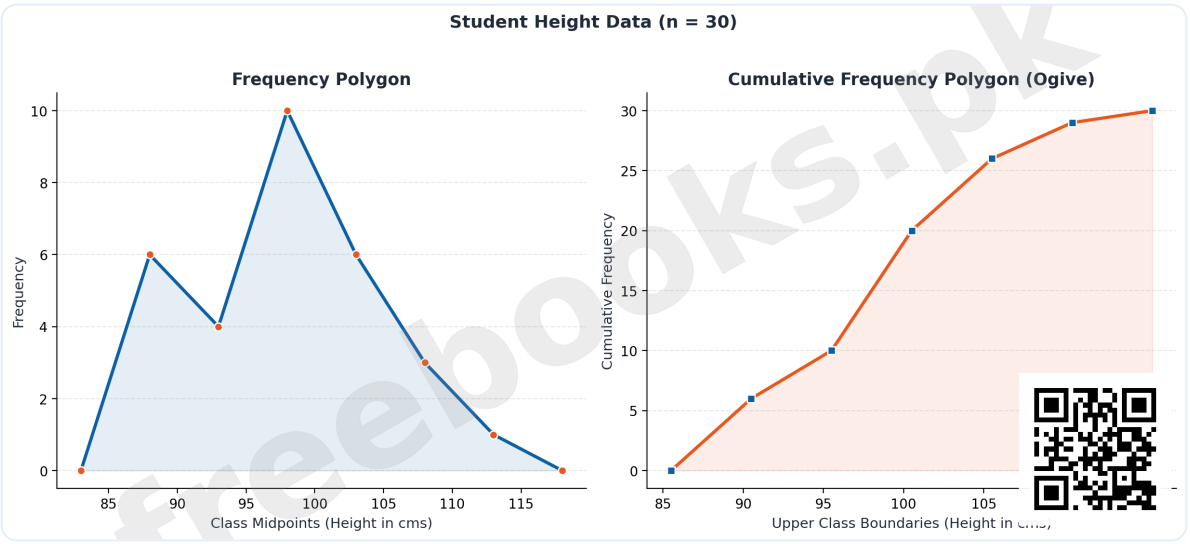
Histogram of a Frequency Distribution

Pie Diagram of Component Parts: A pie chart dividing a circular region into sectors by component share, illustrating the sector-angle formula Q = (part/total) x 360 degrees



Pie Diagram of Component Parts

Frequency Polygon vs. Cumulative Frequency Polygon (Ogive): Two related line graphs from the same grouped data: a frequency polygon joining class midpoints against frequency, and a cumulative frequency polygon (ogive) showing running totals rising to the total frequency



Frequency Polygon vs. Cumulative Frequency Polygon (Ogive)

Short Questions & Answers

What is the difference between one-way and two-way classification?

One-way classification arranges data by a single characteristic at a time, while two-way classification arranges data by two characteristics simultaneously.

Name the main parts of a well-constructed table.

Title, column captions (box-head), row captions (stub), body of the table, source note, and prefatory notes/footnotes.

How is the number of classes for a frequency distribution decided?

Using the formula c = 1 + 3.3 log(n), or approximately c = square root of n, where n is the number of observations.

What is the difference between class limits and class boundaries?

Class limits are the stated end points of a class; class boundaries are obtained by extending the limits one more decimal place so classes become continuous and mutually exclusive, with no observation falling exactly on a boundary.

How does a multiple bar diagram differ from a simple bar diagram?

A simple bar diagram shows one bar per category; a multiple bar diagram groups several related bars (e.g., different years or groups) together at each category for comparison.

What is the formula for the angle of a sector in a pie diagram?

Q = (Component Part / Total) x 360 degrees.

How is the height of a histogram rectangle found when class widths are unequal?

By dividing the class frequency by the class width, giving an adjusted frequency, since the rectangle's area (not height) must be proportional to frequency.

What does a cumulative frequency polygon (ogive) show?

It shows the running (cumulative) frequency up to the upper boundary of each class, rising from zero at the start of the first class to the total frequency at the end of the last class.

Long Questions & Answers

Explain, with the steps involved, how a frequency distribution is constructed for continuous data.

A frequency distribution condenses a raw batch of numbers into a compact table that groups similar values together and shows how many observations fall in each group, making the data far easier to interpret than a long unordered list. Constructing one for continuous data follows five main steps. The first step is to

calculate the range of the data, which is simply the maximum value minus the minimum value; this tells us the total spread that the classes will need to cover. The second step is to decide on the number of classes, commonly denoted c. This is not arbitrary — a widely used guideline is c = 1 + 3.3 log(n), where n is the total number of observations, or approximately c = the square root of n. Getting this number right matters: too few, overly wide classes will lose most of the real variation present in the data, while too many narrow classes will scatter the data so thinly that patterns become hard to see. The third step is to decide the width of each class, usually denoted h, calculated as h = Range divided by the number of classes, and then rounded to a convenient number for ease of use — it is not necessary to follow strict rounding rules here, since this is only a grouping decision. The fourth step is to decide where the first class should start; this is somewhat arbitrary, but it is usually chosen to begin just below the minimum value in the data, in such a way that the midpoint of the first class (the average of its lower and upper limits) falls at a sensible, representative value. An important related idea here is the distinction between class limits and class boundaries: class limits are the stated end points of each class as originally chosen, but because it is best if no actual observation falls exactly on a boundary, class limits are converted into class boundaries by extending each end point by one additional decimal place. This is done so that the upper boundary of one class exactly equals the lower boundary of the next class, which keeps the classes continuous and mutually exclusive, with the convention that an observation falling exactly on a shared boundary is counted in the higher class. The fifth and final step is to go through the data observation by observation, placing a tally mark against whichever class each value belongs to, keeping a running tally until every observation has been placed; the resulting tally counts, once totalled, are the frequencies of each class, and together they form the completed frequency distribution table. Once this base frequency distribution exists, several other useful values can be derived from it directly, such as class centres (the midpoints of each class, used to represent that whole class in later calculations), relative frequencies (each class frequency divided by the total frequency, useful for comparing distributions of different sample sizes), and cumulative frequencies (running totals of frequency up to and including a given class).

Describe the different types of bar diagrams and the pie diagram, explaining when each is the appropriate choice.

Bar diagrams and pie diagrams are among the most common ways of turning a frequency distribution into an easily readable picture, and choosing the right one depends on exactly what the data is meant to show. The simple bar diagram is the most basic form: the categories or values of a discrete or categorical variable are placed along the x-axis, and for each one, a single bar is drawn whose height equals its frequency. Because the underlying variable is discrete or categorical rather than continuous, gaps are deliberately left between the bars, visually emphasizing that the values are separate and distinct rather than part of a continuous scale — this is the key visual difference between a bar diagram and a histogram. The multiple bar diagram is an extension of the simple bar diagram, used specifically when there are two or more related sets of data that need to be compared side by side for each category. Rather than drawing a single bar per category, a small group of bars is drawn together at each category position, with one bar in the group representing each data set — for example, when comparing wheat production across three localities over three different years, one would draw three bars (one per locality) grouped together for each year, allowing the reader to compare the localities directly within each year and across years at a glance. The sub-divided bar diagram, also called a component bar diagram, is used in situations where a simple bar already represents some overall total, but that total can meaningfully be broken down further into its component parts within the same bar — for instance, if a simple bar shows the total number of people in a blood-group category, that same bar can be sub-divided into a male portion and a female portion stacked within it, so a single bar simultaneously shows both the total for that category and how that total splits internally. Moving beyond bar diagrams entirely, the pie diagram takes a different visual approach: instead of using bar heights, it divides an entire circular region into sectors, where each sector's size represents that component's share of the whole. The angle of each sector, out of the full 360 degrees in a circle, is calculated using the formula Q equals the component part divided by the total, multiplied by 360 degrees; once each sector's angle has been calculated this way, the circle can be divided accordingly and each sector shaded differently to keep them visually distinct. A pie diagram is the natural choice specifically when the purpose is to show how a single whole divides into its component parts — for example,

showing how total fertilizer production splits across different crops — and it is also useful for comparing how such a division changes between different times or groups, though it becomes visually cluttered and hard to read if there are too many components to display clearly.

Explain the different graphical tools used to display a frequency distribution — histogram, frequency polygon, and cumulative frequency polygon (ogive) — and describe how each is constructed.

Once a frequency distribution has been tabulated, several standard graphical tools can turn that table into a visual picture that is often far easier to interpret than the raw numbers, and three of the most important of these are the histogram, the frequency polygon, and the cumulative frequency polygon. A histogram is constructed by marking the class boundaries of the distribution along the x-axis and the frequencies along the y-axis, then drawing a rectangle over each class. When all classes have equal width, as is most common, the height of each rectangle is simply made proportional to (in practice, equal to) that class's frequency, so taller rectangles directly represent more observations. However, when classes have unequal widths — which happens in highly skewed distributions or when similar cases are deliberately grouped together — it is the area of each rectangle, not its height, that must remain proportional to frequency; the height in this case is instead calculated as an adjusted frequency, found by dividing each class's frequency by its class width, since a wider class would otherwise misleadingly appear taller than its true frequency warrants if height alone were used. The frequency polygon offers a related but distinct way of visualizing the same distribution using lines rather than bars: the midpoint of each class boundary is calculated and marked along the x-axis, the corresponding frequency is marked along the y-axis, and these plotted points are then joined together with straight lines, with the resulting line extended down to meet the x-axis at both the beginning and the end, turning it into a closed figure. A frequency polygon can equally well be constructed by joining the upper midpoints of the rectangles of an already-drawn histogram and then extending both ends down to the axis; when this polygon is smoothed into a continuous curve rather than straight line segments, it is referred to as a frequency curve, and it is particularly useful for getting a quick visual sense of whether the data's distribution is symmetric or skewed in one direction. The cumulative frequency

polygon, also widely known as an ogive, takes this a step further by plotting running totals rather than individual class frequencies: the upper class boundaries are plotted along the x-axis against their corresponding cumulative frequencies (or, alternatively, cumulative relative or percentage cumulative frequencies) along the y-axis. Because cumulative frequency can only ever increase or stay the same as more classes are included, the resulting ogive is always an increasing curve, starting at a height of zero at the lower boundary of the very first class and rising steadily until it reaches the total frequency at the upper boundary of the very last class. One particularly useful practical application of the ogive is that it allows specific values to be read off the graph corresponding to a chosen cumulative proportion, making it a convenient tool for estimating quartiles, percentiles, or any other position-based summary measure directly from the graph without further calculation. For discrete data specifically, the same cumulative idea is represented slightly differently: rather than a smoothly rising curve, the cumulative frequency polygon makes a distinct vertical jump at each individual data value, with the height of each jump equal to that value's own frequency, and the line stays perfectly flat and horizontal in between consecutive data values, since there is nothing further to accumulate until the next actual data point is reached.

Multiple Choice Questions (MCQs)

The process of arranging data into rows and columns is called: (A) Classification (B) Tabulation (C) Correlation (D) Regression

Correct answer: (B) Tabulation. Tabulation is the process of making tables, or arranging data into rows and columns.

The heading for different columns in a table is called: (A) Stub (B) Box-head/column captions (C) Source note (D) Footnote

Correct answer: (B) Box-head/column captions. The headings for different columns are called column captions, and this part of the table is called the box-head.

The formula for approximate number of classes in a frequency distribution is: (A) c = Range / h (B) c = 1 + 3.3 log(n) (C) c = n / 2 (D) c = h x Range

Correct answer: (B) c = 1 + 3.3 log(n). The number of classes c is approximated by c = 1 + 3.3 log(n), or c = sqrt(n).

Class width h is calculated as: (A) h = Range x number of classes (B) h = Range / number of classes (C) h = number of classes / Range (D) h = Range + number of classes

Correct answer: (B) h = Range / number of classes. Class width h = Range divided by the number of classes (approximately).

Cumulative frequency of a class shows: (A) The frequency of only that class (B) The number of values less than or equal to the upper limit of that class (C) The relative frequency of that class only (D) The class width

Correct answer: (B) The number of values less than or equal to the upper limit of that class. Cumulative frequency indicates the total number of values less than or equal to the upper limit of that class.

In a pie diagram, the angle for a sector is found using: (A) (Part / Total) x 100 (B) (Part / Total) x 360 (C) Part x Total (D) (Total / Part) x 360

Correct answer: (B) (Part / Total) x 360. The sector angle Q = (Component Part / Total) x 360 degrees.

For a histogram with unequal class widths, the height of each rectangle equals: (A) The class frequency (B) The class frequency divided by class width (C) The class width divided by frequency (D) The cumulative frequency

Correct answer: (B) The class frequency divided by class width. For unequal class widths, height = class frequency / class width (the adjusted frequency), so area stays proportional to frequency.

A cumulative frequency polygon is also known as: (A) A histogram (B) An ogive (C) A scatter plot (D) A pie chart

Correct answer: (B) An ogive. The cumulative frequency polygon is also called an ogive.

A scatter plot is used to display: (A) A single variable's frequency distribution (B) The relationship between two variables measured on the same individuals (C) Cumulative frequencies only (D) Sector angles

Correct answer: (B) The relationship between two variables measured on the same individuals. A scatter plot displays bivariate data — pairs of values (xi, yi) for two variables measured together.

A frequency table where the last class is written as '75 and above' is an example of: (A) A bivariate distribution (B) An open-end class (C) A cumulative distribution (D) A pie diagram

Correct answer: (B) An open-end class. An open-end class is one where the lower limit of the first class or the upper limit of the last class is not a fixed number, like '75 and above'.

Quick Revision Summary

- Classification = arranging data into classes/categories; Tabulation = arranging data into rows and columns
- Table parts: title, column captions (box-head), row captions (stub), body, source note, footnotes
- Frequency distribution steps: Range -> number of classes (c = 1+3.3log n) -> class width (h = Range/c) -> starting point -> tally
- Relative frequency = frequency / total; Cumulative frequency = running total of frequencies up to that class
- Simple bar diagram = one bar per category; Multiple bar = grouped bars for comparison; Sub-divided bar = one bar split into segments
- Pie chart sector angle: Q = (part/total) x 360 degrees
- Histogram: height = frequency (equal class widths) or frequency/width (unequal class widths)
- Frequency polygon joins class midpoints to frequency; Ogive (cumulative frequency polygon) is always increasing
- Scatter plot displays bivariate (paired) data as points; never join the dots
- Bivariate frequency distribution: each pair (xi, yi) allocated to the intersection of two variables' classes

Exam Tips

- Remember the 5-step recipe for building a frequency distribution: Range -> classes -> width -> start point -> tally

- Class limits vs. boundaries: boundaries always have one extra decimal place and never coincide with an actual observation
- For histograms: equal class widths -> height = frequency; unequal class widths -> height = frequency/width
- Pie chart angle formula (part/total x 360) is a common numerical question — practice with different totals
- Ogive = cumulative frequency polygon = always rises, never falls
- In scatter plots, never connect the dots with lines — that's a common mistake examiners look for