ICS Part-I · Punjab Board · English Medium

Statistics 1st Year

Complete Notes — All 9 Chapters

freebooks.pk

Table of Contents

Prepared by freebooks.pk — original student notes covering the Punjab Curriculum and Textbook Board (PTB/PCTB) syllabus. Each chapter includes key concepts, definitions, formulas, diagrams, solved examples, short and long questions, MCQs, revision and exam tips.

Chapter 1: Introduction to Statistics

Statistics is the mathematical science of making decisions and drawing conclusions from data in situations of uncertainty. It includes designing experiments, and the collection, organization, summarization, analysis, and interpretation of numerical data. Although the word statistics was first used in the 18th century to describe the collection of data by the state, the modern subject has grown far beyond that, now including probability theory and applied mathematics, and drawing on computers to perform routine statistical analysis.

This chapter introduces the basic vocabulary and ideas that the rest of the subject builds on: population and sample, parameter and statistic, the two broad types of variables (quantitative and qualitative), and the distinction between descriptive statistics (organizing and summarizing data) and inferential statistics (drawing conclusions about a population from a sample). It also introduces the shorthand notations — sigma (sum), product, and factorial — used throughout the subject, and surveys where statistics is applied across social sciences, plant sciences, physical sciences, and medical sciences.

Learning Objectives

- Define statistics and trace briefly how its meaning has evolved since the 18th century
- Distinguish between population and sample, and between parameter and statistic
- Explain ratio, proportion, percentage, and sampling variability with examples
- Distinguish between an experiment and a sample survey
- Describe constants, order statistics, and the linear additive model (Yi = mu + epsilon)
- Use sigma, product, and factorial notation correctly
- Identify quantitative (discrete/continuous) and qualitative (categorical) variables
- Distinguish between descriptive statistics and inferential statistics

Key Concepts

1.1 What Is Statistics?

The word statistics was first used by the German scholar Gotfried Achenwall in the mid-18th century for the science of statecraft concerning the collection and use of data by the state. Through the 19th century, the word acquired a much wider meaning, covering numerical data of almost any subject along with the interpretation of that data through appropriate analysis. Since the early 1920s, the growth of the experimental sciences created a need for reliable scientific methods of analyzing the results of experiments and surveys, and the modern subject of statistics grew largely out of the practical problems faced by experimenters themselves.

Today, statistics is defined as the mathematical science of making decisions and drawing conclusions from data in situations of uncertainty. It includes designing experiments, and the collection, organization, summarization, analysis, and interpretation of numerical data. In everyday usage, the word 'statistics' often just refers to numbers or facts — such as statistics of births or road accidents — while 'a statistic' (singular) is a specific statistical term for a quantity calculated from sample values. Data is the general term for numerical facts, and datum is a single numerical fact.

1.2 Population, Sample, Parameter, and Statistic

Population is the total group under discussion, or the group to which the results will be generalized. Sample is a subset of the population chosen when measurements cannot practically be made on the whole population; for a sample to give meaningful inferences about the population, it must be representative of that population.

Parameter is a quantity computed from a population when the entire population is available — parameters are fixed, constant quantities, though they are usually unknown in practice. Statistic is a quantity computed from a sample, and unlike a parameter, a statistic is variable because it changes from sample to sample. This natural variation across different possible samples from the same population is called sampling variability.

1.3 Ratio, Proportion, and Percentage

Ratio is the fraction A/B, the ratio of quantity A to quantity B. Proportion is a special ratio of a part to its total — for example, if 300 of 500 surveyed students are female, the proportion of females is 300/500. A proportion becomes a percentage when multiplied by 100; in this example, the female percentage is (300/500) x 100 = 60%, meaning 60 out of every 100.

These simple ideas — ratio, proportion, and percentage — recur constantly throughout statistics, from describing sample composition to summarizing survey results, and form the foundation for later measures like relative frequency.

1.4 Experiments and Sample Surveys

An experiment is any study in which the researcher can control the allocation of treatments to the experimental units; every unit must be capable of receiving every treatment, and an allocation mechanism decides which unit receives which treatment. A mechanism where the scientist observes a unit and then decides which treatment to apply is considered unsatisfactory, since it risks subjective bias — a scientifically acceptable mechanism instead assigns treatments according to a fixed rule.

A sample survey, by contrast, has no treatments at all: the units in the population under study are listed in a frame, and a sample of units is selected from that frame using a selection mechanism. The feature that distinguishes an experiment from a sample survey is therefore control over allocation (experiments) versus control over selection (surveys).

1.5 Constants, Order Statistics, and Models

A constant is a quantity that does not vary from individual to individual, such as pi (about 3.14159) or e (about 2.71828). The order statistic of a data set Y1, Y2, ..., Yn is simply that data arranged in order of magnitude, denoted Y(1), Y(2), ..., Y(n), where Y(1) is the minimum and Y(n) is the maximum. Data arranged from smallest to largest is in ascending order; from largest to smallest, it is in descending order.

A model is a mathematical statement used to study the results of an experiment or predict the behaviour of future repetitions of it, and typically involves probability distributions describing the variability we would expect across different samples. The simplest linear additive model describes an observation as a mean plus a random error: Yi = mu + epsilon_i, where mu is the population mean and epsilon_i is the random error (the chance variation in an observational process, not to be confused with a human 'mistake'). Rearranging gives epsilon_i = Yi - mu, the deviation of an observation from the mean; since random errors are assumed to average to zero, there are approximately equal numbers of positive and negative deviations.

1.6 Notations: Sigma, Product, and Factorial

Sigma (the Greek letter Sigma) is shorthand for summation: the sum of Y1 through Y5 is written as the sum, from i=1 to 5, of Yi. It tells us to add all the Yi values starting at the lower limit (i=1) and stopping at the upper limit (i=5), assuming consecutive integer values unless stated otherwise.

Product (the Greek letter Pi) is shorthand for multiplication: the product of Y1 through Yn is written as the product, from i=1 to n, of Yi, meaning Y1 x Y2 x Y3 x ... x Yn. Factorial, written n!, is defined as n x (n-1) x (n-2) x ... x 3 x 2 x 1 — for example, 4! = 4x3x2x1 = 24. By definition, 1! = 1 and, notably, 0! = 1 as well.

1.7 Importance of Statistics in Various Disciplines

Because data collected from any field almost always involves some variability or uncertainty, statistics has applications across virtually every field of research, helping researchers analyze, interpret, and communicate their findings. Statistics applied to economics is called econometrics, applied to biological sciences it is called biometry, and there are similarly named applications like psychometry in other fields.

In social sciences, statistics helps establish relationships between variables through hypothesis testing, with econometric models supporting economic forecasting and planning. In plant sciences, the 'Design of Experiments' technique supports efficient experimentation, such as determining optimum plot sizes for crops like wheat, cotton, and sugarcane. In physical sciences, linear and nonlinear regression models establish cause-and-effect relationships, with

computers now enabling simulation alongside direct experimentation. In medical sciences, statistics supports investigating drug effectiveness and environmental effects, planning future studies, and integrating new data with existing findings.

1.8 Variables and Descriptive vs. Inferential Statistics

A variable is a characteristic that varies from individual to individual in a population, such as plant height or eye colour. Variables are broadly of two types. A quantitative variable is one capable of taking a numerical value, and is further split into continuous variables (which can take any value in an interval, like temperature or plant height) and discrete variables (which take only isolated values, usually positive integers arising from counting, like the number of students in a class). A qualitative or categorical variable is not capable of numerical measurement — observations are instead allocated to mutually exclusive categories, such as sex or colour, and can only be counted.

Descriptive statistics provides procedures for organizing data collected from a sample, summarizing it (through graphs and summary values like mean, proportion, and variance), and presenting those summaries in an understandable form. Inferential statistics is the process of inferring the characteristics of a population based on the characteristics of a sample — since this always involves some risk, inferential conclusions are expressed as probabilistic statements rather than proofs. Probability theory is the discipline that underlies and strengthens inferential statistics.

Important Definitions

What is population in statistics?: The total group under discussion, or the group to which the results of a statistical study will be generalized.

What is a sample?: A subset of the population, chosen because measurements of interest often cannot practically be made on the whole population; a good sample must be representative of the population.

What is a parameter?: A fixed, usually unknown quantity computed from an entire population.

What is a statistic?: A quantity computed from a sample; unlike a parameter, a statistic varies from sample to sample.

What is a quantitative variable?: A variable capable of taking a numerical value, further divided into continuous variables (any value in an interval) and discrete variables (only isolated, usually integer, values).

What is a qualitative (categorical) variable?: A variable that is not capable of numerical measurement, where observations are instead assigned to mutually exclusive categories, such as colour or sex.

What is descriptive statistics?: The branch of statistics concerned with organizing, summarizing, and presenting sample data in an understandable form.

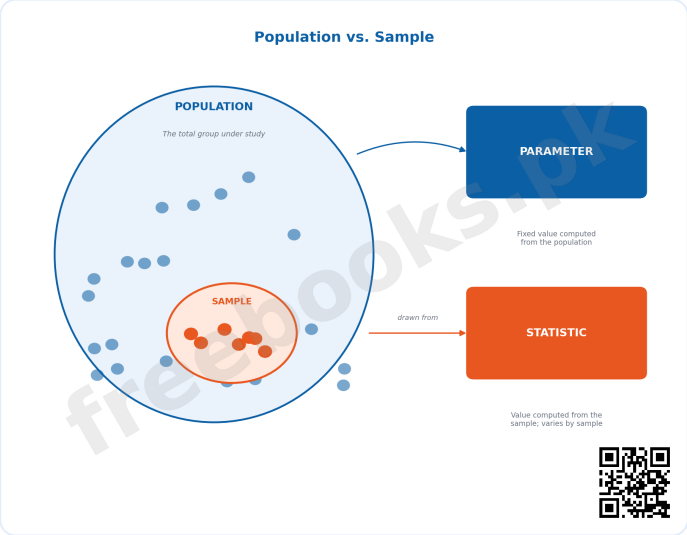
What is inferential statistics?: The process of drawing conclusions about a population's characteristics based on the characteristics observed in a sample, expressed as probabilistic statements.

Key Facts and Relations

Topic	Key Fact / Relation
Sigma (summation) notation	Sum from i=1 to n of Yi = Y1 + Y2 + ... + Yn
Product notation	Product from i=1 to n of Yi = Y1 x Y2 x ... x Yn
Factorial	n! = n(n-1)(n-2)...(2)(1); 0! = 1
Linear additive model	Yi = mu + epsilon_i (observation = mean + random error)
Deviation from the mean	epsilon_i = Yi - mu
Proportion to percentage	Percentage = Proportion x 100
Order statistics	Y(1) = minimum value; Y(n) = maximum value of the ordered data
Statistics applied by field	Economics = Econometrics; Biological Sciences = Biometry

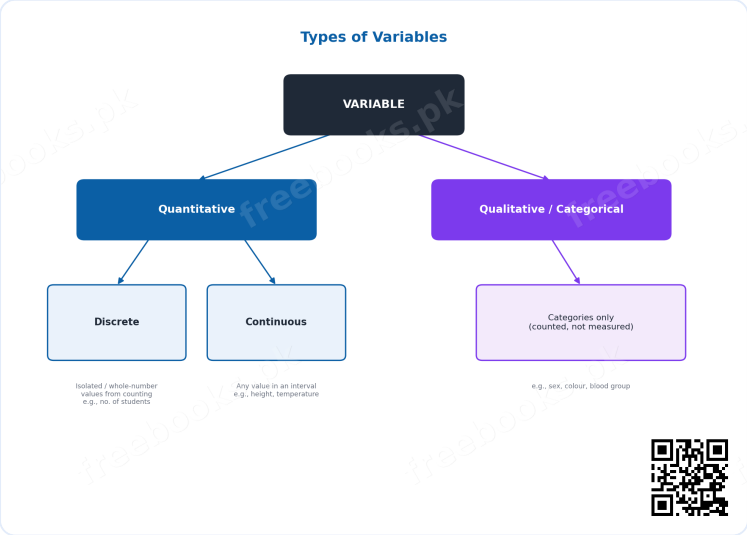
Diagrams

Population vs. Sample: A diagram showing a large population with a representative sample drawn from it, alongside the linked ideas of parameter (population) and statistic (sample)



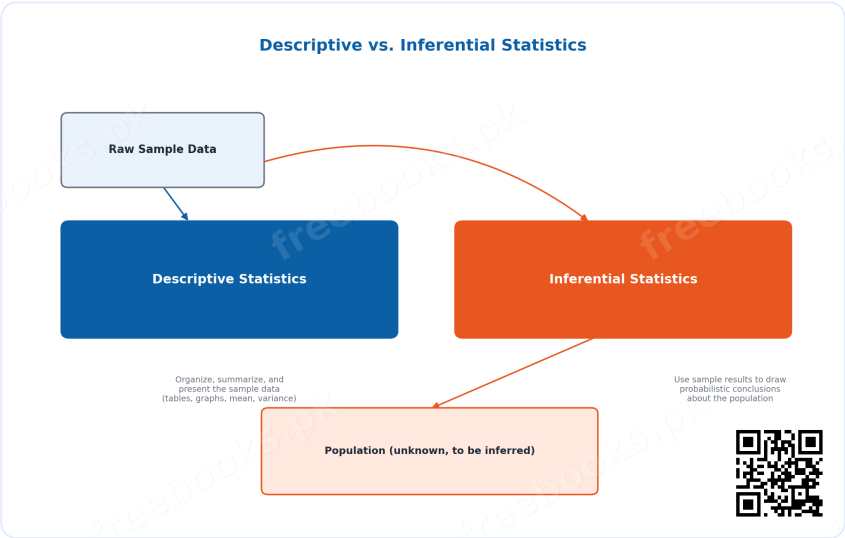
Population vs. Sample

Types of Variables: A classification diagram showing variables split into Quantitative (Discrete, Continuous) and Qualitative/Categorical types, with examples of each



Types of Variables

Descriptive vs. Inferential Statistics: A diagram contrasting descriptive statistics (organizing and summarizing sample data) with inferential statistics (using the sample to draw conclusions about the population)



Descriptive vs. Inferential Statistics

Short Questions & Answers

Define statistics as it is understood today.

Statistics is the mathematical science of making decisions and drawing conclusions from data in situations of uncertainty; it includes designing experiments, and collecting, organizing, summarizing, analyzing, and interpreting numerical data.

Differentiate between population and sample.

Population is the total group under discussion or to which results will be generalized, while a sample is a representative subset of that population chosen when measuring the whole population is not practical.

Differentiate between a parameter and a statistic.

A parameter is a fixed, usually unknown quantity computed from an entire population, while a statistic is a quantity computed from a sample and varies from sample to sample.

What is the difference between an experiment and a sample survey?

An experiment involves the researcher controlling the allocation of treatments to units, while a sample survey has no treatments — it instead involves control over the selection of units from a frame.

What is a discrete variable? Give an example.

A discrete variable is a quantitative variable that can take only isolated values, usually positive integers arising from counting, such as the number of students in a class.

What is a continuous variable? Give an example.

A continuous variable is a quantitative variable that can take any value within an interval on the number line, such as temperature or plant height.

Differentiate between descriptive and inferential statistics.

Descriptive statistics organizes, summarizes, and presents sample data, while inferential statistics uses that sample data to draw probabilistic conclusions about the wider population.

Explain factorial notation with an example.

Factorial, written n!, means n x (n-1) x (n-2) x ... x 2 x 1; for example, 4! = 4x3x2x1 = 24, and by definition 0! = 1.

Long Questions & Answers

Explain the meaning and scope of statistics, tracing briefly how the word's meaning has changed since it was first used, and describe what the modern subject includes.

The word statistics has a history that stretches back to the middle of the 18th century, when the German scholar Gotfried Achenwall first used it to describe the science of statecraft concerned with the collection and use of data by the state. Other early statisticians place its first written appearance even earlier, in a 1770 book called 'The Elements of Universal Erudition,' and the term was used again with a wider definition in 1787, before appearing in the Encyclopedia Britannica by 1797. Through the 19th century, the meaning of the word broadened considerably, coming to cover numerical data on almost any subject, together with the interpretation of that data through appropriate methods of analysis. From the early 1920s onward, the rapid growth of the experimental sciences created a pressing need for reliable, scientific methods of analyzing the results of experiments and surveys, and it was largely out of the practical problems faced by experimenters themselves that the modern subject of statistics developed. Today, statistics is formally defined as the mathematical science of making

decisions and drawing conclusions from data in situations of uncertainty, and its scope now includes the design of experiments, and the collection, organization, summarization, analysis, and interpretation of numerical data. Modern statistics is quite different from its origins as a tool for gathering and presenting government data: it now incorporates probability theory and applied mathematics as core components, and computers have made it possible to perform statistical analysis routinely at a scale and speed that would have been unimaginable in the past, even though computers themselves are simply a tool for applying statistical theory rather than a part of that theory. It is worth noting that in everyday, non-technical usage, the word 'statistics' often simply refers to numbers or facts, such as statistics on births, deaths, or road accidents, and the word is also used as the plural of 'statistic,' a specific technical term referring to a quantity calculated from sample values. Taken together, this history shows that statistics evolved from a narrow, state-focused practice of counting into a broad mathematical science that today underpins research and decision-making across nearly every academic and professional field.

Explain the key statistical terms population, sample, parameter, and statistic, and describe how they relate to one another using a suitable example.

Understanding statistics begins with a small set of core terms that describe how information moves from a large group of interest down to the data that is actually collected and analyzed. The population is the total group under discussion, or the entire group to which the results of a study will eventually be generalized; it represents the full scope of what a researcher ultimately wants to know something about. In many real situations, however, it is impossible or impractical to measure every single member of the population, so a sample is used instead — a chosen subset of the population on which measurements are actually taken. For any inference drawn from a sample to be meaningful and applicable back to the population, it is essential that the sample be genuinely representative of that population, rather than being biased or unusual in some way that does not reflect the population as a whole. Once a population and a sample have been identified, two further terms describe the quantities computed from each: a parameter is a quantity computed from the population when the entire population is available, and because it describes the whole population

directly, a parameter is a fixed, constant value — though in practice it is usually unknown, since measuring an entire population is often infeasible. A statistic, by contrast, is a quantity computed from a sample rather than the whole population, and because different samples drawn from the same population will not be identical to one another, a statistic is inherently variable: its value depends on which particular sample happened to be selected. This natural tendency for different samples from the same population to produce different statistic values is known as sampling variability. A concrete example makes these relationships clear: suppose 500 college students are surveyed and 300 of them are found to be female. Here, the set of all students at the college is the population, while the surveyed group of 500 students is the sample drawn from that population. The true proportion of female students across the entire college is the population's parameter, and it is a fixed but likely unknown value. The proportion of female students found within this particular sample of 500, which works out to 300/500 or 60%, is the sample's statistic — and if a different sample of 500 students were surveyed instead, that second sample might well produce a different proportion, illustrating sampling variability in action. Together, these four terms — population, sample, parameter, and statistic — form the conceptual backbone of the entire subject, since virtually every statistical technique introduced later in the course is ultimately a method for using a sample statistic to say something meaningful and appropriately cautious about an unknown population parameter.

Describe the two broad types of variables used in statistics, with their subtypes, and explain the difference between descriptive and inferential statistics.

A variable, in statistics, is simply a characteristic that varies from one individual to another within a population — since natural phenomena rarely show perfect constancy, almost every measurable characteristic of interest, from plant height to eye colour to exam scores, qualifies as a variable. Statistics divides variables into two broad types based on the kind of measurement involved. A quantitative variable is one that is capable of taking a genuine numerical value, such as the height of a plant, the weight of a bag of grain, or the number of students in a classroom. Quantitative variables are further split into two subtypes depending on what kind of numerical values they can take: a continuous variable can take any value at all within some interval on the number line, such as atmospheric

pressure, plant height, or body temperature, meaning that between any two possible values there are, in principle, infinitely many other possible values; a discrete variable, sometimes called a discontinuous variable, can instead take only isolated, separate points on the number line, and these values are usually positive integers arising from a simple counting process, such as the number of plants in a plot or the number of insects observed in a given area. The second broad type is the qualitative or categorical variable, which is not capable of taking a numerical measurement at all — instead, each observation is simply allocated to one of several mutually exclusive categories, and the resulting data can only be counted rather than measured, as with sex (male or female), general knowledge (poor, moderate, good), or colour (blue, green, red). Once data on a chosen variable has been collected, statistics offers two complementary ways of working with it. Descriptive statistics is concerned with organizing the raw data collected from a sample, summarizing it through graphical representations and calculated summary values such as measures of central value and measures of variability (commonly called statistics, such as the mean, a proportion, or the variance), and then presenting those summaries in a form that other people can readily understand. Inferential statistics goes a meaningful step further: rather than simply describing the sample that was actually collected, it uses the sample's descriptive results to infer or generalize characteristics of the broader population from which that sample was drawn. Because this process of inference is never based on complete information about the population, it is fundamentally supported by probability theory, and as a result, conclusions in inferential statistics are always expressed as probabilistic statements involving some acknowledged degree of risk or uncertainty, rather than as absolute proofs — for example, a researcher might say that the probability is high that an experimental treatment affected an outcome, rather than claiming to have proven that it did with total certainty.

Multiple Choice Questions (MCQs)

Who is credited with first using the word 'statistics' in its modern sense? (A) Sir John Sinclair (B) Gotfried Achenwall (C) E.A.W. Zimmermann (D) Yule

Correct answer: (B) Gotfried Achenwall. Gotfried Achenwall, a German scholar, first used the word statistics in the mid-18th century for the science of statecraft.

The total group to which the results of a study will be generalized is called the: (A) Sample (B) Statistic (C) Population (D) Parameter

Correct answer: (C) Population. The population is the total group under discussion or to which the results will be generalized.

A quantity computed from an entire population is called a: (A) Statistic (B) Parameter (C) Sample (D) Variable

Correct answer: (B) Parameter. A parameter is a fixed quantity computed from a population, usually unknown in practice.

Which of the following best describes a statistic? (A) A fixed, constant population value (B) A quantity computed from a sample that varies from sample to sample (C) A category used for qualitative data (D) A type of experiment

Correct answer: (B) A quantity computed from a sample that varies from sample to sample. A statistic is computed from a sample and varies from sample to sample, unlike a fixed parameter.

Which of these is an example of a discrete variable? (A) Plant height (B) Atmospheric pressure (C) Number of students in a class (D) Body temperature

Correct answer: (C) Number of students in a class. The number of students in a class is a discrete variable, since it takes only isolated, whole-number values arising from counting.

Which of these is an example of a continuous variable? (A) Number of insects per plot (B) Eye colour (C) Number of plants per field (D) Plant height

Correct answer: (D) Plant height. Plant height is a continuous variable, since it can take any value within an interval on the number line.

What distinguishes an experiment from a sample survey? (A) Experiments involve no treatments (B) Experiments control the allocation of treatments; surveys control the selection of units (C) Sample surveys always use the entire population (D) There is no meaningful difference

Correct answer: (B) Experiments control the allocation of treatments; surveys control the selection of units. The distinguishing feature is control over allocation of treatments (experiments) versus control over selection of units (sample surveys).

What is the value of 4!? (A) 4 (B) 16 (C) 24 (D) 12

Correct answer: (C) 24. 4! = 4 x 3 x 2 x 1 = 24.

What does the value of 0! equal? (A) 0 (B) 1 (C) Undefined (D) -1

Correct answer: (B) 1. By definition, 0! = 1.

Descriptive statistics is primarily concerned with: (A) Predicting future population parameters (B) Testing hypotheses about unknown populations (C) Organizing, summarizing, and presenting sample data (D) Calculating probabilities of rare events

Correct answer: (C) Organizing, summarizing, and presenting sample data. Descriptive statistics organizes, summarizes, and presents the data collected from a sample.

Quick Revision Summary

- Statistics = mathematical science of making decisions and drawing conclusions from data under uncertainty
- Population = whole group of interest; Sample = representative subset actually measured
- Parameter = fixed value from population (usually unknown); Statistic = value from sample (varies by sample)
- Proportion x 100 = Percentage
- Experiment = control over allocation of treatments; Sample survey = control over selection of units
- Sigma (sum), Product (multiply), Factorial n! = n(n-1)...(1), with 0! = 1
- Quantitative variables: Continuous (any value in an interval) or Discrete (isolated/integer values)
- Qualitative/categorical variables cannot be measured numerically, only counted by category
- Descriptive statistics organizes/summarizes sample data; Inferential statistics draws conclusions about the population

Exam Tips

- Remember: Parameter = Population (both start with P); Statistic = Sample value that varies
- Discrete data comes from counting (whole numbers); Continuous data comes from measuring (any value)
- 0! = 1 is a common trick question — don't assume it's 0
- Econometrics = Statistics + Economics; Biometry = Statistics + Biology — useful for matching-type questions
- Descriptive statistics describes what you found; Inferential statistics generalizes beyond your sample
- Key distinguishing feature: experiments control allocation of treatments, surveys control selection of units

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.

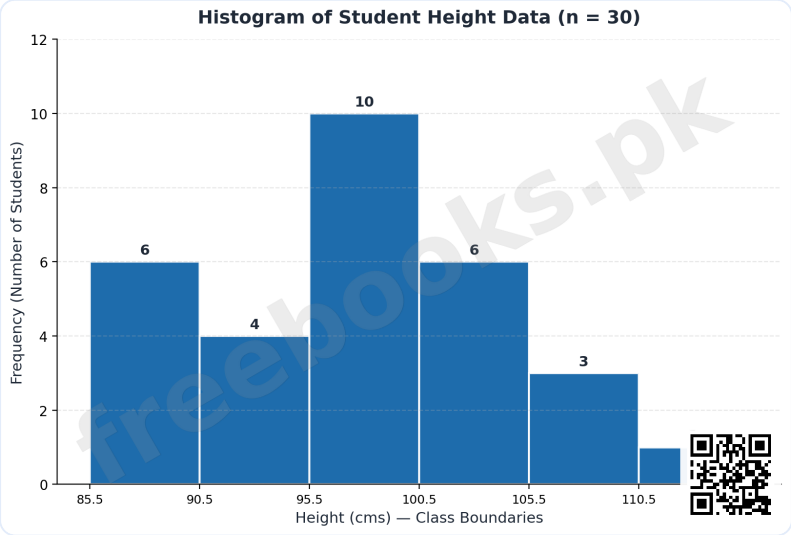
Key Facts and Relations

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

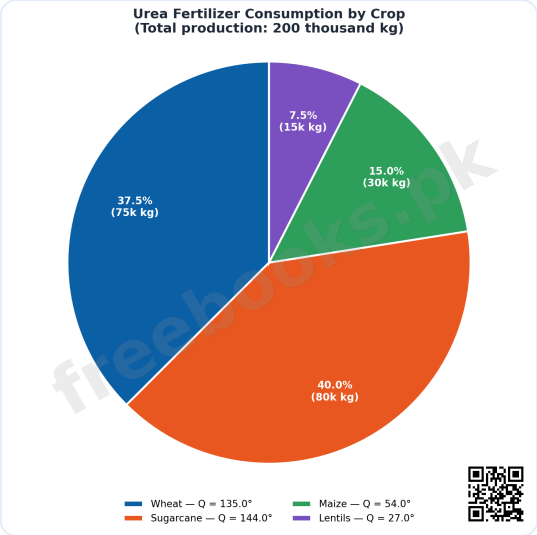
Diagrams

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



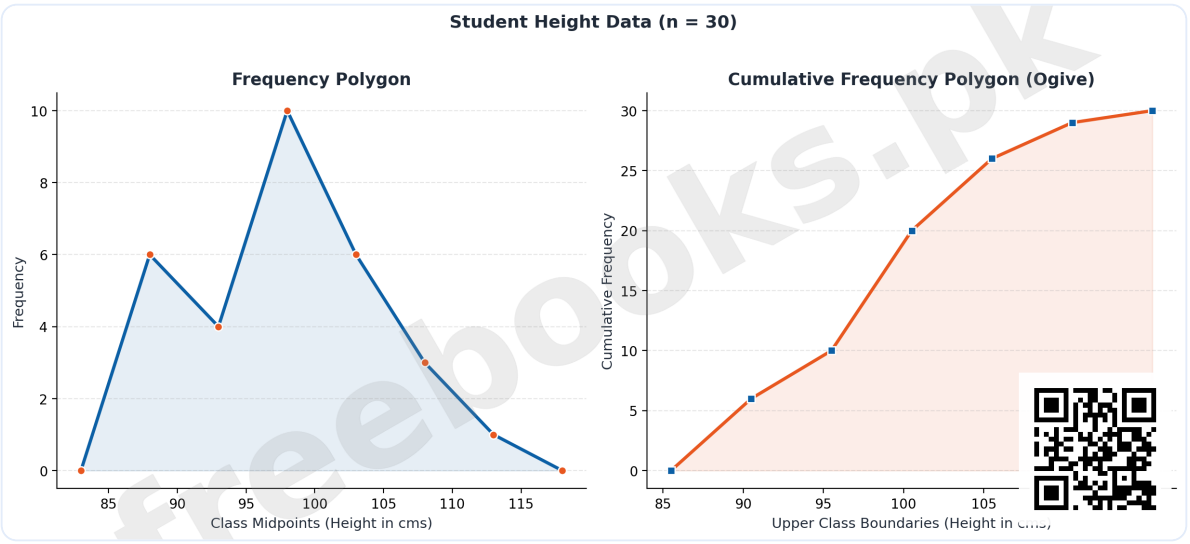
Histogram of a Frequency Distribution

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

Chapter 3: Measures of Location

A diagram or a frequency table gives an overall impression of a data set, but it does not give a single number that tells us where the 'centre' of the data lies. This chapter covers the measures used for exactly that purpose: arithmetic mean, weighted mean, geometric mean, harmonic mean, median and mode. These are collectively called averages, measures of location, or measures of central tendency, because each one gives an indication of where to locate the distribution on the number line.

Along with the measures of central tendency, this chapter also covers quantiles — quartiles, deciles, and percentiles — which describe the position of a particular observation relative to the rest of the data set, rather than just the centre. The chapter closes with the empirical relationship connecting mean, median and mode for skewed distributions, and guidance on selecting the most suitable measure for a given type of data.

Learning Objectives

- Calculate the arithmetic mean for ungrouped and grouped data, including by the short-cut (coding) method
- Explain and calculate the weighted mean, and identify when it is preferred over the ordinary mean
- Calculate the geometric mean and harmonic mean, and identify situations where each is appropriate
- State and apply the general relationship AM >= GM >= HM
- Calculate the median for ungrouped and grouped (continuous and discrete) data
- Calculate quartiles, deciles, and percentiles for ungrouped and grouped data
- Calculate the mode for ungrouped and grouped data
- State the empirical relationship between mean, median and mode for skewed distributions
- Select an appropriate measure of central tendency for a given situation

Key Concepts

3.1 Introduction

A single quantitative measure is needed to indicate the centre of a distribution. The measures commonly used for this are mean, median and mode, with geometric mean and harmonic mean used in specific situations. A good average should be well defined, easy to calculate, easy to understand, based on all the values, and capable of mathematical treatment. The important types of averages covered in this chapter are: arithmetic mean and weighted mean, geometric mean, harmonic mean, median, and mode.

3.2 Arithmetic Mean and Weighted Mean

The arithmetic mean is calculated by adding up all the observations and dividing the sum by the total number of observations. For a population of N observations it is denoted by the Greek letter mu (a parameter, a fixed but usually unknown value); for a sample of n observations it is denoted by Y-bar (a statistic whose value varies from sample to sample) and is a good, unbiased estimate of the population mean. The arithmetic mean carries the same units as the original observations.

For grouped data, the arithmetic mean is calculated as the sum of (frequency x class midpoint) divided by the sum of frequencies, since each observation in a class is assumed equal to that class's midpoint. This assumption causes a small difference between the mean of grouped data and the mean of the same data left ungrouped, called grouping error. The mean can also be calculated by the short-cut (coding) method: an arbitrary provisional mean 'a' is chosen, deviations D = Y - a are computed, and the true mean is recovered as a plus the average deviation — useful for simplifying arithmetic by hand.

Key properties of the arithmetic mean: the algebraic sum of deviations of observations from their mean is always zero; a combined mean of two or more groups can be calculated from each group's mean and size; the sum of squared deviations from the mean is smaller than the sum of squared deviations from any other value (this is why the mean 'minimizes' squared error); the mean shifts by a constant amount when that constant is added to every observation (change of

origin); and the mean scales by a constant factor when every observation is multiplied by that factor (change of scale).

Merits of the arithmetic mean: rigidly defined by a formula, easy to calculate and understand, based on all values, stable across repeated samples, and the total can be recovered if the mean and count are known. Demerits: very sensitive to departures from a bell-shaped distribution (unsuitable for skewed data), can be misleading with high variation, greatly affected by extreme values, and cannot be calculated for open-end classes without assuming their open ends.

The weighted mean is used when different observations should not be given equal importance — each value Yi is given a weight wi, and the weighted mean is the sum of (wi times Yi) divided by the sum of the weights.

3.3 Geometric Mean

The geometric mean (G.M) is defined as the nth root of the product of n positive values, and is appropriate only for positive, ratio-scale data — it is particularly useful for averaging rates, ratios, percentage changes, and index numbers. It is commonly calculated via logarithms: the log of the G.M. equals the average of the logs of the observations, and G.M. is then found by taking the antilog. For grouped data, each value's log is weighted by its class frequency before averaging.

Key properties: a combined geometric mean of several groups can be found from each group's geometric mean and size; and the geometric mean of a ratio of two matched data sets equals the ratio of their individual geometric means. Merits: rigidly defined, based on all observations, capable of further mathematical development, and less affected by extreme values than the arithmetic mean. Demerits: becomes zero if any observation is zero, and is undefined (imaginary) for negative values.

3.4 Harmonic Mean

The harmonic mean (H.M) is defined as the reciprocal of the arithmetic mean of the reciprocals of the values, and is particularly useful for averaging rates and ratios that are expressed 'per unit' of something, such as speed (distance per time) — most notably, average speed over equal distances travelled at different

speeds. For grouped data it is calculated as the sum of frequencies divided by the sum of (frequency divided by the value).

Merits: defined by a formula, based on all observations, capable of further mathematical development. Demerits: cannot be calculated if any observation is zero, more complex to calculate than the arithmetic mean, and gives more relative weight to small values than to large ones. For any set of positive observations, the three means always satisfy AM >= GM >= HM, with equality only when all the observations are identical.

3.5 Median

The median is the value that divides an arranged (ordered) data set into two equal halves. For an odd number of observations n, the median is the value of the ((n+1)/2)th item in the ordered data; for an even number of observations, it is the mean of the (n/2)th and ((n/2)+1)th items. For grouped continuous data, the median is calculated using the formula Median = l + (h/f)(n/2 - c), where l is the lower boundary of the median class, h is its width, f is its frequency, n is the total frequency, and c is the cumulative frequency of the class preceding the median class; the median class itself is located by finding which class contains the (n/2)th observation using the cumulative frequency column. For discrete grouped data, the median is found directly from the formal definition using a cumulative frequency column.

Key properties: adding a constant to every observation shifts the median by that same constant, and multiplying every observation by a constant scales the median by that factor; the sum of absolute deviations from the median is smaller than from any other value; and for a symmetrical distribution, the median is equidistant from the first and third quartiles. Merits: quick to find, not much affected by extreme values, and suitable for skewed distributions. Demerits: not rigidly defined, not readily suitable for further algebraic development, less stable across repeated samples than the mean, and not based on all the observations.

3.6 Quantiles: Quartiles, Deciles and Percentiles

Quantiles (also called fractiles) describe the position of an observation relative to the rest of a data set. Quartiles divide the ordered data into four equal parts: Q1 (lower quartile) exceeds one quarter of the observations, Q2 is the median, and

Q3 (upper quartile) exceeds three quarters of the observations. For grouped data, Qk = l + (h/f)(k.n/4 - c) for k = 1, 2, 3, using the same lower-boundary/ width/frequency/cumulative-frequency logic as the median formula, just targeting a different fraction of n.

Deciles divide the ordered data into ten equal parts (D1 through D9), with D5 equal to the median; for grouped data, Dm = l + (h/f)(m.n/10 - c). Percentiles describe finer relative standing, dividing the data into 100 parts (P1 through P99); for grouped data, Pm = l + (h/f)(m.n/100 - c). By definition, P50 is the median, P25 is Q1, and P75 is Q3 — quartiles, deciles and percentiles are simply different granularities of the same underlying idea, and can also be read directly off a cumulative frequency (ogive) graph.

3.7 Mode

The mode is the most frequently occurring value in a data set. For ungrouped data, it is found directly by inspecting which value occurs most often. Data with exactly one mode is called unimodal; data can also be bimodal (two modes), multimodal (more than two modes), or have no mode at all if every value occurs equally often. For grouped continuous data, the mode is calculated using Mode = l + (fm - f1)/((fm - f1)+(fm - f2)) x h, where l is the lower boundary of the modal class (the class with the highest frequency), fm is the modal class's frequency, f1 and f2 are the frequencies of the classes immediately before and after the modal class, and h is the modal class width.

Merits of the mode: very quick to find, and not affected by extreme values. Demerits: not rigidly defined, not easily suited to further mathematical development, uses only a few members of the data set (so it can be misleading in small samples), unstable across repeated samples like the median, may have more than one value, and may not exist at all for some data sets.

3.7.4 Empirical Relationship Between Mean, Median and Mode

This relationship depends on the shape of the distribution. A distribution is symmetrical if the frequency curve's left side, mirrored about the mean, matches its right side; otherwise it is called skewed (to the right or to the left). For a single-peaked symmetrical distribution, Mean = Median = Mode. For a moderately positively skewed distribution, Mean > Median > Mode. For a

moderately negatively skewed distribution, Mean < Median < Mode. For moderately skewed distributions generally, the median divides the distance between the mean and the mode in the ratio 1:2, which gives the useful empirical shortcut: Mode = 3 x Median - 2 x Mean.

3.8 Selecting a Suitable Measure of Central Tendency

Choosing the right measure depends on the type of variable, the purpose of the statistic, and the shape of the distribution. For quantitative variables, the arithmetic mean is usually appropriate. For categorical variables, median or mode is preferred depending on the type of category — for example, mode suits eye colour, while median suits ordered income groups. If a distribution is symmetrical, mean, median and mode are equal and equally good; for skewed distributions, the median is preferred since it is not affected by extreme values, and medians are also preferred over means when the sample is only a small part of the population. Geometric and harmonic means are reserved specifically for averaging rates and ratios.

Important Definitions

What is a measure of central tendency?: A single value that represents an entire data set by indicating where the centre of its distribution lies; common examples are the mean, median, and mode.

What is the arithmetic mean?: The sum of all observations divided by the total number of observations; denoted mu for a population and Y-bar for a sample.

What is grouping error?: The small difference between the mean calculated from grouped data (using class midpoints) and the mean calculated from the same data left ungrouped.

When is the geometric mean preferred over the arithmetic mean?: When averaging rates, ratios, percentage changes, or index numbers involving positive, ratio-scale data.

When is the harmonic mean preferred?: When averaging rates or ratios expressed per unit of something, such as speeds covered over equal distances.

What is the median?: The value that divides an ordered data set into two equal halves — the middle value for an odd number of observations, or the mean of the two middle values for an even number.

What are quartiles?: Values that divide an ordered data set into four equal parts: Q1 (lower quartile), Q2 (median), and Q3 (upper quartile).

What is the mode?: The most frequently occurring value in a data set; data can be unimodal, bimodal, multimodal, or have no mode at all.

What is the empirical relationship between mean, median and mode?: For moderately skewed distributions, Mode = 3 x Median - 2 x Mean; for symmetrical distributions, Mean = Median = Mode.

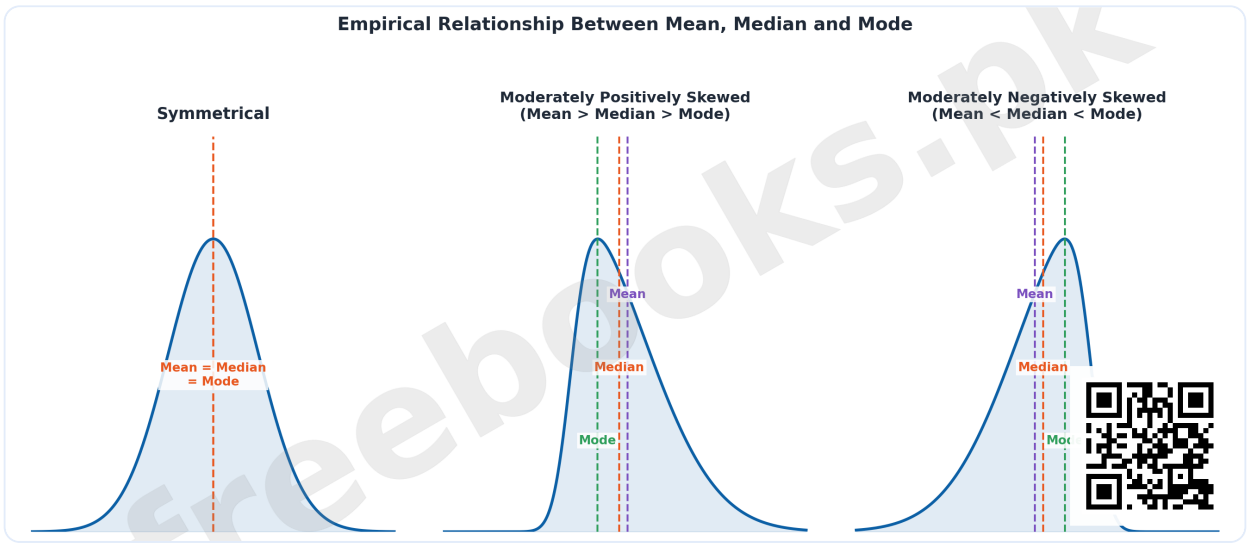
Key Facts and Relations

Topic	Key Fact / Relation
Population mean	mu = (sum of Y) / N
Sample mean (grouped data)	Y-bar = (sum of fi * yi) / (sum of fi)
Short-cut (coding) method	Y-bar = a + (sum of fi*Di) / (sum of fi), where Di = Yi - a
Weighted mean	Y-bar(w) = (sum of wi*Yi) / (sum of wi)
Geometric mean (ungrouped)	G.M = (Y1 x Y2 x ... x Yn)^(1/n) = antilog[(sum of log Yi)/n]
Harmonic mean (ungrouped)	H.M = n / (sum of 1/Yi)
General relationship	A.M >= G.M >= H.M (equal only if all observations are identical)
Median (grouped data)	Median = l + (h/f)(n/2 - c)
Quartiles (grouped data)	Qk = l + (h/f)(k.n/4 - c), for k = 1, 2, 3
Deciles (grouped data)	Dm = l + (h/f)(m.n/10 - c)

Topic	Key Fact / Relation
Percentiles (grouped data)	Pm = l + (h/f)(m.n/100 - c)
Mode (grouped data)	Mode = l + [(fm - f1) / ((fm - f1)+(fm - f2))] x h
Empirical relation (skewed data)	Mode = 3 x Median - 2 x Mean

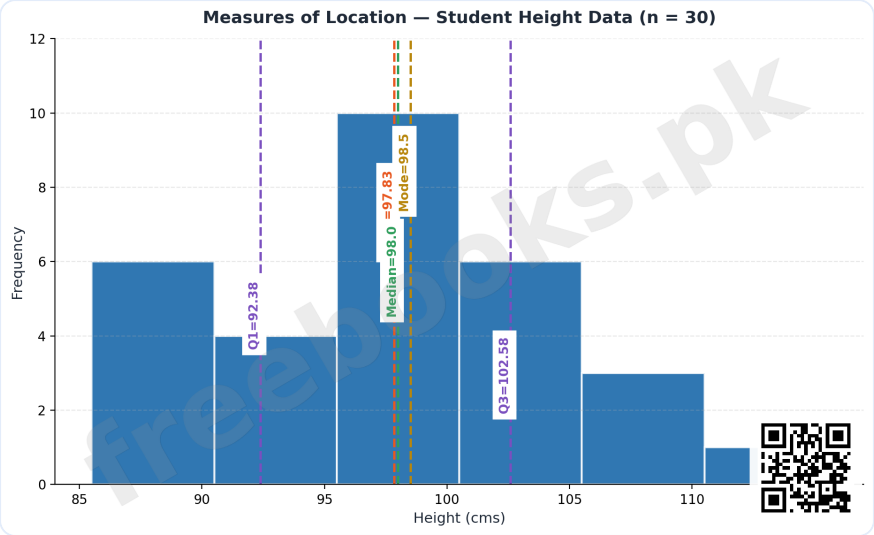
Diagrams

Symmetrical vs. Skewed Distributions: Three frequency curves comparing the relative positions of mean, median and mode: symmetrical (all three equal), moderately positively skewed (Mean > Median > Mode), and moderately negatively skewed (Mean < Median < Mode)



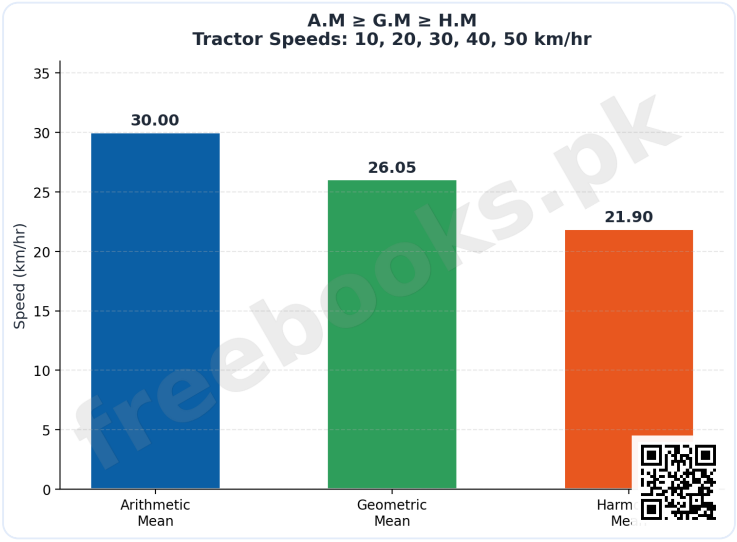
Symmetrical vs. Skewed Distributions

Measures of Location on the Student Height Data: The student-height histogram (from Chapter 2) with the calculated mean, median, and mode marked, alongside Q1 and Q3, showing how these measures locate the centre and spread of the same data set



Measures of Location on the Student Height Data

A.M >= G.M >= H.M for the Tractor Speed Example: A bar comparison of the arithmetic mean, geometric mean, and harmonic mean calculated from the same worked example (tractor speeds of 10, 20, 30, 40, 50 km/hr over equal distances), illustrating the general relationship A.M >= G.M >= H.M



A.M >= G.M >= H.M for the Tractor Speed Example

Short Questions & Answers

What is the difference between a population mean and a sample mean?

The population mean (mu) is a fixed but usually unknown parameter for the entire population; the sample mean (Y-bar) is a statistic calculated from a sample and varies from sample to sample.

Why is the arithmetic mean not suitable for skewed distributions?

Because it is very sensitive to extreme values, which pull it away from the true centre of a skewed distribution, making it a misleading measure of location in such cases.

When would you use the weighted mean instead of the ordinary mean?

When different observations should not be treated with equal importance, so each is given a weight reflecting its relative significance before averaging.

Why does the geometric mean become undefined for negative values?

Because taking roots of negative products can produce imaginary numbers, so the geometric mean is defined only for positive, ratio-scale data.

What real-world quantity is the harmonic mean most useful for averaging?

Speeds or rates covered over equal distances or equal amounts, such as average speed on a multi-stage journey.

How is the median class located for grouped data?

By finding which class contains the (n/2)th observation, using the cumulative frequency column of the grouped frequency table.

What is the relationship between mode and median-mean spacing in a moderately skewed distribution?

The median divides the distance between the mean and the mode in the ratio 1:2, giving Mode = 3 x Median - 2 x Mean.

Why is the median often preferred over the mean for income data?

Because income distributions are typically skewed by a small number of very high incomes, and the median is not affected by such extreme values the way the mean is.

Long Questions & Answers

Explain the arithmetic mean, geometric mean, and harmonic mean, including when each is the appropriate choice, and state the relationship between them.

The arithmetic mean, geometric mean, and harmonic mean are the three classical 'Pythagorean' averages, and each is suited to a different kind of data. The arithmetic mean is calculated by adding up all the observations and dividing by their count; for a population of N observations it is denoted by the Greek letter mu, and for a sample of n observations it is denoted Y-bar, which serves as a good, unbiased estimate of the unknown population mean. For grouped data, it is calculated as the sum of each class's frequency multiplied by its midpoint, divided by the total frequency, under the assumption that every observation within a class equals that class's midpoint — an assumption that introduces a small amount of grouping error compared to calculating the mean from the original ungrouped data. The arithmetic mean is the natural default choice for quantitative data that is roughly symmetrically distributed, since it uses every observation and is easy to calculate and interpret, but it is also the measure most distorted by extreme values or a strongly skewed shape, since a few very large or very small values can pull it noticeably away from where most of the data actually lies. The geometric mean, by contrast, is defined as the nth root of the product of n positive values, and is calculated in practice by averaging the logarithms of the observations and then taking the antilog of that average. Because it works multiplicatively rather than additively, the geometric mean is the appropriate choice specifically for averaging rates of change, ratios, percentage growth figures, and index numbers — quantities where a doubling and a halving should, in some sense, cancel out, which the arithmetic mean does not achieve but the geometric mean does. It is only defined for positive values, becoming zero if any single observation is zero and undefined (imaginary) if any observation is negative, which limits its applicability to strictly positive, ratio-scale data. The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of the observations, and it is the correct choice specifically when averaging rates or ratios that are expressed 'per unit' of some other varying quantity — the classic example being average speed, where a vehicle covers equal distances at different speeds; naively averaging the speeds

with the arithmetic mean would give a misleadingly high answer, because it fails to account for the fact that more time is spent travelling at the slower speeds. Like the geometric mean, the harmonic mean is undefined if any observation is zero, since it would require dividing by zero. Crucially, for any set of positive observations that are not all identical, these three means always satisfy the ordering A.M is greater than or equal to G.M, which is in turn greater than or equal to H.M, with the three becoming exactly equal only in the special case where every observation in the data set is the same value; this ordering is a general mathematical fact that holds regardless of what the underlying data represents, and understanding it helps explain why choosing the wrong type of mean for a given kind of data — for instance, using the arithmetic mean where the harmonic mean is called for — will systematically produce an inflated answer.

Describe how the median, quartiles, deciles and percentiles are calculated for grouped data, and explain what each of these measures tells us about a data set.

The median, quartiles, deciles, and percentiles are all members of the same family of measures, generally called quantiles or fractiles, and they all share a common underlying idea: rather than summarizing an entire data set with a single 'typical' value the way the mean does, they identify positions within the ordered data set that split it into equal-sized groups. The median is the simplest and most familiar of these: it is the value that splits an ordered data set exactly in half, so that half of the observations lie below it and half lie above it. For grouped, continuous data arranged in ascending order, the median is calculated with the formula Median equals the lower boundary of the median class, plus the class width divided by the class frequency, multiplied by the difference between half the total frequency and the cumulative frequency of the class immediately before the median class. Before this formula can be applied, the median class itself must first be located, which is done by scanning down the cumulative frequency column of the grouped frequency table to find which class contains the observation at position n divided by 2. Quartiles extend this same idea to dividing the data into four equal parts rather than two: the first quartile Q1 is the value below which one quarter of the observations fall, the second quartile Q2 is simply the median itself, and the third quartile Q3 is the value below which three quarters of the observations fall. For grouped data, each quartile is calculated

with essentially the same formula used for the median, except that the fraction of the total frequency being targeted changes to reflect which quartile is being calculated — one quarter of n for Q1, two quarters (equivalent to one half) of n for Q2, and three quarters of n for Q3 — with the appropriate class boundary, width, frequency, and preceding cumulative frequency substituted in each time. Deciles carry this subdivision further, splitting the ordered data into ten equal parts denoted D1 through D9, with the fifth decile D5 always being identical to the median; the formula again follows the identical pattern, this time targeting m tenths of the total frequency for the mth decile. Percentiles go furthest of all, dividing the data into one hundred equal parts denoted P1 through P99, and by definition several of these named quantiles coincide exactly: the 50th percentile P50 is the median, the 25th percentile P25 is the first quartile Q1, and the 75th percentile P75 is the third quartile Q3. What all of these measures give us, that a simple average cannot, is a sense of relative standing — where a particular value or a particular class sits relative to the rest of the data set, rather than just where the overall centre of the data happens to be. This makes quantiles especially valuable for describing spread and position simultaneously; for example, knowing that a student's height falls at the 90th percentile immediately conveys that the student is taller than roughly ninety percent of their peers, which is a far more specific and useful statement than simply knowing the average height of the whole group. As a practical convenience, all of these quantiles can also be read directly off a cumulative frequency graph (an ogive), without needing to apply the formula at all, by locating the desired cumulative frequency or percentage on the vertical axis and reading across to find the corresponding value on the horizontal axis.

Multiple Choice Questions (MCQs)

The population mean is denoted by: (A) Y-bar (B) mu (C) M (D) sigma
Correct answer: (B) mu. The population mean is denoted by the Greek letter mu; Y-bar denotes the sample mean.

Which mean is most appropriate for averaging percentage growth rates? (A) Arithmetic mean (B) Geometric mean (C) Harmonic mean (D) Median

Correct answer: (B) Geometric mean. The geometric mean is appropriate for averaging rates, ratios, and percentage changes on positive, ratio-scale data.

The harmonic mean is best suited for averaging: (A) Categorical data (B) Rates such as speed over equal distances (C) Skewed income data (D) Modal categories

Correct answer: (B) Rates such as speed over equal distances. The harmonic mean is the reciprocal of the mean of reciprocals, and is ideal for averaging rates like speed over equal distances.

For any positive data set that is not all identical, which relationship always holds? (A) H.M >= G.M >= A.M (B) A.M >= G.M >= H.M (C) G.M >= A.M >= H.M (D) A.M = G.M = H.M

Correct answer: (B) A.M >= G.M >= H.M. The general relationship is A.M >= G.M >= H.M, with equality only when all observations are identical.

For an even number of ordered observations, the median is: (A) The single middle value (B) The mean of the two middle values (C) The most frequent value (D) The largest value

Correct answer: (B) The mean of the two middle values. For an even n, the median is the mean of the (n/2)th and ((n/2)+1)th ordered observations.

The second quartile Q2 is the same as: (A) The mode (B) The median (C) The 90th percentile (D) The first decile

Correct answer: (B) The median. Q2 (second quartile) is identical to the median, and also equals the 50th percentile (P50).

In the grouped-data mode formula, fm refers to: (A) The frequency of the class preceding the modal class (B) The frequency of the modal class (C) The frequency of the class following the modal class (D) The total frequency

Correct answer: (B) The frequency of the modal class. fm is the frequency of the modal class itself (the class with the highest frequency).

For a moderately positively skewed distribution, which relationship holds? (A) Mean = Median = Mode (B) Mean > Median > Mode (C) Mean < Median < Mode (D) Mode > Mean > Median

Correct answer: (B) Mean > Median > Mode. For moderate positive skew, Mean > Median > Mode.

The empirical relationship Mode = 3 Median - 2 Mean applies to: (A) Perfectly symmetrical distributions only (B) Moderately skewed distributions (C) Bimodal distributions only (D) Only ungrouped data
Correct answer: (B) Moderately skewed distributions. This empirical shortcut applies to moderately skewed distributions, where the median divides the mean-mode distance in a 1:2 ratio.

Which measure of central tendency is least affected by extreme values? (A) Arithmetic mean (B) Geometric mean (C) Median (D) Weighted mean
Correct answer: (C) Median. The median is not much affected by exceptionally large or small values, making it suitable for skewed data.

Quick Revision Summary

- Arithmetic mean = sum of values / count; sensitive to extreme values and skew
- Weighted mean used when observations carry different importance (weights wi)
- Geometric mean: nth root of product of positive values; best for rates, ratios, percentage changes
- Harmonic mean: reciprocal of mean of reciprocals; best for speeds/rates over equal distances
- General relation: A.M >= G.M >= H.M (equal only if all values identical)
- Median splits ordered data in half: middle value (odd n) or mean of two middle values (even n)
- Grouped-data formulas (median, quartiles, deciles, percentiles) all use: l + (h/f)(target - c)
- Quartiles (4 parts), Deciles (10 parts), Percentiles (100 parts) — P50=Median=Q2, P25=Q1, P75=Q3
- Mode = most frequent value; grouped mode formula uses fm, f1 (preceding), f2 (following)
- Empirical relation for skewed data: Mode = 3 Median - 2 Mean

Exam Tips

- Memorize the grouped-data pattern l + (h/f)(target - c) — it's identical for median, quartiles, deciles, and percentiles, only the target fraction of n changes
- For AM/GM/HM word problems, identify whether the question involves rates (HM), ratios/growth (GM), or plain totals (AM) before choosing a formula
- Always order the data first before finding median, quartiles, or percentiles by the direct/formal definition
- Remember P50 = Median = Q2 and P25 = Q1, P75 = Q3 — these equivalences are common exam questions
- For the empirical relation, remember the ratio is 1:2 (mean-median gap : median-mode gap), giving Mode = 3 Median - 2 Mean
- Mode can fail to exist or have multiple values — always check ungrouped data by inspection first

Chapter 4: Measures of Dispersion

A measure of central tendency alone does not tell us anything about how spread out the values in a data set are. Two data sets can share the exact same mean while differing enormously in their variability -- for example, the sets 8, 7, 5, 8, 6 and 1, 4, 7, 10, 12 both have a mean of 6.8, yet the second set is clearly far more spread out than the first. To give a fuller description of data, we need a numerical quantity called a measure of dispersion (or variability) that captures this spread.

This chapter covers the common absolute measures of dispersion -- range, quartile deviation, mean deviation, variance, and standard deviation -- which carry the same units as the original data, along with relative measures such as the coefficient of variation, which are unit-free ratios useful for comparing the variability of different data sets. It also introduces moments as a more general tool for describing a distribution's shape, and uses them to define skewness (lack of symmetry) and kurtosis (peakedness and tail length).

Learning Objectives

- Explain why a measure of central tendency alone is insufficient to describe a data set
- Calculate range and quartile deviation for ungrouped and grouped data
- Calculate mean deviation from the mean and from the median
- Calculate variance and standard deviation using the direct and short-cut (coding) formulas
- State and apply the key properties of variance and standard deviation, including for combined data sets
- Calculate the coefficient of variation and other relative measures of dispersion
- Define moments about the mean, about an arbitrary origin, and about zero, and relate them to variance
- Explain skewness and calculate Karl Pearson's and Bowley's coefficients of skewness

- Explain kurtosis and distinguish mesokurtic, platykurtic, and leptokurtic distributions

Key Concepts

4.1 Introduction

Two data sets with vastly different variability can still share the same measure of central tendency, which shows that central tendency alone is not sufficient to describe a data set. A measure of dispersion (or variability) is a numerical quantity that describes how spread out the values in a data set are. There are two broad types: absolute measures, which carry the same units as the original data (range, quartile deviation, mean deviation, variance, and standard deviation), and relative measures, which are unitless ratios (discussed under coefficient of variation).

4.1.1 Range

The range of n values is the difference between the largest and smallest observation: R = Y(n) - Y(1). For grouped data, the range is calculated as the mid-value of the highest class minus the mid-value of the lowest class. Merits: very easy to calculate, and useful for small samples. Demerits: not based on all the observations, and depends only on the two most extreme values, making it highly sensitive to outliers.

4.1.2 Quartile Deviation

Quartile deviation (Q.D), also called the semi-interquartile range, is based on the first and third quartiles: Q.D = (Q3 - Q1) / 2. It can never be negative, since Q3 is always at least as large as Q1; a small Q.D indicates low variability, and a large Q.D indicates high variability. The coefficient of quartile deviation, a relative (unit-free) measure, is calculated as (Q3 - Q1) / (Q3 + Q1). Merits: easy to calculate, and not affected by extreme observations. Demerits: not based on all the observations, and two different distributions can share identical quartiles and therefore identical Q.D despite otherwise differing.

4.1.3 Mean Deviation

Mean deviation (M.D) is the mean of the absolute deviations of observations from the mean, median, or mode: M.D = (sum of |Y - M|) / n, where absolute deviations means all deviations are treated as positive. For grouped data, M.D = (sum of fi|Yi - M|) / (sum of fi). Key properties: mean deviation from the median is always the smallest possible mean deviation (smaller than from any other value); M.D is always greater than or equal to zero; and for symmetrical distributions, M.D = (4/5) times the standard deviation. Merits: easy to calculate, and based on all the observations. Demerits: affected by extreme values, not readily suited to further mathematical development, and it discards the sign of each deviation, losing information about direction.

4.1.4 Variance and Standard Deviation

Variance is the mean of the squared deviations of all observations from their mean. The population variance is denoted sigma-squared; the sample variance, denoted S-squared, is calculated as the sum of squared deviations from the sample mean divided by n (or, for grouped data, weighted by frequency). A short-cut formula avoids computing deviations directly: S-squared equals the mean of the squared values minus the square of the mean. Variance is based on all the observations, easy to calculate and understand, but is affected by extreme values.

Standard deviation (S) is the positive square root of variance, restoring the original units of measurement that variance's squaring operation removes. It has several important properties: the variance and standard deviation of a constant are zero; they are unaffected by adding or subtracting a constant from every observation (independent of origin); multiplying every observation by a constant a multiplies the variance by a-squared and the standard deviation by |a|; for two independent variables, the variance of their sum or difference equals the sum of their individual variances; and the combined variance of several groups can be calculated from each group's own variance, mean, and size, and the overall combined mean.

4.2 Coefficient of Variation and Other Relative Measures

The coefficient of variation (C.V) is the most widely used relative measure of dispersion: C.V = (S / mean) x 100 for a sample, or (sigma / mu) x 100 for a population. Because it is a ratio of two quantities with the same units, C.V is dimensionless and unaffected by the unit of measurement used -- the same data will give the same C.V whether measured in millimeters, centimeters, or meters. A lower C.V indicates a comparatively more consistent (less variable relative to its mean) group, which makes C.V especially useful for comparing the variability of two or more data sets that may have very different means or units. Other relative measures include the coefficient of range, coefficient of quartile deviation, mean coefficient of dispersion (M.D divided by mean), median coefficient of dispersion (M.D divided by median), and coefficient of standard deviation (S divided by mean).

4.3 Moments

Measures of location and dispersion describe a data set's centre and spread, but tell us nothing about its shape. Moments are a family of measures used to describe shape. The rth moment about the mean (a central moment), denoted mr, is the mean of the rth powers of deviations from the mean. The first moment m1 is always zero; the second moment m2 is identical to the variance; the third moment m3 relates to skewness; and the fourth moment m4 relates to kurtosis (peakedness). Moments about an arbitrary origin a (raw or non-central moments), denoted mr-prime, are often easier to calculate directly and can be converted into moments about the mean using standard conversion relations -- this is especially useful for grouped data with equal class widths, where a coding variable u = (y - a)/h simplifies the arithmetic considerably. When calculating moments from grouped data, choosing class midpoints introduces a small grouping error in the second and fourth moments in particular; Sheppard's corrections adjust for this when the frequency curve tapers gradually at both ends.

4.5 Skewness

Skewness refers to a lack of symmetry in a distribution. A distribution is symmetrical when mean, median and mode coincide and the frequency curve

mirrors itself around the mean; otherwise it is skewed. A positively skewed distribution has a longer tail extending to the right; a negatively skewed distribution has a longer tail extending to the left. Extreme skewness produces J-shaped distributions. Several numerical measures quantify skewness: Karl Pearson's first coefficient, (Mean - Mode) / S; Karl Pearson's second coefficient, 3(Mean - Median) / S; and Bowley's coefficient, based on quartiles, (Q1 + Q3 - 2 x Median) / (Q3 - Q1), which lies between -1 and +1. All of these coefficients are zero for a perfectly symmetrical distribution, negative for negative skew, and positive for positive skew. The moment-based measure root-beta-1 (= m3 / (m2 raised to the power 1.5)) serves the same purpose: zero indicates symmetry, negative indicates negative skew, positive indicates positive skew.

4.5.1 Kurtosis

Kurtosis describes the peakedness of a distribution and the length of its tails. A mesokurtic distribution is the standard, normally-peaked shape. A leptokurtic distribution is more sharply peaked, with more values clustered close to the mean and also more values out in the tails. A platykurtic distribution is comparatively flatter, with more values spread between the mean and the tails than a mesokurtic distribution would have. Kurtosis is commonly assessed using the fourth moment about the mean relative to the square of the second moment (the variance).

Important Definitions

What is a measure of dispersion?: A numerical quantity that describes how spread out or variable the values in a data set are, complementing a measure of central tendency.

What is the difference between absolute and relative measures of dispersion?: Absolute measures carry the same units as the original data (e.g., range, standard deviation); relative measures are unit-free ratios (e.g., coefficient of variation), useful for comparing variability across different units or scales.

What is the range?: The difference between the largest and smallest observation in a data set: R = Y(n) - Y(1).

What is quartile deviation?: Half the difference between the third and first quartiles: Q.D = (Q3 - Q1) / 2, also called the semi-interquartile range.

What is mean deviation?: The mean of the absolute deviations of all observations from the mean, median, or mode.

What is variance?: The mean of the squared deviations of all observations from their mean; its square root is the standard deviation.

What is the coefficient of variation?: A relative, unit-free measure of dispersion equal to (standard deviation / mean) x 100, used to compare variability across data sets.

What are moments in statistics?: A family of measures -- the mean of deviations from a reference point raised to integer powers -- used to describe the shape of a distribution, including its skewness and kurtosis.

What is skewness?: A lack of symmetry in a distribution; positive skew has a longer right tail, negative skew has a longer left tail.

What is kurtosis?: A measure of a distribution's peakedness and tail length, classifying it as mesokurtic (normal), leptokurtic (more peaked), or platykurtic (flatter).

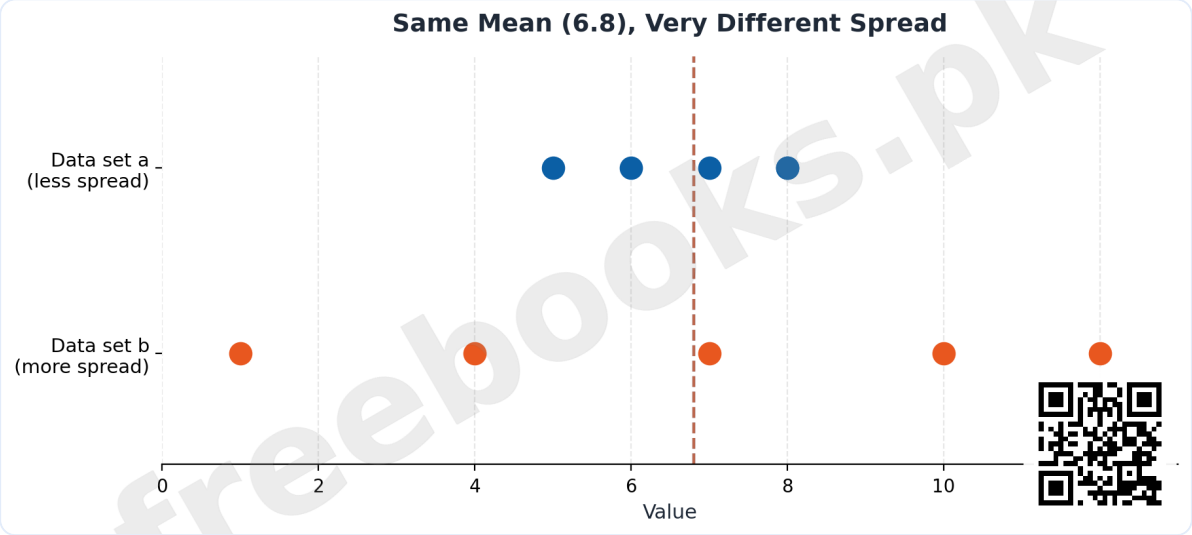
Key Facts and Relations

Topic	Key Fact / Relation
Range (ungrouped)	R = Y(n) - Y(1)
Range (grouped)	R = mid-value of highest class - mid-value of lowest class
Quartile deviation	Q.D = (Q3 - Q1) / 2
Coefficient of Q.D.	Coefficient of Q.D = (Q3 - Q1) / (Q3 + Q1)
Mean deviation	M.D = (sum of |Y - M|) / n, where M = mean, median, or mode
Sample variance (short-cut)	S^2 = (sum of Y^2)/n - (sum of Y / n)^2

Topic	Key Fact / Relation
Standard deviation	S = square root of variance
Coefficient of variation	C.V = (S / Y-bar) x 100
Combined variance (k groups)	Sc^2 = sum[ni(Si^2 + (Yi-bar - Yc-bar)^2)] / sum(ni)
Karl Pearson's 1st coefficient of skewness	(Mean - Mode) / S
Bowley's coefficient of skewness	(Q1 + Q3 - 2 x Median) / (Q3 - Q1)
Moment-based skewness	root(beta1) = m3 / (m2)^1.5

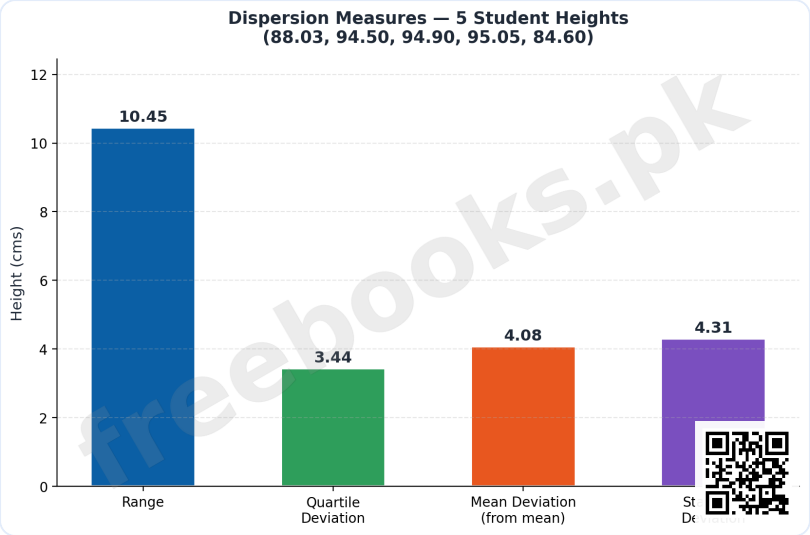
Diagrams

Same Mean, Different Spread: Two data sets, 8,7,5,8,6 and 1,4,7,10,12, both with mean 6.8 but very different variability, shown as dot plots along a shared number line -- the motivating example for why dispersion measures are needed



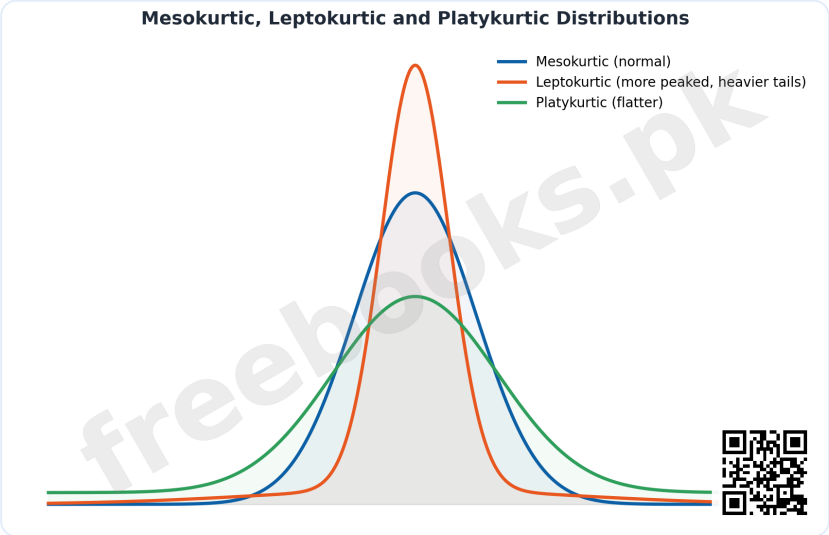
Same Mean, Different Spread

Dispersion Measures Compared: Range, Quartile Deviation, Mean Deviation (from mean), and Standard Deviation calculated for the same 5-value student height data used throughout this chapter's worked examples, shown side by side as a bar chart



Dispersion Measures Compared

Mesokurtic, Platykurtic and Leptokurtic Distributions: Three distribution curves with equal variance illustrating kurtosis: mesokurtic (normal peak), leptokurtic (sharper peak, heavier tails), and platykurtic (flatter peak, lighter tails)



Mesokurtic, Platykurtic and Leptokurtic Distributions

Short Questions & Answers

Why is a measure of central tendency alone insufficient to describe a data set?

Because two data sets can share the exact same mean while having very different amounts of spread, so a single central value hides important information about variability.

What is the key difference between an absolute and a relative measure of dispersion?

An absolute measure keeps the same units as the original data (e.g., cm), while a relative measure is a unit-free ratio, making it useful for comparing variability across data sets measured in different units or scales.

Why can quartile deviation never be negative?

Because the third quartile Q3 is always at least as large as the first quartile Q1, so their difference (and hence Q.D) is always zero or positive.

Why is mean deviation from the median described as the 'smallest possible' mean deviation?

Because the sum of absolute deviations from the median is mathematically smaller than the sum of absolute deviations from any other single value, including the mean.

Why is standard deviation preferred over variance for reporting results?

Because standard deviation is expressed in the original units of the data, while variance is expressed in squared units, which are harder to interpret directly.

Why is the coefficient of variation useful when comparing two data sets with very different means?

Because it expresses variability as a percentage of the mean, allowing fair comparison between groups whose means (and possibly units) differ substantially.

What does it mean if root-beta-1 is negative?

It indicates the distribution is negatively skewed, meaning it has a longer tail extending toward smaller values.

How does a leptokurtic distribution differ visually from a platykurtic one?

A leptokurtic distribution is more sharply peaked with heavier tails, while a platykurtic distribution is flatter, with more values spread between the centre and the tails.

Long Questions & Answers

Explain range, quartile deviation, mean deviation, and standard deviation as measures of dispersion, comparing their merits and demerits.

Range, quartile deviation, mean deviation, and standard deviation form a natural progression of measures of dispersion, each addressing a limitation of the one before it, and understanding this progression helps explain why standard deviation is the most widely used of the four despite being the most complex to calculate by hand. The range is the simplest possible measure: it is nothing more than the difference between the largest and smallest observation in the data set. Its great virtue is how easy it is to calculate, requiring only that the two extreme values be identified, which makes it a genuinely useful quick check, particularly for small samples where a fuller calculation may not be worth the effort. Its weakness, however, is severe: because it depends on only two values out of the entire data set, it completely ignores everything that happens in between, and it is extremely sensitive to a single unusually large or small observation, which can inflate the range dramatically without reflecting the typical spread of the bulk of the data at all. Quartile deviation addresses this weakness by looking at the middle half of the data instead of the two most extreme points: it is defined as half the difference between the third quartile and the first quartile, effectively describing how spread out the central fifty percent of the data is. Because it deliberately excludes the most extreme twenty-five percent of values on either end, quartile deviation is far less sensitive to outliers than the range, and it is still reasonably easy to calculate once the quartiles themselves are known. Its remaining weakness is that, like the range, it is not based on all of the observations -- it only uses two specific points in the ordered data, the first and third quartiles, and so two data sets that happen to share the same quartiles will report identical quartile deviations even if they differ substantially in how the remaining data points are arranged. Mean deviation moves further still, becoming the first of these four measures to actually incorporate every single observation in the data set: it is calculated as the average of the absolute deviations of every observation from a chosen central value, typically the mean or median, with 'absolute' meaning that every deviation is treated as a positive distance regardless of whether the original observation fell above or below that

central value. This makes mean deviation considerably more representative of the data set as a whole than either the range or quartile deviation, since no observation is excluded from the calculation. Its distinctive weakness is a mathematical one: by forcing every deviation to be positive through the absolute value operation, mean deviation becomes very difficult to use in further algebraic or mathematical development, since absolute value functions do not have the same clean differentiability and combination properties that squared quantities do. Standard deviation solves this final problem by squaring the deviations instead of taking their absolute value -- squaring, like the absolute value, guarantees a positive result, but unlike absolute value, squaring is a smooth, well-behaved mathematical operation that supports extensive further development, which is precisely why standard deviation (and the closely related variance that precedes it) underlies the vast majority of more advanced statistical techniques taught afterward. Standard deviation is based on all the observations, like mean deviation, but is additionally far more mathematically tractable, at the cost of being calculated by first squaring deviations (producing variance, in squared units) and then taking a square root at the end to return to the original units -- a two-step process that is somewhat more involved than the direct averaging used for mean deviation, but which is precisely what makes standard deviation so much more useful in later statistical work.

Explain the coefficient of variation, why it is needed in addition to standard deviation, and describe how skewness and kurtosis extend the description of a distribution beyond location and spread.

Standard deviation is an excellent measure of absolute spread, but it has one significant practical limitation: because it is expressed in the same units as the original data, it cannot be used directly to compare the variability of two data sets that are measured in different units, or even two data sets measured in the same units but with very different typical magnitudes. For instance, a standard deviation of five kilograms might represent enormous relative variability if the data set consists of newborn baby weights averaging around three kilograms, but the very same standard deviation of five kilograms would represent barely any relative variability at all if the data set consists of adult body weights averaging around seventy kilograms -- the raw number five kilograms simply does not carry enough context on its own to judge whether that amount of spread is large or

small relative to the data. The coefficient of variation solves exactly this problem by expressing the standard deviation as a percentage of the mean: C.V equals the standard deviation divided by the mean, multiplied by one hundred. Because this calculation divides one quantity by another quantity measured in the same units, the units themselves cancel out entirely, leaving a single dimensionless percentage that can be compared meaningfully across data sets regardless of what units or scales they were originally measured in. A lower coefficient of variation indicates that a data set is comparatively more consistent, meaning its spread is small relative to its own average size, while a higher coefficient of variation indicates comparatively more variability relative to the average. This single property -- unit independence -- is what makes the coefficient of variation indispensable whenever a genuine comparison of variability between different groups, different measurement scales, or even different variables entirely is required. Beyond location and spread, however, a data set's distribution has further shape characteristics that neither the mean nor the standard deviation captures at all, and this is where skewness and kurtosis come in. Skewness measures the lack of symmetry in a distribution: a perfectly symmetrical distribution has its mean, median, and mode all coinciding at the same central point, with the frequency curve on one side of that centre being a mirror image of the curve on the other side, whereas a skewed distribution has one tail stretched out noticeably longer than the other, either toward larger values (positive skew) or toward smaller values (negative skew). Numerical measures of skewness, such as Karl Pearson's coefficients or Bowley's quartile-based coefficient, condense this shape information into a single number that is zero for symmetry, negative for a longer left tail, and positive for a longer right tail, allowing the asymmetry of a distribution to be quantified and compared numerically rather than judged only by eye from a graph. Kurtosis, meanwhile, describes a different aspect of shape entirely: rather than asymmetry, it describes how sharply peaked a distribution is around its centre and how heavy or light its tails are, relative to the standard, moderately-peaked mesokurtic (normal) shape. A leptokurtic distribution is more sharply peaked than normal, with unusually many observations clustered tightly around the mean but also, somewhat counterintuitively, an unusually large number of observations out in the extreme tails far from the mean; a platykurtic distribution is comparatively flatter than normal, with observations more evenly spread between the centre

and the tails rather than being concentrated in either region. Together, skewness and kurtosis extend the description of a data set well beyond what location and spread alone can offer, capturing the overall shape of the distribution's curve and enabling far richer comparisons between different data sets or between a data set and a theoretical reference distribution.

Multiple Choice Questions (MCQs)

Which of the following is a relative (unit-free) measure of dispersion? (A) Range (B) Standard deviation (C) Coefficient of variation (D) Variance

Correct answer: (C) Coefficient of variation. The coefficient of variation is a ratio (S / mean x 100) and is therefore unit-free; the others carry the original units.

Quartile deviation is calculated as: (A) Q3 - Q1 (B) (Q3 - Q1) / 2 (C) (Q3 + Q1) / 2 (D) Q3 x Q1

Correct answer: (B) (Q3 - Q1) / 2. Quartile deviation (semi-interquartile range) is Q.D = (Q3 - Q1) / 2.

Mean deviation is smallest when calculated from: (A) The mean (B) The mode (C) The median (D) The range

Correct answer: (C) The median. Mean deviation from the median is always less than or equal to mean deviation from any other value.

Standard deviation is: (A) The mean of squared deviations (B) The positive square root of variance (C) The same as mean deviation (D) Always negative

Correct answer: (B) The positive square root of variance. Standard deviation is defined as the positive square root of the variance.

If every observation in a data set is multiplied by a constant a, the variance is multiplied by: (A) a (B) a^2 (C) 2a (D) 1/a

Correct answer: (B) a^2. Variance scales by the square of the constant: var(aY) = a^2 * var(Y).

The coefficient of variation is useful because it: (A) Has the same units as the data (B) Is always equal to zero (C) Is unit-free and allows comparison across data sets (D) Cannot be expressed as a percentage

Correct answer: (C) Is unit-free and allows comparison across data sets. C.V is a dimensionless ratio expressed as a percentage, making it ideal for comparing variability across different units or scales.

The second moment about the mean (m2) is identical to: (A) The range (B) The variance (C) The mode (D) The coefficient of variation
Correct answer: (B) The variance. The second central moment m2 equals the variance S^2.

A positively skewed distribution has: (A) A longer tail to the left (B) A longer tail to the right (C) No tail at all (D) Mean equal to mode
Correct answer: (B) A longer tail to the right. A positively skewed distribution has a longer tail extending to the right-hand side.

Bowley's coefficient of skewness is based on: (A) Mean and mode only (B) Range only (C) Quartiles and the median (D) Standard deviation only
Correct answer: (C) Quartiles and the median. Bowley's coefficient uses (Q1 + Q3 - 2*Median) / (Q3 - Q1).

A leptokurtic distribution compared to a mesokurtic (normal) one is: (A) Flatter with lighter tails (B) More sharply peaked with heavier tails (C) Identical in shape (D) Always negatively skewed
Correct answer: (B) More sharply peaked with heavier tails. Leptokurtic distributions are more peaked with heavier tails than the normal (mesokurtic) shape.

Quick Revision Summary

- Absolute measures (same units as data): Range, Q.D, M.D, Variance, S.D. Relative measures (unit-free): C.V and coefficients of dispersion
- Range R = Y(n) - Y(1); simple but ignores all but the two extreme values
- Quartile deviation Q.D = (Q3 - Q1)/2; not affected by extremes but ignores most of the data
- Mean deviation M.D = mean of |Y - M|; based on all values but hard to develop mathematically
- Variance = mean of squared deviations from mean; S.D = positive square root of variance

- Variance is unaffected by adding a constant (change of origin); scales by a^2 under multiplication (change of scale)
- Coefficient of variation C.V = (S / mean) x 100; unit-free, ideal for comparing different data sets
- Moments: m1=0 always, m2=variance, m3 relates to skewness, m4 relates to kurtosis
- Skewness: zero=symmetric, negative=left tail longer, positive=right tail longer
- Kurtosis: mesokurtic=normal peak, leptokurtic=more peaked/heavier tails, platykurtic=flatter

Exam Tips

- When a question asks to 'compare variability' between two data sets with different units or means, that's a signal to use coefficient of variation, not raw standard deviation
- Remember the short-cut variance formula S^2 = (sum Y^2)/n - (mean)^2 — much faster than computing every deviation by hand
- Quartile deviation and mean deviation both ignore the sign of spread (never negative) — only variance/S.D. use squaring to achieve this
- For grouped data with equal class width, use the coding variable u=(y-a)/h to simplify moment calculations before converting back
- Remember: variance is unaffected by shifting all data by a constant, but scales by the square of any multiplier
- Bowley's coefficient of skewness only needs Q1, Q2 (median), and Q3 — no need to calculate the mean at all

Chapter 5: Index Numbers

The buying power of a rupee changes over time -- something that cost 10 rupees in 1960 could easily cost 60 rupees decades later. To make meaningful comparisons of prices, wages, quantities, or the general cost of living across different time periods, statisticians use index numbers: ratios or averages of ratios, usually expressed as percentages, that measure the relative change in a variable over time.

This chapter covers how to construct simple and composite (aggregate) index numbers, the difference between unweighted and weighted index numbers, the major named index formulas (Laspeyre's, Paasche's, and Fisher's), and how real-world price indices like the Consumer Price Index (CPI) and Wholesale Price Index (WPI) are built and interpreted, including how they're used to measure inflation and the purchasing power of money.

Learning Objectives

- Define an index number and calculate a simple index number for a single variable
- Distinguish simple index numbers from composite (aggregate) index numbers
- Calculate unweighted index numbers: simple aggregate index and average relative index
- Explain link relatives and calculate chain indices
- Calculate weighted index numbers: Laspeyre's, Paasche's, and Fisher's ideal index
- Explain how the Consumer Price Index (CPI) and Wholesale Price Index (WPI) are constructed and used
- Calculate the rate of inflation from CPI values across years
- State the main limitations and uses of index numbers

Key Concepts

5.1 Introduction

An index number measures the relative change in a variable over time, usually expressed as a percentage. To construct one, two or more time periods are considered, with one period chosen as the base. The index number for the current year n is calculated as In = (price in current year / price in base year) x 100. By convention, the index number for the base year itself is always 100. This chapter uses standard notation: p0i and q0i for the price and quantity of the ith commodity in the base year, and pni and qni for the current year.

5.1.1 Types of Index Numbers

A simple index number measures the relative change in a single variable with respect to a base year -- for example, an index of wheat prices alone or wages alone, calculated as Pon = (price in year n / price in base year) x 100. A composite (aggregate) index number measures the relative change in two or more variables together, such as a whole basket of commodity prices, and can be calculated as either unweighted or weighted index numbers, each of which may in turn be a price index, quantity index, or value index. The value index number is Von = (sum of pn*qn) / (sum of p0*q0) x 100.

5.1.2 and 5.1.3 Limitations and Uses of Index Numbers

Index numbers have real limitations: it is not possible to account for every change in a product, no single index suits every purpose, the choice of base period can introduce error, they are only rough indications of relative change, and different construction methods can give different results for the same data. Despite this, index numbers are widely used: price indices measure average price change and the buying power of money over time; the CPI helps cancel out the effect of inflation or deflation; the WPI helps adjust contract prices and payments for industrial organizations; quantity indices track changes in production, consumption, and trade; and import/export price indices measure a country's terms of trade.

5.2 Construction of Price Index Numbers

Building a price index involves several deliberate steps. First, the purpose and scope must be clearly defined -- why, where, and what changes are being measured. Second, the components (commodities and their prices) must be selected carefully, with at least twenty items recommended for practical purposes, precisely defined and consistently available over time. Third, the base year must be chosen with care: it should reflect normal prices and not be too far removed from the current year, and may even be an average of several years rather than a single year. Fourth, weights must be chosen to reflect the relative importance of each commodity -- for example, wheat should carry more weight than tea -- commonly formulated as Woi = Voi / (sum of Voi), where Voi = poi x qoi is the value of the commodity in the base year. Fifth, an averaging method must be chosen; arithmetic mean is used in practice for convenience, even though geometric mean is theoretically more appropriate for averaging ratios.

5.3 Unweighted Index Numbers

Unweighted index numbers give equal importance to every item. The simple aggregate index is the ratio of the sum of current-year prices to the sum of base-year prices, expressed as a percentage: Pon = (sum of Pn) / (sum of P0) x 100. Its main drawback is that it is sensitive to the measuring units used and does not reflect that commodities differ in importance. The average relative index instead averages the individual price relatives (Pni/P0i) for each commodity: Pon = (1/k) x sum of (Pni/P0i). This avoids the units problem but still fails to weight commodities by their real-world importance.

When the relative importance of commodities shifts over time, a link relative can be calculated instead, using the previous year (rather than a single fixed base year) as the comparison point: Link relative = (price in current year / price in previous year) x 100. Because link relatives have no fixed base, they are converted into chain indices for comparability -- the chain index for the first year is set to 100, and each subsequent year's chain index is obtained by multiplying its link relative by the previous year's chain index and dividing by 100. Chain indices computed this way turn out to be mathematically identical to price relatives computed from the first year as a fixed base. The chain base method is rarely used in practice for this reason, but has some advantages: it makes year-

to-year comparisons easy, allows substituting new items for obsolete ones, allows the weighting system to be updated over time, and can accommodate changes in geographic coverage.

5.4 Weighted Index Numbers

Weighted index numbers assign each commodity a weight proportional to its real importance, most commonly using its quantity as the weight. The weighted aggregate price index is Pon = (sum of poi*qoi in current year) / (sum of poi*qoi in base year) x 100. Three major named formulas differ in which year's quantities they use as weights. Laspeyre's index uses base year quantities as weights: Pon = (sum of pni*q0i) / (sum of p0i*q0i) x 100 -- convenient since base year quantities don't need to be re-collected, though this tends to overweight commodities whose prices have risen. Paasche's index instead uses current year quantities as weights: Pon = (sum of pni*qni) / (sum of p0i*qni) x 100 -- this gives comparatively less weight to commodities whose prices have increased. Fisher's ideal index is the geometric mean of the Laspeyre and Paasche indices, splitting the difference between the two: Pon = square root(Laspeyre x Paasche) x 100.

5.5 Consumer Price Index (CPI) and Wholesale Price Index (WPI)

The Consumer Price Index (CPI), also called the cost of living index, measures the aggregate change in the cost of a fixed basket of goods and services purchased at current prices compared with a base period (always set to 100). It is calculated separately for different income and occupational groups across many cities, covering hundreds of consumption items reflecting local tastes and habits. Constructing the CPI involves five main steps: deciding the category of people it represents, conducting a family budget inquiry (via random sampling) to learn what is consumed and at what prices, selecting the items to include, collecting retail price quotations, and choosing weights -- most commonly via the Aggregative Expenditure Method (Laspeyre's formula using base year quantities) or the Family Budget Method, which computes Pon = (sum of I*W) / (sum of W) x 100, where I is the price relative and W = p0*q0 is the base year value weight for each item.

The CPI is the standard way of measuring inflation: the rate of inflation for a given year is calculated as [(current year CPI - previous year CPI) / previous year CPI] x 100, and an average annual inflation rate over several years is simply the average of the individual yearly rates. Because the CPI's base is set to 100 (and currency subdivides into 100 smaller units), the purchasing power of money can be measured as its inverse: (100 / CPI) x 100. The Sensitive Price Indicator (SPI) is calculated identically to the CPI but covers only a small set of essential commodities rather than the full consumption basket. The Wholesale Price Index (WPI), despite its name, actually measures changes in producers' selling prices (not wholesale prices), using weights derived from the value of marketable surplus, and is computed using the same formulas as the CPI.

Important Definitions

What is an index number?: A ratio or average of ratios, usually expressed as a percentage, that measures the relative change in a variable (such as price, wages, or quantity) over time, relative to a base period.

What is a simple index number?: An index number that measures the relative change in a single variable with respect to a base year, such as an index of wheat prices alone.

What is a composite (aggregate) index number?: An index number that measures the relative change in two or more variables together with respect to a base year, such as a basket of many commodity prices.

What is a link relative?: The price of an item in the current year expressed as a percentage of its price in the immediately preceding year, rather than a fixed base year.

What is a chain index?: An index built from link relatives by setting the first year to 100 and multiplying each subsequent year's chain index by that year's link relative, divided by 100.

What is Laspeyre's index?: A weighted price index that uses base year quantities as weights for both the base and current year prices.

What is Paasche's index?: A weighted price index that uses current year quantities as weights for both the base and current year prices.

What is Fisher's ideal index?: The geometric mean of Laspeyre's and Paasche's indices, used because it balances the tendency of each to over- or under-weight commodities with rising prices.

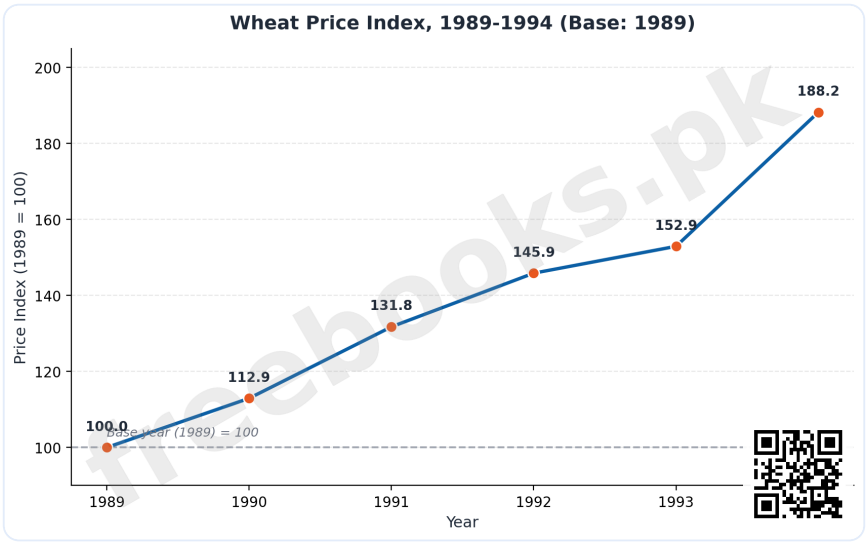
What is the Consumer Price Index (CPI)?: An index measuring the aggregate change in the cost of a fixed basket of consumer goods and services at current prices compared to a base period, also called the cost of living index.

Key Facts and Relations

Topic	Key Fact / Relation
Simple index number	In = (price in current year / price in base year) x 100
Simple aggregate index	Pon = (sum of Pn) / (sum of P0) x 100
Average relative index	Pon = (1/k) x sum of (Pni / P0i)
Link relative	Link relative = (Pn / Pn-1) x 100
Chain index	Chain index(year n) = link relative(n) x chain index(n-1) / 100
Laspeyre's price index	Pon = (sum of pni*q0i) / (sum of p0i*q0i) x 100
Paasche's price index	Pon = (sum of pni*qni) / (sum of p0i*qni) x 100
Fisher's ideal index	Pon = sqrt(Laspeyre x Paasche) x 100
Family Budget Method (CPI)	Pon = (sum of I*W) / (sum of W) x 100, where I = (Pn/P0)x100, W = p0*q0
Rate of inflation	Rate = [(current CPI - previous CPI) / previous CPI] x 100
Purchasing power of money	Purchasing power = (100 / CPI) x 100

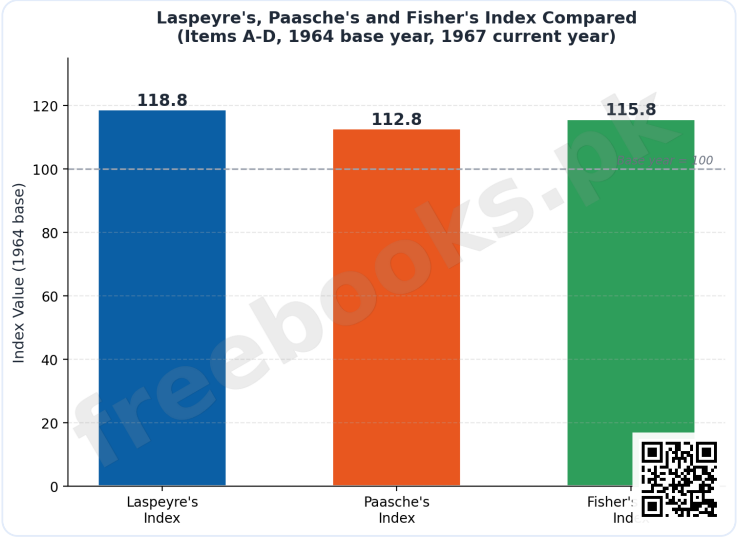
Diagrams

Wheat Price Index, 1989-1994 (Base: 1989): A line chart of the simple price index for wheat from 1989 to 1994, using 1989 as the base year (index = 100), showing the rising trend calculated in the chapter's opening worked example



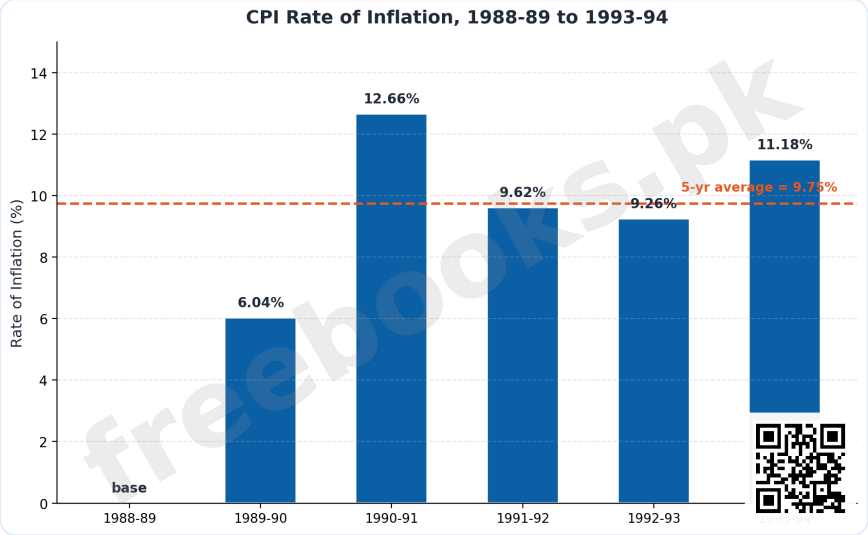
Wheat Price Index, 1989-1994 (Base: 1989)

Laspeyre's, Paasche's and Fisher's Index Compared: A bar chart comparing the three weighted index numbers calculated from the same worked example data (items A-D, 1964 base year, 1967 current year): Laspeyre's 118.8, Paasche's 112.8, and Fisher's ideal index 115.8



Laspeyre's, Paasche's and Fisher's Index Compared

CPI Rate of Inflation, 1988-89 to 1993-94: A bar chart of Pakistan's annual inflation rate calculated from CPI values across six years (1988-89 through 1993-94), with the average annual rate of 9.75% marked as a reference line



CPI Rate of Inflation, 1988-89 to 1993-94

Short Questions & Answers

Why is the base year's index number always 100?

Because the index number formula compares each year's value to the base year's own value, and any value divided by itself, expressed as a percentage, equals 100.

What is the key difference between a simple and a composite index number?

A simple index measures the change in a single variable, while a composite (aggregate) index measures the combined relative change across two or more variables.

Why do simple aggregate indices suffer from a 'units' problem?

Because they sum raw prices across different commodities, changing the units of measurement for any one commodity can distort the index, since the sum mixes quantities that aren't naturally comparable.

Why is a chain index mathematically equivalent to a fixed-base price relative?

Because multiplying each year's link relative by the accumulated chain index (and dividing by 100 each time) telescopes down to the same ratio as comparing each year directly back to the very first year.

Why does Laspeyre's index tend to overweight commodities with rising prices?

Because it fixes quantities at base year levels, ignoring that consumers typically buy less of a commodity once its price rises, so the calculation continues crediting the old, higher consumption level to price increases.

Why is Fisher's index called an 'ideal' index?

Because it takes the geometric mean of Laspeyre's and Paasche's indices, balancing Laspeyre's tendency to overweight price increases against Paasche's tendency to underweight them.

How is the annual rate of inflation calculated from CPI values?

By taking the percentage change from the previous year's CPI to the current year's CPI: [(current - previous) / previous] x 100.

Long Questions & Answers

Explain the difference between unweighted and weighted index numbers, describing the simple aggregate index, average relative index, Laspeyre's index, Paasche's index, and Fisher's ideal index.

Index numbers used to track composite change across many commodities at once fall into two broad families depending on whether every commodity is treated as equally important or not, and understanding this distinction is central to understanding why several different named formulas exist for what seems like the same basic task. Unweighted index numbers deliberately give every commodity in the basket equal importance, regardless of how significant that commodity actually is in real spending or production. The simplest form of this is the simple aggregate index, calculated by summing the current-year prices of every commodity in the basket, summing their base-year prices separately, and expressing the ratio of these two totals as a percentage. While straightforward to calculate, this approach has a serious flaw: because it directly sums raw prices measured in whatever units each commodity happens to use, the resulting index can be distorted simply by changing the unit of measurement for any single

commodity, even though nothing about the real economic situation has changed. A related unweighted approach, the average relative index, avoids this units problem by first calculating an individual price relative for each commodity separately -- the ratio of its current price to its base price -- and only then averaging these already-unitless ratios across all commodities, typically using the arithmetic mean, though geometric mean or median can also be used. This solves the units issue but retains the other core weakness of unweighted indices: every commodity, whether it's a staple like wheat or a minor item like matches, contributes equally to the final index, which rarely reflects how people actually experience price change in their day-to-day spending. Weighted index numbers correct this by deliberately assigning each commodity a weight proportional to its real importance, most commonly based on the quantity of that commodity being bought, sold, or produced. The question that immediately follows is: importance measured using whose quantities, and from which year? Laspeyre's index answers this by using the base year's quantities as the weight for both the base and current year prices, which has a practical convenience -- the base year quantities only need to be collected once and can then be reused for every subsequent year's calculation without needing fresh quantity data each time. However, this convenience comes at a cost: because real consumers typically buy less of a commodity once its price rises significantly (and more of substitutes that have become relatively cheaper), continuing to use the old, pre-price-increase quantities as weights tends to systematically overstate the true impact of price increases on the overall index. Paasche's index takes the opposite approach, using current year quantities as the weight instead. This corrects Laspeyre's overweighting problem, since it reflects how consumption has actually adjusted in response to the new prices, but it introduces the opposite bias: because it uses quantities from the same year as the prices being measured, it tends to understate the true impact of price increases, since heavily-affected commodities have already had their consumption reduced by the time they're being weighted. Fisher's ideal index resolves this tension elegantly rather than trying to argue for one bias over the other: since Laspeyre's index systematically runs a bit too high and Paasche's index systematically runs a bit too low, Fisher's index simply takes the geometric mean of the two, landing at a value that sits meaningfully between them and is widely regarded as a more balanced measure of overall price change than either individual formula on its own.

Explain how the Consumer Price Index (CPI) is constructed, how it is used to measure inflation and purchasing power, and how it differs from the Sensitive Price Indicator (SPI) and Wholesale Price Index (WPI).

The Consumer Price Index, often simply called the cost of living index, is one of the most practically important applications of index number theory, because it is the standard tool governments and economists use to track how the cost of everyday life changes over time for ordinary consumers. Constructing a reliable CPI is a deliberate, multi-step process rather than a single calculation. The first step is deciding exactly which category of people the index is meant to represent -- industrial workers, clerks, a particular income bracket, or some other defined group -- since spending patterns vary enormously across different segments of a population, and a single index cannot meaningfully represent everyone at once. Once the target group is defined, a family budget inquiry is carried out, typically through random sampling of an adequate number of households from that group during a normal (non-crisis) period; this inquiry reveals what quantities and qualities of different goods people actually consume, across categories such as food, clothing, housing, fuel, and so on, along with the retail prices they pay and how much they spend on each category. This budget data then informs the third step, selecting which specific items to include in the index: only items that are widely used by the target group and that aren't subject to extreme, erratic swings in quantity, supply, or price are chosen, since highly volatile or rarely-purchased items would make the index unstable or unrepresentative. The fourth step is collecting price quotations, which must specifically be retail prices actually paid by consumers, gathered from real shops and official sources across the relevant geographic area, rather than wholesale or list prices that don't reflect what people are truly paying. Finally, weights must be chosen to reflect how important each category of spending is to the group's overall household budget; this is most commonly done either through the Aggregative Expenditure Method, which is simply Laspeyre's index applied using base year quantities as weights, or through the Family Budget Method, which instead computes a weighted average of individual price relatives, where each commodity's weight is based on how much of the family's total base year spending it represented. Once constructed, the CPI's primary practical use is measuring inflation: the annual

rate of inflation for any given year is calculated as the percentage change from the previous year's CPI value to the current year's CPI value, and an average inflation rate spanning several years can then be found simply by averaging these individual year-over-year rates. A second major practical use follows directly from the CPI's own construction: because the CPI's base period is always fixed at exactly 100, and currency itself typically subdivides into 100 smaller units, the purchasing power of money at any later point can be estimated as one hundred divided by the current CPI, multiplied by one hundred -- effectively showing how many of the original 100 base-period units of buying power remain. The CPI has two close relatives that are constructed using the exact same underlying formulas and methodology but differ in scope. The Sensitive Price Indicator (SPI) is calculated identically to the CPI, but rather than tracking the full basket of several hundred consumption items, it tracks only a small handful of essential commodities, making it a faster, more responsive gauge of short-term price movements in daily necessities. The Wholesale Price Index (WPI), despite what its name might suggest, does not actually measure wholesale prices at all -- it measures changes in producers' own selling prices, with its weights derived from the value of each commodity's marketable surplus (the portion actually available for sale rather than kept for personal use), and it is calculated using the same underlying index formulas as the CPI, just applied to a completely different stage of the supply chain.

Multiple Choice Questions (MCQs)

An index number is best described as: (A) A single fixed price (B) A ratio or average of ratios expressed as a percentage (C) A type of frequency distribution (D) A measure of central tendency

Correct answer: (B) A ratio or average of ratios expressed as a percentage. An index number is a ratio or an average of ratios, usually expressed as a percentage, measuring relative change over time.

The index number for the base year is always: (A) 0 (B) 50 (C) 100 (D) It varies each time

Correct answer: (C) 100. By definition and convention, the base year's own index number is always 100.

A composite (aggregate) index number differs from a simple index number in that it: (A) Uses only one commodity (B) Measures relative change across two or more variables at once (C) Cannot be expressed as a percentage (D) Ignores the base year

Correct answer: (B) Measures relative change across two or more variables at once. A composite/aggregate index measures relative change in two or more variables together, unlike a simple index which tracks just one.

The main weakness of a simple aggregate index is that it: (A) Cannot be calculated by hand (B) Is distorted by changes in measuring units and ignores relative importance (C) Always equals exactly 100 (D) Only applies to quantities, never prices

Correct answer: (B) Is distorted by changes in measuring units and ignores relative importance. Simple aggregate indices are sensitive to units of measurement and treat every commodity as equally important.

Laspeyre's index uses which year's quantities as weights? (A) Current year (B) Base year (C) The average of all years (D) It uses no quantities at all

Correct answer: (B) Base year. Laspeyre's index uses base year quantities as weights for both the base and current year prices.

Paasche's index uses which year's quantities as weights? (A) Base year (B) Current year (C) A randomly chosen year (D) It doesn't use quantities

Correct answer: (B) Current year. Paasche's index uses current year quantities as weights, unlike Laspeyre's which uses base year quantities.

Fisher's ideal index is calculated as: (A) The sum of Laspeyre's and Paasche's index (B) The arithmetic mean of Laspeyre's and Paasche's index (C) The geometric mean of Laspeyre's and Paasche's index (D) Laspeyre's index divided by Paasche's index

Correct answer: (C) The geometric mean of Laspeyre's and Paasche's index. Fisher's ideal index is the geometric mean (square root of the product) of Laspeyre's and Paasche's indices.

The Consumer Price Index (CPI) is also known as the: (A) Wholesale price index (B) Cost of living index (C) Sensitive price indicator (D) Value index

Correct answer: (B) Cost of living index. CPI is also called the cost of living index, since it tracks the changing cost of a fixed consumer basket.

The rate of inflation for a given year is calculated using: (A) The current year's CPI alone (B) The percentage change between current and previous year's CPI (C) The base year's index only (D) The Wholesale Price Index only

Correct answer: (B) The percentage change between current and previous year's CPI. Rate of inflation = [(current CPI - previous CPI) / previous CPI] x 100.

The Wholesale Price Index (WPI) actually measures: (A) Retail consumer prices (B) Producers' selling prices (C) Only import prices (D) Only export prices

Correct answer: (B) Producers' selling prices. Despite its name, the WPI measures changes in producers' selling prices, not literal wholesale prices.

Quick Revision Summary

- Index number In = (current year value / base year value) x 100; base year index is always 100
- Simple index = single variable; Composite/aggregate index = two or more variables combined
- Unweighted: Simple aggregate index (sum of prices ratio) and Average relative index (average of ratios)
- Link relative uses previous year as base; Chain index converts link relatives to a fixed-base-equivalent series
- Laspeyre's index: base year quantities as weights (tends to overweight price rises)
- Paasche's index: current year quantities as weights (tends to underweight price rises)
- Fisher's ideal index = geometric mean of Laspeyre's and Paasche's index
- CPI = cost of living index; constructed via 5 steps: category, budget inquiry, item selection, price quotations, weights

- Rate of inflation = % change in CPI year over year; Purchasing power of money = (100/CPI) x 100
- SPI = CPI formula but fewer essential items; WPI = measures producers' selling prices, not literal wholesale prices

Exam Tips

- Remember: base year index is ALWAYS 100 by definition — a very common quick-check on any calculation
- Laspeyre's = base year quantities; Paasche's = current year quantities; Fisher's = geometric mean of both
- For 'compare two commodities across time' word problems, check if quantities are given — if yes, it's likely a weighted index question
- Rate of inflation always compares CONSECUTIVE years' CPI, not a year against the fixed base
- Chain index and fixed-base price relative give the same final answer — don't be confused by the different-looking calculation steps
- CPI, SPI and WPI all use the same underlying formulas — what differs is only the basket of items and whose prices are tracked

Chapter 6: Probability

If you bought 7 tickets out of 700 sold for a raffle, you'd say you have 1 chance in 100 of winning -- probability is exactly this: a measure of the likelihood that something will happen. It cannot predict whether a specific event will actually occur, only how likely it is, yet most decisions in daily life are based on likelihood rather than absolute certainty.

This chapter builds the mathematical toolkit needed to work with chance events: essential set theory and Venn diagrams, counting techniques (factorials, permutations, and combinations), the formal definitions of random experiments, sample spaces, and events, the classical and axiomatic definitions of probability, and the two central theorems of the subject -- the addition theorem (for finding the probability that at least one of two events occurs) and the multiplication theorem (for finding the probability that two events occur together), including conditional probability.

Learning Objectives

- Use basic set notation and operations: union, intersection, difference, and complement
- Calculate factorials, permutations, and combinations for counting problems
- Define random experiment, sample space, sample point, and event, and classify events as simple or compound
- Distinguish mutually exclusive, independent, dependent, equally likely, and exhaustive events
- State the classical, relative frequency, and axiomatic (mathematical) definitions of probability
- Apply the addition theorem for mutually exclusive and not mutually exclusive events
- Calculate conditional probability and apply the multiplication theorem for independent and dependent events
- Solve applied probability problems involving cards, dice, and drawing balls from a bag

Key Concepts

6.1 Sets and Venn Diagrams

A set is a well-defined collection of distinct objects, called its elements; sets are denoted by capital letters (A, B, C) and elements by small letters. A null (empty) set contains no elements. Set A is a subset of B if every element of A is also in B; it is a proper subset if B also contains at least one element not in A. A universal set U contains all elements under consideration. Two sets are disjoint if they share no elements (A intersect B = empty set); they are overlapping if they share at least one element but neither is a subset of the other.

A Venn diagram represents the universal set U as a rectangle and its subsets as circles, visually showing relationships between sets. The union of A and B (A union B) contains every element belonging to A, to B, or to both. The intersection of A and B (A intersect B) contains only elements belonging to both A and B. The difference A minus B contains elements of A that are not in B. The complement of A (written A-bar or A with a superscript c) is U minus A -- everything in the universal set that is not in A.

6.1.1 Factorial, Permutations and Combinations

For a positive integer n, the factorial n! is the product of the first n positive integers: n! = n(n-1)(n-2)...3.2.1, which can also be written recursively as n! = n(n-1)!. A permutation is an arrangement of a finite number of objects in a definite order; the number of ways of arranging n objects taken r at a time is nPr = n!/(n-r)!. When n objects include repeated items (n1 alike of one kind, n2 alike of another, and so on), the number of distinct permutations is n!/(n1! n2! ... nk!).

A combination is a selection of objects made without regard to order; the number of combinations of n things taken r at a time is nCr = n!/(r!(n-r)!). The key distinction: permutations count arrangements (order matters), while combinations count selections (order doesn't matter), so nCr is always smaller than or equal to nPr for the same n and r.

6.1.2 Random Experiments, Sample Space and Events

A random experiment produces different outcomes even when repeated many times under similar conditions, and can be repeated any number of times with at least two possible outcomes each trial. The sample space S is the set of all possible outcomes of a random experiment, and each individual outcome is a sample point -- for a single coin toss, S = {H, T}; for two coin tosses, S = {HH, HT, TH, TT}; for a single die, S = {1,2,3,4,5,6}; and for two dice, S contains all 36 ordered pairs (i,j) with i,j from 1 to 6.

An event is any subset of the sample space. A simple event consists of exactly one sample point; a compound event consists of more than one. Two events are independent if the occurrence of one does not affect the occurrence of the other, and dependent if it does. Two events are mutually exclusive if they cannot occur together (their intersection is empty). Two events are equally likely if they have the same chance of occurring. When a sample space is partitioned into mutually exclusive events whose union is the entire sample space, those events are called exhaustive.

6.2 Definitions of Probability

The classical (a priori) definition: if there are n equally likely, mutually exclusive, and exhaustive outcomes, and m of them are favourable to event A, then P(A) = m/n (favourable outcomes divided by possible outcomes). The relative frequency (a posteriori) definition: if a random experiment is repeated n times under uniform conditions and event A occurs m times, then P(A) is the limiting value of m/n as n grows very large. The mathematical (axiomatic) definition expresses probability using sample points: P(A) = n(A)/n(S), the number of sample points in A divided by the total number of sample points in S.

Every valid probability must satisfy these axioms: P(A) is never negative; P(A) always lies between 0 and 1 inclusive; the probability of the entire sample space is P(S) = 1; and for two mutually exclusive events, P(A union B) = P(A) + P(B).

6.3 Addition Theorem of Probability

For two events that are not mutually exclusive, the probability that at least one of them occurs is P(A union B) = P(A) + P(B) - P(A intersect B) -- the subtraction

corrects for double-counting the outcomes that belong to both events. For two mutually exclusive events (where A intersect B is empty, so P(A intersect B) = 0), this simplifies to the special case P(A union B) = P(A) + P(B).

6.4 Conditional Probability and the Multiplication Theorem

The conditional probability of event A given that event B has already occurred, written P(A/B), is defined as P(A intersect B) / P(B), provided P(B) is greater than zero; if P(B) = 0, the conditional probability is undefined. This lets us update the probability of one event based on knowing another event has happened.

For two independent events, the multiplication theorem states that the probability both occur is simply the product of their individual probabilities: P(A intersect B) = P(A) x P(B). For events that are not independent (dependent events), the joint probability instead uses conditional probability: P(A intersect B) = P(A) x P(B/A), or equivalently P(A intersect B) = P(B) x P(A/B) -- the probability of the first event happening, multiplied by the conditional probability of the second given that the first has already occurred.

Important Definitions

What is a sample space?: The set of all possible outcomes of a random experiment, usually denoted S; each individual outcome within it is called a sample point.

What is an event?: Any subset of the sample space; a simple event contains exactly one sample point, while a compound event contains more than one.

What are mutually exclusive events?: Two events that cannot occur together, meaning their intersection is the empty set (A intersect B = empty set).

What are independent events?: Two events where the occurrence of one does not affect the probability of the other occurring.

What is the classical definition of probability?: P(A) = (number of favourable outcomes) / (number of equally likely, mutually exclusive, exhaustive possible outcomes), i.e., m/n.

What is conditional probability?: The probability that event A occurs given that event B has already occurred, calculated as P(A/B) = P(A intersect B) / P(B), for P(B) greater than 0.

What is the addition theorem for events that are not mutually exclusive?: P(A union B) = P(A) + P(B) - P(A intersect B), subtracting the overlap so it isn't counted twice.

What is the multiplication theorem for independent events?: P(A intersect B) = P(A) x P(B), the probability that both independent events occur equals the product of their individual probabilities.

What is the difference between a permutation and a combination?: A permutation counts arrangements where order matters; a combination counts selections where order does not matter.

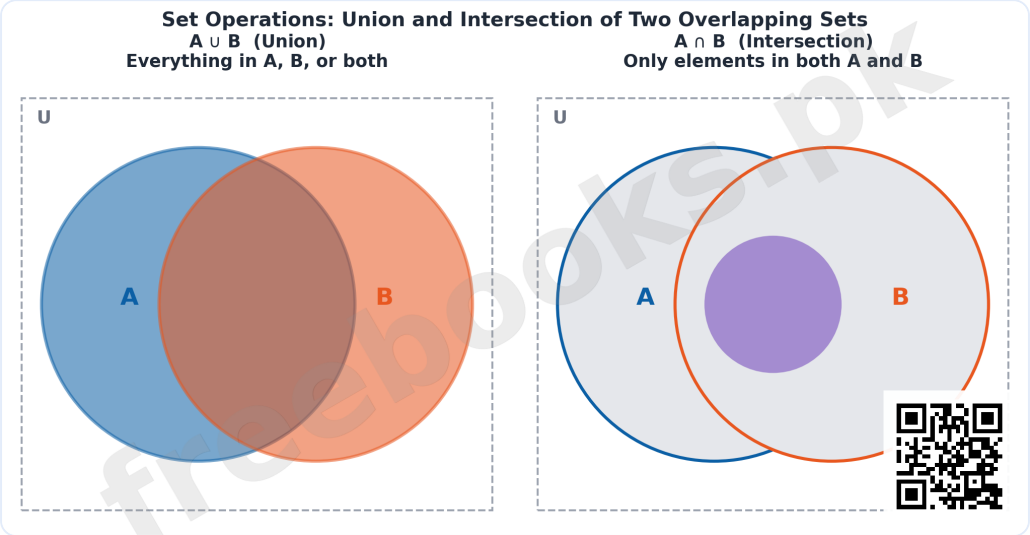
Key Facts and Relations

Topic	Key Fact / Relation
Factorial	n! = n(n-1)(n-2)...3.2.1
Permutations	nPr = n! / (n-r)!
Permutations with repeated items	n! / (n1! n2! ... nk!)
Combinations	nCr = n! / (r!(n-r)!)
Classical probability	P(A) = m / n = favourable outcomes / possible outcomes
Mathematical (axiomatic) probability	P(A) = n(A) / n(S)
Complement rule	P(A-bar) = 1 - P(A)
Addition theorem (general)	P(A union B) = P(A) + P(B) - P(A intersect B)
Addition theorem (mutually exclusive)	P(A union B) = P(A) + P(B)
Conditional probability	P(A/B) = P(A intersect B) / P(B), for P(B) > 0
	P(A intersect B) = P(A) x P(B)

Topic	Key Fact / Relation
Multiplication theorem (independent events)	
Multiplication theorem (dependent events)	P(A intersect B) = P(A) x P(B/A) = P(B) x P(A/B)

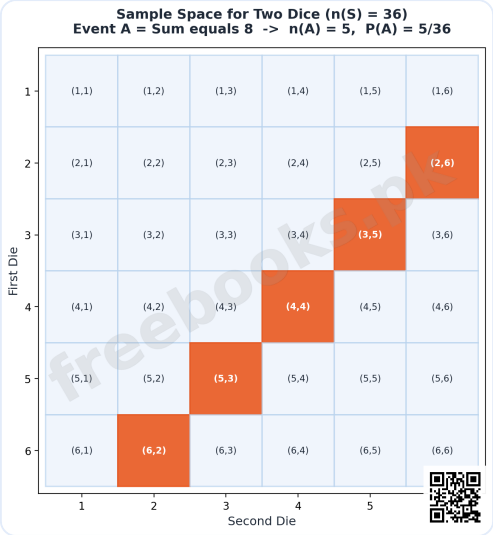
Diagrams

Union and Intersection of Two Overlapping Sets: A Venn diagram of two overlapping sets A and B inside universal set S, with the union (A or B) and intersection (A and B) regions shaded and labelled, matching the set operations introduced in this chapter



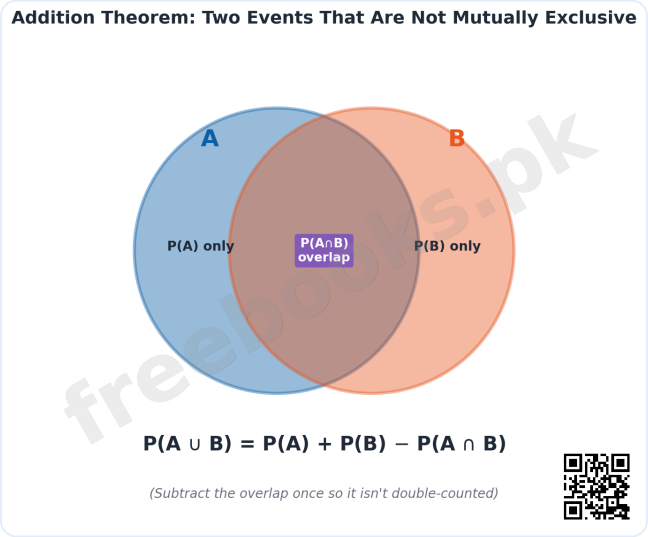
Union and Intersection of Two Overlapping Sets

Sample Space for Two Dice: Sum Equals 8: The full 6x6 grid of 36 equally likely outcomes for rolling two dice, with the 5 outcomes summing to 8 highlighted, illustrating the classical definition of probability P(A) = n(A)/n(S)



Sample Space for Two Dice: Sum Equals 8

Addition Theorem: P(A union B) = P(A) + P(B) - P(A intersect B): A Venn diagram showing two overlapping events A and B with their individual probabilities and overlap labelled, illustrating why the intersection must be subtracted once to avoid double-counting



Addition Theorem: P(A union B) = P(A) + P(B) - P(A intersect B)

Short Questions & Answers

What is the difference between a random experiment and a deterministic one?

A random experiment produces different outcomes even under identical repeated conditions, while a deterministic process always produces the same outcome given the same conditions.

Why must probability always lie between 0 and 1?

Because probability is defined as a ratio of favourable outcomes to total possible outcomes, and the number of favourable outcomes can never be negative or exceed the total number of possible outcomes.

Give an example of mutually exclusive events.

In a single coin toss, getting heads and getting tails are mutually exclusive, since both cannot happen on the same toss.

Why is nPr always greater than or equal to nCr for the same n and r?

Because nPr counts every distinct ordering of a selection separately, while nCr counts each unique selection only once regardless of order, so nPr = nCr multiplied by r! (the number of orderings).

Why is conditional probability undefined when P(B) = 0?

Because the conditional probability formula P(A intersect B)/P(B) requires dividing by P(B), and division by zero is undefined.

How does the addition theorem simplify for mutually exclusive events?

Since mutually exclusive events cannot overlap, P(A intersect B) = 0, so the general formula P(A union B) = P(A) + P(B) - P(A intersect B) simplifies to just P(A union B) = P(A) + P(B).

What does it mean for two events to be exhaustive?

That together, their union covers the entire sample space, with no possible outcome left out.

Long Questions & Answers

Explain the classical, relative frequency, and axiomatic (mathematical) definitions of probability, and describe how they relate to each other.

Probability can be approached from three different but ultimately compatible directions, and understanding how each one arrives at essentially the same idea helps clarify what probability actually measures. The classical, or a priori, definition applies when an experiment has a fixed number of possible outcomes that are all equally likely to occur, all mutually exclusive of one another, and collectively exhaustive of every possibility. Under these conditions, if there are n

total possible outcomes and m of them count as favourable to some event A, the classical definition simply sets the probability of A equal to the ratio m over n. This definition works purely on logical symmetry -- it doesn't require actually running the experiment even once, since it reasons directly from the structure of the situation itself, such as knowing in advance that a fair coin has exactly two equally likely sides or that a standard die has exactly six equally likely faces. Its limitation is equally direct: it only applies cleanly when we can be confident the outcomes really are equally likely, which isn't always true in real-world situations, such as a biased coin or a horse race where the competitors clearly don't have identical chances. The relative frequency, or a posteriori, definition takes the opposite approach, relying entirely on repeated observation rather than logical symmetry. Here, an experiment is actually carried out a large number of times, say n times, under conditions that are kept as uniform as possible, and the number of times event A actually occurs, m, is recorded. The probability of A is then defined as the value that the ratio m over n settles down to, or approaches, as n is allowed to grow larger and larger -- technically expressed as the limit of m/n as n approaches infinity. This definition is far more flexible than the classical one, since it doesn't require any assumption of equal likelihood beforehand and can be applied to virtually any repeatable situation, but it comes with its own practical cost: it requires genuinely large amounts of real, repeated data to produce a stable, trustworthy estimate, and with only a small number of trials the observed ratio can fluctuate considerably before settling toward its true long-run value. The mathematical, or axiomatic, definition unifies these two ideas into a single, more abstract and rigorous framework built directly on top of the sample space concept. Here, probability is defined as the ratio of the number of sample points contained in event A to the total number of sample points contained in the entire sample space S, written as n(A) divided by n(S). This framework is considered mathematical because it doesn't just offer a calculation method -- it also lays out a small set of axioms that any valid assignment of probabilities must satisfy no matter how those probabilities were originally derived: the probability of any event must never be negative, every probability must fall somewhere between 0 and 1 inclusive, the probability of the entire sample space itself must equal exactly 1 (since something in the sample space is guaranteed to happen), and for any two mutually exclusive events, the probability that either one occurs must equal the simple sum of their individual

probabilities. These axioms don't compete with the classical or relative frequency definitions -- rather, both of those earlier definitions can be shown to satisfy these same axioms under the right conditions, which is exactly what makes the axiomatic definition the more general and mathematically foundational of the three: it captures the essential logical structure that both practical definitions were already following, whether that structure came from equally likely outcomes reasoned out in advance, or from a stable pattern observed empirically after many repeated trials.

Explain the addition theorem and the multiplication theorem of probability, including the role conditional probability plays in the multiplication theorem for dependent events.

The addition theorem and the multiplication theorem are the two central tools used to calculate the probability of combined events -- the addition theorem handles the question of whether at least one of two events happens, while the multiplication theorem handles the question of whether both events happen together, and understanding the logic behind each explains why their formulas take the specific shape that they do. The addition theorem, in its general form, states that for any two events A and B, the probability that at least one of them occurs is given by P(A union B) equals P(A) plus P(B) minus P(A intersect B). The reasoning behind subtracting the intersection term becomes clear when the calculation is thought of in terms of counting sample points rather than probabilities directly: if event A contains a certain number of sample points and event B contains another number of sample points, and some of those sample points happen to belong to both events at once, then simply adding the two counts together would count every one of those shared, overlapping sample points twice over -- once as part of A's total, and again as part of B's total. Subtracting the probability of the intersection corrects for exactly this double-counting, ensuring each sample point belonging to the combined event A union B is counted exactly once, no matter how many of the original events it happened to belong to. When the two events happen to be mutually exclusive, meaning they cannot possibly occur together and therefore have no sample points in common at all, this correction term disappears entirely, since P(A intersect B) is simply zero in that case, leaving the simpler special-case formula P(A union B) equals P(A) plus P(B) directly. The multiplication theorem addresses an entirely

different question: not whether at least one event happens, but whether both events happen together, and its correct form depends critically on whether the two events are independent of one another or not. When events A and B are independent -- meaning that knowing whether A has occurred tells us absolutely nothing about whether B is any more or less likely to occur -- the probability that both happen simultaneously is simply the straightforward product of their individual probabilities, P(A intersect B) equals P(A) times P(B). This multiplicative relationship falls directly out of counting the total possible outcomes of the combined experiment: if event A has a favourable-to-total outcome ratio of m over n, and independent event B has its own separate favourable-to-total ratio of M over N, then every one of A's n possible outcomes can be freely combined with every one of B's N possible outcomes, giving nN total possible combined outcomes, of which exactly mM combinations represent both A and B occurring together, and mM over nN reduces algebraically to exactly (m/n) times (M/N). However, when the two events are not independent of each other -- meaning the occurrence of one event genuinely does shift the likelihood of the other -- this simple multiplication no longer holds, and this is precisely the situation where conditional probability becomes essential. Conditional probability, written P(A given B), captures exactly how the probability of event A changes once we already know for certain that event B has occurred, and is formally defined as the probability of both events happening together divided by the probability of the known, already-occurred event: P(A intersect B) divided by P(B). Rearranging this definition algebraically produces the multiplication theorem for dependent events: P(A intersect B) equals P(A) times the conditional probability P(B given A), or equivalently, and just as validly, P(A intersect B) equals P(B) times the conditional probability P(A given B). Both versions express precisely the same underlying logic from two different starting directions: the probability that both events occur together is the probability of whichever event is considered to happen 'first,' multiplied by the updated, conditional probability of the second event occurring given that the first one is now already known to have happened -- and this conditional adjustment is exactly what distinguishes the dependent-events version of the multiplication theorem from the simpler, unconditional product used when the two events are independent of one another.

Multiple Choice Questions (MCQs)

The number of ways of arranging n objects taken r at a time (order matters) is given by: (A) nCr (B) nPr (C) n! (D) n/r

Correct answer: (B) nPr. nPr = n!/(n-r)! counts ordered arrangements (permutations), where order matters.

A combination differs from a permutation in that a combination: (A) Considers order important (B) Does not consider order (C) Only applies to two objects (D) Is always larger than the permutation

Correct answer: (B) Does not consider order. A combination is a selection made without regard to order, unlike a permutation.

Two events that cannot occur together are called: (A) Independent events (B) Exhaustive events (C) Mutually exclusive events (D) Equally likely events

Correct answer: (C) Mutually exclusive events. Mutually exclusive events have an empty intersection -- they cannot both occur on the same trial.

The classical definition of probability requires outcomes to be: (A) Random and unpredictable only (B) Equally likely, mutually exclusive, and exhaustive (C) Based on repeated trials only (D) Independent of the sample space

Correct answer: (B) Equally likely, mutually exclusive, and exhaustive. The classical (a priori) definition specifically requires equally likely, mutually exclusive, exhaustive outcomes.

For any event A, which of the following is always true? (A) P(A) can be negative (B) 0 <= P(A) <= 1 (C) P(A) > 1 (D) P(A) = n(S)

Correct answer: (B) 0 <= P(A) <= 1. One of the basic axioms of probability is that P(A) always lies between 0 and 1 inclusive.

If A and B are mutually exclusive, P(A union B) equals: (A) P(A) x P(B) (B) P(A) - P(B) (C) P(A) + P(B) (D) P(A) / P(B)

Correct answer: (C) P(A) + P(B). For mutually exclusive events, P(A intersect B) = 0, so the addition theorem simplifies to P(A union B) = P(A) + P(B).

Conditional probability P(A/B) is defined as: (A) P(A) x P(B) (B) P(A intersect B) / P(B) (C) P(A) + P(B) (D) P(B) / P(A)

Correct answer: (B) P(A intersect B) / P(B). P(A/B) = P(A intersect B) / P(B), provided P(B) > 0.

If A and B are independent events, P(A intersect B) equals: (A) P(A) + P(B) (B) P(A) - P(B) (C) P(A) x P(B) (D) P(A/B)

Correct answer: (C) P(A) x P(B). For independent events, the multiplication theorem gives P(A intersect B) = P(A) x P(B).

The general addition theorem for two events not mutually exclusive is: (A) P(A) + P(B) (B) P(A) + P(B) - P(A intersect B) (C) P(A) x P(B) (D) P(A) - P(B)

Correct answer: (B) P(A) + P(B) - P(A intersect B). The general form subtracts the overlap: P(A union B) = P(A) + P(B) - P(A intersect B).

A sample space is best described as: (A) Any subset of possible outcomes (B) The set of all possible outcomes of a random experiment (C) A single favourable outcome (D) The complement of an event

Correct answer: (B) The set of all possible outcomes of a random experiment. The sample space S is the complete set of all possible outcomes of a random experiment.

Quick Revision Summary

- Set operations: Union (A or B), Intersection (A and B), Difference (A minus B), Complement (U minus A)
- Permutation nPr = n!/(n-r)! (order matters); Combination nCr = n!/(r!(n-r)!) (order doesn't matter)
- Sample space S = all possible outcomes; Event = any subset of S; Simple event = 1 point, Compound = more than 1
- Mutually exclusive: A intersect B = empty set. Independent: occurrence of one doesn't affect the other
- Classical probability: P(A) = m/n (favourable/possible, when outcomes equally likely)
- Axiomatic: 0 <= P(A) <= 1; P(S) = 1; P(A union B) = P(A)+P(B) for mutually exclusive events
- Addition theorem (general): P(A union B) = P(A) + P(B) - P(A intersect B)
- Conditional probability: P(A/B) = P(A intersect B) / P(B), for P(B) > 0

- Multiplication (independent): P(A intersect B) = P(A) x P(B)
- Multiplication (dependent): P(A intersect B) = P(A) x P(B/A) = P(B) x P(A/B)

Exam Tips

- Always double-check whether a word problem needs a permutation (order matters, e.g. 'arrange', 'rank') or a combination (order doesn't matter, e.g. 'select', 'choose')
- Before applying the addition theorem, check if events are mutually exclusive — if yes, skip the subtraction term entirely
- For 'and' probability questions, check independence first: independent events just multiply; dependent events need conditional probability
- Remember P(A/B) is NOT the same as P(B/A) in general — always double-check which event is 'given'
- Complement trick: P(at least one) is often easier via 1 - P(none), especially for 'at least' word problems
- When drawing balls/cards without replacement, events become dependent — use conditional probability, not simple multiplication

Chapter 7: Random Variables

When a pair of dice is thrown, we're rarely interested in the raw outcome like (3, 5) -- we care about a number derived from it, such as the total of the two upturned faces, which could be anywhere from 2 to 12. That derived number is a random variable: a rule that assigns a real number to every outcome of a random experiment. This chapter formalizes that idea, introduces how random numbers are generated and used in simulation (Monte Carlo methods), and draws the essential distinction between discrete random variables (which take isolated values, like counts) and continuous random variables (which take any value in an interval, like measurements).

Random variables are the bridge between the abstract sample spaces of Chapter 6 and the probability distributions built in Chapter 8 -- once outcomes are converted into numbers, we can describe their behaviour using tables, formulas, means, and variances, the same tools used throughout the rest of statistics.

Learning Objectives

- Define a random variable and identify examples from real experiments
- Understand what random numbers are and how the pseudo-random number generation formula works
- Use a random number table to simulate outcomes of simple experiments (coin tosses, etc.)
- Explain the role of random numbers in simulation (Monte Carlo methods)
- Construct random variables from experiments in different fields and list their possible values
- Distinguish between discrete and continuous random variables with examples

Key Concepts

7.1 Introduction to Random Variables

Every random experiment produces two or more outcomes, and usually we are interested in some particular numerical aspect of those outcomes rather than the

raw outcome itself. When a pair of dice is thrown, for instance, the interest may lie in the total of the two upturned faces -- a value that could be 2, 3, 4, and so on up to 12. Likewise, if we record the number of children in each of fifty randomly chosen families (assuming no family has more than 5), the values of interest are 0, 1, 2, 3, 4, or 5.

A variable whose values depend upon the outcomes of a random experiment is called a random variable. Random variables are denoted by capital letters such as X, Y, or Z, and the specific values they take are denoted by the corresponding small letters x, y, or z. For example, if a pair of dice is thrown and Y denotes the sum of the two upturned values, then Y assigns the outcome (1,1) the value 1+1=2, the outcome (2,1) the value 2+1=3, and so on up to the outcome (6,6), which gets the value 6+6=12 -- so Y can take any of the values 2 through 12.

7.2 Random Numbers and Their Generation

Random numbers are a sequence of digits from the set {0,1,2,...,9} such that at every position in the sequence, each digit has the same probability of 1/10 (0.1) of being selected, regardless of the digits already generated. These are also called random digits. The simplest way to produce them is through games of chance -- dice, coins, cards, or drawing numbered slips from a hat -- but this becomes tedious for long sequences, so printed tables of random digits are widely used instead.

For computer implementation, the most common approach is the pseudo-random technique. These numbers are called 'pseudo' because they are not truly random -- they are generated by a completely deterministic recursive formula, yet they exhibit most of the statistical properties of genuinely random digits. Eventually the sequence repeats itself (a cycle), but for a good generator this cycle can be tens of thousands of digits long. The standard recursive formula is x(n+1) = [a.x(n) + b] mod m, where a, b, and m are suitably chosen integer constants and x(1) (the seed) is a chosen starting integer. 'mod m' means: if the result exceeds m, divide by m and keep only the remainder. For the method to be useful, m, a, and b should be chosen reasonably large.

7.3 Application of Random Numbers

Random numbers are widely used in simulation techniques, also called Monte Carlo methods, which are applied across the sciences whenever direct experimentation is impossible, too costly, or too time-consuming. Random number tables group digits (in 2, 3, 4, or 5-digit sequences) purely for ease of reading, and can be read row-wise or column-wise depending on how many digits are needed for a given problem.

To simulate a coin toss using a random number table: assign even digits (0,2,4,6,8) to represent heads and odd digits (1,3,5,7,9) to represent tails, then read single digits from an arbitrarily chosen row and column of the table, converting each digit to H or T according to the assigned rule. To simulate two coins at once, two-digit numbers are read: both digits even means both heads, both odd means both tails, and one of each means exactly one head. When outcomes have unequal probabilities (rather than the equally likely 50-50 split of a fair coin), a related method called probability proportional to size (PPS) is used: cumulative probabilities are computed for each possible outcome, and a corresponding range of random numbers is assigned to each outcome in proportion to its probability, so that a randomly selected number automatically falls into the correct class the right proportion of the time.

7.4 Constructing Random Variables from Different Fields

Random variables can be built from experiments across almost any field, and the values they take depend entirely on how the variable is defined. Consider three students each choosing between Biology (B) and Computer Science (C): if Y is defined as the number of students taking computer science, the eight equally likely outcomes (BBB, CBB, BCB, BBC, CCB, CBC, BCC, CCC) get mapped to the values 0, 1, 1, 1, 2, 2, 2, and 3 respectively -- so Y takes only the isolated values 0, 1, 2, and 3.

By contrast, consider recording the time (in minutes) a customer waits in a queue at a utility store: this random variable can take any value in an interval -- 5.0 minutes, 5.3 minutes, 12.0 minutes, or anything in between -- rather than a fixed set of separate values. This contrast between a variable that jumps

between isolated points and one that can take any value in a range is exactly the distinction formalized in the next section.

7.5 Discrete and Continuous Random Variables

A random variable is called discrete if the set of values it can take is a collection of isolated points on the real number line -- that is, its sample space is discrete. The outcomes of the experiment are noted, and a number is assigned to each outcome by some rule. For example, if three coins are tossed and Y is the number of heads, then Y takes only the whole-number values 0, 1, 2, or 3 -- a discrete random variable. The 'number of students taking computer science' example above is also discrete.

A random variable is called continuous if the set of values it can take is an entire interval on the number line -- its sample space is continuous. The outcomes are represented by points on a line, and a number is assigned to each point by some rule. For example, if the heights of students in a statistics class range from a minimum of 5.0 feet to a maximum of 5.8 feet, the height variable Y can take any value in the interval 5.0 to 5.8 feet -- a continuous random variable. The customer-waiting-time example above is also continuous. In short: discrete random variables almost always arise from counting, while continuous random variables arise from measuring.

Important Definitions

What is a random variable?: A variable whose values depend on the outcomes of a random experiment; denoted by capital letters (X, Y, Z) with specific values denoted by the corresponding small letters.

What are random numbers?: A sequence of digits from {0,1,...,9} such that each digit at each position has the same probability, 1/10, of being selected, regardless of prior digits.

What is a pseudo-random number?: A number generated by a deterministic recursive formula (such as x(n+1) = [a.x(n)+b] mod m) that is not truly random but exhibits the statistical properties of random digits, eventually repeating in a long cycle.

What is a Monte Carlo method?: A simulation technique that uses random numbers to study problems where direct experimentation is impossible, too costly, or too time-consuming.

What is a discrete random variable?: A random variable whose possible values form a collection of isolated points on the number line, typically arising from counting.

What is a continuous random variable?: A random variable whose possible values form an entire interval on the number line, typically arising from measurement.

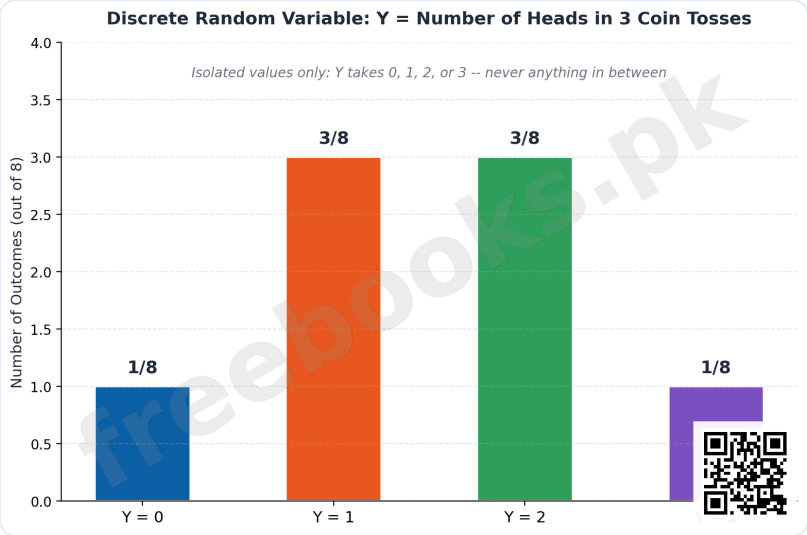
What is probability proportional to size (PPS)?: A simulation method that assigns a range of random numbers to each possible outcome in proportion to its probability, using cumulative probabilities to define the ranges.

Key Facts and Relations

Topic	Key Fact / Relation
Pseudo-random number generator	x(n+1) = [a.x(n) + b] mod m
Random digit probability	P(any single digit 0-9) = 1/10 = 0.1
Discrete random variable	Values are isolated points (arise from counting)
Continuous random variable	Values fill an entire interval (arise from measuring)
PPS cumulative range rule	Random number range assigned to a class is proportional to its cumulative probability

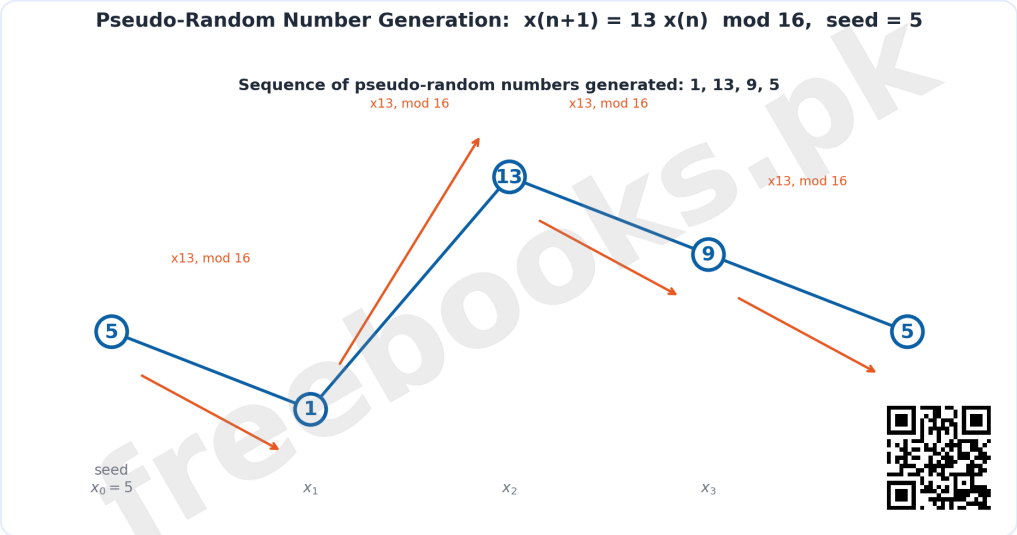
Diagrams

Discrete Random Variable: Number of Heads in 3 Coin Tosses: A probability mass bar chart for Y = number of heads when 3 coins are tossed, showing the 8 equally likely outcomes mapping to isolated values 0, 1, 2, 3 with their frequencies, illustrating a discrete random variable



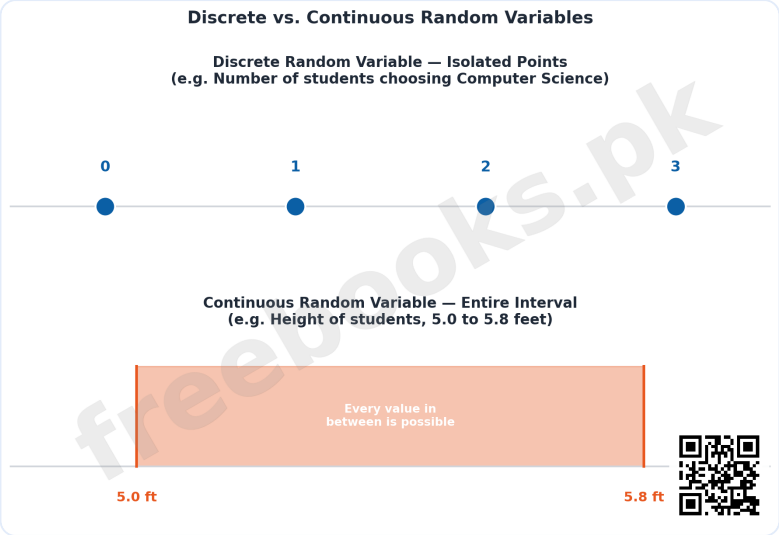
Discrete Random Variable: Number of Heads in 3 Coin Tosses

Pseudo-Random Number Generation (LCG Sequence): A step-by-step visualization of the linear congruential generator x(n+1) = 13x(n) mod 16 starting from seed 5, tracing the sequence 5 -> 1 -> 13 -> 9 -> 5, matching Example 7.2 from the textbook



Pseudo-Random Number Generation (LCG Sequence)

Discrete vs Continuous Random Variables: A side-by-side comparison: isolated dots at 0,1,2,3 representing a discrete random variable (number of computer-science students) versus a shaded continuous interval from 5.0 to 5.8 representing a continuous random variable (student heights)



Discrete vs Continuous Random Variables

Short Questions & Answers

Why are pseudo-random numbers called 'pseudo'?

Because they are produced by a completely deterministic recursive formula rather than a truly random physical process, even though they display most of the statistical properties of genuine random digits.

Give an example of a discrete random variable and a continuous random variable.

Discrete: the number of heads in three coin tosses (values 0,1,2,3). Continuous: the height of students in a class, which can take any value in an interval such as 5.0 to 5.8 feet.

Why is the number of children in a family a discrete random variable?

Because its possible values (0,1,2,3,...) are isolated whole numbers arising from counting, not from measurement over a continuous interval.

What role do random numbers play in Monte Carlo simulation?

They allow researchers to imitate the outcomes of a random experiment on paper or by computer, which is useful when direct experimentation is impossible, too expensive, or too time-consuming.

In the pseudo-random formula x(n+1) = [a.x(n)+b] mod m, what does 'mod m' mean?

It means that if a.x(n)+b exceeds m, you divide the result by m and keep only the remainder as the next random number in the sequence.

Why must a discrete random variable's sample space consist of isolated points?

Because the variable arises from counting outcomes, which can only take specific whole-number (or otherwise separated) values, with no possible values existing between them.

Long Questions & Answers

Explain what a random variable is, and describe with examples how random variables can be constructed from experiments in different fields.

A random variable is best understood as a bridge between the raw outcomes of a random experiment and the numerical values we actually want to analyze, since most of the time we are not directly interested in an outcome itself but in some measurable quantity that outcome produces. Formally, a random variable is defined as a variable whose values depend upon the outcomes of a random experiment, and by convention it is represented using a capital letter such as X, Y, or Z, while its specific realized values are written using the corresponding lowercase letter x, y, or z. The reason this concept matters so much in statistics is that random experiments on their own -- rolling dice, tossing coins, or interviewing families -- don't come with built-in numbers; it is the random variable that supplies the rule for converting each outcome into a number we can then summarize, average, or build a probability distribution around. Consider, for example, an experiment in which three students in a class are each asked to choose between two courses, Biology or Computer Science. On its own, an outcome of this experiment is simply a string like 'first student picks Biology, second picks Computer Science, third picks Biology' -- not yet a number. To turn this into a random variable, we might define Y as the number of students who choose computer science. Once this rule is fixed, every one of the eight equally likely outcomes of the experiment can be mapped onto a specific value of Y: the all-Biology outcome maps to Y=0, each of the three outcomes with exactly one computer-science student maps to Y=1, each of the three outcomes with exactly two computer-science students maps to Y=2, and the all-computer-science outcome maps to Y=3. Notice that the random variable has compressed eight distinct outcomes down into just four possible numerical values, which is

precisely what makes it useful: instead of tracking every possible combination of student choices, we can now describe the situation purely in terms of the values 0, 1, 2, and 3 and how likely each one is. A completely different field can produce an equally valid but very differently shaped random variable. Consider an experiment where we record, for each customer arriving at a utility store, the number of minutes they wait in a queue before being served. Here, we might define Y as the waiting time of a given customer. Unlike the course-selection example, this random variable cannot be reduced to a small fixed list of possible values -- a customer might wait 5.0 minutes, or 5.3 minutes, or 12.0 minutes, or any other value that a stopwatch could register, and there is no natural gap between one possible waiting time and the next. In both cases, notice that the underlying logic of constructing a random variable is identical regardless of the field: first, identify the random experiment; second, decide precisely what numerical aspect of the outcome is of interest; and third, define the rule that assigns exactly one number to every possible outcome of that experiment. What differs entirely from field to field, and indeed from definition to definition even within the same experiment, is the resulting set of possible values the random variable can take -- sometimes a short list of isolated whole numbers, as with counting how many students chose a particular course, and sometimes an entire continuous range of values, as with measuring how long a customer waits. This very difference in the character of the resulting values is what leads directly into the distinction between discrete and continuous random variables, since the same three-step construction process applied to a counting-based question versus a measurement-based question naturally produces these two fundamentally different types of variables.

Explain the difference between discrete and continuous random variables, and describe how random numbers are generated and used in simulation.

Once a random variable has been defined for a given experiment, the next essential step is to understand what kind of values it can actually take, and this is where the distinction between discrete and continuous random variables becomes essential to the rest of the subject. A random variable is called discrete if the complete set of values it can take forms a collection of isolated points on the real number line -- in other words, its underlying sample space is itself

discrete, and the outcomes of the experiment are noted and then converted into a number through some assignment rule. A classic illustration is tossing three coins and letting Y represent the number of heads observed: working through all eight possible outcomes (HHH, HHT, HTH, HTT, THH, THT, TTH, TTT) and counting the heads in each shows that Y can only ever equal 0, 1, 2, or 3 -- four separated whole-number values with nothing meaningful existing in between them, such as 1.5 heads. Because these values arise from a simple counting process, discrete random variables almost always show up whenever a problem asks 'how many' of something occurred. In sharp contrast, a random variable is called continuous if the complete set of values it can take forms an entire interval on the number line rather than a set of separated points -- here, the outcomes of the experiment correspond to points along a continuous line, and a number is assigned to each point by some appropriate rule. A clear illustration is measuring the heights of students in a statistics class: if the shortest student measures 5.0 feet and the tallest measures 5.8 feet, then the height variable can, in principle, take on any value at all within that interval -- 5.1, 5.23, 5.457, or any other real number the measuring instrument is capable of registering -- with no natural gaps between one achievable value and the next. Because continuous random variables arise from measurement rather than counting, they show up whenever a problem asks 'how much' or 'how long' rather than 'how many.' This distinction matters immensely for everything that follows in statistics, because discrete and continuous random variables are eventually described using entirely different mathematical tools -- probability mass functions for the discrete case and probability density functions for the continuous case. Turning to how experiments involving randomness are actually simulated on paper or by computer, this relies on the concept of random numbers: a sequence of digits drawn from the set 0 through 9, chosen so that at every single position in the sequence, each of the ten digits has exactly the same probability, one-tenth, of appearing, completely independent of whatever digits came immediately before it. While physical devices such as dice, coins, or numbered slips drawn from a hat can generate short bursts of genuinely random digits, this quickly becomes impractical for producing the long sequences that real simulation work requires, which is why computers instead rely on what are called pseudo-random numbers. These numbers are termed 'pseudo' precisely because they are not, in the strictest sense, actually random at all -- they are produced by a completely deterministic

mathematical formula that will, given the exact same starting point, always reproduce the exact same sequence of numbers every single time it is run. Despite this underlying determinism, however, a well-designed pseudo-random number generator produces sequences that convincingly display all of the important statistical properties expected of true randomness, and the sequence will only begin visibly repeating itself after an extremely long stretch, or cycle, of generated digits -- for a properly designed generator this cycle can run into the tens of thousands of digits before any repetition becomes apparent. The specific recursive formula almost universally used for this purpose is x(n+1) equals the quantity a times x(n) plus b, all taken modulo m, where a, b, and m are constants chosen carefully in advance by whoever designs the generator, and x(1), referred to as the seed, is simply the starting integer that kicks the entire sequence into motion. The 'modulo m' operation appearing in this formula means that whenever the raw calculated result of a.x(n)+b happens to exceed m, that result is divided by m and only the remainder of that division is retained and carried forward as the next number in the generated sequence. Once a long sequence of these pseudo-random digits has been generated, it becomes the essential raw material for Monte Carlo simulation methods -- techniques that use random numbers to imitate, and thereby study, the behaviour of random experiments in situations where actually carrying out the real physical experiment would be genuinely impossible, prohibitively expensive, or simply take far too long to be practical, such as simulating decades of customer arrivals at a business in a matter of seconds on a computer rather than genuinely waiting for that much real time to pass.

Multiple Choice Questions (MCQs)

A random variable is best described as: (A) A fixed constant used in every experiment (B) A variable whose values depend on the outcomes of a random experiment (C) Only the outcomes of a coin toss (D) A type of sample space

Correct answer: (B) A variable whose values depend on the outcomes of a random experiment. A random variable assigns a numerical value to each outcome of a random experiment.

Random numbers are digits from the set {0,...,9} where each digit has probability: (A) 1 (B) 0.5 (C) 0.1 (D) 0

Correct answer: (C) 0.1. Each of the 10 digits has an equal probability of 1/10 = 0.1 of being selected at any position.

Pseudo-random numbers are called 'pseudo' because they are: (A) Truly random (B) Generated by a deterministic formula, not a truly random process (C) Always irrational numbers (D) Only usable once

Correct answer: (B) Generated by a deterministic formula, not a truly random process. Pseudo-random numbers come from a deterministic recursive formula, yet they mimic the statistical behaviour of true random digits.

The general pseudo-random number generator formula is: (A) x(n+1) = a + b.x(n) (B) x(n+1) = [a.x(n) + b] mod m (C) x(n+1) = x(n)/m (D) x(n+1) = a.m + b

Correct answer: (B) x(n+1) = [a.x(n) + b] mod m. The standard linear congruential formula is x(n+1) = [a.x(n) + b] mod m.

Simulation techniques that use random numbers are also known as: (A) Linear regression methods (B) Monte Carlo methods (C) Least squares methods (D) Index number methods

Correct answer: (B) Monte Carlo methods. Simulation using random numbers is widely known as the Monte Carlo method.

The number of heads obtained in 3 coin tosses is an example of a: (A) Continuous random variable (B) Discrete random variable (C) Constant (D) Sample space

Correct answer: (B) Discrete random variable. It takes only isolated whole-number values (0,1,2,3), making it a discrete random variable.

The height of students in a class, measured in feet, is an example of a: (A) Discrete random variable (B) Continuous random variable (C) Random number (D) Pseudo-random variable

Correct answer: (B) Continuous random variable. Height can take any value within an interval, making it a continuous random variable.

Discrete random variables typically arise from: (A) Measuring (B) Counting (C) Rounding (D) Sampling error

Correct answer: (B) Counting. Discrete random variables almost always arise in connection with counting.

Continuous random variables typically arise from: (A) Counting (B) Measurement (C) Guessing (D) Random digit tables only

Correct answer: (B) Measurement. Continuous random variables are typically obtained by measurement, taking any value in an interval.

In probability proportional to size (PPS), random numbers are assigned to classes based on: (A) Alphabetical order (B) Cumulative probabilities of each class (C) Random guessing (D) The size of the sample space only

Correct answer: (B) Cumulative probabilities of each class. PPS assigns a range of random numbers to each class proportional to its cumulative probability.

Quick Revision Summary

- Random variable: assigns a real number to each outcome of a random experiment; denoted X, Y, Z (values: x, y, z)
- Random numbers: digits 0-9, each with probability 0.1 at every position, independent of prior digits
- Pseudo-random generator: $x(n+1) = [a.x(n) + b] \bmod m$, deterministic but statistically random-like
- Random number tables read row-wise or column-wise; digit grouping (1,2,3...digit) depends on the problem
- Monte Carlo methods: simulate random experiments using random numbers when real experimentation is costly/impossible
- PPS method: cumulative probabilities define a range of random numbers assigned to each outcome class
- Discrete random variable: isolated values, arises from counting (e.g., number of heads)
- Continuous random variable: any value in an interval, arises from measuring (e.g., height, waiting time)

Exam Tips

- To classify a random variable quickly, ask: 'Am I counting or measuring?' Counting = discrete, measuring = continuous
- In LCG problems, always compute step by step: multiply, add b, then take mod m (divide and keep remainder)
- When reading random number tables, note whether the problem needs 1-digit, 2-digit, or 3-digit groups before you start reading
- For PPS problems, build the cumulative probability column first, then convert cumulative probabilities into number ranges (0 to 999 for 3-decimal probabilities)
- Remember: a random variable is a function/rule, not the outcome itself — always double-check what rule assigns values to outcomes

Chapter 8: Probability Distributions

Chapter 7 introduced the random variable as a rule that assigns a number to every outcome of a random experiment. But knowing the possible values isn't enough -- to actually use a random variable for prediction and decision-making, we need to know how likely each value (or range of values) is. That complete description -- every possible value paired with its probability -- is called a probability distribution.

This chapter builds the two core tools for describing probability distributions: the probability mass function (PMF) for discrete random variables and the probability density function (PDF) for continuous random variables. It also introduces the expected value (mean) and variance of a random variable, ways to graph a distribution, and the cumulative distribution function, which answers 'what is the probability that Y is at or below some value?' -- laying the groundwork for the specific distributions (binomial, hypergeometric) covered in Chapter 9.

Learning Objectives

- Define probability mass function (PMF) and probability density function (PDF)
- State and verify the two conditions a function must satisfy to be a valid PMF or PDF
- Construct probability distributions for discrete random variables from real experiments
- Calculate probabilities for continuous random variables as areas under a density curve
- Draw a probability histogram and a bar chart for a discrete probability distribution
- Calculate the expected value (mean) and variance of a discrete random variable
- Apply the properties of expectation to find E(constant), E(bY plus a), E(X+Y), and E(XY)
- Define and construct the cumulative distribution function (CDF) of a discrete random variable

Key Concepts

8.1 Introduction

Whenever we work with random experiments, we need to associate a numerical value with each outcome in order to study it -- this gives rise to two types of random variables: discrete and continuous. A discrete random variable almost always arises from counting, while a continuous random variable arises from measurement. For a discrete random variable, its probability distribution describes how much probability is placed on each possible value, with all these probabilities summing to 1; this is called its probability mass function. For a continuous random variable, we cannot talk about probability at a single point -- instead we talk about probability over an interval, with total probability again equal to 1; this is called its probability density function.

A probability distribution can be described either by a formula (function) or with a two-column table, similar to a frequency distribution -- one column lists the values (or intervals of values, for continuous variables) and the other lists the corresponding probabilities.

8.2 Probability Mass Function

The probability mass function (PMF) of a discrete random variable Y describes the values Y can take and the probability associated with each value, usually presented in a two-column table. For example, if three coins are tossed and Y denotes the number of heads, there are 8 equally likely outcomes: no head occurs once (TTT), one head occurs 3 times (HTT, THT, TTH), two heads occurs 3 times (HHT, HTH, THH), and three heads occurs once (HHH). This gives the probability function P(Y=0)=1/8, P(Y=1)=3/8, P(Y=2)=3/8, P(Y=3)=1/8.

8.3 Probability Density Function

The probability density function (PDF) of a continuous random variable Y is specified by a smooth curve such that the total area under the curve equals 1. The probability that Y falls within a particular interval is the area under the curve above that interval -- unlike the discrete case, we never ask for the probability at a single exact point (that probability is always zero for a continuous variable), only for probability over a range.

8.4 Discrete and Continuous Distributions, and Their Properties

A discrete uniform distribution arises when every possible value of a discrete random variable has the same probability -- for example, when a fair die is thrown, each face 1 through 6 has probability 1/6, giving f(y) = 1/6 for y=1,2,...,6 and 0 otherwise. A continuous uniform distribution is defined by f(y) = 1/(b-a) for a < y < b, meaning the probability is spread evenly across the interval from a to b; when a=0 and b=1, this simplifies to f(y)=1 for 0<y<1.

For a probability mass function P(y) of a discrete random variable Y to be valid, it must satisfy two conditions: first, 0 <= P(y) <= 1 for every possible value of Y (probabilities are always between 0 and 1); second, the sum of P(y) over all possible values of Y must equal exactly 1. For a probability density function f(y) of a continuous random variable Y to be valid, it must satisfy: first, f(y) >= 0 for all y; second, the total area under the curve, P(-infinity < Y < infinity), must equal exactly 1. Not every function defined over the values of a random variable automatically qualifies as a probability distribution -- both conditions must be checked and confirmed.

An important distinction: for discrete distributions, P(a <= Y <= b) is NOT equal to P(a < Y < b), because the endpoint probabilities P(Y=a) and P(Y=b) are real, nonzero quantities that must be included or excluded depending on the inequality used. For continuous distributions, however, these two probabilities ARE equal, because the probability at any single exact point is always zero (the 'area' at a single point has zero width).

8.5 Drawing the PMF and PDF

A discrete probability distribution can be drawn as a probability histogram: values of the random variable go on the x-axis, probabilities on the y-axis, and adjacent rectangles are drawn with width 1 (extending 0.5 units to each side of the value) and height equal to the probability at that point -- since the width is 1, the area of each rectangle equals its probability. Alternatively, a bar chart can be drawn with thin bars (rather than full-width rectangles) whose height still equals the probability of the corresponding value. When the values of a discrete random variable become very closely spaced, the probability histogram can be

approximated by a smooth curve, which is exactly the idea behind a continuous probability density function.

8.6 Expectation and Variance of a Discrete Random Variable

The mathematical expectation, or expected value, of a discrete random variable Y with probability function P(y) is defined as E(Y) = sum of y.P(y) over all values of Y. This expectation is simply the mean of the probability distribution -- E(Y) is an alternative notation for the population mean, mu. Similarly, E(Y-squared) = sum of y-squared.P(y) over all values of Y.

The variance of a random variable Y is defined as sigma-squared = E[(Y - mu)-squared] = sum of (Y-mu)-squared.P(y) over all Y. Through algebraic expansion, this simplifies to the more practical computing formula: Var(Y) = E(Y-squared) - [E(Y)]-squared -- exactly analogous to the shortcut variance formula used for raw data in Chapter 4.

Four key properties of expectation make calculations much easier: (i) if c is a constant, E(c) = c; (ii) if a and b are constants, E(bY plus-or-minus a) = b.E(Y) plus-or-minus a (setting a = -mu and b = 1 gives the useful result E(Y - mu) = 0); (iii) for two random variables X and Y, the expectation of their sum or difference is the sum or difference of their expectations, E(X plus-or-minus Y) = E(X) plus-or-minus E(Y); (iv) if X and Y are independent, the expectation of their product is the product of their expectations, E(XY) = E(X).E(Y).

8.9 Distribution Function (Cumulative Distribution Function)

Very often we want the probability that a random variable takes a value at or below some fixed point -- for example, the chance a student scores no more than 80%, or that 5 coin tosses produce no more than 3 heads. This probability, P(Y <= y), is called the distribution function, or cumulative distribution function (CDF), and is written F(y) = P(Y <= y). For a continuous random variable, F(y) is the integral of the density function from negative infinity up to y.

Every valid CDF satisfies four properties: (i) F(-infinity) = 0 (there's no probability of being below the smallest possible value); (ii) F(+infinity) = 1 (there's certainty of being at or below the largest possible value); (iii) F(y1) <= F(y2) whenever y1 <= y2 (the CDF is non-decreasing); (iv) F(y) is continuous at least from the right

at every point. For a discrete random variable, the CDF is a step function: it jumps up by P(Y=y) at each possible value of y and stays flat (constant) in between consecutive possible values.

Important Definitions

What is a probability distribution?: A complete description of a random variable's possible values together with their associated probabilities, presented as a formula or a two-column table.

What is a probability mass function (PMF)?: A function P(y) that gives the probability associated with each possible value of a discrete random variable Y, satisfying 0<=P(y)<=1 and the sum of all P(y) equal to 1.

What is a probability density function (PDF)?: A smooth curve f(y) for a continuous random variable such that f(y)>=0 everywhere and the total area under the curve equals 1; probability corresponds to area over an interval, not a value at a single point.

What is the expected value E(Y) of a discrete random variable?: The mean of its probability distribution, calculated as the sum of y.P(y) over all possible values of Y.

What is the variance of a random variable?: The expected value of the squared deviation from the mean, Var(Y) = E[(Y-mu)^2], which simplifies to E(Y^2) - [E(Y)]^2.

What is a discrete uniform distribution?: A distribution in which every possible value of a discrete random variable has the same probability, such as each face of a fair die having probability 1/6.

What is a cumulative distribution function (CDF)?: The function F(y) = P(Y<=y) giving the probability that a random variable takes a value at or below y; for discrete variables it is a step function.

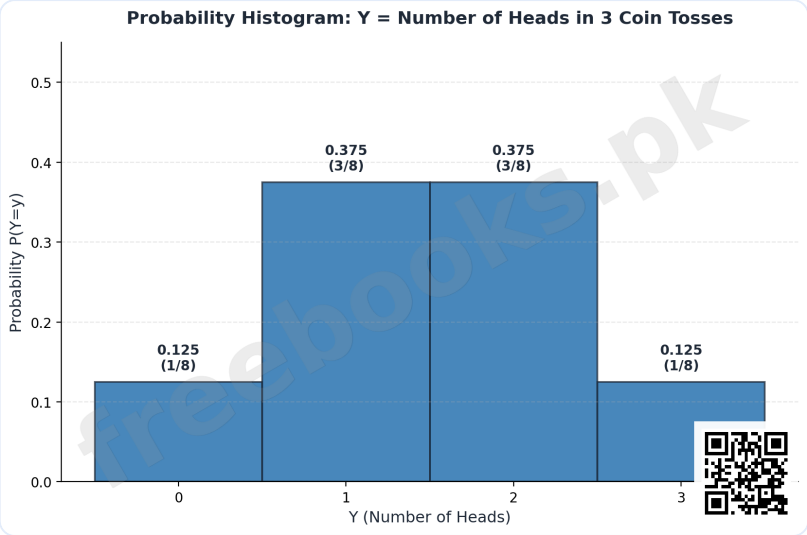
Why is P(Y=a) always zero for a continuous random variable?: Because probability for a continuous variable corresponds to area under the density curve, and the area over a single point (zero width) is always zero.

Key Facts and Relations

Topic	Key Fact / Relation
PMF validity conditions	0 <= P(y) <= 1 for each y, and sum of P(y) over all y = 1
PDF validity conditions	f(y) >= 0 for all y, and total area under curve = 1
Discrete uniform distribution (die)	f(y) = 1/6 for y = 1,2,...,6
Continuous uniform distribution	f(y) = 1/(b-a) for a < y < b
Expected value (discrete)	E(Y) = sum y.P(y)
Variance (discrete, computing formula)	Var(Y) = E(Y^2) - [E(Y)]^2
Expectation of a constant	E(c) = c
Linear transformation of expectation	E(bY + a) = b.E(Y) + a
Expectation of sum/difference	E(X +/- Y) = E(X) +/- E(Y)
Expectation of product (independent X, Y)	E(XY) = E(X).E(Y)
Cumulative distribution function	F(y) = P(Y <= y)
Trapezoidal area rule (continuous probability)	P(a<Y<b) = [f(a)+f(b)]/2 x (b-a) (for linear density functions)

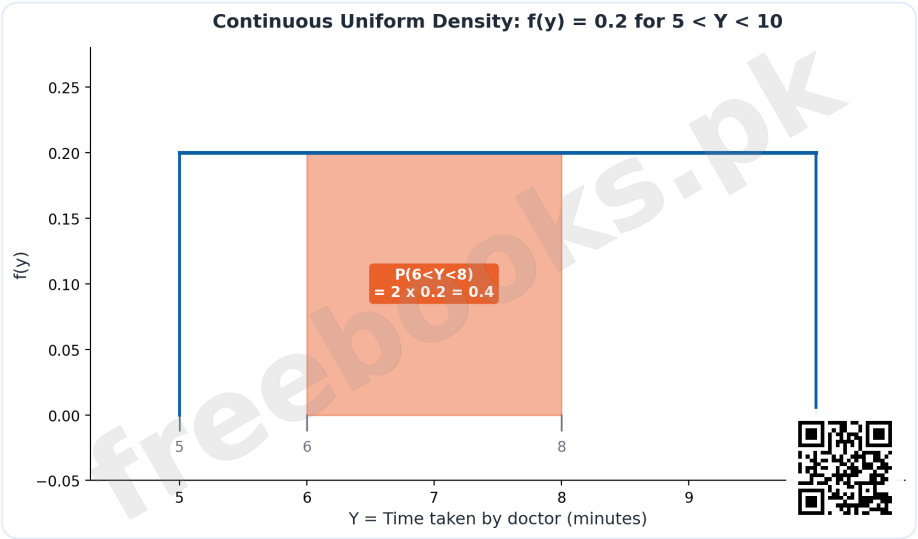
Diagrams

Probability Histogram: Number of Heads in 3 Coin Tosses: A probability histogram for the discrete random variable Y = number of heads in 3 coin tosses, with rectangles of width 1 at Y=0,1,2,3 and heights 1/8, 3/8, 3/8, 1/8 respectively, illustrating a valid probability mass function



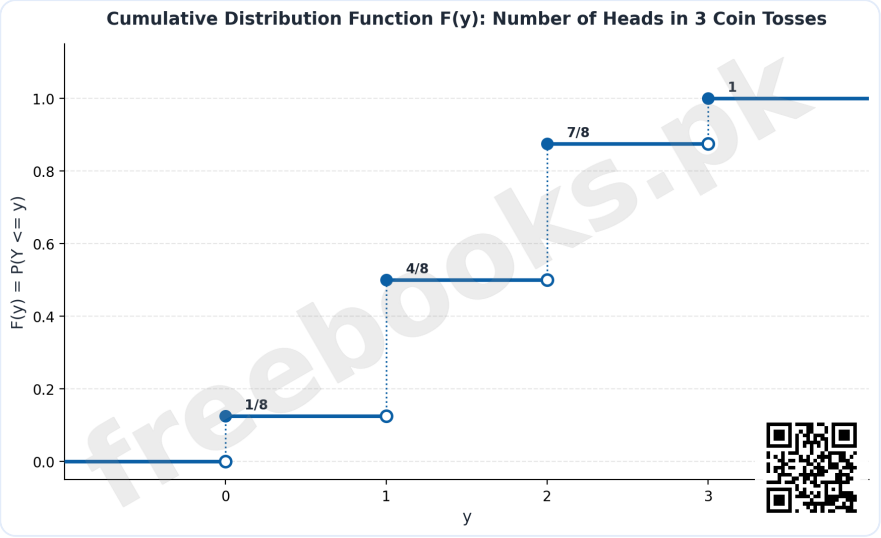
Probability Histogram: Number of Heads in 3 Coin Tosses

Continuous Uniform Density: Doctor's Waiting Time: The continuous uniform density function f(y)=0.2 for 5<y<10 minutes (doctor's time with a patient), with the area between y=6 and y=8 shaded to show P(6<Y<8)=0.4, illustrating that continuous probabilities are areas under the curve



Continuous Uniform Density: Doctor's Waiting Time

Cumulative Distribution Function: 3 Coin Tosses: The step-function cumulative distribution function F(y)=P(Y<=y) for the number of heads in 3 coin tosses, jumping from 0 to 1/8 to 4/8 to 7/8 to 1 at Y=0,1,2,3, illustrating the CDF of a discrete random variable



Cumulative Distribution Function: 3 Coin Tosses

Short Questions & Answers

What two conditions must a function satisfy to be a valid probability mass function?

It must give a probability between 0 and 1 for each possible value of the random variable, and the sum of all these probabilities over every possible value must equal exactly 1.

Why can't we ask for the exact probability that a continuous random variable equals a specific value?

Because probability for continuous variables is measured as area under the density curve, and the area above a single point has zero width, so that probability is always exactly zero.

What is the difference between a probability histogram and a bar chart for a discrete distribution?

In a probability histogram, adjacent rectangles of width 1 are drawn so that area equals probability; in a bar chart, thin bars are drawn whose height alone (not area) represents the probability.

State the formula that relates variance to E(Y) and E(Y^2).

Var(Y) = E(Y^2) - [E(Y)]^2, the computing (shortcut) formula for variance.

If E(X) = 5, what is E(3X + 2)?

Using E(bY+a) = b.E(Y)+a: E(3X+2) = 3(5)+2 = 17.

What does the cumulative distribution function F(y) represent?

F(y) = P(Y<=y), the probability that the random variable takes a value at or below y.

Why is a discrete random variable's CDF a step function rather than a smooth curve?

Because probability only accumulates at the isolated points the discrete variable can take, so F(y) jumps up at each possible value and stays flat in between.

Long Questions & Answers

Explain the concept of a probability mass function and a probability density function, including the conditions each must satisfy, and explain why the treatment of probability differs between discrete and continuous random variables.

Once a random variable has been defined for a given experiment, the natural next question is not just what values it can take, but how likely each of those values actually is -- and answering that question completely, for every possible value at once, is exactly what a probability distribution provides. For a discrete random variable, this complete description is called the probability mass function, commonly abbreviated as PMF, and it works by pairing up each possible value of the random variable with its own specific probability, usually organized into a simple two-column table with the values in one column and their corresponding probabilities in the other. A textbook illustration of building a PMF from scratch involves tossing three fair coins and letting Y represent the number of heads observed. Working carefully through the eight equally likely outcomes of this experiment -- TTT, HTT, THT, TTH, HHT, HTH, THH, and HHH -- shows that exactly one outcome produces zero heads, exactly three outcomes produce one head, exactly three outcomes produce two heads, and exactly one outcome produces three heads. Applying the classical definition of probability, this translates directly into P(Y=0)=1/8, P(Y=1)=3/8, P(Y=2)=3/8, and P(Y=3)=1/8, and together these four probability statements constitute the complete probability mass function for this experiment. Not every mathematical function that happens to be defined over a random variable's possible values, however, is automatically entitled to be called a valid probability mass function -- two strict

conditions must always hold before a function earns that label. First, every single probability the function produces must fall somewhere between 0 and 1 inclusive, since a negative probability or a probability exceeding 1 has no meaningful interpretation in the real world. Second, and equally essential, when all of the probabilities across every single possible value of the random variable are added together, that total sum must come out to exactly 1, reflecting the simple logical fact that the random variable is guaranteed to take on some value among all of its listed possibilities. For continuous random variables, the situation requires a fundamentally different mathematical tool, called the probability density function, or PDF, precisely because a continuous random variable can take on infinitely many possible values within any given interval, no matter how small that interval is drawn. Because of this, it no longer makes any sense to ask for the probability that the random variable equals one single, exact value -- that probability is mathematically always zero, since a single point on a continuous interval has literally zero width and therefore zero area beneath the curve directly above it. Instead, a probability density function is visualized as a smooth curve drawn over the entire range of values the random variable can take, and probability is redefined entirely in terms of area: specifically, the probability that the random variable falls somewhere within a given interval is precisely equal to the area enclosed underneath the density curve and directly above that particular interval on the horizontal axis. Just as with the discrete case, an arbitrary mathematical function cannot automatically be treated as a legitimate probability density function without first satisfying its own pair of validity conditions, which closely mirror the discrete conditions in spirit even though they are phrased somewhat differently. First, the function must never dip below zero at any point across its entire domain, since a curve with any negative height would imply an impossible negative area, and therefore an impossible negative probability, over whatever interval contained that dip. Second, when the total area beneath the entire curve is calculated across the complete range of values the random variable is capable of taking, that grand total area must work out to exactly 1, which is simply the continuous-case restatement of the same underlying certainty that the random variable must take on some value from within its full range of possibilities. This fundamental difference in how probability is defined and measured -- discrete probabilities read directly off a table of individual values, versus continuous probabilities computed as areas beneath a

curve over a range -- also explains a subtle but important distinction that trips up many students encountering this material for the first time: for a discrete random variable, the probability P(a <= Y <= b) is genuinely and numerically different from the probability P(a < Y < b), because the individual endpoint probabilities P(Y=a) and P(Y=b) are real, calculable, nonzero quantities that must be deliberately included when the inequality is written with 'or equal to' and just as deliberately excluded when it is written with strict inequality symbols alone. For a continuous random variable, by beautiful contrast, these two probabilities always turn out to be exactly equal to one another, and the reason traces directly back to everything already established above: since the probability of a continuous random variable landing on any single exact point is always precisely zero regardless of which particular point is chosen, including or excluding those two single endpoints from the calculation literally changes nothing about the total area being measured, and therefore changes nothing about the final probability that gets computed.

Explain the concept of expectation (mean) and variance for a discrete random variable, state the properties of expectation, and explain the purpose and construction of the cumulative distribution function.

Once a complete probability distribution has been established for a discrete random variable, one of the most immediately useful things we can do with it is summarize the entire distribution using just two single numbers, in exactly the same spirit as computing a mean and a variance for an ordinary set of raw data back in earlier chapters -- except that here, instead of averaging over data points that were actually observed, we are averaging over every theoretically possible value the random variable could produce, weighted according to how likely each value actually is. This theoretical average is called the mathematical expectation, or expected value, of the random variable, conventionally written E(Y), and it is calculated by taking each possible value y that the random variable Y can assume, multiplying it by its own corresponding probability P(y), and then summing all of these individual products together across every single possible value in the distribution. What makes this expectation genuinely important, rather than merely an abstract computational exercise, is that E(Y) turns out to be nothing more than an alternative, more formal notation for exactly the same

population mean, conventionally denoted by the Greek letter mu, that appeared constantly throughout the earlier chapters covering measures of location -- the crucial difference is only that here the mean is being computed directly from a theoretical probability distribution rather than estimated from a finite batch of collected sample data. Building directly on this same underlying logic, the variance of a discrete random variable is formally defined as the expected value of the squared deviation of the random variable away from its own mean, written out fully as sigma-squared equals E of the quantity (Y minus mu) squared, which unpacks computationally into the sum, taken over every possible value of Y, of (Y minus mu) squared multiplied by P(y). While this defining formula is conceptually the clearest way to understand exactly what variance measures -- namely, the average squared spread of the distribution around its own center -- it is rarely the most convenient version of the formula to actually use for computation by hand. Through straightforward algebraic expansion and simplification of the squared term, this defining formula reduces neatly down to the far more practical computing formula Var(Y) equals E(Y-squared) minus the quantity E(Y) squared, which is structurally identical in spirit to the computational shortcut formula for variance that was already introduced for raw, ungrouped data all the way back in the dispersion chapter, just now expressed in terms of expectations rather than raw sums. Working with expectations directly, rather than being forced to recompute an entire distribution completely from scratch every single time some new random variable is introduced, becomes dramatically easier once four essential and very general properties of expectation are firmly understood and readily available for use. The first and most basic property states that the expectation of any constant is simply that same constant itself, written E(c) equals c, which makes intuitive sense since a constant, by its very definition, never actually varies at all and therefore trivially averages out to itself no matter how many times it might hypothetically be observed. The second property, arguably the single most useful and most frequently applied of the four in practice, states that for any two fixed constants a and b, the expectation of the linear transformation bY plus or minus a works out to exactly b times E(Y), plus or minus that same constant a -- in other words, constants can always be pulled directly outside of the expectation operation according to entirely ordinary and predictable algebraic rules, without needing to redo the underlying summation from scratch. The third property extends this same convenient logic to

combinations of two entirely separate random variables X and Y, stating that the expectation of their sum, or equally of their difference, is simply equal to the sum, or difference, of each variable's own individual expectation calculated completely separately -- a property that, notably, holds true universally regardless of whether X and Y happen to be independent of one another or not. The fourth and final property, however, does specifically require independence between the two variables involved: when X and Y are genuinely independent random variables, the expectation of their product, E(XY), can be calculated simply as the ordinary product of each variable's own separate individual expectation, E(X) multiplied by E(Y), which importantly does not hold true in general whenever the two variables in question happen to be dependent on one another in any way. Turning now to an entirely different but closely related question that a completed probability distribution is perfectly positioned to answer, we frequently find ourselves needing to know not the probability that a random variable takes on one single specific value, but rather the accumulated probability that it takes on some value at or below a given fixed threshold -- questions phrased exactly like 'what are the chances a particular student scores no more than 80 percent' or 'what is the probability that five tosses of a coin produce no more than three heads total.' This particular type of accumulated, running-total probability is formally captured by what is called the distribution function, more commonly referred to as the cumulative distribution function or simply the CDF for short, written using the notation F(y) and formally defined as F(y) equals P(Y is less than or equal to y). For a discrete random variable specifically, this cumulative distribution function takes on the very distinctive visual shape of a step function when it is graphed out: the function value remains perfectly flat and completely unchanged across any stretch of the number line lying strictly between two consecutive possible values of the random variable, and then it abruptly jumps straight upward by an amount precisely equal to that particular value's own individual probability P(y) at the exact instant the graph passes over each new possible value the random variable is capable of taking. Every legitimate cumulative distribution function, without exception, must satisfy four defining properties that together fully characterize its overall shape and behavior: it must approach exactly 0 as y approaches negative infinity, since there is certainly zero accumulated probability of landing at or below any value lying below the very smallest value the random variable could ever possibly take;

it must likewise approach exactly 1 as y approaches positive infinity, since by that point every single possible value the random variable could ever take has already been fully accounted for and included in the running cumulative total; the function must be non-decreasing throughout its entire domain, meaning F(y1) is always less than or equal to F(y2) whenever y1 is less than or equal to y2, since accumulated probability can only ever increase or stay flat as more possible values get progressively included, and can certainly never decrease as we move further along the number line; and finally, the function must remain continuous at the very least from the right-hand side at every single point along its domain, which is simply the more technical and mathematically rigorous way of describing that characteristic staircase-like step shape already described above.

Multiple Choice Questions (MCQs)

The probability mass function (PMF) applies to: (A) Continuous random variables only (B) Discrete random variables (C) Constants only (D) Sample spaces only

Correct answer: (B) Discrete random variables. The PMF describes the probability of each specific value of a discrete random variable.

For a valid PMF, the sum of all probabilities P(y) must equal: (A) 0 (B) 0.5 (C) 1 (D) Any positive number

Correct answer: (C) 1. One of the two defining conditions of a valid PMF is that all probabilities sum to exactly 1.

For a continuous random variable, P(Y = a) for any specific value a is: (A) 1 (B) 0.5 (C) Undefined (D) Always 0

Correct answer: (D) Always 0. Probability at a single point has zero width under the density curve, so it is always exactly 0 for continuous variables.

The area under a valid probability density function (PDF) must equal: (A) 0 (B) 1 (C) Infinity (D) The mean

Correct answer: (B) 1. A defining condition of a valid PDF is that the total area under the curve equals 1.

The expected value E(Y) of a discrete random variable is calculated as: (A) Sum of Y only (B) Sum of y.P(y) over all values (C) Sum of P(y) only (D) Maximum value of Y

Correct answer: (B) Sum of y.P(y) over all values. E(Y) = sum of y.P(y), the probability-weighted average of all possible values.

The computing (shortcut) formula for variance of a discrete random variable is: (A) Var(Y) = E(Y) - E(Y^2) (B) Var(Y) = E(Y^2) - [E(Y)]^2 (C) Var(Y) = [E(Y)]^2 (D) Var(Y) = E(Y)^2 + E(Y^2)

Correct answer: (B) Var(Y) = E(Y^2) - [E(Y)]^2. Var(Y) = E(Y^2) - [E(Y)]^2 is the standard computing formula, derived by expanding E[(Y-mu)^2].

If E(X) = 4, then E(2X + 3) equals: (A) 7 (B) 8 (C) 11 (D) 14

Correct answer: (C) 11. Using E(bY+a) = b.E(Y)+a: E(2X+3) = 2(4)+3 = 11.

If X and Y are independent random variables, E(XY) equals: (A) E(X) + E(Y) (B) E(X) - E(Y) (C) E(X).E(Y) (D) E(X)/E(Y)

Correct answer: (C) E(X).E(Y). For independent random variables, the expectation of the product equals the product of the expectations.

The cumulative distribution function F(y) is defined as: (A) P(Y = y) (B) P(Y >= y) (C) P(Y <= y) (D) P(Y < 0)

Correct answer: (C) P(Y <= y). F(y) = P(Y <= y), the probability of the random variable being at or below y.

The graph of the CDF of a discrete random variable looks like a: (A) Smooth curve (B) Straight line (C) Step function (D) Circle

Correct answer: (C) Step function. Since probability accumulates only at isolated points for a discrete random variable, its CDF forms a step function.

Quick Revision Summary

- PMF (discrete): P(y), with 0<=P(y)<=1 for each y and sum of P(y)=1
- PDF (continuous): f(y)>=0 everywhere, total area under curve = 1; probability = area over an interval
- Discrete: P(a<=Y<=b) != P(a<Y<b). Continuous: these two ARE equal (P(Y=a)=0 always)
- Probability histogram: rectangles of width 1, height = probability, area = probability

- E(Y) = sum y.P(y) (the mean); Var(Y) = E(Y^2) - [E(Y)]^2
- Properties: E(c)=c; E(bY+a)=bE(Y)+a; E(X+/-Y)=E(X)+/-E(Y); E(XY)=E(X)E(Y) if independent
- CDF: F(y) = P(Y<=y); F(-inf)=0, F(+inf)=1, non-decreasing, step function for discrete variables

Exam Tips

- Before treating any function as a probability distribution, always check BOTH validity conditions: non-negativity and total probability/area = 1
- For continuous PDF area problems with a linear f(x), use the trapezoid formula: [f(a)+f(b)]/2 x base, rather than full integration when possible
- Remember discrete P(a<=Y<=b) includes both endpoints separately — don't confuse it with the strict-inequality version
- When computing variance, build a table with columns y, P(y), yP(y), $y^2P(y)$ to avoid arithmetic slips
- To sketch a discrete CDF, list cumulative sums of P(y) at each value, then draw a step that jumps at each value and stays flat in between

Chapter 9: Binomial and Hypergeometric Probability Distribution

Chapter 8 built the general tools for describing any probability distribution -- the PMF, PDF, expectation, and variance. This chapter applies those tools to two of the most important and widely used discrete probability distributions in statistics: the binomial distribution, which models the number of successes in a fixed number of independent trials (like counting heads in repeated coin tosses, or defective items when sampling with replacement), and the hypergeometric distribution, which models the same kind of counting problem but when sampling is done without replacement from a finite population, making successive trials dependent rather than independent.

Both distributions answer the same basic question -- 'how many successes out of n trials?' -- but they apply to different real-world sampling situations, and recognizing which one fits a given problem (replacement vs. no replacement, fixed vs. changing probability of success) is one of the most practical skills this chapter builds.

Learning Objectives

- Define a Bernoulli trial and list the four properties of a binomial experiment
- Apply the binomial probability formula P(X=x) = C(n,x) p^x q^(n-x) to compute exact, cumulative, and range probabilities
- Identify whether a binomial distribution is symmetric, positively skewed, or negatively skewed based on p
- Calculate expected (binomial) frequencies for a fixed number of repeated experiments
- Derive and apply the mean (np) and variance (npq) of the binomial distribution
- Define a hypergeometric experiment and list its four defining properties
- Apply the hypergeometric probability formula and calculate its mean and variance

- Distinguish when to use the binomial distribution versus the hypergeometric distribution for a given sampling problem

Key Concepts

9.1 Introduction: Bernoulli Trials and the Binomial Experiment

In experiments like tossing a coin or repeatedly drawing a card, each individual repetition is called a trial, and the result of each trial is classified as either a success or a failure. The probability of success is denoted p and the probability of failure is denoted q, where q = 1-p (so p+q=1). If an experiment is repeated n times, the number of successes is denoted x and the number of failures is n-x.

A trial with exactly two possible outcomes -- success or failure -- is called a Bernoulli trial; for each Bernoulli trial the probability of success stays the same, and successive trials are independent. An experiment made up of repeated Bernoulli trials, where the probability of success stays constant from trial to trial, is called a binomial experiment. A binomial experiment has four defining properties: (i) each trial results in an outcome classified as success or failure; (ii) the probability of success remains constant from one trial to the next; (iii) the repeated trials are independent of each other; (iv) the experiment is repeated a fixed number of times.

9.2 Binomial Probability Distribution

For n independent trials, each with probability of success p and probability of failure q (q+p=1), the probability of exactly x successes is P(X=x) = C(n,x).p^x.q^(n-x), for x=0,1,2,...,n. The random variable X is called the binomial variable, its distribution is the binomial distribution, and n and p are the two parameters of the distribution, often written b(x,n,p). These probabilities are exactly the individual terms produced by expanding the binomial expression (q+p)^n -- which is why the distribution is named 'binomial' and why the coefficients C(n,x) are called binomial coefficients.

The shape of the binomial distribution depends entirely on p: if p=q=1/2, the distribution is perfectly symmetrical. If p is not equal to q, the distribution is skewed -- specifically, if p is greater than 1/2 the distribution is negatively

skewed (tail toward lower values), and if p is less than 1/2 the distribution is positively skewed (tail toward higher values).

9.2.1 Binomial Frequency Distribution

When a binomial probability distribution is multiplied by the number of experiments N (repetitions of the whole n-trial experiment), the result is called the binomial frequency distribution. The expected frequency of exactly x successes across N experiments is f(X=x) = N.P(X=x) = N.C(n,x).q^(n-x).p^x -- this is simply the theoretical probability scaled up to a real count, useful for predicting how many out of N groups (families, batches, samples) will show a particular number of successes.

9.2.2 Mean and Variance of the Binomial Distribution

Through algebraic derivation starting from E(X) = sum of x.P(x) and expanding using the binomial theorem, the mean of the binomial distribution simplifies elegantly to E(X) = np -- the number of trials multiplied by the probability of success on each trial, which matches the intuitive idea that if you flip a coin n times with probability p of heads, you'd 'expect' about np heads on average.

Following the same style of derivation for E(X^2), and using the computing formula Var(X) = E(X^2) - [E(X)]^2, the variance of the binomial distribution simplifies to Var(X) = npq, and consequently the standard deviation is sigma = square-root(npq). These two compact formulas, mean = np and variance = npq, are among the most frequently used results in this chapter and let you find the mean and spread of a binomial distribution instantly, without building a full probability table.

9.3 Hypergeometric Distribution

When successive trials are conducted without replacement, they are no longer independent -- the probability of success changes from one trial to the next depending on what was already drawn. An experiment where a random sample is chosen without replacement from a finite population is called a hypergeometric experiment. It has four defining properties: (i) the experiment is repeated a fixed number of times; (ii) the successive trials are dependent; (iii) the probability of success varies from trial to trial (it is not fixed); (iv) each outcome is still

classified as success or failure. The random variable representing the number of successes is called the hypergeometric variable, and its distribution is the hypergeometric distribution.

Suppose there are N total items, of which k are classified as successes and (N-k) as failures, and n items are selected at random without replacement (n <= N). The probability of getting exactly x successes (and n-x failures) is P(X=x) = [C(k,x).C(N-k,n-x)] / C(N,n), for x=0,1,2,...,n. The hypergeometric distribution has three parameters: n, k, and N. Its mean is nk/N, and its variance is (nk/N).(1 - k/N).((N-n)/(N-1)) -- notice the extra factor (N-n)/(N-1), called the finite population correction factor, which distinguishes it from the simpler binomial variance and accounts for the fact that sampling is without replacement from a limited population.

Important Definitions

What is a Bernoulli trial?: A trial with exactly two possible outcomes -- success and failure -- where the probability of success stays constant and successive trials are independent.

What is a binomial experiment?: An experiment made up of a fixed number of independent Bernoulli trials, each with the same probability of success p.

What is the binomial probability formula?: P(X=x) = C(n,x).p^x.q^(n-x), giving the probability of exactly x successes in n independent trials.

What are the two parameters of the binomial distribution?: n (the number of trials) and p (the probability of success on each trial).

What is the mean and variance of the binomial distribution?: Mean = np; Variance = npq (where q = 1-p); Standard deviation = square-root(npq).

What is a hypergeometric experiment?: An experiment in which a random sample is drawn without replacement from a finite population, making successive trials dependent and the probability of success variable.

What is the hypergeometric probability formula?: P(X=x) = [C(k,x).C(N-k,n-x)] / C(N,n), where N is the population size, k is the number of successes in the population, and n is the sample size.

What is the finite population correction factor?: The term (N-n)/(N-1) appearing in the hypergeometric variance formula, which adjusts for sampling without replacement from a finite population of size N.

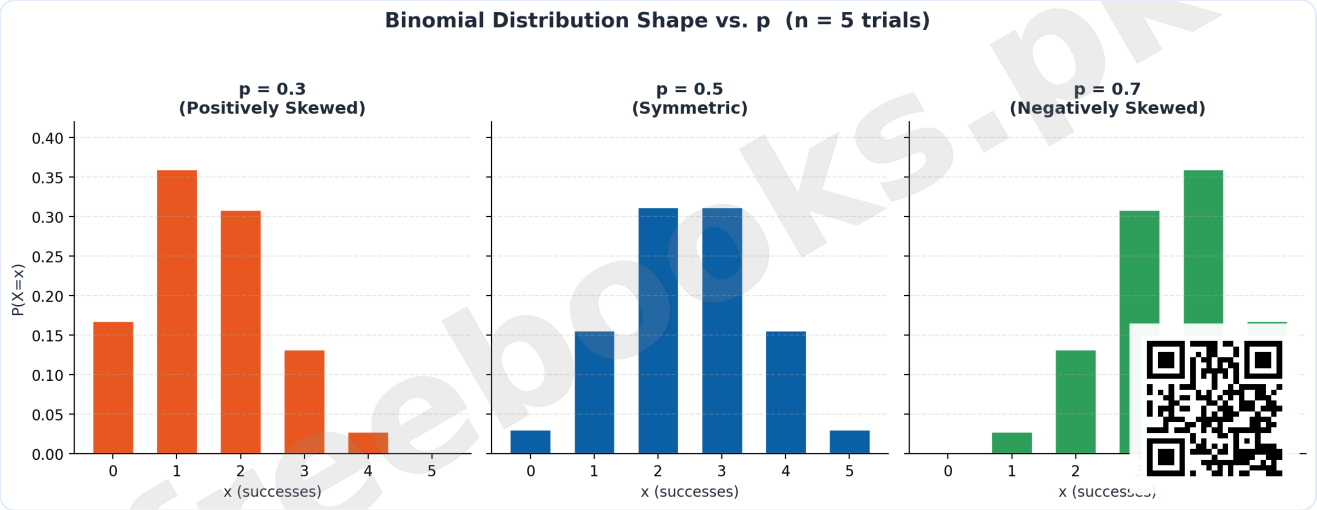
How do you decide whether to use the binomial or hypergeometric distribution?: Use the binomial distribution when sampling is with replacement (or from an effectively infinite population) so trials are independent and p is constant; use the hypergeometric distribution when sampling is without replacement from a small, finite population so trials are dependent and the probability of success changes.

Key Facts and Relations

Topic	Key Fact / Relation
Binomial probability	P(X=x) = C(n,x) p^x q^(n-x), x = 0,1,...,n
Binomial frequency distribution	f(X=x) = N.P(X=x) = N.C(n,x) q^(n-x) p^x
Binomial mean	E(X) = np
Binomial variance	Var(X) = npq
Binomial standard deviation	sigma = sqrt(npq)
Binomial distribution shape	p=q=0.5: symmetric | p>0.5: negatively skewed | p<0.5: positively skewed
Hypergeometric probability	P(X=x) = [C(k,x).C(N-k,n-x)] / C(N,n)
Hypergeometric mean	Mean = nk/N
Hypergeometric variance	Var(X) = (nk/N)(1 - k/N)((N-n)/(N-1))

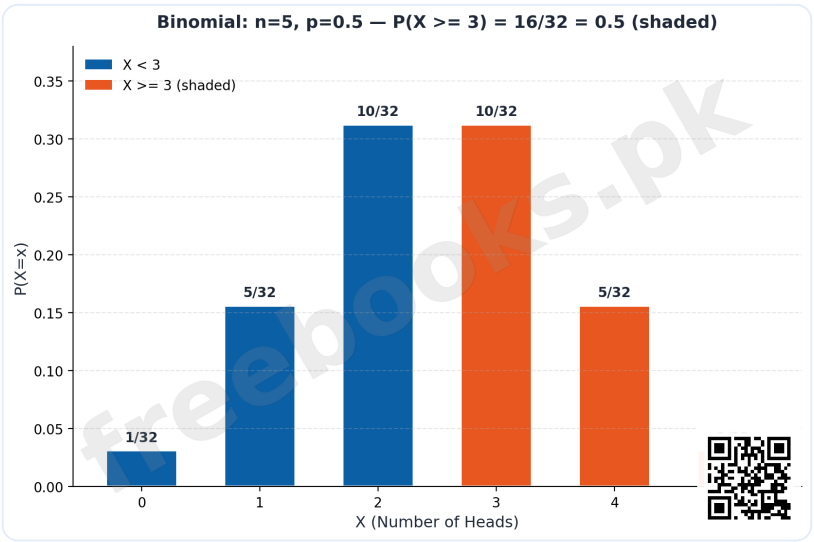
Diagrams

Binomial Distribution Shape vs. p (n=5): Three side-by-side binomial probability bar charts for n=5 with p=0.3, p=0.5, and p=0.7, illustrating positively skewed, symmetric, and negatively skewed binomial distributions respectively



Binomial Distribution Shape vs. p (n=5)

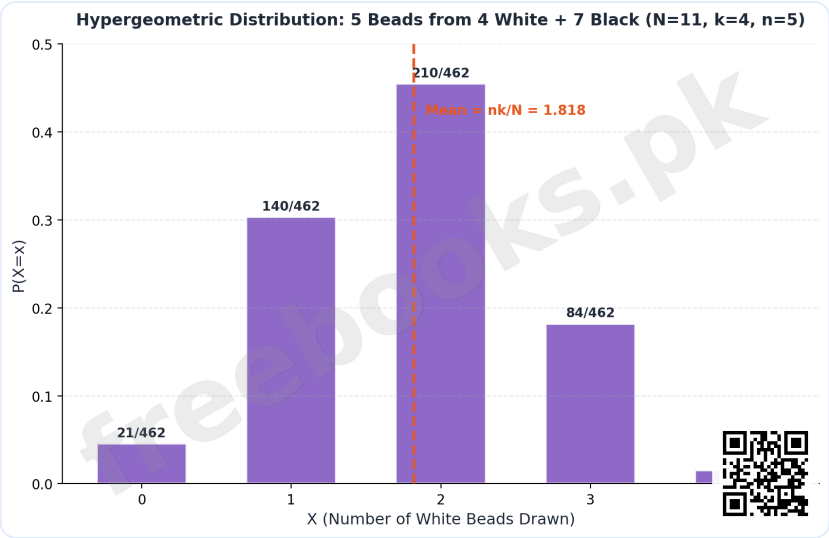
Binomial Probability: At Least 3 Heads in 5 Coin Tosses: A binomial bar chart for n=5, p=0.5 (5 fair coin tosses) with the bars for X=3, X=4, X=5 shaded to show P(X>=3) = 16/32 = 0.5, matching Example 9.2 from the textbook



Binomial Probability: At Least 3 Heads in 5 Coin Tosses

Hypergeometric Probability Distribution: White Beads Example: A probability bar chart for the hypergeometric distribution of white beads drawn (5

beads from a bowl of 4 white and 7 black, N=11, k=4, n=5), with the mean (1.818) marked, matching Example 9.9 from the textbook



Hypergeometric Probability Distribution: White Beads Example

Short Questions & Answers

List the four properties of a binomial experiment.

Each trial results in success or failure; the probability of success stays constant across trials; the trials are independent; the experiment is repeated a fixed number of times.

Why is the binomial distribution symmetric when p = 0.5?

Because when p equals q (both 0.5), the probability of getting x successes equals the probability of getting (n-x) successes, making the distribution perfectly mirror-symmetric around its center.

State the mean and variance of a binomial distribution with n=20 trials and p=0.4.

Mean = np = 20(0.4) = 8; Variance = npq = 20(0.4)(0.6) = 4.8.

Why are successive trials dependent in a hypergeometric experiment but independent in a binomial experiment?

Because a hypergeometric experiment samples without replacement, so removing an item changes the composition of what remains and shifts the probability of success on the next draw; a binomial experiment samples with replacement (or from an effectively infinite population), so the probability of success never changes.

What does the finite population correction factor (N-n)/(N-1) account for in the hypergeometric variance?

It adjusts the variance downward to reflect that sampling without replacement from a finite population produces less variability than sampling with replacement would.

How many parameters does the binomial distribution have, and what are they? How about the hypergeometric?

Binomial: two parameters, n and p. Hypergeometric: three parameters, n, k, and N.

Long Questions & Answers

Explain the binomial probability distribution: define a binomial experiment, state its formula, and explain how its mean and variance are used, including how the shape of the distribution depends on p.

The binomial distribution is one of the most widely applied discrete probability distributions in all of statistics, precisely because it models an extremely common real-world situation: repeating the same simple yes-or-no experiment a fixed number of times under conditions that stay constant throughout. To understand why the binomial distribution takes the particular mathematical form it does, it helps to start from its most basic building block, the Bernoulli trial, which is simply any single trial with exactly two possible outcomes, conventionally labeled success and failure, where the probability of success, denoted p, stays fixed and where each trial has no influence whatsoever on any other trial. A binomial experiment is then nothing more than a fixed number, n, of these identical, independent Bernoulli trials strung together, and for a sequence of trials to genuinely qualify as a binomial experiment, four conditions must all hold simultaneously: every individual trial must produce an outcome that can be cleanly classified as either a success or a failure; the probability of success, p, must remain exactly the same from the very first trial all the way through to the very last trial; every trial must be completely independent of every other trial, meaning the outcome of one trial provides absolutely no information about the outcome of any other; and the total number of trials, n, must be fixed and known in advance before the experiment begins, rather than

being allowed to vary. Given these conditions, the probability of observing exactly x successes across all n trials is given by the binomial probability formula, P(X=x) equals C(n,x) multiplied by p raised to the power x, multiplied by q raised to the power (n-x), where q simply represents the probability of failure and is always calculated as 1 minus p. The combinatorial term C(n,x) appearing at the front of this formula exists for a very specific and important reason: it counts the number of distinct ways that exactly x successes could be arranged among the n total trials, since a run of x successes could occur in many different orders across the sequence of trials, and every one of those equally likely orderings needs to be included in the total probability. This entire formula, and indeed the very name 'binomial' distribution itself, traces directly back to the algebraic expansion of the binomial expression (q+p) raised to the power n, since expanding this expression using the binomial theorem produces exactly these same terms, C(n,x).q^(n-x).p^x, as its individual components -- which is also precisely why the combinatorial coefficients C(n,x) are traditionally referred to as binomial coefficients. Beyond simply calculating the probability of one specific exact number of successes, the binomial formula also directly enables answering more complex 'at least,' 'at most,' or 'between' style questions, simply by calculating the individual probabilities for every relevant value of x and adding them together, since these composite events are always mutually exclusive when x is restricted to whole numbers. Once a specific binomial distribution has been fully defined by choosing its two parameters, n and p, two remarkably clean summary formulas become available without needing to build out an entire probability table from scratch. Through a somewhat lengthy but ultimately elegant algebraic derivation that expands E(X) as a weighted sum and repeatedly applies the underlying binomial theorem, the mean of any binomial distribution simplifies down to the strikingly simple result E(X) equals n times p -- a result that lines up perfectly with basic intuition, since if a fair coin with a 50% chance of heads is flipped 10 times, we would naturally expect around 10 times 0.5, or 5 heads, on average across many repetitions of that same 10-flip experiment. Following essentially the same style of derivation for E(X-squared), and then substituting the results into the standard computing formula Var(X) equals E(X-squared) minus [E(X)] squared, the variance of the binomial distribution likewise reduces down to the compact and equally memorable result Var(X) equals n.p.q, with the standard deviation then simply following as the

square root of n.p.q. Beyond their obvious practical usefulness for quickly summarizing a binomial distribution's center and spread without laboriously constructing a full probability table, the parameter p specifically also governs the overall visual shape of the distribution whenever it gets plotted out as a bar chart or histogram. Whenever p happens to equal q, which necessarily means both must equal exactly 0.5, the resulting binomial distribution comes out perfectly symmetrical, since under these particular conditions the probability of observing x successes always exactly equals the probability of observing (n minus x) successes, causing the entire distribution to mirror itself precisely around its own central peak. Whenever p and q are not equal to each other, however, the distribution instead becomes skewed in one particular direction or the other, and the specific direction of that skew depends entirely and predictably on whether p happens to be larger or smaller than 0.5: specifically, whenever p is greater than 0.5, meaning success is inherently the comparatively more likely outcome on any single individual trial, the resulting distribution becomes negatively skewed, with its longer statistical tail stretching out toward the lower, smaller possible values of x; conversely, whenever p is less than 0.5, meaning success is instead the comparatively less likely outcome on any individual trial, the distribution becomes positively skewed instead, with its longer tail now stretching out toward the higher, larger possible values of x.

Explain the hypergeometric distribution, including how it differs from the binomial distribution, state its probability formula and its mean and variance, and explain when each distribution should be used.

While the binomial distribution elegantly handles situations where repeated trials are genuinely independent of one another, an enormous number of real-world sampling situations instead involve drawing items one after another from some fixed, limited collection without ever putting each drawn item back before drawing the next one -- picking playing cards from a single deck without reshuffling them back in, selecting members for a committee from a fixed pool of candidates, or pulling components from a finite batch to test for defects, are all everyday examples of exactly this kind of situation. This particular style of sampling breaks one of the four essential conditions required for a genuinely binomial experiment, namely the strict requirement that the probability of success must remain perfectly constant across every single trial, because as

soon as one item has been removed from a finite population and is not returned before the next draw, the remaining composition of that population has necessarily and permanently changed, which in turn shifts the probability of success on whatever draw comes immediately afterward. An experiment specifically built around drawing a genuinely random sample without replacement from some finite population is formally called a hypergeometric experiment, and like the binomial experiment discussed earlier, it too is fully characterized by exactly four defining properties, though these four properties differ from the binomial case in several crucially important respects. First, exactly as with the binomial experiment, the hypergeometric experiment must still be repeated some fixed, predetermined number of times, which is conventionally denoted n. Second, and here is where the two experiment types begin to diverge sharply, the successive trials of a hypergeometric experiment are inherently dependent on one another rather than independent, precisely because each individual draw permanently and irreversibly alters what specifically remains available to be drawn on every subsequent trial. Third, following directly and logically from that same dependency between trials, the probability of success is explicitly allowed to vary meaningfully from one trial to the very next trial, rather than staying artificially fixed and unchanging throughout the entire experiment as the binomial model insists it must. Fourth and finally, exactly as with the binomial case, each individual trial outcome must still be cleanly classifiable as either a success or a failure, with no other possible categories or classifications permitted. To fully specify any particular hypergeometric distribution mathematically, we generally begin by defining three separate quantities: N, representing the total overall size of the entire finite population being sampled from; k, representing how many of those N total items specifically qualify as successes within that same population; and n, representing the specific size of the particular sample being randomly drawn out from that larger population, all without any form of replacement whatsoever as items are successively drawn. Given these three carefully defined quantities, the probability of observing exactly x successes appearing within that particular random sample of size n is then formally given by the hypergeometric probability formula, P(X=x) equals the quantity C(k,x) multiplied by C(N-k, n-x), with that entire product then divided by C(N,n). The underlying logical structure embedded within this particular formula makes excellent intuitive sense once each individual piece is

examined carefully in turn: the numerator's first term, C(k,x), counts every possible distinct way of choosing exactly x specific successes out from the total pool of k available successes present somewhere within the overall population; the numerator's second term, C(N-k, n-x), separately counts every possible distinct way of simultaneously choosing the remaining (n-x) failures needed to complete the sample, drawn out specifically from the (N-k) failures present within the broader population; and the single denominator term, C(N,n), simply counts the total number of ways that any sample of size n whatsoever, containing absolutely any conceivable mixture of successes and failures, could theoretically be selected from that same overall population of N total items, thereby correctly normalizing the entire calculated probability. Just as the binomial distribution offers convenient shortcut formulas for its mean and variance that avoid the tedium of constructing an entire probability table from scratch, the hypergeometric distribution similarly offers its own comparable shortcut formulas, though these particular formulas carry one additional term not present anywhere in the binomial case. The mean of the hypergeometric distribution is given simply by n.k divided by N, which bears an unmistakably close conceptual resemblance to the familiar binomial mean formula n.p, since the ratio k/N here essentially plays the exact same conceptual role that p plays over in the binomial formula, representing the underlying overall proportion of successes present within the total population being sampled from. The variance of the hypergeometric distribution, however, is given by the more elaborate expression (n.k/N) multiplied by (1 minus k/N), which itself is then further multiplied by an additional correction term, (N-n) divided by (N-1). This particular extra multiplicative term, conventionally referred to as the finite population correction factor, exists precisely because sampling without any replacement from a strictly finite, limited population mechanically produces systematically less overall variability in the resulting outcomes than would otherwise occur under comparable independent sampling with replacement, and this specific correction factor mathematically adjusts the variance downward in order to properly and accurately account for that underlying structural difference. In practical terms, deciding which of these two closely related distributions is the genuinely correct one to apply to any particular sampling problem essentially always comes down to answering one single, deceptively simple diagnostic question: is the sampling actually being conducted with replacement, or effectively from some sufficiently

large, practically unlimited population where removing individual items makes no meaningful difference to the underlying probabilities involved -- in which case the binomial distribution remains the entirely appropriate choice -- or is the sampling instead being conducted explicitly without replacement from some comparatively small, genuinely finite population, where each individual item removed does meaningfully and measurably shift the probabilities that govern every subsequent draw -- in which case the hypergeometric distribution becomes the distribution that must instead be used in order to arrive at correct and reliable probability calculations.

Multiple Choice Questions (MCQs)

A trial with exactly two possible outcomes, success and failure, is called a: (A) Binomial experiment (B) Bernoulli trial (C) Hypergeometric trial (D) Random trial

Correct answer: (B) Bernoulli trial. A Bernoulli trial has exactly two outcomes with constant success probability and independence between trials.

Which of these is NOT a property of a binomial experiment? (A) Fixed number of trials (B) Constant probability of success (C) Independent trials (D) Probability of success changes each trial

Correct answer: (D) Probability of success changes each trial. A changing probability of success is a property of the hypergeometric experiment, not the binomial experiment.

The binomial probability formula is: (A) P(X=x) = np (B) P(X=x) = C(n,x) p^x q^(n-x) (C) P(X=x) = npq (D) P(X=x) = p/q

Correct answer: (B) P(X=x) = C(n,x) p^x q^(n-x). P(X=x) = C(n,x) p^x q^(n-x) gives the probability of exactly x successes in n independent trials.

A binomial distribution is symmetric when: (A) p = 0 (B) p = 1 (C) p = q = 0.5 (D) n = 1

Correct answer: (C) p = q = 0.5. When p equals q (both 0.5), the binomial distribution is perfectly symmetric.

If p > 0.5 in a binomial distribution, the distribution is: (A) Symmetric (B) Positively skewed (C) Negatively skewed (D) Uniform

Correct answer: (C) Negatively skewed. When p exceeds 0.5, success is more likely, and the distribution becomes negatively skewed.

The mean of a binomial distribution is given by: (A) npq (B) np (C) sqrt(npq) (D) n/p

Correct answer: (B) np. The mean of the binomial distribution is E(X) = np.

The variance of a binomial distribution is given by: (A) np (B) npq (C) nq (D) sqrt(np)

Correct answer: (B) npq. The variance of the binomial distribution is Var(X) = npq.

A hypergeometric experiment involves sampling: (A) With replacement from an infinite population (B) Without replacement from a finite population (C) With replacement from a finite population (D) Without any population

Correct answer: (B) Without replacement from a finite population. The hypergeometric distribution specifically applies to sampling without replacement from a finite population.

The hypergeometric probability formula is: (A) P(X=x) = C(n,x)p^x q^(n-x) (B) P(X=x) = [C(k,x).C(N-k,n-x)] / C(N,n) (C) P(X=x) = nk/N (D) P(X=x) = np

Correct answer: (B) P(X=x) = [C(k,x).C(N-k,n-x)] / C(N,n). The hypergeometric formula divides the product of two combinations by C(N,n).

The finite population correction factor in the hypergeometric variance formula is: (A) np (B) npq (C) (N-n)/(N-1) (D) k/N

Correct answer: (C) (N-n)/(N-1). The factor (N-n)/(N-1) adjusts the variance to account for sampling without replacement from a finite population.

Quick Revision Summary

- Binomial experiment: fixed n trials, constant p, independent trials, success/failure outcomes
- Binomial formula: P(X=x) = C(n,x) p^x q^(n-x)
- Binomial shape: p=0.5 symmetric | p>0.5 negatively skewed | p<0.5 positively skewed
- Binomial mean = np; Binomial variance = npq; SD = sqrt(npq)

- Binomial frequency distribution: f(X=x) = N.P(X=x) for N repetitions of the experiment
- Hypergeometric experiment: fixed n, dependent trials, changing probability of success, finite population
- Hypergeometric formula: P(X=x) = [C(k,x).C(N-k,n-x)] / C(N,n)
- Hypergeometric mean = nk/N; variance = (nk/N)(1-k/N)((N-n)/(N-1)) — includes finite population correction
- Use binomial when sampling WITH replacement (or huge population); use hypergeometric WITHOUT replacement from a small finite population

Exam Tips

- Always check the four binomial conditions before applying the binomial formula — 'with replacement' or 'independent' are the key trigger words
- 'Without replacement' from a small finite population is the trigger phrase for hypergeometric — don't default to binomial
- For 'at least' or 'at most' binomial problems, list out and sum the individual P(X=x) terms rather than trying to shortcut it
- Memorize mean=np and variance=npq for binomial — they save enormous time versus building a full probability table
- In hypergeometric problems, carefully identify N (population), k (successes in population), and n (sample size) before plugging into the formula