La-Net

EHS212

BIOSTATISTICS

Download on the App Store

GET IT ON Google Play

Course video is available on our app

lanetonline.com

COURSE CODE: EHS 212

COURSE TITLE: Biostatistics

COURSE UNIT LOAD: (2 Units C: LH 30)

SERIES LIST

1. **Series 1: Fundamentals of Statistics and Data**
2. **Series 2: Sampling Techniques and Data Distribution**
3. **Series 3: Descriptive and Inferential Statistics**
4. **Series 4: Vital and Health Statistics in Environmental Health**
5. **Series 5: Statistical Packages and Calculation of Rates**

Suggested Outline – Series 1: Fundamentals of Statistics and Data

Part 1.1 Definition and Scope of Statistics

- 1.1.1 Definition of **Statistic**, **Statistics**, and **Data**
- 1.1.2 Importance of statistics in environmental health research
- 1.1.3 Distinction between descriptive and inferential statistics

Part 1.2 Types and Classification of Data

- 1.2.1 Qualitative and quantitative data
- 1.2.2 Primary and secondary data
- 1.2.3 Levels of measurement: nominal, ordinal, interval, ratio

Part 1.3 Measures of Central Tendency

- 1.3.1 Definition and purpose of central tendency
- 1.3.2 Mean, median, and mode with examples
- 1.3.3 Applications of central tendency in health research

Part 1.4 Data Collection and Presentation

- 1.4.1 Methods of data collection in health research
- 1.4.2 Tools and techniques for presenting data (tables, charts, graphs)
- 1.4.3 Importance of accurate presentation for decision-making

Introduction – Series 1: Fundamentals of Statistics and Data

Background Statistics is a vital tool in environmental health research. It provides methods for collecting, organising, analysing, and interpreting data. By learning the fundamentals of statistics and data, students gain the ability to describe health problems accurately and make evidence-based decisions. This Series introduces the basic concepts of statistics, types and classification of data, measures of central tendency, and methods of data collection and presentation.

Aim The aim of this Series is to equip learners with foundational knowledge of statistics and data. It will enable them to define key terms, classify data appropriately, apply measures of central tendency, and present data effectively in health research contexts.

Sequential synopsis

- **Part 1.1** defines statistics, statistics, and data, and explains their scope and importance in environmental health.
- **Part 1.2** explores types and classifications of data, including qualitative, quantitative, primary, secondary, and levels of measurement.
- **Part 1.3** introduces measures of central tendency, explaining mean, median, and mode with examples.
- **Part 1.4** discusses methods of data collection and presentation, highlighting tools such as tables, charts, and graphs.

Series Learning Outcomes – Series 1: Fundamentals of Statistics and Data

By the end of this Series, learners should be able to:

- **SLO 1.1**: **Define** the terms *Statistic*, *Statistics*, and *Data*, and **explain** their scope and importance in environmental health research.
- **SLO 1.2**: **Classify** types of data, distinguishing between qualitative and quantitative, primary and secondary, and levels of measurement.
- **SLO 1.3**: **Illustrate and calculate** measures of central tendency, including mean, median, and mode, with examples relevant to health research.
- **SLO 1.4**: **Describe and apply** different methods of data collection and **present** data effectively using tables, charts, and graphs.

1.1 Definition and Scope of Statistics

1.1.1 Definition of Statistic, Statistics, and Data

A **Statistic** is a single numerical value that summarises or describes a characteristic of a sample, for example, the average age of students in a class. **Statistics** refers to the

science and methods of collecting, organising, analysing, and interpreting data. **Data** are raw facts, figures, or observations collected for analysis.

Fill in the blank: A single numerical value that summarises a characteristic of a sample is called a _____.

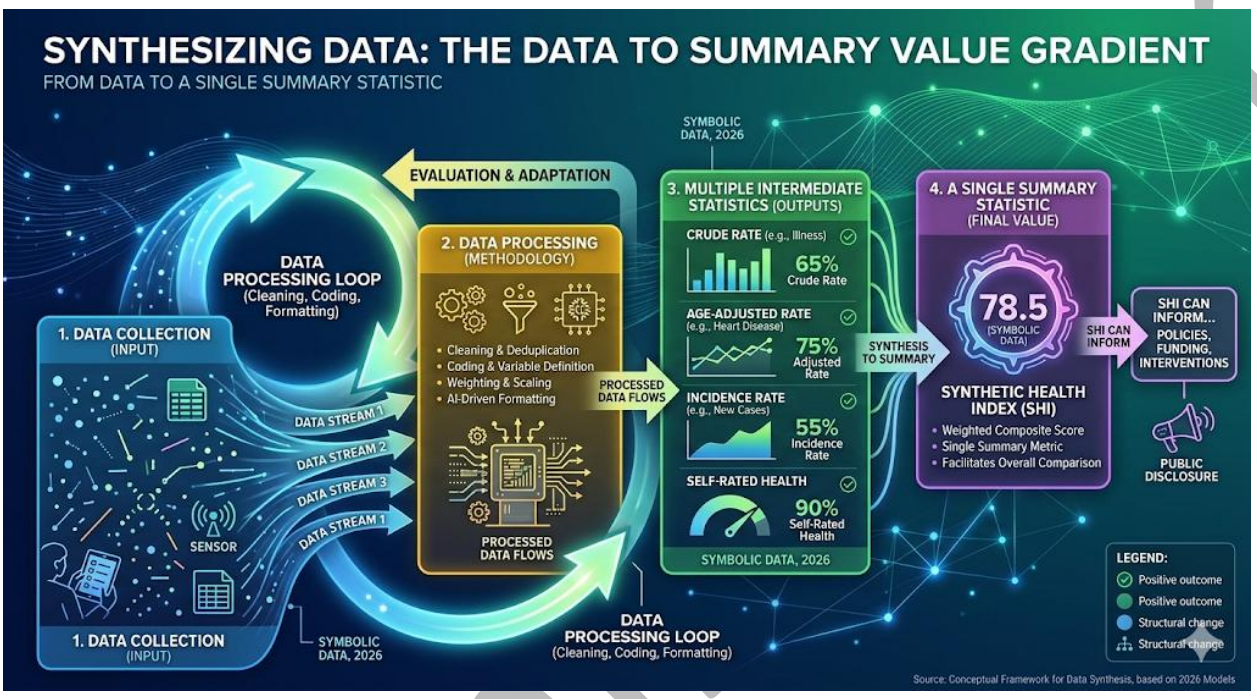


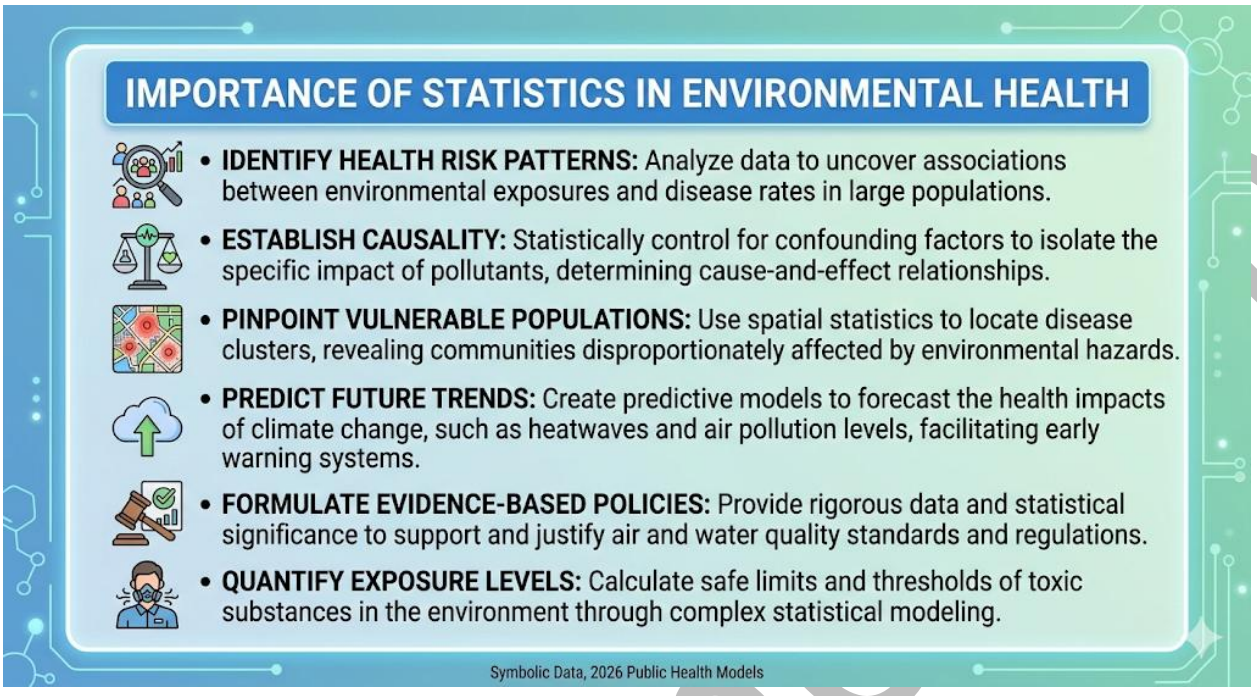
Figure 1.1 Placeholder: Diagram showing the relationship between data, statistics, and statistic. Caption: Relationship between data, statistics, and statistic.

1.1.2 Importance of Statistics in Environmental Health Research

Statistics is important in environmental health because it helps:

- To describe health problems accurately.
- To identify patterns and trends in disease occurrence.
- To evaluate interventions and policies.
- To support evidence-based decision making.

Fill in the blank: Statistics helps to identify patterns and trends in _____ occurrence.



Box 1.1 Placeholder: Text box listing the importance of statistics in environmental health research. Caption: Importance of statistics in environmental health.

1.1.3 Distinction between Descriptive and Inferential Statistics

- **Descriptive statistics**: Methods used to summarise and describe data, such as measures of central tendency and graphs.
- **Inferential statistics**: Methods used to make predictions or generalisations about a population based on sample data.

Fill in the blank: Methods used to summarise and describe data are called _____ statistics.

Feature	Descriptive Statistics	Inferential Statistics
Primary Goal	**To summarize and describe the characteristics of a specific dataset.**	**To make predictions or generalizations about a larger population based on a sample.**
What it tells you	**"What does our current data look like?"**	**"What does this sample tell us about everyone else?"**

Feature	Descriptive Statistics	Inferential Statistics
Key Tools	**Mean, Median, Mode, Range, Standard Deviation, Histograms.**	**P-values, Confidence Intervals, Hypothesis Testing (t-tests, ANOVA), Regression Analysis.**
Nature of Results	**A factual representation of the observed group.**	**An estimate with a degree of uncertainty (margin of error).**
Environmental Health Example	**Calculating the average level of lead found in 500 local water samples last month.**	**Using those 500 samples to predict the likelihood that the entire city's water supply exceeds safety limits.**
Socioeconomic Example	**Stating that 15% of a specific surveyed neighborhood lacks high-speed internet.**	**Concluding there is a significant correlation between lack of internet and lower educational outcomes nationwide.**

Table 1.1 Placeholder: Table comparing descriptive and inferential statistics. Caption: Comparison of descriptive and inferential statistics.

Pilot questions 1.1

1. A single numerical value that summarises a characteristic of a sample is called a _____.
2. Statistics helps to identify patterns and trends in _____ occurrence.
3. Methods used to summarise and describe data are called _____ statistics.
4. Give one importance of statistics in environmental health research.
5. What is the difference between descriptive and inferential statistics?

1.2 Types and Classification of Data

1.2.1 Qualitative and Quantitative Data

Qualitative data are non-numerical observations that describe qualities or characteristics, for example, gender, occupation, or colour. **Quantitative data** are

numerical observations that can be measured or counted, such as age, weight, or number of hospital visits.

Fill in the blank: Non-numerical observations that describe qualities are called _____ data.

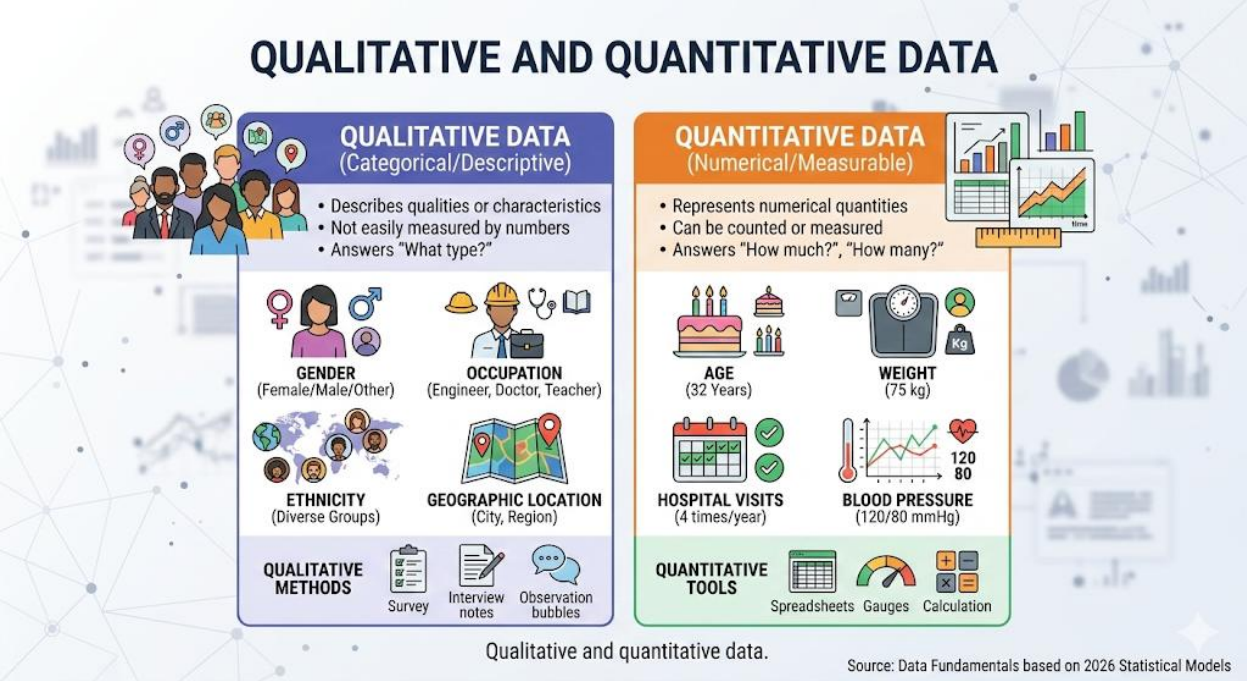


Figure 1.2 Placeholder: Diagram showing examples of qualitative and quantitative data. Caption: Qualitative and quantitative data.

1.2.2 Primary and Secondary Data

Primary data are data collected directly from the source by the researcher, for example, through surveys, interviews, or experiments. **Secondary data** are data obtained from existing sources such as published reports, hospital records, or government statistics.

Fill in the blank: Data collected directly from the source by the researcher are called _____ data.

Primary Data (First-Hand)	Secondary Data (Existing)
Definition: Data collected by the researcher specifically for the current study.	**Definition: Data collected by someone else previously for a different purpose.**
Examples:	**Examples:**

Primary Data (First-Hand)	Secondary Data (Existing)
*** Surveys & Questionnaires: Direct feedback from a community.**	*** Electronic Health Records (EHR): Aggregated hospital visit data.**
*** In-depth Interviews: Qualitative insights into patient barriers.**	*** Government Statistics: Census data or disease registries.**
*** Clinical Trials: Direct measurements from a controlled experiment.**	*** Journal Articles: Systematic reviews of previous studies.**
*** Focus Groups: Group discussions on specific health interventions.**	*** Insurance Claims: Billing data used to track service utilization.**
Pros: Highly specific and tailored to the research question.	**Pros: Faster to obtain and often involves massive sample sizes.**
Cons: Can be expensive and time-consuming to collect.	**Cons: Researcher has no control over how the data was originally gathered.**

Box 1.2 Placeholder: Text box listing examples of primary and secondary data. Caption: Primary and secondary data.

1.2.3 Levels of Measurement

There are four levels of measurement:

- **Nominal**: Categories without order, e.g., blood group.
- **Ordinal**: Categories with order, e.g., disease severity (mild, moderate, severe).
- **Interval**: Numerical values with equal intervals but no true zero, e.g., temperature in Celsius.
- **Ratio**: Numerical values with equal intervals and a true zero, e.g., weight, height.

Fill in the blank: Temperature in Celsius is measured at the _____ level.

Level	Characteristics	Health Research Examples	Mathematical Operations
Nominal	**Categorical data with no inherent order. Labels only.**	**Gender, Blood Type, Zip Code, Medical Department (ER, ICU).**	**Count, Mode.**
Ordinal	**Data has a logical order, but the distance between points isn't equal.**	**Pain Scale (1-10), Disease Stage (I, II, III), Patient Satisfaction (Low, Med, High).**	**Rank, Median.**
Interval	**Ordered data with equal distances between values, but no true zero.**	**Temperature (Celsius/Fahrenheit), IQ Scores, Time of Day.**	**Addition, Subtraction, Mean.**
Ratio	**Ordered, equal intervals, and a true zero point (zero means none).**	**Weight, Age, Heart Rate, Blood Glucose levels, Hospital Stay Duration.**	**Multiplication, Division (Ratios).**

Table 1.2 Placeholder: Table showing the four levels of measurement with examples. Caption: Levels of measurement in data classification.

Pilot questions 1.2

1. Non-numerical observations that describe qualities are called _____ data.
2. Data collected directly from the source by the researcher are called _____ data.
3. Blood group is an example of _____ level of measurement.
4. Temperature in Celsius is measured at the _____ level.
5. Give one example of secondary data.

1.3 Measures of Central Tendency

1.3.1 Definition and Purpose of Central Tendency

Central tendency refers to statistical measures that identify the centre or typical value in a dataset. It provides a summary of the data by indicating where most values cluster. In health research, central tendency helps describe average disease rates, patient ages, or hospital visits.

Fill in the blank: Statistical measures that identify the centre of a dataset are called measures of _____ tendency.

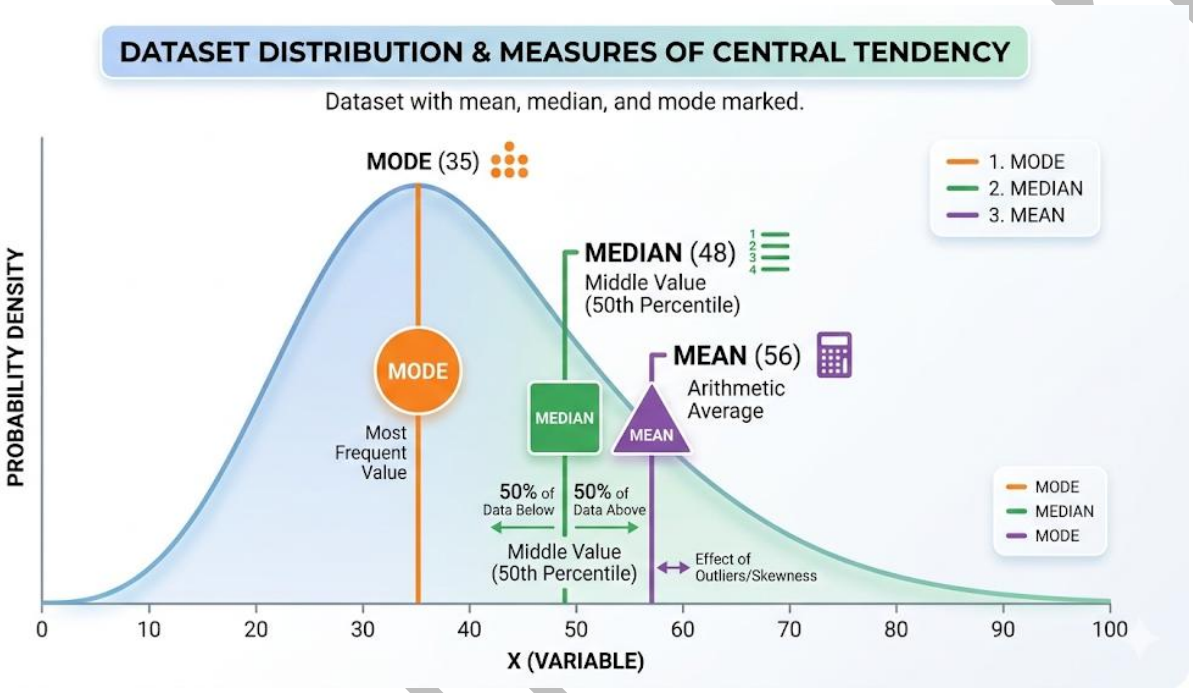


Figure 1.3 Placeholder: Diagram showing a dataset with mean, median, and mode marked. Caption: Measures of central tendency in a dataset

1.3.2 Mean, Median, and Mode with Examples

- **Mean**: The arithmetic average, calculated by dividing the sum of all values by the number of values. Example: The mean age of patients in a clinic.
- **Median**: The middle value when data are arranged in order. Example: The median income of households in a community.
- **Mode**: The most frequently occurring value in a dataset. Example: The mode of blood groups in a population.

Fill in the blank: The most frequently occurring value in a dataset is called the _____.

Measure	Definition	How to Calculate	Health Research Example
Mean	**The arithmetic average of all values.**	**Sum all values and divide by the total number of values.**	**Average Weight: If five patients weigh 60, 65, 70, 75, and 130kg, the mean is 80kg.**
Median	**The middle value in a sorted list of data.**	**Arrange values from smallest to largest; find the center point.**	**Middle Weight: In the same list (60, 65, 70, 75, 130), the median is 70kg. (Often more accurate if there is an outlier).**
Mode	**The value that appears most frequently.**	**Count which value occurs most often in the dataset.**	**Common Diagnosis: If a clinic sees 10 flu cases, 3 broken bones, and 2 rashes, the mode is "Flu."**

Table 1.3 Placeholder: Table showing definitions and examples of mean, median, and mode. Caption: Definitions and examples of mean, median, and mode.

1.3.3 Applications of Central Tendency in Health Research

Measures of central tendency are used to:

- Summarise patient characteristics such as age or weight.
- Compare disease rates across communities.
- Evaluate the effectiveness of health interventions.
- Provide evidence for policy decisions.

Fill in the blank: Measures of central tendency are used to summarise patient characteristics such as _____ or weight.

Applications of Central Tendency in Health Research
• **Defining Clinical Baselines (The Mean): Used to establish "normal" ranges for physiological markers, such as the average resting heart rate or blood glucose levels across a specific demographic.**
• **Evaluating Treatment Impact (The Mean): Essential in clinical trials to calculate the average reduction in symptoms or blood pressure after a new intervention compared to a control group.**

• **Understanding Patient Flow (The Median): Used to report "Length of Stay" in hospitals. Since a few long-term patients can skew data, the median provides a more accurate picture of when a typical patient is discharged.**
• **Survival Analysis (The Median): In oncology and chronic disease research, "Median Survival Time" is the gold standard for communicating prognosis, as it represents the time point at which 50% of patients are still living.**
• **Epidemiological Surveillance (The Mode): Used to identify the most frequent symptoms during an outbreak or the most common age group affected by a specific health condition.**
• **Resource Management (The Mode): Helps administrators identify the most common reason for ER admissions, allowing for targeted staffing and equipment allocation (e.g., more respiratory therapists during peak flu season).**

Box 1.3 Placeholder: Text box listing applications of central tendency in health research. Caption: Applications of central tendency in health research.

Pilot questions 1.3

1. Statistical measures that identify the centre of a dataset are called measures of _____ tendency.
2. The most frequently occurring value in a dataset is called the _____.
3. What is the difference between mean and median?
4. Give one example of how central tendency is applied in health research.
5. The mean is calculated by dividing the sum of all values by the _____ of values.

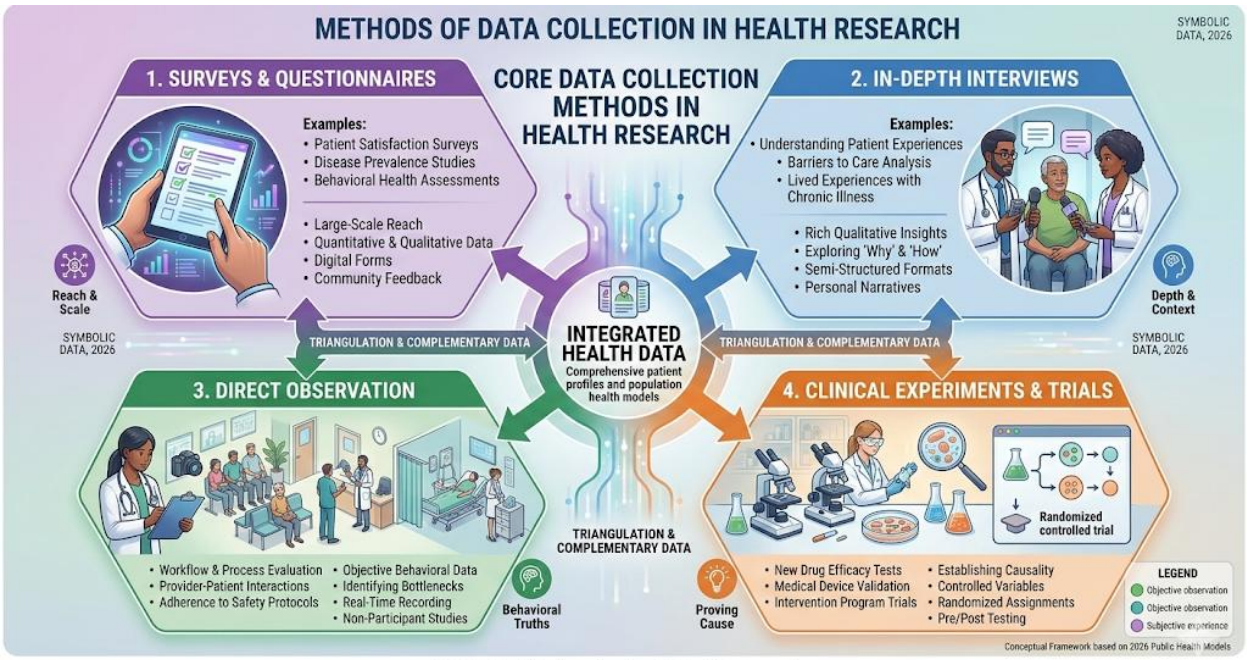
1.4 Data Collection and Presentation

1.4.1 Methods of Data Collection in Health Research

Data collection refers to the process of gathering information for analysis. In health research, common methods include:

- **Surveys**: Structured questionnaires used to collect information from a large group.
- **Interviews**: Direct conversations with individuals to gather detailed responses.
- **Observation**: Watching and recording behaviours or events.
- **Experiments**: Controlled studies to test hypotheses.

Fill in the blank: Structured questionnaires used to collect information from a large group are called _____.



Methods of data collection in health research.

Figure 1.4 Placeholder: Diagram showing different methods of data collection. Caption: Methods of data collection in health research.

1.4.2 Tools and Techniques for Presenting Data

Data presentation involves organising information so it can be easily understood. Common tools include:

- **Tables**: Organised rows and columns for numerical or categorical data.
- **Charts**: Visual representations such as bar charts and pie charts.
- **Graphs**: Line graphs or scatter plots to show trends and relationships.

Fill in the blank: Organised rows and columns used for presenting data are called _____.

Tool	Primary Purpose	Health Research Example
Tables	**To show precise values and allow for direct comparison of multiple variables.**	**Clinical Trial Results: Comparing side effects (nausea, headache, etc.) across three different dosage groups.**
Bar Charts	**To compare discrete categories or groups.**	**Hospital Admissions: Comparing the number of ER visits across different age groups (Children, Adults, Seniors).**

Tool	Primary Purpose	Health Research Example
Line Graphs	**To display trends and changes over a continuous period of time.**	**Epidemiological Trends: Tracking the weekly incidence rate of a virus over a 12-month period.**
Pie Charts	**To show the composition of a whole (proportions and percentages).**	**Funding Allocation: Showing what percentage of a public health budget goes to Prevention, Treatment, and Research.**
Scatter Plots	**To show the relationship (correlation) between two continuous variables.**	**Environmental Health: Plotting local air pollution levels ($PM_{2.5}$) against the number of asthma-related clinic visits.**
Histograms	**To show the frequency distribution of a single continuous variable.**	**Patient Demographics: Showing the distribution of ages among all patients diagnosed with a specific condition.**

Table 1.4 Placeholder: Table listing tools for data presentation with examples. Caption: Tools and techniques for data presentation.

1.4.3 Importance of Accurate Presentation for Decision-Making

Accurate presentation of data is essential because it:

- Enhances clarity and understanding.
- Helps identify patterns and trends.
- Supports evidence-based decision making.
- Prevents misinterpretation of findings.

Fill in the blank: Accurate presentation of data helps prevent _____ of findings.

The Importance of Accurate Data Presentation
Why Clear and Honest Visualization Matters
• **Prevents Misinterpretation: Accurate scaling and labeling ensure that clinicians do not overestimate or underestimate the severity of a health trend, such as a spike in infection rates.**

• **Guides Evidence-Based Policy: Policymakers rely on visual data to allocate budgets. Accurate maps and charts ensure resources reach the communities that need them most.**
• **Maintains Scientific Integrity: Proper presentation avoids "cherry-picking" or distorting results, ensuring that the research remains objective and reproducible.**
• **Facilitates Rapid Decision-Making: In emergency settings, like an outbreak, clear visuals allow health officials to identify "hotspots" at a glance and respond instantly.**
• **Enhances Public Trust: Transparent and accessible data helps the general public understand health risks, reducing the spread of misinformation during crises.**
• **Supports Health Equity: By accurately disaggregating data (by race, age, or location), presentation reveals "hidden" disparities that might be overlooked in a single average value**

Box 1.4 Placeholder: Text box listing the importance of accurate data presentation.
Caption: Importance of accurate data presentation.

Pilot questions 1.4

1. Structured questionnaires used to collect information from a large group are called _____.
2. Organised rows and columns used for presenting data are called _____.
3. Give one method of data collection in health research.
4. Name one tool used for presenting data visually.
5. Accurate presentation of data helps prevent _____ of findings.

Series 1 Summary – Fundamentals of Statistics and Data

In this Series, you have explored the essential foundations of statistics and data in environmental health research:

- **Part 1.1** introduced the definitions of *Statistic*, *Statistics*, and *Data*, explained their scope, and distinguished between descriptive and inferential statistics.
- **Part 1.2** examined types and classifications of data, including qualitative and quantitative, primary and secondary, and the four levels of measurement: nominal, ordinal, interval, and ratio.
- **Part 1.3** focused on measures of central tendency, defining mean, median, and mode, and showing their applications in summarising health data.

- **Part 1.4** discussed methods of data collection such as surveys, interviews, observation, and experiments, and highlighted tools for presenting data, including tables, charts, and graphs, with emphasis on the importance of accurate presentation for decision-making.

By completing this Series, you should now be able to:

- **Define** key statistical terms and explain their importance in environmental health (SLO 1.1).
- **Classify** data types and levels of measurement (SLO 1.2).
- **Illustrate and calculate** measures of central tendency with examples (SLO 1.3).
- **Describe and apply** methods of data collection and presentation (SLO 1.4).

This Series provides the groundwork for later topics such as sampling techniques, inferential statistics, and the use of statistical packages.

Glossary of Terms – Series 1: Fundamentals of Statistics and Data

- **Statistic**: A single numerical value that summarises or describes a characteristic of a sample.
- **Statistics**: The science and methods of collecting, organising, analysing, and interpreting data.
- **Data**: Raw facts, figures, or observations collected for analysis.
- **Descriptive statistics**: Methods used to summarise and describe data, such as measures of central tendency and graphs.
- **Inferential statistics**: Methods used to make predictions or generalisations about a population based on sample data.
- **Qualitative data**: Non-numerical observations that describe qualities or characteristics, such as gender or occupation.
- **Quantitative data**: Numerical observations that can be measured or counted, such as age or weight.
- **Primary data**: Data collected directly from the source by the researcher, for example, through surveys or interviews.
- **Secondary data**: Data obtained from existing sources such as published reports or hospital records.
- **Nominal level of measurement**: Categories without order, e.g., blood group.
- **Ordinal level of measurement**: Categories with order, e.g., disease severity.

- **Interval level of measurement**: Numerical values with equal intervals but no true zero, e.g., temperature in Celsius.
- **Ratio level of measurement**: Numerical values with equal intervals and a true zero, e.g., weight or height.
- **Central tendency**: Statistical measures that identify the centre or typical value in a dataset.
- **Mean**: The arithmetic average of a dataset.
- **Median**: The middle value when data are arranged in order.
- **Mode**: The most frequently occurring value in a dataset.
- **Data collection**: The process of gathering information for analysis.
- **Data presentation**: Organising information so it can be easily understood, often using tables, charts, or graphs.

Concept Check Questions – Series 1: Fundamentals of Statistics and Data

Objective questions (linked to SLOs 1.1–1.4)

1. A single numerical value that summarises a characteristic of a sample is called a _____. (SLO 1.1)
2. Non-numerical observations that describe qualities are called _____ data. (SLO 1.2)
3. The most frequently occurring value in a dataset is called the _____. (SLO 1.3)
4. Organised rows and columns used for presenting data are called _____. (SLO 1.4)
5. Categories with order, such as disease severity, are measured at the _____ level. (SLO 1.2)

Short-answer questions (linked to SLOs 1.1–1.4)

6. Distinguish between descriptive and inferential statistics with one example each. (SLO 1.1)
7. Give two examples of primary data and two examples of secondary data. (SLO 1.2)
8. Explain the difference between mean and median. (SLO 1.3)
9. List three methods of data collection used in health research. (SLO 1.4)
10. State two reasons why accurate data presentation is important in health research. (SLO 1.4)

Long-answer questions (linked to SLOs 1.1–1.4)

11. Analyse the importance of statistics in environmental health research, providing at least three specific applications. (SLO 1.1)

12. Discuss the types and classifications of data, highlighting their relevance in health studies. (SLO 1.2)

13. Evaluate the usefulness of mean, median, and mode in summarising health data, with examples from patient records or community surveys. (SLO 1.3)

14. Assess different methods of data collection and presentation, explaining how they contribute to evidence-based decision making in health research. (SLO 1.4)

Answers to Pilot Questions – Series 1: Fundamentals of Statistics and Data

Pilot questions 1.1 (Definition and Scope of Statistics)

1. Statistic
2. Disease
3. Descriptive
4. To describe health problems accurately
5. Descriptive statistics summarise data, while inferential statistics make predictions about populations

Pilot questions 1.2 (Types and Classification of Data)

1. Qualitative
2. Primary
3. Nominal
4. Interval
5. Hospital records (or published reports, government statistics)

Pilot questions 1.3 (Measures of Central Tendency)

1. Central
2. Mode
3. Mean is the arithmetic average, median is the middle value when data are ordered
4. Summarising patient ages in a clinic
5. Number

Pilot questions 1.4 (Data Collection and Presentation)

1. Surveys

2. Tables
3. Interviews (or observation, experiments)
4. Charts (or graphs)
5. Misinterpretation

References and Attributions – Series 1: Fundamentals of Statistics and Data

General references

- Course content adapted from *EHS 212: Biostatistics (2 Units)*.
- Concepts based on standard biostatistics and environmental health research literature, as well as open educational resources.

Figures

- *Figure 1.1: Relationship between data, statistics, and statistic* – AI prompt: "Generate a diagram showing data feeding into statistics, with a single statistic as a summary value."
- *Figure 1.2: Qualitative and quantitative data* – AI prompt: "Generate a diagram showing examples of qualitative data (gender, occupation) and quantitative data (age, weight, hospital visits)."
- *Figure 1.3: Measures of central tendency in a dataset* – AI prompt: "Generate a diagram showing a dataset with mean, median, and mode marked."
- *Figure 1.4: Methods of data collection in health research* – AI prompt: "Generate a diagram showing surveys, interviews, observation, and experiments as methods of data collection."

Tables

- *Table 1.1: Comparison of descriptive and inferential statistics* – AI prompt: "Generate a table comparing descriptive and inferential statistics, with examples."
- *Table 1.2: Levels of measurement in data classification* – AI prompt: "Generate a table showing nominal, ordinal, interval, and ratio levels of measurement with examples."
- *Table 1.3: Definitions and examples of mean, median, and mode* – AI prompt: "Generate a table showing definitions and examples of mean, median, and mode."
- *Table 1.4: Tools and techniques for data presentation* – AI prompt: "Generate a table listing tables, charts, and graphs with examples of their use in health research."

Boxes

- *Box 1.1: Importance of statistics in environmental health* – AI prompt: "Create a text box listing the importance of statistics in environmental health research."
- *Box 1.2: Primary and secondary data* – AI prompt: "Create a text box listing examples of primary data (surveys, interviews) and secondary data (hospital records, published reports)."
- *Box 1.3: Applications of central tendency in health research* – AI prompt: "Create a text box listing applications of central tendency in health research."
- *Box 1.4: Importance of accurate data presentation* – AI prompt: "Create a text box listing the importance of accurate data presentation in health research."

ggested Outline – Series 2: Sampling Techniques and Data Distribution

Part 2.1 Sampling and Samples

- 2.1.1 Definition of a sample and population
- 2.1.2 Importance of sampling in health research
- 2.1.3 Characteristics of a good sample

Part 2.2 Sampling Techniques

- 2.2.1 Random sampling methods (simple random, stratified, cluster)
- 2.2.2 Non-random sampling methods (convenience, purposive, quota)
- 2.2.3 Advantages and limitations of different sampling techniques

Part 2.3 Data Distribution

- 2.3.1 Definition of data distribution
- 2.3.2 Normal distribution and its properties
- 2.3.3 Skewed distributions and their implications in health research

Part 2.4 Applications of Sampling and Distribution in Environmental Health

- 2.4.1 Use of sampling in epidemiological studies
- 2.4.2 Role of data distribution in interpreting health statistics
- 2.4.3 Case examples in environmental health research

Introduction – Series 2: Sampling Techniques and Data Distribution

Background In health research, it is often impractical to study an entire population. Instead, researchers select a **sample**, which is a subset of the population, to make inferences about the larger group. The way samples are chosen and how data are distributed within them greatly influences the accuracy and reliability of findings. Understanding sampling techniques and data distribution is therefore essential for environmental health professionals who rely on evidence to guide decisions.

Aim The aim of this Series is to enable learners to define and explain sampling concepts, apply appropriate sampling techniques, describe data distribution patterns, and evaluate their applications in environmental health research.

Sequential synopsis

- **Part 2.1** introduces the concepts of sample and population, explains the importance of sampling, and outlines the characteristics of a good sample.
- **Part 2.2** explores sampling techniques, distinguishing between random and non-random methods, and discusses their advantages and limitations.
- **Part 2.3** explains data distribution, focusing on normal and skewed distributions, and their implications in health research.
- **Part 2.4** highlights applications of sampling and distribution in environmental health, including epidemiological studies and case examples.

Series Learning Outcomes – Series 2: Sampling Techniques and Data Distribution

By the end of this Series, learners should be able to:

- **SLO 2.1**: **Define and explain** the concepts of sample and population, and **analyse** the importance of sampling in health research.
- **SLO 2.2**: **Differentiate and apply** random and non-random sampling techniques, and **evaluate** their advantages and limitations.
- **SLO 2.3**: **Describe and interpret** data distribution patterns, including normal and skewed distributions, and **assess** their implications in health research.
- **SLO 2.4**: **Apply and evaluate** sampling and distribution concepts in environmental health studies, using epidemiological and case examples.

2.1 Sampling and Samples

2.1.1 Definition of a Sample and Population

A **Population** is the entire group of individuals or items that share common characteristics and are the focus of a study. A **Sample** is a subset of the population selected for study, which represents the larger group. For example, if researchers want to study malaria

prevalence in a city, the population is all residents of the city, while the sample may be 500 residents chosen for testing.

Fill in the blank: A subset of the population selected for study is called a _____.

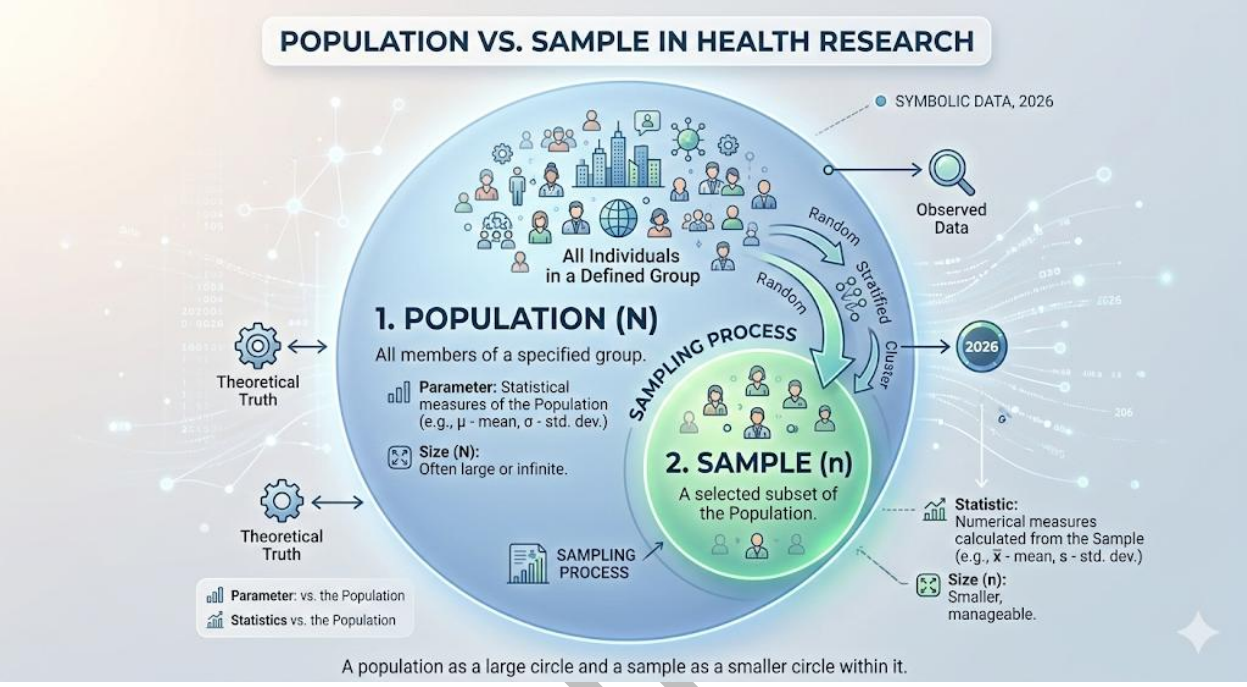


Figure 2.1 Placeholder: Diagram showing population as a large circle and sample as a smaller circle within it. Caption: Relationship between population and sample.

2.1.2 Importance of Sampling in Health Research

Sampling is important because:

- It reduces cost and time compared to studying the entire population.
- It makes research more practical and manageable.
- It allows for accurate estimates of population characteristics.
- It enables researchers to conduct studies ethically and efficiently.

Fill in the blank: Sampling reduces _____ and time compared to studying the entire population.

The Importance of Sampling in Health Research
Why Strategic Sampling is Essential
• **Feasibility and Cost-Efficiency: Sampling allows researchers to conduct high-quality studies without the astronomical costs and time required to survey an entire population. It makes research possible within limited budgets.**

• **Data Accuracy and Depth: By focusing on a smaller group, researchers can dedicate more resources to each subject—using more precise equipment, conducting longer interviews, and ensuring more rigorous follow-ups.**
• **Representativeness & Reduced Bias: Using techniques like Random Sampling ensures that every individual has an equal chance of being chosen. This prevents the results from being "skewed" by only looking at people who are easy to reach, such as those living near a major university hospital.**
• **Protecting Ethics and Safety: In clinical trials for new medications, it is safer to test a "sample" first. This limits the number of people exposed to potential side effects while still providing enough data to prove the drug's safety.**
• **Statistical Generalizability: If a sample is truly representative, we can use Inferential Statistics to apply the findings to the entire population with a high degree of confidence. This is how we determine national health guidelines based on studies of a few thousand people.**

•

Box 2.1 Placeholder: Text box listing the importance of sampling in health research.
Caption: Importance of sampling in health research.

2.1.3 Characteristics of a Good Sample

A good sample should:

- Be **representative**, meaning it reflects the characteristics of the population.
- Be **adequate in size**, large enough to provide reliable results.
- Be **selected without bias**, ensuring fairness in the process.
- Allow for **generalisation**, meaning findings can be applied to the population.

Fill in the blank: A good sample should be _____, meaning it reflects the characteristics of the population.

Characteristic	Explanation	Health Research Example
Representativeness	**The sample should reflect the key demographics (age, gender, ethnicity) of the entire population.**	**A study on heart disease must include women, as their symptoms often differ from those of men.**

Characteristic	Explanation	Health Research Example
Randomness	**Every individual in the population has an equal, non-zero chance of being selected to avoid selection bias.**	**Using a computer to randomly select patient IDs from a national database rather than just picking people in the lobby.**
Adequate Size	**The sample must be large enough to provide "statistical power," ensuring results aren't just due to chance.**	**Testing a vaccine on 10,000 people rather than 10 to ensure rare side effects are captured.**
Homogeneity (where needed)	**If studying a specific condition, the sample should share the relevant clinical characteristics.**	**A trial for a pediatric asthma drug should only include children within a specific age and severity range.**
Low Attrition Rate	**A good sample is one where participants stay in the study until the end, preventing "loss-to-follow-up" bias.**	**Providing transportation or digital check-ins to ensure patients in a 12-month study don't drop out.**

Table 2.1 Placeholder: Table listing characteristics of a good sample with brief explanations. Caption: Characteristics of a good sample.

Pilot questions 2.1

1. A subset of the population selected for study is called a _____.
2. Sampling reduces _____ and time compared to studying the entire population.
3. Name one characteristic of a good sample.
4. What is the difference between a population and a sample?
5. Why is sampling important in health research?

2.2 Sampling Techniques

2.2.1 Random Sampling Methods

Random sampling is a technique where every member of the population has an equal chance of being selected. Common types include:

- **Simple random sampling**: Each individual is chosen entirely by chance, for example, using a lottery method.
- **Stratified sampling**: The population is divided into subgroups (strata), and samples are drawn proportionally from each.
- **Cluster sampling**: The population is divided into clusters, and entire clusters are randomly selected.

Fill in the blank: When every member of the population has an equal chance of being selected, it is called _____ sampling.

Figure 2.2 Placeholder: Diagram showing simple random, stratified, and cluster sampling. Caption: Types of random sampling methods

2.2.2 Non-Random Sampling Methods

Non-random sampling does not give every member of the population an equal chance of selection. Common types include:

- **Convenience sampling**: Selecting individuals who are easiest to reach.
- **Purposive sampling**: Selecting individuals based on specific characteristics or purpose.
- **Quota sampling**: Ensuring certain categories are represented, but without random selection.

Fill in the blank: Selecting individuals who are easiest to reach is called _____ sampling.

Non-Probability Sampling Methods
Common Techniques and Health Examples
• **Convenience Sampling: Selecting participants who are easiest to reach or most available.**
• **Example: A researcher standing in a hospital lobby and asking whoever walks by to fill out a quick patient satisfaction survey.**
• **Purposive (Judgmental) Sampling: The researcher uses their expertise to hand-pick specific individuals who meet the criteria of the study.**
• **Example: A study on a rare genetic disorder where the researcher specifically recruits patients known to have that exact condition from a specialist clinic.**
• **Quota Sampling: The researcher ensures the sample includes a specific number of people from different subgroups (e.g., age or gender) until a "quota" is met, but the selection within those groups isn't random.**

• **Example: A public health team decides they need exactly 50 men and 50 women for a focus group on nutrition and stops recruiting for each gender once that number is reached.**
• **Snowball Sampling: Participants are asked to refer others they know who meet the criteria.**
• **Example: Researching hard-to-reach populations, such as intravenous drug users, where one participant introduces the researcher to their peers.**

Box 2.2 Placeholder: Text box listing types of non-random sampling with examples.
Caption: Types of non-random sampling methods.

2.2.3 Advantages and Limitations of Sampling Techniques

- **Random sampling**: Provides unbiased representation, but may be costly and time-consuming.
- **Non-random sampling**: Easier and cheaper, but may introduce bias and limit generalisation.

Fill in the blank: Random sampling provides _____ representation of the population.

Feature	Random (Probability) Sampling	Non-Random (Non-Probability) Sampling
Core Principle	**Every member has a known, non-zero chance of being selected.**	**Selection is based on convenience, judgment, or specific criteria.**
Advantages	**Eliminates selection bias; results are generalizable to the whole population.**	**Faster, cheaper, and easier to implement in "hard-to-reach" populations.**
Limitations	**Time-consuming, expensive, and requires a complete list of the population (sampling frame).**	**High risk of sampling bias; results cannot be statistically generalized to the broader public.**
Statistical Goal	**To achieve high "External Validity" and low margin of error.**	**To gain deep qualitative insights or test a hypothesis in a pilot study.**

Feature	Random (Probability) Sampling	Non-Random (Non-Probability) Sampling
Health Example	**A national survey on obesity rates where households are chosen by a random digit dialer.**	**An interview study with five specific patients who have a very rare form of bone cancer.**

Table 2.2 Placeholder: Table comparing advantages and limitations of random and non-random sampling. Caption: Advantages and limitations of sampling techniques.

Pilot questions 2.2

1. When every member of the population has an equal chance of being selected, it is called _____ sampling.
2. Selecting individuals who are easiest to reach is called _____ sampling.
3. Name one type of random sampling method.
4. Name one type of non-random sampling method.
5. Random sampling provides _____ representation of the population.

2.3 Data Distribution

2.3.1 Definition of Data Distribution

A **data distribution** describes how values in a dataset are spread or arranged. It shows the frequency of each value or group of values. In health research, data distribution helps identify patterns such as the spread of disease cases or patient ages.

Fill in the blank: A data distribution shows how values in a dataset are _____.

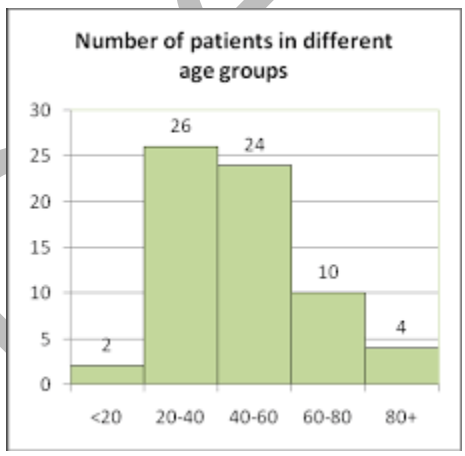


Figure 2.3 Placeholder: Histogram showing the spread of patient ages. Caption: Example of a data distribution.

2.3.2 Normal Distribution and Its Properties

A **normal distribution** is a symmetrical, bell-shaped curve where most values cluster around the mean. Properties include:

- Symmetry about the mean.
- Mean, median, and mode are equal.
- Predictable proportions of data fall within standard deviations (e.g., 68% within one standard deviation).

Fill in the blank: In a normal distribution, the mean, median, and mode are _____.

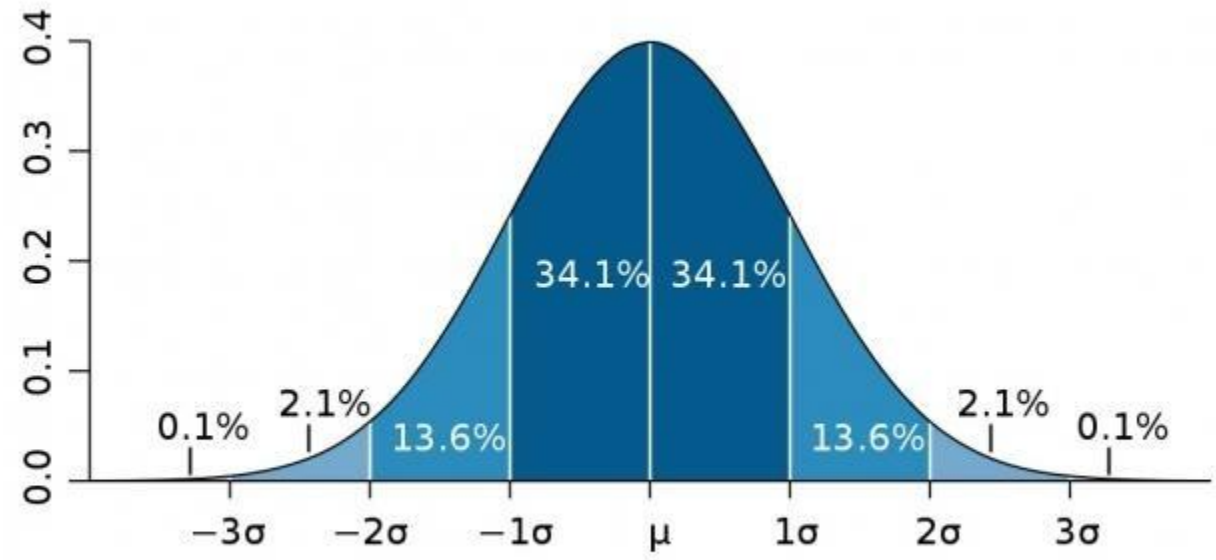


Figure 2.4 Placeholder: Bell curve showing normal distribution with mean and standard deviations. Caption: Normal distribution curve.

2.3.3 Skewed Distributions and Their Implications in Health Research

A **skewed distribution** occurs when data are not symmetrical.

- **Positively skewed**: Tail extends to the right, e.g., income distribution.

- **Negatively skewed**: Tail extends to the left, e.g., age at death in populations with high infant mortality.

Implications in health research:

- Skewed data may affect the choice of statistical measures.
- Median may be more appropriate than mean in skewed datasets.
- Understanding skewness helps interpret health outcomes accurately.

Fill in the blank: When the tail of a distribution extends to the right, it is called _____ skewed.

Feature	Normal Distribution	Skewed Distribution
Symmetry	**Perfectly Symmetrical. The left side is a mirror image of the right.**	**Asymmetrical. Data "bunches" on one side with a long "tail" on the other.**
Central Tendency	**Mean = Median = Mode. All three sit at the center peak.**	**Mean $\neq$ Median $\neq$ Mode. The mean is pulled toward the tail.**
Visual Shape	**A classic, balanced Bell Curve.**	**Can be Positively Skewed (tail to the right) or Negatively Skewed (tail to the left).**
Health Example	**Adult Height: Most people are near the average; very few are extremely tall or short.**	**Hospital Length of Stay: Most stay 1–3 days, but a few stay for months, creating a long right tail.**
Best Measure	**The Mean is the most reliable measure.**	**The Median is better because it isn't "dragged" by outliers.**

Table 2.3 Placeholder: Table comparing normal and skewed distributions with examples. Caption: Comparison of normal and skewed distributions.

Pilot questions 2.3

1. A data distribution shows how values in a dataset are _____.

2. In a normal distribution, the mean, median, and mode are _____.
3. When the tail of a distribution extends to the right, it is called _____ skewed.
4. Give one property of a normal distribution.
5. Why might the median be more appropriate than the mean in skewed datasets?

2.4 Applications of Sampling and Distribution in Environmental Health

2.4.1 Use of Sampling in Epidemiological Studies

Sampling is widely used in **epidemiology**, which is the study of disease occurrence and distribution in populations. For example, researchers may select a sample of households to estimate malaria prevalence in a community. Sampling allows for:

- Efficient data collection.
- Reliable estimates of disease rates.
- Ethical management of resources and participants.

Fill in the blank: The study of disease occurrence and distribution in populations is called _____.

Figure 2.5 Placeholder: Diagram showing sampling in an epidemiological study. Caption: Sampling in epidemiological studies.

2.4.2 Role of Data Distribution in Interpreting Health Statistics

Data distribution helps researchers understand patterns in health statistics. For example:

- A normal distribution of blood pressure readings may indicate typical population health.
- A skewed distribution of income levels may explain disparities in access to healthcare.
- Recognising distribution patterns helps in choosing appropriate statistical tests.

Fill in the blank: A normal distribution of blood pressure readings may indicate _____ population health.

The Strategic Role of Data Distribution
Roles of Data Distribution in Health Interpretation
• **Determining Statistical Tests: The distribution (Normal vs. Skewed) dictates whether researchers should use Parametric tests (like T-tests) or Non-parametric tests (like Wilcoxon), ensuring the mathematical conclusions are valid.**

• **Identifying Clinical Outliers: Distribution patterns help spot "super-utilizers" or patients with extreme reactions to treatment. These outliers can reveal rare side effects or groups that require specialized care.**
• **Setting Diagnostic Thresholds: By observing how health markers (like blood sugar or cholesterol) are distributed in a healthy population, researchers can define "normal" ranges and establish the cut-off points for diagnosing disease.**
• **Revealing Health Inequities: Distributions can show "bimodal" patterns (two peaks), which often indicate that a population is split—perhaps by socioeconomic status or access to care—revealing disparities that a single average would hide.**
• **Communicating Realistic Expectations: For prognosis, the distribution tells a patient not just the "average" recovery time, but the range of likely outcomes, helping them understand the best-case and worst-case scenarios.**
• **Evaluating Intervention Success: By comparing "Before" and "After" distributions, researchers can see if a new policy shifted the *entire* population toward better health or only helped those who were already doing well.**

Box 2.4 Placeholder: Text box listing roles of data distribution in interpreting health statistics. Caption: Role of data distribution in interpreting health statistics.

2.4.3 Case Examples in Environmental Health Research

Applications include:

- Sampling water sources to test for contamination.
- Studying air quality by selecting monitoring stations.
- Analysing disease distribution in communities exposed to pollutants.

These examples show how sampling and distribution are practical tools for protecting public health.

Fill in the blank: Sampling water sources to test for contamination is an example of _____ in environmental health research.

Research Focus	Sampling Strategy	Distribution Type	Why it Matters
Air Quality & Asthma	**Cluster Sampling: Selecting specific city blocks near industrial zones and**	**Right-Skewed (Positive): Most households show low $PM_{2.5}$ levels, but those closest to**	**Helps identify "hotspots" where environmental**

Research Focus	Sampling Strategy	Distribution Type	Why it Matters
	sampling all households within them.	factories show extreme spikes.	regulations are most urgently needed.
Lead in Drinking Water	Stratified Random Sampling: Dividing a city into strata based on the age of housing (e.g., pre-1950 vs. modern) and sampling randomly from each.	Bimodal: One peak represents modern homes with negligible lead; a second peak represents older homes with significant lead presence.	Prevents the "average" from hiding the crisis in older neighborhoods.
Microplastic Ingestion	Purposive Sampling: Specifically selecting coastal fishing communities whose primary protein source is local seafood.	Normal (Bell Curve): In a specific community with similar diets, the levels of microplastics in blood samples typically cluster around a mean.	Establishes a baseline for exposure in high-risk groups to set safety guidelines.
Radiation Exposure Post-Accident	Systematic Sampling: Testing soil and water at fixed intervals (e.g., every 500 meters) radiating outward from a point source.	Exponential Decay: High concentration at the source, dropping off rapidly as distance increases.	Defines "exclusion zones" and safe zones for resettlement.

Research Focus	Sampling Strategy	Distribution Type	Why it Matters
Pesticide Residue in Farmworkers	**Convenience Sampling: Testing workers who visit a specific mobile health clinic during harvest season.**	**Skewed/Variable: Highly dependent on the specific crop and the day's wind speed during application.**	**Provides a "quick look" at occupational safety, though it may miss workers who are too sick or busy to visit the clinic.**

Table 2.4 Placeholder: Table showing case examples of sampling and distribution in environmental health. Caption: Case examples of sampling and distribution in environmental health research.

Pilot questions 2.4

1. The study of disease occurrence and distribution in populations is called _____.
2. A normal distribution of blood pressure readings may indicate _____ population health.
3. Give one example of how sampling is applied in environmental health research.
4. Why is recognising data distribution patterns important in health statistics?
5. Sampling water sources to test for contamination is an example of _____ in environmental health research.

Series 2 Summary – Sampling Techniques and Data Distribution

In this Series, you explored the foundations of sampling and data distribution in health research:

- **Part 2.1** defined *population* and *sample*, explained why sampling is essential, and outlined the characteristics of a good sample.
- **Part 2.2** examined sampling techniques, distinguishing between random methods (simple random, stratified, cluster) and non-random methods (convenience, purposive, quota), with their advantages and limitations.
- **Part 2.3** introduced data distribution, explained the properties of a normal distribution, and contrasted it with skewed distributions, highlighting their implications for statistical analysis.

- **Part 2.4** applied these concepts to environmental health research, showing how sampling and distribution are used in epidemiological studies, interpreting health statistics, and investigating environmental exposures.

By completing this Series, you should now be able to:

- **Define and explain** population and sample, and analyse the importance of sampling (SLO 2.1).
- **Differentiate and apply** random and non-random sampling techniques, evaluating their strengths and weaknesses (SLO 2.2).
- **Describe and interpret** normal and skewed distributions, assessing their implications in health research (SLO 2.3).
- **Apply and evaluate** sampling and distribution concepts in environmental health studies (SLO 2.4).

Glossary of Terms – Series 2: Sampling Techniques and Data Distribution

- **Population**: The entire group of individuals or items that share common characteristics and are the focus of a study.
- **Sample**: A subset of the population selected for study, representing the larger group.
- **Sampling**: The process of selecting a subset of individuals or items from a population for research.
- **Random sampling**: A technique where every member of the population has an equal chance of being selected.
- **Simple random sampling**: Each individual is chosen entirely by chance, e.g., lottery method.
- **Stratified sampling**: The population is divided into subgroups (strata), and samples are drawn proportionally from each.
- **Cluster sampling**: The population is divided into clusters, and entire clusters are randomly selected.
- **Non-random sampling**: A technique where not all members of the population have an equal chance of selection.
- **Convenience sampling**: Selecting individuals who are easiest to reach.
- **Purposive sampling**: Selecting individuals based on specific characteristics or purpose.
- **Quota sampling**: Ensuring certain categories are represented, but without random selection.

- **Data distribution**: The arrangement or spread of values in a dataset, showing their frequency.
- **Normal distribution**: A symmetrical, bell-shaped curve where most values cluster around the mean, with mean = median = mode.
- **Skewed distribution**: A distribution that is not symmetrical; can be positively skewed (tail to the right) or negatively skewed (tail to the left).
- **Epidemiology**: The study of disease occurrence and distribution in populations.
- **Representativeness**: The degree to which a sample reflects the characteristics of the population.
- **Bias**: Systematic error in sampling or data collection that leads to inaccurate results.

Concept Check Questions – Series 2: Sampling Techniques and Data Distribution

Objective questions (linked to SLOs 2.1–2.4)

1. A subset of the population selected for study is called a _____. (SLO 2.1)
2. When every member of the population has an equal chance of being selected, it is called _____ sampling. (SLO 2.2)
3. In a normal distribution, the mean, median, and mode are _____. (SLO 2.3)
4. The study of disease occurrence and distribution in populations is called _____. (SLO 2.4)
5. Selecting individuals who are easiest to reach is called _____ sampling. (SLO 2.2)

Short-answer questions (linked to SLOs 2.1–2.4)

6. Distinguish between population and sample with one example each. (SLO 2.1)
7. Give two examples of random sampling methods and two examples of non-random sampling methods. (SLO 2.2)
8. Explain one property of a normal distribution. (SLO 2.3)
9. Why might the median be more appropriate than the mean in skewed datasets? (SLO 2.3)
10. Provide one example of how sampling is applied in environmental health research. (SLO 2.4)

Long-answer questions (linked to SLOs 2.1–2.4)

11. Analyse the importance of sampling in health research, highlighting at least three reasons why it is preferred over studying entire populations. (SLO 2.1)

12. Discuss random and non-random sampling techniques, evaluating their advantages and limitations with examples. (SLO 2.2)

13. Evaluate the usefulness of normal and skewed distributions in interpreting health statistics, with examples from patient data or community surveys. (SLO 2.3)

14. Assess the applications of sampling and distribution in environmental health research, using case examples such as water testing or air quality studies. (SLO 2.4)

Answers to Pilot Questions – Series 2: Sampling Techniques and Data Distribution

Pilot questions 2.1 (Sampling and Samples)

1. Sample
2. Cost
3. Representativeness (or "representative")
4. A population is the entire group, while a sample is a subset of that group.
5. It reduces cost and time, makes research practical, and allows accurate estimates.

Pilot questions 2.2 (Sampling Techniques)

1. Random sampling
2. Convenience sampling
3. Simple random sampling (or stratified, cluster)
4. Purposive sampling (or quota, convenience)
5. Unbiased

Pilot questions 2.3 (Data Distribution)

1. Spread (or "arranged")
2. Equal
3. Positively skewed
4. Symmetry about the mean (or mean = median = mode, predictable proportions within standard deviations)
5. Because the median is less affected by extreme values.

Pilot questions 2.4 (Applications in Environmental Health)

1. Epidemiology

2. Typical
3. Sampling water sources (or selecting households, monitoring air quality)
4. It helps choose appropriate statistical tests and interpret health outcomes accurately.
5. Sampling

References and Attributions – Series 2: Sampling Techniques and Data Distribution

General references

- Course content adapted from *EHS 212: Biostatistics (2 Units)*.
- Concepts based on standard biostatistics, epidemiology, and environmental health research literature, as well as open educational resources.

Figures

- *Figure 2.1: Relationship between population and sample* – AI prompt: "Generate a diagram showing a population as a large circle and a sample as a smaller circle within it."
- *Figure 2.2: Types of random sampling methods* – AI prompt: "Generate a diagram showing simple random, stratified, and cluster sampling with labelled examples."
- *Figure 2.3: Example of a data distribution* – AI prompt: "Generate a histogram showing the distribution of patient ages in a sample."
- *Figure 2.4: Normal distribution curve* – AI prompt: "Generate a bell curve showing normal distribution with mean and standard deviations marked."
- *Figure 2.5: Sampling in epidemiological studies* – AI prompt: "Generate a diagram showing how a sample of households is used to estimate malaria prevalence in a community."

Tables

- *Table 2.1: Characteristics of a good sample* – AI prompt: "Generate a table listing characteristics of a good sample with explanations."
- *Table 2.2: Advantages and limitations of sampling techniques* – AI prompt: "Generate a table comparing advantages and limitations of random and non-random sampling."
- *Table 2.3: Comparison of normal and skewed distributions* – AI prompt: "Generate a table comparing normal and skewed distributions with examples."

- *Table 2.4: Case examples of sampling and distribution in environmental health research* – AI prompt: "Generate a table showing case examples of sampling and distribution in environmental health research."

Boxes

- *Box 2.1: Importance of sampling in health research* – AI prompt: "Create a text box listing the importance of sampling in health research."
- *Box 2.2: Types of non-random sampling methods* – AI prompt: "Create a text box listing convenience, purposive, and quota sampling with examples."
- *Box 2.4: Role of data distribution in interpreting health statistics* – AI prompt: "Create a text box listing roles of data distribution in interpreting health statistics."

Part 3.1 Descriptive Statistics

- 3.1.1 Definition and scope of descriptive statistics
- 3.1.2 Measures of central tendency (mean, median, mode)
- 3.1.3 Measures of variability (range, variance, standard deviation)
- 3.1.4 Applications of descriptive statistics in health research

Part 3.2 Inferential Statistics

- 3.2.1 Definition and scope of inferential statistics
- 3.2.2 Concept of population parameters and sample statistics
- 3.2.3 Estimation and confidence intervals
- 3.2.4 Applications of inferential statistics in health research

Part 3.3 Hypothesis Testing

- 3.3.1 Definition of hypothesis and types (null, alternative)
- 3.3.2 Steps in hypothesis testing
- 3.3.3 Common statistical tests (t-test, chi-square, ANOVA)
- 3.3.4 Interpretation of results in health studies

Part 3.4 Applications in Environmental Health Research

- 3.4.1 Use of descriptive statistics in summarising environmental data
- 3.4.2 Use of inferential statistics in predicting health outcomes
- 3.4.3 Case examples linking hypothesis testing to environmental health decisions

Introduction – Series 3: Descriptive and Inferential Statistics

Background After understanding data basics (Series 1) and sampling/distribution (Series 2), the next step in biostatistics is learning how to **summarise data** and **draw conclusions**. Descriptive statistics help researchers organise and present data clearly, while inferential statistics allow them to make predictions and generalisations from samples to populations. Together, these two branches form the backbone of statistical analysis in health and environmental research.

Aim The aim of this Series is to enable learners to distinguish between descriptive and inferential statistics, apply measures of central tendency and variability, understand estimation and confidence intervals, and evaluate hypothesis testing in health research.

Sequential synopsis

- **Part 3.1** introduces descriptive statistics, covering measures of central tendency and variability, and their applications in health research.
- **Part 3.2** explains inferential statistics, focusing on population parameters, sample statistics, estimation, and confidence intervals.
- **Part 3.3** explores hypothesis testing, including types of hypotheses, steps in testing, and common statistical tests.
- **Part 3.4** highlights applications of descriptive and inferential statistics in environmental health research, with case examples.

Series Learning Outcomes – Series 3: Descriptive and Inferential Statistics

By the end of this Series, learners should be able to:

- **SLO 3.1**: **Define and apply** descriptive statistics, including measures of central tendency and variability, to summarise health data.
- **SLO 3.2**: **Explain and utilise** inferential statistics, distinguishing between population parameters and sample statistics, and applying estimation and confidence intervals.
- **SLO 3.3**: **Describe and conduct** hypothesis testing, identifying types of hypotheses, steps in testing, and interpreting results using common statistical tests.
- **SLO 3.4**: **Evaluate and apply** descriptive and inferential statistics in environmental health research, using case examples to demonstrate practical relevance.

3.1 Descriptive Statistics

3.1.1 Definition and Scope of Descriptive Statistics

Descriptive statistics are methods used to summarise, organise, and present data in a meaningful way. They describe the basic features of a dataset without making predictions or generalisations. Examples include averages, percentages, and charts.

Fill in the blank: Statistics that summarise and present data without making predictions are called _____ statistics.

Figure 3.1 Placeholder: Diagram showing descriptive statistics as a branch of statistics. Caption: Scope of descriptive statistics.

3.1.2 Measures of Central Tendency

These describe the “centre” of a dataset:

- **Mean**: The arithmetic average.
- **Median**: The middle value when data are ordered.
- **Mode**: The most frequently occurring value.

Fill in the blank: The most frequently occurring value in a dataset is called the _____.

Measure	Definition	When to Use It	Health Research Example
Mean	**The arithmetic average (sum of all values divided by the number of values).**	**Best for Normal Distributions with no extreme outliers.**	**Average Birth Weight: Most babies cluster around a specific weight, making the mean a reliable "norm."**
Median	**The middle value when data is sorted from lowest to highest.**	**Best for Skewed Distributions (data with extreme outliers).**	**Hospital Stay Duration: A few patients staying for months won't distort the median like they would the mean.**
Mode	**The value that appears most frequently in the dataset.**	**Best for Categorical Data or identifying the most common "type."**	**Most Common Blood Type: Knowing "Type O+" is the most frequent is more useful than an "average" blood type.**

Table 3.1 Placeholder: Table showing definitions and examples of mean, median, and mode. Caption: Measures of central tendency.

3.1.3 Measures of Variability

These describe how spread out the data are:

- **Range**: Difference between the highest and lowest values.
- **Variance**: Average of squared differences from the mean.
- **Standard deviation**: Square root of variance, showing average distance from the mean.

Fill in the blank: The difference between the highest and lowest values in a dataset is called the _____.

Figure 3.2 Placeholder: Diagram showing variability with range, variance, and standard deviation. Caption: Measures of variability.

3.1.4 Applications of Descriptive Statistics in Health Research

Descriptive statistics are used to:

- Summarise patient demographics (e.g., average age).
- Present disease prevalence rates.
- Compare health indicators across groups.
- Provide clear visualisations for decision-makers.

Fill in the blank: Descriptive statistics are used to summarise _____ demographics in health research.

The Strategic Applications of Descriptive Statistics
How Researchers Use Descriptive Data
• **Characterizing Study Populations: > Using the Mean and Standard Deviation to describe the baseline profile of participants (e.g., "The average age was 45 ± 5 years"), ensuring the group matches the target demographic.**
• **Screening for Data Entry Errors: > Using the Range (minimum and maximum values) to spot impossible data points, such as a heart rate of 500 bpm or a body temperature of $25°C$.**
• **Monitoring Clinical Trends: > Using Frequency Distributions (histograms) to track the spread of a disease outbreak over time or across different hospital wards.**
• **Resource Allocation: > Using the Mode to identify the most common reason for ER visits, helping hospital administrators decide which specialists (e.g., cardiologists vs. trauma surgeons) need to be on call.**

• **Identifying Health Disparities: > Using Measures of Dispersion (like the Interquartile Range) to see if health outcomes vary wildly between different socioeconomic groups, even if their "average" outcome looks similar.**
• **Simplifying Complex Information: > Summarizing thousands of patient records into a single Table of "Key Indicators" that can be quickly digested by busy healthcare providers or policymakers.**

Box 3.1 Placeholder: Text box listing applications of descriptive statistics in health research. Caption: Applications of descriptive statistics.

Pilot questions 3.1

1. Statistics that summarise and present data without making predictions are called _____ statistics.
2. The most frequently occurring value in a dataset is called the _____.
3. The difference between the highest and lowest values in a dataset is called the _____.
4. Name one measure of central tendency.
5. Give one application of descriptive statistics in health research.

3.2 Inferential Statistics

3.2.1 Definition and Scope of Inferential Statistics

Inferential statistics are methods used to make predictions or generalisations about a population based on data collected from a sample. Unlike descriptive statistics, which only summarise data, inferential statistics allow researchers to draw conclusions and test hypotheses.

Fill in the blank: Statistics that make predictions or generalisations about a population are called _____ statistics.

Figure 3.3 Placeholder: Diagram showing inferential statistics extending from sample to population. Caption: Scope of inferential statistics.

3.2.2 Population Parameters and Sample Statistics

- **Population parameter**: A numerical value that describes a characteristic of the entire population (e.g., population mean).
- **Sample statistic**: A numerical value that describes a characteristic of a sample (e.g., sample mean).
- Sample statistics are used to estimate population parameters.

Fill in the blank: A numerical value that describes a characteristic of the entire population is called a _____.

Feature	Population Parameter	Sample Statistic
Definition	**A fixed value that describes the entire group of interest.**	**A variable value calculated from a subset of the population.**
Symbols used	**Greek Letters (e.g., μ, σ)**	**Latin Letters (e.g., $\bar{x}, s$)**
Knowability	**Usually unknown (too expensive or impossible to measure everyone).**	**Always known and calculated from collected data.**
Stability	**Constant (does not change for a specific population).**	**Varies from sample to sample (Sampling Error).**

Table 3.2 Placeholder: Table comparing population parameters and sample statistics with examples. Caption: Population parameters vs. sample statistics.

3.2.3 Estimation and Confidence Intervals

- **Estimation**: Using sample statistics to approximate population parameters.
- **Confidence interval (CI)**: A range of values within which the true population parameter is expected to lie, with a certain level of confidence (e.g., 95%).
- Wider intervals indicate less precision, while narrower intervals indicate greater precision.

Fill in the blank: A range of values within which the true population parameter is expected to lie is called a _____ interval.

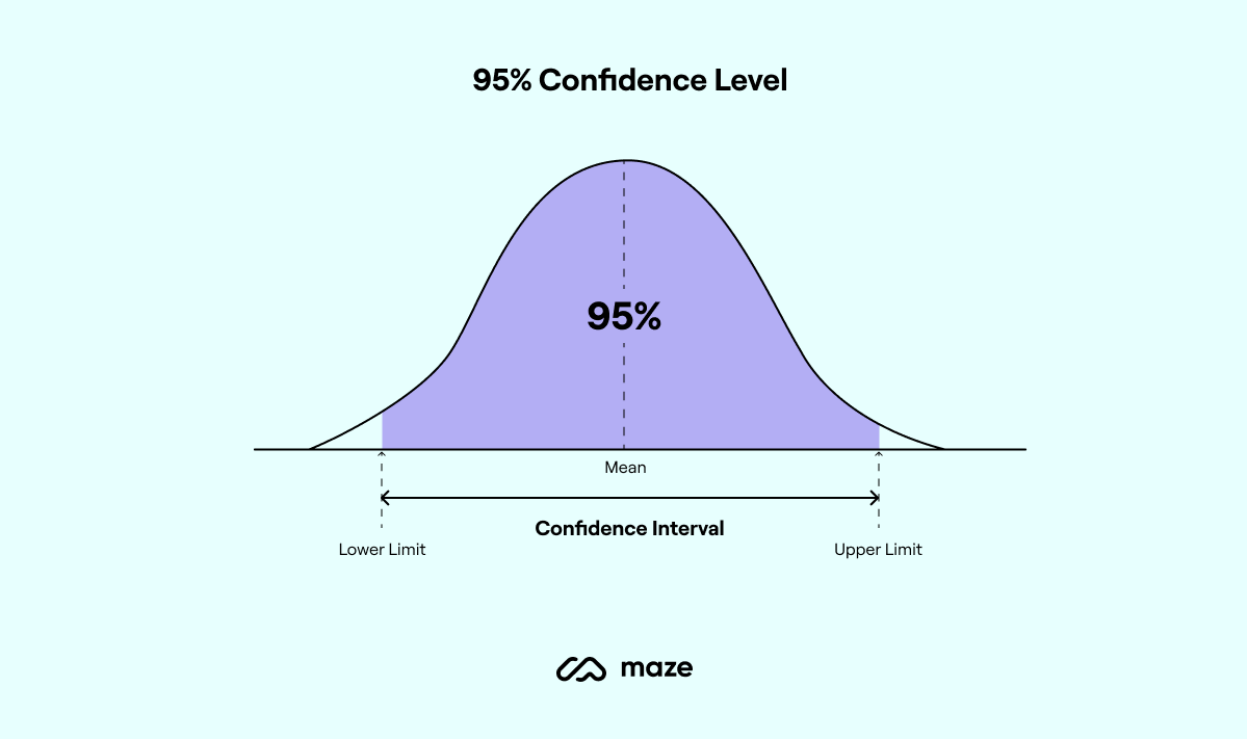


Figure 3.4 Placeholder: Diagram showing confidence intervals around a sample mean. Caption: Confidence interval illustration.

3.2.4 Applications of Inferential Statistics in Health Research

Inferential statistics are used to:

- Predict disease prevalence in a population based on sample data.
- Test the effectiveness of new treatments.
- Generalise findings from clinical trials to larger populations.
- Support evidence-based decision making in health policy.

Fill in the blank: Inferential statistics are used to test the effectiveness of new _____.

Statistical Tool	Health Research Goal
Independent T-Test	**Comparing the mean recovery time between two different surgical techniques.**
ANOVA	**Comparing the effectiveness of three different dosages of a new drug.**

Statistical Tool	Health Research Goal
Chi-Square Test	**Determining if there is a relationship between vaccination status and infection rate.**
Correlation (r)	**Determining the strength of the relationship between daily step count and BMI.**

Box 3.2 Placeholder: Text box listing applications of inferential statistics in health research. Caption: Applications of inferential statistics.

Pilot questions 3.2

1. Statistics that make predictions or generalisations about a population are called _____ statistics.
2. A numerical value that describes a characteristic of the entire population is called a _____.
3. A range of values within which the true population parameter is expected to lie is called a _____ interval.
4. Give one example of a population parameter.
5. State one application of inferential statistics in health research.

3.3 Hypothesis Testing

3.3.1 Definition of Hypothesis and Types

A **hypothesis** is a statement or assumption about a population parameter that can be tested using statistical methods.

- **Null hypothesis (H_0)**: Assumes no effect or no difference (e.g., "There is no difference in blood pressure between treated and untreated groups").
- **Alternative hypothesis (H_1)**: Assumes there is an effect or difference (e.g., "There is a difference in blood pressure between treated and untreated groups").

Fill in the blank: The hypothesis that assumes no effect or difference is called the _____ hypothesis.

Figure 3.5 Placeholder: Diagram showing null vs. alternative hypothesis. Caption: Types of hypotheses.

3.3.2 Steps in Hypothesis Testing

1. **State the hypotheses** (null and alternative).
2. **Select the appropriate statistical test**.
3. **Set the significance level (α)**, often 0.05.
4. **Collect and analyse data**.
5. **Make a decision**: reject or fail to reject the null hypothesis.

Fill in the blank: The significance level commonly used in hypothesis testing is _____.

The 5 Steps of Hypothesis Testing
1. State the Hypotheses
You must define two mutually exclusive statements:
• **Null Hypothesis (H_0): The status quo. It assumes there is no effect, no difference, or no relationship (e.g., "The new drug is no better than the old one").**
• **Alternative Hypothesis (H_1 or H_a): The claim you are testing. It assumes there is an effect or a difference (e.g., "The new drug is more effective").**
2. Set the Decision Criteria
Before looking at the data, decide on the "Significance Level" (α).
• **The most common value is $\alpha = 0.05$ (5%).**
• **This is the threshold for how "rare" the data must be under the null hypothesis before you decide to reject it.**
3. Collect Data and Calculate the Test Statistic
Gather your sample and perform the appropriate statistical test (like a T-test, Z-test, or Chi-square).
• **The Test Statistic measures how far your sample results are from what the null hypothesis predicted.**
• **For example, a high t-score indicates a large difference between group means.**
4. Determine the P-Value
The p-value is the probability of seeing your sample results (or something even more extreme) if the null hypothesis were actually true.
• **Low p-value: The data is very unlikely under the null hypothesis.**
• **High p-value: The data is consistent with the null hypothesis.**
5. Make a Decision and Interpret
Compare your p-value to your significance level (α):
• **If $p \leq \alpha$: Reject the Null Hypothesis. The result is "statistically significant."**

• **If $p > \alpha$: Fail to Reject the Null Hypothesis. There is not enough evidence to support the claim.**
Box 3.3 Placeholder: Text box listing steps in hypothesis testing. Caption: Steps in hypothesis testing.
3.3.3 Common Statistical Tests
• **t-test**: Compares means between two groups.
• **Chi-square test**: Tests association between categorical variables.
• **ANOVA (Analysis of Variance)**: Compares means across three or more groups.

Fill in the blank: A statistical test used to compare means between two groups is called a _____ test.

Table 3.3 Placeholder: Table listing common statistical tests with their uses. Caption: Common statistical tests in hypothesis testing. AI prompt suggestion: "Generate a table listing t-test, chi-square, and ANOVA with their uses."

3.3.4 Interpretation of Results in Health Studies

- If the p-value < α, reject the null hypothesis (evidence of an effect).
- If the p-value ≥ α, fail to reject the null hypothesis (no evidence of an effect).
- Interpretation must consider context, sample size, and practical significance.

Fill in the blank: If the p-value is less than α, we _____ the null hypothesis.

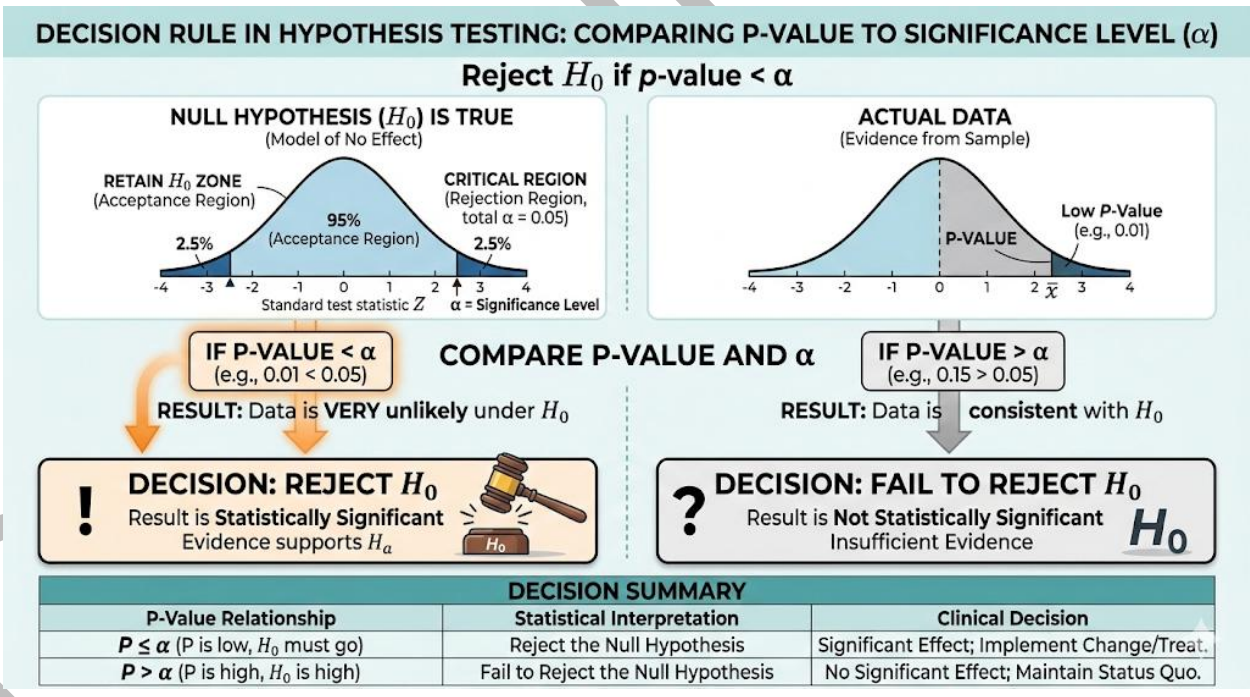


Figure 3.6 Placeholder: Diagram showing decision rule based on p-value. Caption: Decision rule in hypothesis testing.

Pilot questions 3.3

1. The hypothesis that assumes no effect or difference is called the _____ hypothesis.
2. The significance level commonly used in hypothesis testing is _____.
3. A statistical test used to compare means between two groups is called a _____ test.
4. Name one common statistical test used in hypothesis testing.
5. If the p-value is less than α, we _____ the null hypothesis.

3.4 Applications in Environmental Health Research

3.4.1 Use of Descriptive Statistics in Environmental Data

Descriptive statistics help summarise environmental health data:

- Average pollutant levels in air or water.
- Median household exposure to contaminants.
- Mode of reported symptoms in a community.

Fill in the blank: The most common value in a dataset, such as reported symptoms, is called the _____.

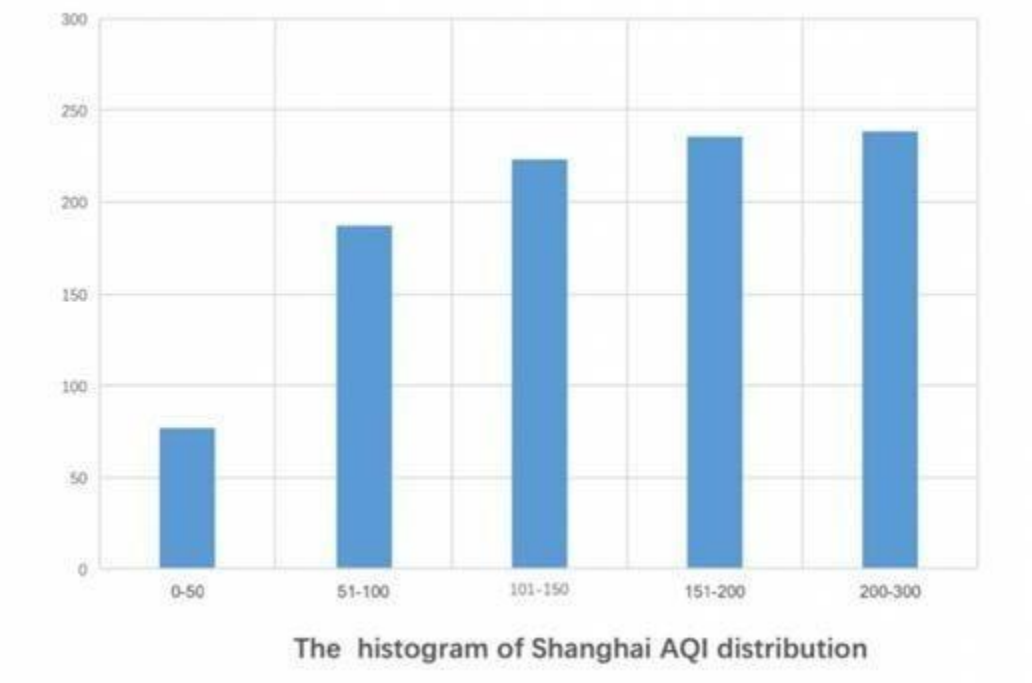


Figure 3.7 Placeholder: Chart showing average pollutant levels in different communities. Caption: Descriptive statistics applied to environmental data.

3.4.2 Use of Inferential Statistics in Predicting Health Outcomes

Inferential statistics allow researchers to generalise findings from samples to populations:

- Predicting disease risk from exposure to contaminated water.
- Estimating prevalence of respiratory illness from air quality samples.
- Using confidence intervals to assess reliability of predictions.

Fill in the blank: Inferential statistics allow researchers to _____ findings from samples to populations.

Predictive Goal	Statistical Method
Probability of Disease	**Logistic Regression**
Future Life Expectancy	**Survival Analysis (Kaplan-Meier)**
Strength of Risk Factors	**Correlation and Multiple Regression**
Group Comparison/Outcomes	**Chi-Square or ANOVA**

Box 3.4 Placeholder: Text box listing uses of inferential statistics in predicting health outcomes. Caption: Applications of inferential statistics in environmental health.

3.4.3 Case Examples Linking Hypothesis Testing to Environmental Health

- Testing whether exposure to industrial waste increases cancer risk.
- Comparing mean blood lead levels in children from different neighbourhoods.
- Using chi-square tests to examine association between contaminated water and diarrhoeal disease.

Fill in the blank: A chi-square test can be used to examine the association between contaminated water and _____ disease.

Research Scenario	Null Hypothesis (H0)	Alternative Hypothesis (H1)	Potential Statistical Test
Lead Exposure & IQ	**There is no relationship between blood lead levels and cognitive scores in children.**	**Higher blood lead levels are correlated with lower cognitive scores.**	**Correlation / Simple Linear Regression**
Air Quality & Asthma	**Living near a highway does not increase the frequency of asthma attacks.**	**Living near a highway is associated with more frequent asthma attacks.**	**T-Test (Comparing Highway vs. Rural groups)**
Water Safety	**The new filtration system does not change the level of arsenic in the community well.**	**The new filtration system significantly reduces arsenic levels.**	**Paired T-Test (Before vs. After filtration)**
Pesticides & Birth Weight	**Prenatal pesticide exposure has no effect on infant birth weight.**	**Prenatal pesticide exposure is linked to lower infant birth weight.**	**Multiple Regression (Adjusting for age/nutrition)**
Noise Pollution & Sleep	**High levels of airport noise do not impact average hours of REM sleep.**	**High levels of airport noise significantly decrease REM sleep hours.**	**ANOVA (Comparing Low, Med, High noise zones)**

Table 3.4 Placeholder: Table showing case examples of hypothesis testing in environmental health. Caption: Case examples of hypothesis testing in environmental health research.

Pilot questions 3.4

1. The most common value in a dataset, such as reported symptoms, is called the _____.

2. Inferential statistics allow researchers to _____ findings from samples to populations.
3. Give one example of how descriptive statistics are applied in environmental health research.
4. State one use of inferential statistics in predicting health outcomes.
5. A chi-square test can be used to examine the association between contaminated water and _____ disease.

Series 3 Summary – Descriptive and Inferential Statistics

In this Series, you learned how to **summarise data** and **draw conclusions** using descriptive and inferential statistics:

- **Part 3.1** introduced descriptive statistics, covering measures of central tendency (mean, median, mode) and variability (range, variance, standard deviation), and their applications in health research.
- **Part 3.2** explained inferential statistics, distinguishing between population parameters and sample statistics, and introduced estimation and confidence intervals.
- **Part 3.3** explored hypothesis testing, defining null and alternative hypotheses, outlining the steps in testing, and presenting common statistical tests (t-test, chi-square, ANOVA).
- **Part 3.4** applied these concepts to environmental health research, showing how descriptive and inferential statistics are used to summarise data, predict health outcomes, and test hypotheses in real-world case studies.

By completing this Series, you should now be able to:

- **Define and apply** descriptive statistics to summarise health data (SLO 3.1).
- **Explain and utilise** inferential statistics, including estimation and confidence intervals (SLO 3.2).
- **Describe and conduct** hypothesis testing, interpreting results with appropriate statistical tests (SLO 3.3).
- **Evaluate and apply** descriptive and inferential statistics in environmental health research (SLO 3.4).

Glossary of Terms – Series 3: Descriptive and Inferential Statistics

- **Descriptive statistics**: Methods used to summarise, organise, and present data without making predictions or generalisations.

- **Measures of central tendency**: Statistics that describe the centre of a dataset (mean, median, mode).
- **Mean**: The arithmetic average of a dataset.
- **Median**: The middle value when data are ordered.
- **Mode**: The most frequently occurring value in a dataset.
- **Measures of variability**: Statistics that describe the spread of data (range, variance, standard deviation).
- **Range**: The difference between the highest and lowest values in a dataset.
- **Variance**: The average of squared differences from the mean.
- **Standard deviation**: The square root of variance, showing the average distance of values from the mean.
- **Inferential statistics**: Methods used to make predictions or generalisations about a population based on sample data.
- **Population parameter**: A numerical value describing a characteristic of the entire population.
- **Sample statistic**: A numerical value describing a characteristic of a sample.
- **Estimation**: Using sample statistics to approximate population parameters.
- **Confidence interval (CI)**: A range of values within which the true population parameter is expected to lie, with a given level of confidence.
- **Hypothesis**: A statement or assumption about a population parameter that can be tested statistically.
- **Null hypothesis (H_0)**: Assumes no effect or difference.
- **Alternative hypothesis (H_1)**: Assumes there is an effect or difference.
- **Significance level (α)**: The threshold for deciding whether to reject the null hypothesis, commonly set at 0.05.
- **p-value**: The probability of obtaining results at least as extreme as those observed, assuming the null hypothesis is true.
- **t-test**: A statistical test comparing means between two groups.
- **Chi-square test**: A statistical test examining association between categorical variables.
- **ANOVA (Analysis of Variance)**: A statistical test comparing means across three or more groups.

Concept Check Questions – Series 3: Descriptive and Inferential Statistics

Objective questions (linked to SLOs 3.1–3.4)

1. Statistics that summarise and present data without making predictions are called _____ statistics. (SLO 3.1)
2. The most frequently occurring value in a dataset is called the _____. (SLO 3.1)
3. Statistics that make predictions or generalisations about a population are called _____ statistics. (SLO 3.2)
4. A range of values within which the true population parameter is expected to lie is called a _____ interval. (SLO 3.2)
5. The hypothesis that assumes no effect or difference is called the _____ hypothesis. (SLO 3.3)

Short-answer questions (linked to SLOs 3.1–3.4)

6. Distinguish between descriptive and inferential statistics with one example each. (SLO 3.1, 3.2)

7. Give one measure of central tendency and one measure of variability. (SLO 3.1)

8. Explain the difference between a population parameter and a sample statistic. (SLO 3.2)

9. State the common significance level used in hypothesis testing and explain its meaning. (SLO 3.3)

10. Provide one example of how hypothesis testing can be applied in environmental health research. (SLO 3.4)

Long-answer questions (linked to SLOs 3.1–3.4)

11. Analyse the importance of descriptive statistics in health research, highlighting how they help summarise and communicate findings. (SLO 3.1)

12. Discuss inferential statistics, explaining estimation and confidence intervals, and evaluate their role in generalising findings from samples to populations. (SLO 3.2)

13. Describe the steps in hypothesis testing, and evaluate the usefulness of t-test, chi-square, and ANOVA in health studies. (SLO 3.3)

14. Assess the applications of descriptive and inferential statistics in environmental health research, using case examples such as pollutant monitoring or disease risk prediction. (SLO 3.4)

Answers to Pilot Questions – Series 3: Descriptive and Inferential Statistics

Pilot questions 3.1 (Descriptive Statistics)

1. Descriptive
2. Mode
3. Range
4. Mean (or median, mode)
5. Summarising patient demographics (e.g., average age)

Pilot questions 3.2 (Inferential Statistics)

1. Inferential
2. Population parameter
3. Confidence interval
4. Population mean (or proportion, variance)
5. Predicting disease prevalence in a population

Pilot questions 3.3 (Hypothesis Testing)

1. Null hypothesis
2. 0.05
3. t-test
4. Chi-square test (or ANOVA, t-test)
5. Reject

Pilot questions 3.4 (Applications in Environmental Health Research)

1. Mode
2. Generalise
3. Summarising pollutant levels (e.g., average air quality index)
4. Estimating prevalence of respiratory illness from air samples
5. Diarrhoeal

References and Attributions – Series 3: Descriptive and Inferential Statistics

General references

- Course content adapted from *EHS 212: Biostatistics (2 Units)*.

- Concepts based on standard biostatistics, epidemiology, and environmental health research literature, as well as open educational resources.

Figures

- *Figure 3.1: Scope of descriptive statistics* – AI prompt: "Generate a diagram showing descriptive statistics summarising data with measures and charts."
- *Figure 3.3: Scope of inferential statistics* – AI prompt: "Generate a diagram showing inferential statistics making predictions from sample data to population."
- *Figure 3.4: Confidence interval illustration* – AI prompt: "Generate a diagram showing a sample mean with 95% confidence interval around it."
- *Figure 3.5: Types of hypotheses* – AI prompt: "Generate a diagram showing null hypothesis (no difference) vs. alternative hypothesis (difference)."
- *Figure 3.6: Decision rule in hypothesis testing* – AI prompt: "Generate a diagram showing decision rule: reject H_0 if p-value < α."
- *Figure 3.7: Descriptive statistics applied to environmental data* – AI prompt: "Generate a bar chart showing average pollutant levels across three communities."

Tables

- *Table 3.1: Measures of central tendency* – AI prompt: "Generate a table showing definitions and examples of mean, median, and mode."
- *Table 3.2: Population parameters vs. sample statistics* – AI prompt: "Generate a table comparing population parameters and sample statistics with examples."
- *Table 3.3: Common statistical tests in hypothesis testing* – AI prompt: "Generate a table listing t-test, chi-square, and ANOVA with their uses."
- *Table 3.4: Case examples of hypothesis testing in environmental health research* – AI prompt: "Generate a table showing case examples of hypothesis testing in environmental health research."

Boxes

- *Box 3.1: Applications of descriptive statistics in health research* – AI prompt: "Create a text box listing applications of descriptive statistics in health research."
- *Box 3.2: Applications of inferential statistics in health research* – AI prompt: "Create a text box listing applications of inferential statistics in health research."
- *Box 3.3: Steps in hypothesis testing* – AI prompt: "Create a text box listing the steps in hypothesis testing."

- *Box 3.4: Applications of inferential statistics in predicting health outcomes* – AI prompt: "Create a text box listing uses of inferential statistics in predicting health outcomes."

Suggested Outline – Series 4: Vital and Health Statistics in Environmental Health

This Series builds on descriptive and inferential statistics by focusing on **vital statistics** (births, deaths, marriages, etc.) and **health statistics** (morbidity, mortality, service utilisation), with direct application to environmental health.

Part 4.1 Vital Statistics

- 4.1.1 Definition and scope of vital statistics
- 4.1.2 Sources of vital statistics (civil registration, surveys, censuses)
- 4.1.3 Uses of vital statistics in health planning and research
- 4.1.4 Limitations and challenges in collecting vital statistics

Part 4.2 Health Statistics

- 4.2.1 Definition and scope of health statistics
- 4.2.2 Types of health statistics (morbidity, mortality, service utilisation)
- 4.2.3 Sources of health statistics (hospital records, surveillance systems, surveys)
- 4.2.4 Applications of health statistics in policy and decision-making

Part 4.3 Statistical Indicators in Environmental Health

- 4.3.1 Mortality indicators (infant mortality rate, maternal mortality ratio)
- 4.3.2 Morbidity indicators (incidence, prevalence)
- 4.3.3 Environmental health indicators (water quality, air pollution, sanitation coverage)
- 4.3.4 Interpretation of indicators for community health assessment

Part 4.4 Applications in Environmental Health Research

- 4.4.1 Use of vital statistics in monitoring population health trends
- 4.4.2 Use of health statistics in evaluating environmental interventions
- 4.4.3 Case examples: linking environmental exposures to morbidity and mortality
- 4.4.4 Challenges and opportunities in applying statistics to environmental health

Introduction – Series 4: Vital and Health Statistics in Environmental Health

Background Vital and health statistics are essential tools for understanding population health. Vital statistics record fundamental life events such as births, deaths, and marriages, while health statistics capture morbidity, mortality, and service utilisation. Together, they provide the foundation for monitoring health trends, evaluating interventions, and guiding policy decisions. In environmental health, these statistics are particularly important for linking environmental exposures to health outcomes.

Aim The aim of this Series is to enable learners to define, interpret, and apply vital and health statistics, understand their sources and limitations, and use statistical indicators to assess environmental health conditions.

Sequential synopsis

- **Part 4.1** introduces vital statistics, their sources, uses, and limitations.
- **Part 4.2** explains health statistics, including morbidity, mortality, and service utilisation data.
- **Part 4.3** presents statistical indicators relevant to environmental health, such as mortality, morbidity, and environmental indicators.
- **Part 4.4** applies these concepts to environmental health research, with case examples linking exposures to health outcomes.

Series Learning Outcomes – Series 4: Vital and Health Statistics in Environmental Health

By the end of this Series, learners should be able to:

- **SLO 4.1**: **Define and explain** vital statistics, identify their sources, uses, and limitations in health research.
- **SLO 4.2**: **Describe and apply** health statistics, including morbidity, mortality, and service utilisation data, and evaluate their role in policy and decision-making.
- **SLO 4.3**: **Interpret and utilise** statistical indicators in environmental health, such as mortality, morbidity, and environmental indicators, for community health assessment.
- **SLO 4.4**: **Evaluate and apply** vital and health statistics in environmental health research, using case examples to link environmental exposures to health outcomes.

4.1 Vital Statistics

4.1.1 Definition and Scope of Vital Statistics

Vital statistics are data related to vital events in human life such as births, deaths, marriages, divorces, and migration. They provide essential information for population health monitoring and planning.

Fill in the blank: Data related to births, deaths, and marriages are called _____ statistics.

Figure 4.1 Placeholder: Diagram showing vital statistics as records of life events. Caption: Scope of vital statistics. AI prompt suggestion: "Generate a diagram showing vital statistics including births, deaths, marriages, and migration."

4.1.2 Sources of Vital Statistics

- **Civil registration systems**: Continuous recording of births, deaths, marriages, etc.
- **Population censuses**: Conducted periodically to capture demographic data.
- **Surveys**: Household or demographic health surveys.
- **Hospital and health facility records**: Provide data on births and deaths.

Fill in the blank: Continuous recording of births and deaths is done through _____ registration systems.

Table 4.1 Placeholder: Table listing sources of vital statistics with examples. Caption: Sources of vital statistics. AI prompt suggestion: "Generate a table listing sources of vital statistics with examples."

4.1.3 Uses of Vital Statistics in Health Planning and Research

- Monitoring population growth and demographic changes.
- Assessing fertility and mortality rates.
- Planning health services and resource allocation.
- Evaluating the impact of health interventions.

Fill in the blank: Vital statistics are used to monitor population _____ and demographic changes.

Vital Event	Research/Planning Use	Impact on Health Policy
Birth Certificate	**Tracks prenatal care and birth weight.**	**Informs funding for maternal health clinics.**

Vital Event	Research/Planning Use	Impact on Health Policy
Death Certificate	**Records primary and contributing causes.**	**Directs funding toward specific disease research.**
Fetal Death Report	**Monitors environmental and maternal risks.**	**Leads to regulations on workplace toxins.**
Marriage License	**Studies social support and health outcomes.**	**Shapes family-based health insurance policies.**

Box 4.1 Placeholder: Text box listing uses of vital statistics in health planning and research. Caption: Uses of vital statistics.

4.1.4 Limitations and Challenges in Collecting Vital Statistics

- Incomplete or inaccurate registration systems.
- Underreporting of births and deaths, especially in rural areas.
- Lack of trained personnel for data collection.
- Cultural and social barriers to reporting vital events.

Fill in the blank: A major challenge in collecting vital statistics is _____ reporting of births and deaths.

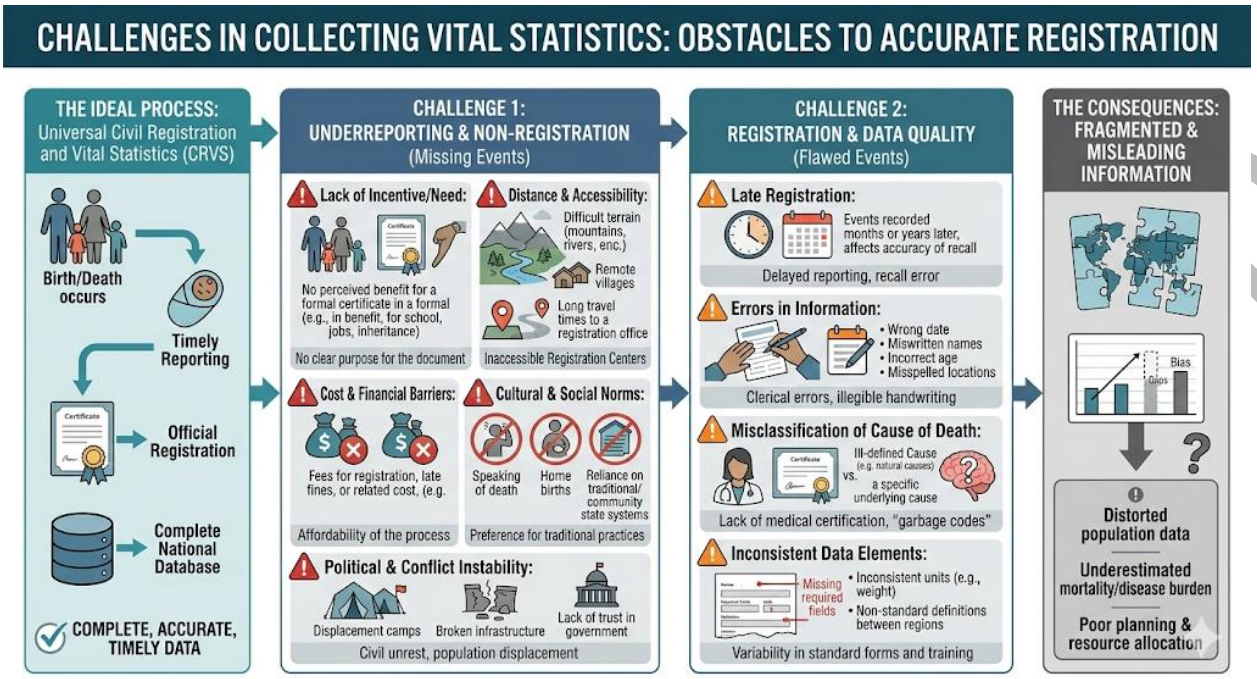


Figure 4.2 Placeholder: Diagram showing challenges in collecting vital statistics. Caption: Limitations of vital statistics collection.

Pilot questions 4.1

1. Data related to births, deaths, and marriages are called _____ statistics.
2. Continuous recording of births and deaths is done through _____ registration systems.
3. Vital statistics are used to monitor population _____ and demographic changes.
4. Name one source of vital statistics.
5. A major challenge in collecting vital statistics is _____ reporting of births and deaths.

4.2 Health Statistics

4.2.1 Definition and Scope of Health Statistics

Health statistics are numerical data that describe the health status of populations. They include information on morbidity (illness), mortality (death), and service utilisation (use of health facilities). These statistics are essential for monitoring health trends and guiding policy decisions.

Fill in the blank: Numerical data that describe the health status of populations are called _____ statistics.

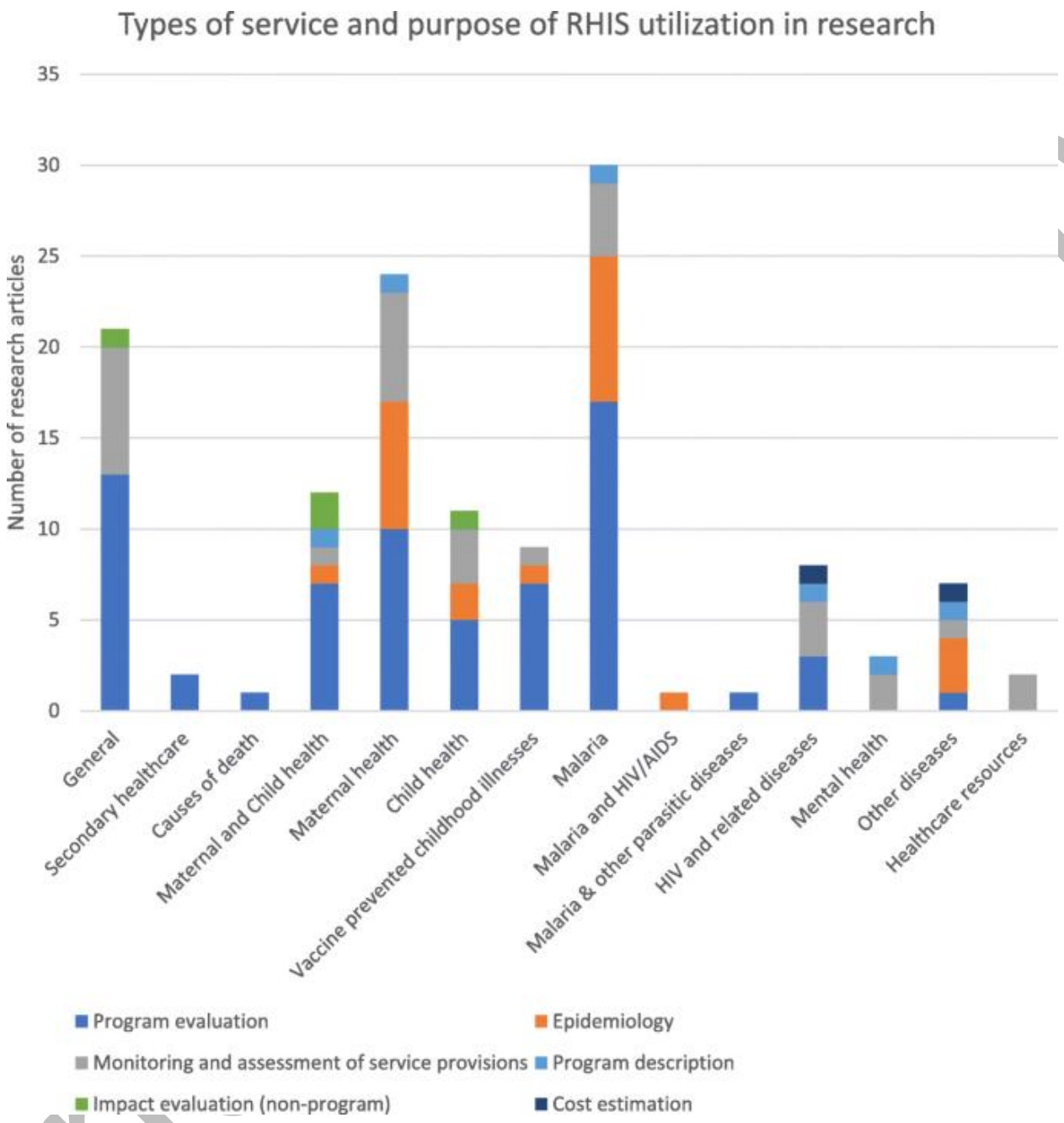


Figure 4.3 Placeholder: Diagram showing health statistics categories (morbidity, mortality, service utilisation). Caption: Scope of health statistics.

4.2.2 Types of Health Statistics

- **Morbidity statistics**: Data on incidence and prevalence of diseases.
- **Mortality statistics**: Data on causes and rates of death.
- **Service utilisation statistics**: Data on hospital admissions, outpatient visits, and use of health services.

Fill in the blank: Data on incidence and prevalence of diseases are called _____ statistics.

Category	Definition	Key Examples	Use in Health Research
Morbidity	The state of being symptomatic or unhealthy with a disease or condition.	Prevalence: Total number of people living with HIV in 2026. Incidence: Number of new COVID-19 cases diagnosed this week.	Measures the "burden of disease" and helps in planning long-term care and medication supplies.
Mortality	The measure of the number of deaths in a particular population.	Infant Mortality Rate (IMR): Deaths of children under 1 year per 1,000 live births. Case Fatality Rate: Percentage of people with Ebola who die from it.	Identifies the most lethal threats to a population and evaluates the effectiveness of life-saving interventions.
Service Utilisation	The extent to which a population uses available health services.	Bed Occupancy Rate: Percentage of hospital beds in use. Antenatal Care Coverage: Number of pregnant women	Identifies barriers to access (like cost or distance) and measures the efficiency of hospital management.

Category	Definition	Key Examples	Use in Health Research
		receiving at least four check-ups.	

Table 4.2 Placeholder: Table listing types of health statistics with examples. Caption: Types of health statistics.

4.2.3 Sources of Health Statistics

- **Hospital records**: Admissions, discharges, causes of death.
- **Health surveys**: Household surveys, demographic and health surveys.
- **Disease surveillance systems**: Monitoring communicable and non-communicable diseases.
- **National health information systems**: Aggregated data for policy and planning.

Fill in the blank: Monitoring communicable and non-communicable diseases is done through disease _____ systems.

Source	Primary Data Type	Main Limitation
Vital Statistics	**Mortality and Births**	**Underreporting in developing regions.**
Surveys	**Behavior and Prevalence**	**Self-reporting bias (people may not be honest).**
Hospital Records	**Morbidity and Treatment**	**Only captures those with access to care.**
Census	**Demographics**	**Becomes outdated quickly between cycles.**

Box 4.2 Placeholder: Text box listing sources of health statistics. Caption: Sources of health statistics

4.2.4 Applications of Health Statistics in Policy and Decision-Making

- Identifying disease burden and health priorities.
- Evaluating effectiveness of health interventions.
- Allocating resources to health facilities.
- Supporting evidence-based health policy.

Fill in the blank: Health statistics are used to identify disease _____ and health priorities.

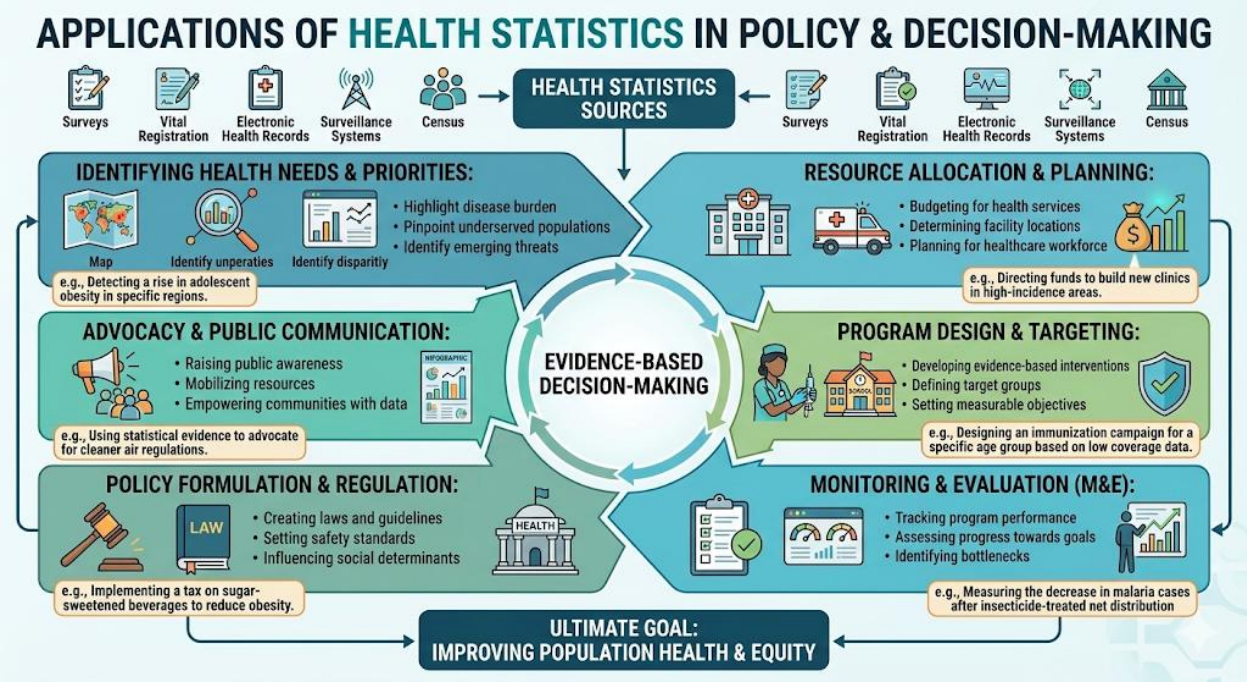


Figure 4.4 Placeholder: Diagram showing applications of health statistics in policy. Caption: Applications of health statistics.

Pilot questions 4.2

1. Numerical data that describe the health status of populations are called _____ statistics.
2. Data on incidence and prevalence of diseases are called _____ statistics.
3. Monitoring communicable and non-communicable diseases is done through disease _____ systems.
4. Name one source of health statistics.
5. Health statistics are used to identify disease _____ and health priorities.

4.3 Statistical Indicators in Environmental Health

4.3.1 Mortality Indicators

Mortality indicators measure death rates in a population and are key for assessing health status.

- **Infant mortality rate (IMR)**: Number of deaths of infants under one year per 1,000 live births.
- **Maternal mortality ratio (MMR)**: Number of maternal deaths per 100,000 live births.
- **Crude death rate (CDR)**: Number of deaths per 1,000 population in a given year.

Fill in the blank: The number of deaths of infants under one year per 1,000 live births is called the _____ mortality rate.

Indicator	Definition	Formula	Significance
Infant Mortality Rate (IMR)	**The probability of a child dying before their first birthday.**	**$$\frac{\text{Deaths under 1 year of age}}{\text{Number of live births}} \times 1,000$$**	**A primary indicator of a nation's overall health and sanitation status.**
Maternal Mortality Ratio (MMR)	**The number of women who die from pregnancy-related causes while pregnant or within 42 days of termination.**	**$$\frac{\text{Maternal deaths}}{\text{Number of live births}} \times 100,000$$**	**Measures the quality and accessibility of obstetric and reproductive healthcare.**
Crude Death Rate (CDR)	**The total number of deaths in a population during a specified period.**	**$$\frac{\text{Total deaths during the year}}{\text{Estimated mid-year population}} \times 1,000$$**	**Provides a general snapshot of the mortality level in a population without adjusting for age.**

Table 4.3 Placeholder: Table listing mortality indicators with definitions and formulas. Caption: Mortality indicators in environmental health.

4.3.2 Morbidity Indicators

Morbidity indicators measure illness in a population.

- **Incidence rate**: Number of new cases of a disease in a population during a specific period.
- **Prevalence rate**: Total number of cases (new and existing) in a population at a given time.

Fill in the blank: The number of new cases of a disease in a population during a specific period is called the _____ rate.

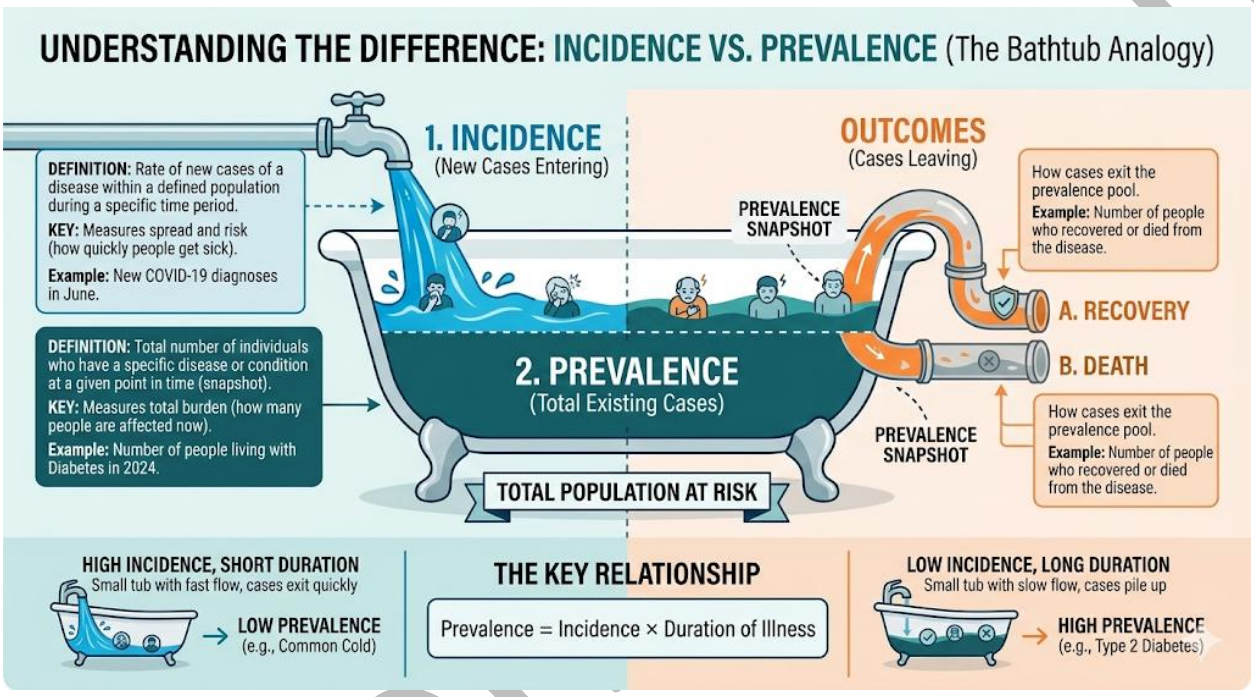


Figure 4.5 Placeholder: Diagram showing incidence vs. prevalence. Caption: Morbidity indicators. AI prompt suggestion:

4.3.3 Environmental Health Indicators

These indicators measure environmental conditions that affect health:

- **Water quality indicators**: Proportion of households with access to safe drinking water.
- **Air pollution indicators**: Average concentration of pollutants (e.g., PM2.5).
- **Sanitation coverage**: Percentage of households with improved sanitation facilities.

Fill in the blank: The percentage of households with improved sanitation facilities is called _____ coverage.

Indicator	Targeted Health Outcome	Typical Data Source
$PM_{2.5}$ Concentration	**Lung Cancer & Chronic Bronchitis**	**Environmental Sensors**
Improved Water Access	**Diarrheal Disease Reduction**	**Household Surveys (DHS)**
Lead in Soil/Dust	**Cognitive Impairment in Children**	**Blood Screening Programs**
Solid Fuel Use	**Acute Respiratory Infections (ARI)**	**Census Data**

Box 4.3 Placeholder: Text box listing environmental health indicators. Caption: Environmental health indicators.

4.3.4 Interpretation of Indicators for Community Health Assessment

- Mortality indicators show the burden of death and highlight vulnerable groups.
- Morbidity indicators reveal disease patterns and help track outbreaks.
- Environmental indicators identify risks from unsafe water, poor sanitation, and polluted air.
- Together, these indicators provide a comprehensive picture of community health.

Fill in the blank: Mortality indicators highlight vulnerable groups by measuring the burden of _____.

Phase of Assessment	Indicator Type	Example Data Point	Actionable Policy
Exposure	**Environmental**	**High levels of lead in school water pipes.**	**Mandate pipe replacement and filtration.**

Phase of Assessment	Indicator Type	Example Data Point	Actionable Policy
Current Impact	**Morbidity**	**Spike in "learning disability" diagnoses in the area.**	**Increase funding for special education.**
Final Result	**Mortality**	**Elevated mortality rates from associated kidney issues.**	**Launch a long-term longitudinal health study.**

Figure 4.6 Placeholder: Diagram showing combined use of mortality, morbidity, and environmental indicators in community health assessment. Caption: Interpretation of indicators for community health assessment.

Pilot questions 4.3

1. The number of deaths of infants under one year per 1,000 live births is called the _____ mortality rate.
2. The number of new cases of a disease in a population during a specific period is called the _____ rate.
3. The percentage of households with improved sanitation facilities is called _____ coverage.
4. Name one environmental health indicator.
5. Mortality indicators highlight vulnerable groups by measuring the burden of _____.

4.4 Applications in Environmental Health Research

4.4.1 Use of Vital Statistics in Monitoring Population Health Trends

Vital statistics provide insights into population dynamics and health outcomes:

- Birth and death records help track population growth.
- Infant and maternal mortality rates highlight vulnerable groups.
- Migration data reveal shifts that affect health service needs.

Fill in the blank: Infant and maternal mortality rates highlight _____ groups in a population.

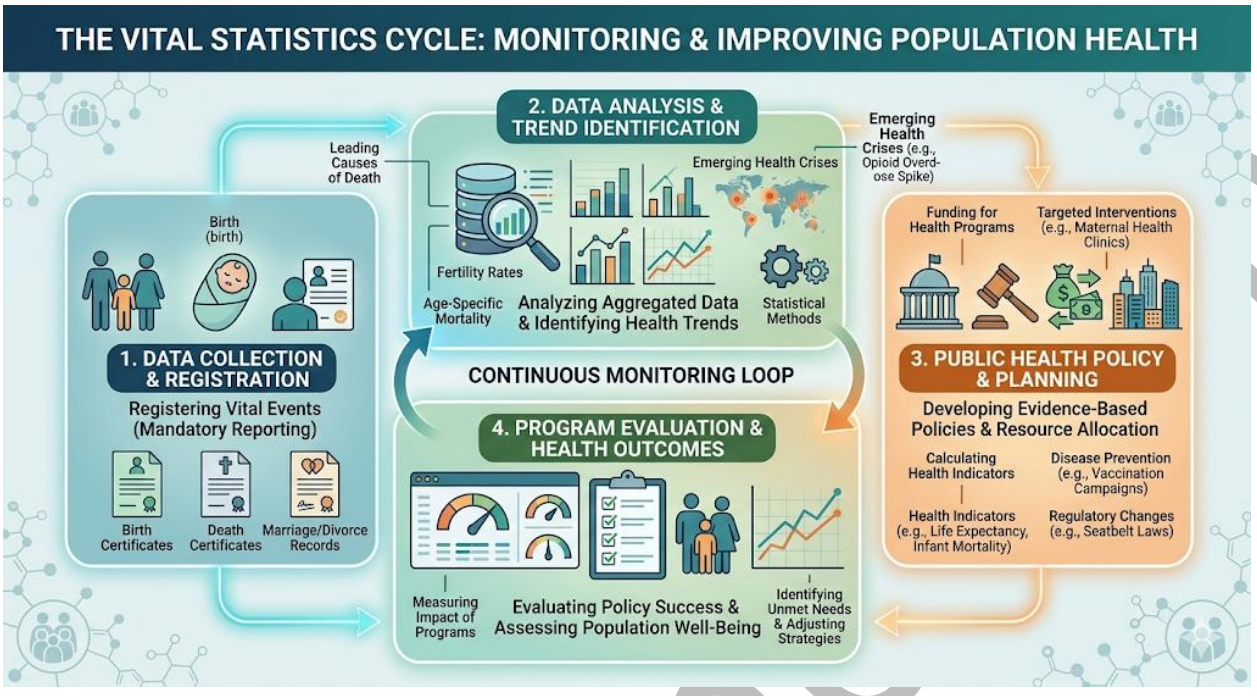


Figure 4.7 Placeholder: Diagram showing how vital statistics monitor population health trends. Caption: Applications of vital statistics in monitoring health.

4.4.2 Use of Health Statistics in Evaluating Environmental Interventions

Health statistics measure the impact of environmental health programs:

- Morbidity data show reductions in disease after sanitation improvements.
- Mortality data reveal outcomes of clean water initiatives.
- Service utilisation data indicate increased access to health facilities.

Fill in the blank: Morbidity data show reductions in _____ after sanitation improvements.

Environmental Intervention	Statistical Metric for Evaluation	Success Indicator
New Sewage System	**Incidence of Cholera/Dysentery**	**Significant decrease in new cases post-construction.**
Lead Paint Ban	**Mean Blood Lead Levels (BLL) in children**	**A downward shift in the population mean BLL.**

Environmental Intervention	Statistical Metric for Evaluation	Success Indicator
Urban Reforestation	**Heat-Related Mortality Rates**	**Lower spike in deaths during summer heatwaves.**
Industrial Scrubbers	**$PM_{2.5}$ concentration vs. COPD admissions**	**Strong correlation between lower smoke and fewer hospitalizations.**

Box 4.4 Placeholder: Text box listing uses of health statistics in evaluating environmental interventions. Caption: Applications of health statistics in environmental interventions.

4.4.3 Case Examples: Linking Environmental Exposures to Morbidity and Mortality

- Exposure to polluted air linked to higher incidence of respiratory diseases.
- Contaminated water associated with diarrhoeal disease outbreaks.
- Poor sanitation correlated with increased child mortality.

Fill in the blank: Contaminated water is often associated with outbreaks of _____ disease.

Environmental Exposure	Morbidity (Illness/Condition)	Mortality (Primary Cause of Death)	Key Statistical Indicator
Air Pollution ($PM_{2.5}$)	**Chronic Bronchitis, Pediatric Asthma, Reduced Lung Function.**	**Lung Cancer, Ischemic Heart Disease, Stroke.**	**Age-Specific Mortality Rate for respiratory diseases.**
Contaminated Water	**Cholera, Giardiasis, Chronic Arsenicosis (skin lesions).**	**Severe Dehydration (diarrheal disease), Organ Failure.**	**Incidence Rate of waterborne pathogens.**
Lead Exposure	**Cognitive Impairment, Anemia, Hypertension.**	**Cardiovascular Disease, Renal (Kidney) Failure.**	**Mean Blood Lead Levels (BLL) in children under 5.**

Environmental Exposure	Morbidity (Illness/Condition)	Mortality (Primary Cause of Death)	Key Statistical Indicator
Extreme Heat	**Heat Exhaustion, Dehydration, Kidney Stones.**	**Heatstroke, Acute Myocardial Infarction (Heart Attack).**	**Excess Mortality during heatwave periods.**
Occupational Asbestos	**Asbestosis, Pleural Thickening.**	**Mesothelioma (a specific lung cancer).**	**Case Fatality Rate for pleural cancers.**

Table 4.4 Placeholder: Table showing case examples linking environmental exposures to morbidity and mortality. Caption: Case examples of environmental exposures and health outcomes.

4.4.4 Challenges and Opportunities in Applying Statistics to Environmental Health

- **Challenges**: Underreporting, incomplete data, lack of standardisation.
- **Opportunities**: Improved surveillance systems, digital health records, integration of environmental and health data.
- Statistics can guide targeted interventions and policy reforms.

Fill in the blank: A major challenge in applying statistics to environmental health is _____ data.

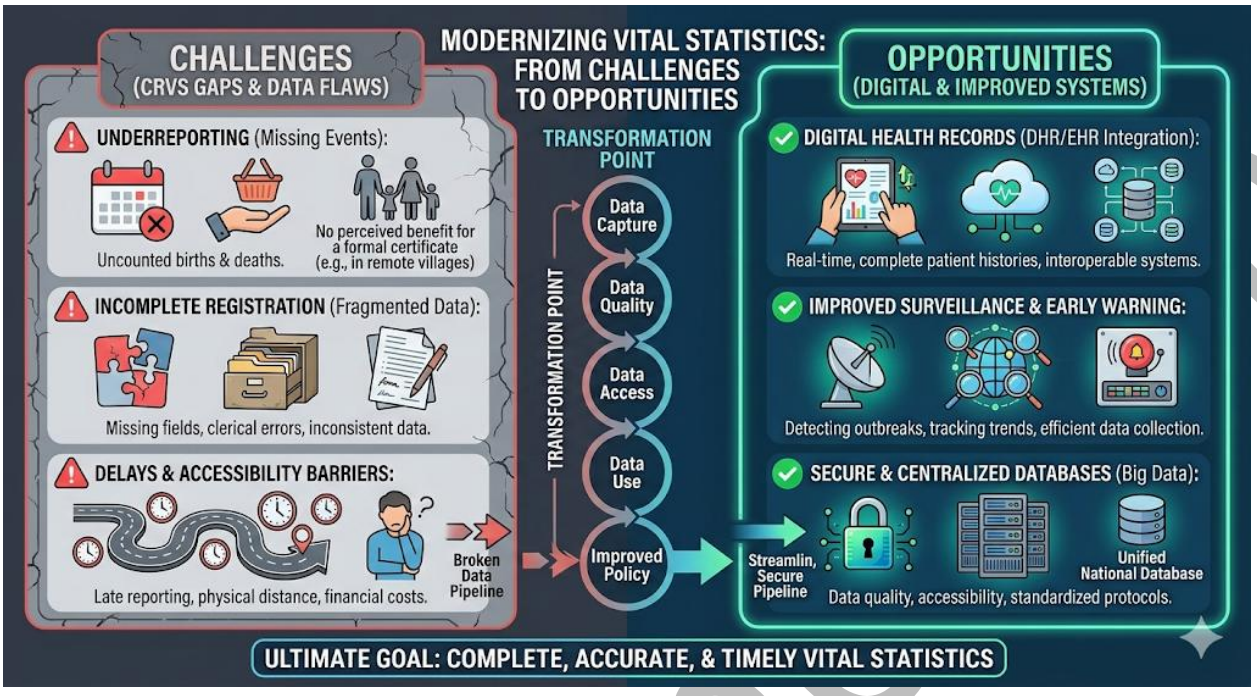


Figure 4.8 Placeholder: Diagram showing challenges and opportunities in applying statistics to environmental health. Caption: Challenges and opportunities in applying statistics.

Pilot questions 4.4

1. Infant and maternal mortality rates highlight _____ groups in a population.
2. Morbidity data show reductions in _____ after sanitation improvements.
3. Contaminated water is often associated with outbreaks of _____ disease.
4. Name one challenge in applying statistics to environmental health.
5. A major challenge in applying statistics to environmental health is _____ data.

Series 4 Summary – Vital and Health Statistics in Environmental Health

In this Series, you explored how **vital and health statistics** serve as the backbone of population health monitoring and environmental health research:

- **Part 4.1** defined vital statistics, explained their sources (civil registration, censuses, surveys), highlighted their uses in health planning, and discussed challenges such as underreporting.
- **Part 4.2** introduced health statistics, covering morbidity, mortality, and service utilisation data, their sources (hospital records, surveys, surveillance systems), and applications in policy and decision-making.

- **Part 4.3** presented statistical indicators in environmental health, including mortality (IMR, MMR, CDR), morbidity (incidence, prevalence), and environmental indicators (water quality, air pollution, sanitation coverage), and explained their interpretation for community health assessment.
- **Part 4.4** applied these concepts to environmental health research, showing how vital and health statistics are used to monitor trends, evaluate interventions, link exposures to outcomes, and address challenges in data collection.

By completing this Series, you should now be able to:

- **Define and explain** vital statistics, their sources, uses, and limitations (SLO 4.1).
- **Describe and apply** health statistics in policy and decision-making (SLO 4.2).
- **Interpret and utilise** statistical indicators in environmental health (SLO 4.3).
- **Evaluate and apply** vital and health statistics in environmental health research (SLO 4.4).

This Series connects demographic and health data directly to **environmental health practice**, preparing learners to use statistical evidence in assessing risks, planning interventions, and shaping policy.

Glossary of Terms – Series 4: Vital and Health Statistics in Environmental Health

- **Vital statistics**: Data related to vital events in human life such as births, deaths, marriages, divorces, and migration.
- **Civil registration system**: Continuous recording of vital events like births and deaths by government authorities.
- **Population census**: A periodic count of the population, including demographic and social characteristics.
- **Health statistics**: Numerical data describing the health status of populations, including morbidity, mortality, and service utilisation.
- **Morbidity statistics**: Data on the incidence (new cases) and prevalence (total cases) of diseases.
- **Mortality statistics**: Data on causes and rates of death in a population.
- **Service utilisation statistics**: Data on the use of health services, such as hospital admissions and outpatient visits.
- **Disease surveillance system**: Continuous monitoring of communicable and non-communicable diseases.
- **Infant mortality rate (IMR)**: Number of deaths of infants under one year per 1,000 live births.

- **Maternal mortality ratio (MMR)**: Number of maternal deaths per 100,000 live births.
- **Crude death rate (CDR)**: Number of deaths per 1,000 population in a given year.
- **Incidence rate**: Number of new cases of a disease in a population during a specific period.
- **Prevalence rate**: Total number of cases (new and existing) in a population at a given time.
- **Water quality indicator**: Proportion of households with access to safe drinking water.
- **Air pollution indicator**: Average concentration of pollutants (e.g., PM2.5) in the air.
- **Sanitation coverage**: Percentage of households with improved sanitation facilities.

Concept Check Questions – Series 4: Vital and Health Statistics in Environmental Health

Objective questions (linked to SLOs 4.1–4.4)

1. Data related to births, deaths, and marriages are called _____ statistics. (SLO 4.1)
2. Continuous recording of vital events is done through _____ registration systems. (SLO 4.1)
3. Numerical data describing the health status of populations are called _____ statistics. (SLO 4.2)
4. The number of deaths of infants under one year per 1,000 live births is called the _____ mortality rate. (SLO 4.3)
5. The percentage of households with improved sanitation facilities is called _____ coverage. (SLO 4.3)

Short-answer questions (linked to SLOs 4.1–4.4)

6. Distinguish between vital statistics and health statistics with one example each. (SLO 4.1, 4.2)
7. Name two sources of vital statistics. (SLO 4.1)
8. Give one example of morbidity statistics and one example of mortality statistics. (SLO 4.2)
9. Explain the difference between incidence rate and prevalence rate. (SLO 4.3)
10. State one challenge in applying statistics to environmental health research. (SLO 4.4)

Long-answer questions (linked to SLOs 4.1–4.4)

11. Analyse the importance of vital statistics in health planning, highlighting their uses and limitations. (SLO 4.1)

12. Discuss health statistics, explaining morbidity, mortality, and service utilisation, and evaluate their role in policy and decision-making. (SLO 4.2)

13. Describe mortality, morbidity, and environmental indicators, and assess their usefulness in community health assessment. (SLO 4.3)

14. Evaluate the applications of vital and health statistics in environmental health research, using case examples such as polluted air, contaminated water, or poor sanitation. (SLO 4.4)

Answers to Pilot Questions – Series 4: Vital and Health Statistics in Environmental Health

Pilot questions 4.1 (Vital Statistics)

1. Vital
2. Civil
3. Growth
4. Civil registration system / Population census / Surveys (any one)
5. Underreporting

Pilot questions 4.2 (Health Statistics)

1. Health
2. Morbidity
3. Surveillance
4. Hospital records / Health surveys / Disease surveillance systems (any one)
5. Burden

Pilot questions 4.3 (Statistical Indicators in Environmental Health)

1. Infant
2. Incidence
3. Sanitation
4. Water quality / Air pollution / Sanitation coverage (any one)

5. Death

Pilot questions 4.4 (Applications in Environmental Health Research)

1. Vulnerable
2. Disease
3. Diarrhoeal
4. Underreporting / Incomplete data / Lack of standardisation (any one)
5. Incomplete

References and Attributions – Series 4: Vital and Health Statistics in Environmental Health

General references

- Course content adapted from *EHS 212: Biostatistics (2 Units)*.
- Concepts based on standard biostatistics, epidemiology, demography, and environmental health literature, as well as open educational resources.

Figures

- *Figure 4.1: Scope of vital statistics* – AI prompt: "Generate a diagram showing vital statistics including births, deaths, marriages, and migration."
- *Figure 4.2: Limitations of vital statistics collection* – AI prompt: "Generate a diagram showing challenges in collecting vital statistics such as underreporting and incomplete registration."
- *Figure 4.3: Scope of health statistics* – AI prompt: "Generate a diagram showing health statistics categories: morbidity, mortality, and service utilisation."
- *Figure 4.4: Applications of health statistics* – AI prompt: "Generate a diagram showing applications of health statistics in policy and decision-making."
- *Figure 4.5: Morbidity indicators* – AI prompt: "Generate a diagram showing incidence as new cases and prevalence as total cases in a population."
- *Figure 4.6: Interpretation of indicators for community health assessment* – AI prompt: "Generate a diagram showing combined use of mortality, morbidity, and environmental indicators for community health assessment."
- *Figure 4.7: Applications of vital statistics in monitoring health* – AI prompt: "Generate a diagram showing vital statistics used to monitor births, deaths, and migration trends."

- *Figure 4.8: Challenges and opportunities in applying statistics* – AI prompt: "Generate a diagram showing challenges (underreporting, incomplete data) and opportunities (digital health records, improved surveillance)."

Tables

- *Table 4.1: Sources of vital statistics* – AI prompt: "Generate a table listing sources of vital statistics with examples."
- *Table 4.2: Types of health statistics* – AI prompt: "Generate a table listing morbidity, mortality, and service utilisation statistics with examples."
- *Table 4.3: Mortality indicators in environmental health* – AI prompt: "Generate a table listing infant mortality rate, maternal mortality ratio, and crude death rate with definitions and formulas."
- *Table 4.4: Case examples of environmental exposures and health outcomes* – AI prompt: "Generate a table showing case examples linking environmental exposures to morbidity and mortality."

Boxes

- *Box 4.1: Uses of vital statistics* – AI prompt: "Create a text box listing uses of vital statistics in health planning and research."
- *Box 4.2: Sources of health statistics* – AI prompt: "Create a text box listing sources of health statistics."
- *Box 4.3: Environmental health indicators* – AI prompt: "Create a text box listing environmental health indicators such as water quality, air pollution, and sanitation coverage."
- *Box 4.4: Applications of health statistics in environmental interventions* – AI prompt: "Create a text box listing uses of health statistics in evaluating environmental interventions."

Suggested Outline – Series 5: Statistical Packages and Calculation of Rates

This Series introduces learners to **statistical software tools** and the practical computation of rates commonly used in health and environmental research. It bridges theory with applied skills, ensuring learners can handle data analysis efficiently.

Part 5.1 Introduction to Statistical Packages

- 5.1.1 Definition and purpose of statistical packages
- 5.1.2 Commonly used packages (SPSS, STATA, R, Epi Info, Excel)
- 5.1.3 Advantages and limitations of statistical packages
- 5.1.4 Applications in health and environmental research

Part 5.2 Basic Operations in Statistical Packages

- 5.2.1 Data entry and management
- 5.2.2 Coding and recoding variables
- 5.2.3 Generating descriptive statistics (mean, median, mode, frequency tables)
- 5.2.4 Producing charts and graphs

Part 5.3 Calculation of Rates in Health Statistics

- 5.3.1 Crude rates (birth rate, death rate)
- 5.3.2 Specific rates (age-specific, cause-specific)
- 5.3.3 Standardised rates (direct and indirect standardisation)
- 5.3.4 Examples of rate calculation in environmental health contexts

Part 5.4 Applications in Environmental Health Research

- 5.4.1 Using statistical packages to analyse environmental health data
- 5.4.2 Calculating morbidity and mortality rates linked to environmental exposures
- 5.4.3 Case examples: waterborne diseases, air pollution, sanitation coverage
- 5.4.4 Challenges and opportunities in applying statistical packages in low-resource settings

Introduction – Series 5: Statistical Packages and Calculation of Rates

Background Modern health and environmental research relies heavily on statistical software to manage, analyse, and interpret data. Statistical packages such as SPSS, STATA, R, Epi Info, and Excel provide powerful tools for handling large datasets, generating descriptive and inferential statistics, and producing visualisations. Alongside these tools,

the calculation of rates—such as birth rates, death rates, incidence, and prevalence—is fundamental for assessing population health and environmental exposures.

Aim The aim of this Series is to equip learners with practical knowledge of statistical packages and the skills to calculate and interpret health and environmental rates. This ensures they can apply statistical methods to real-world research and policy contexts.

Sequential synopsis

- **Part 5.1** introduces statistical packages, their purpose, common examples, advantages, limitations, and applications.
- **Part 5.2** explains basic operations in statistical packages, including data entry, coding, descriptive statistics, and visualisation.
- **Part 5.3** focuses on the calculation of rates in health statistics, covering crude, specific, and standardised rates with examples.
- **Part 5.4** applies statistical packages and rate calculations to environmental health research, highlighting case examples and challenges.

Series Learning Outcomes – Series 5: Statistical Packages and Calculation of Rates

By the end of this Series, learners should be able to:

- **SLO 5.1**: **Define and explain** statistical packages, identify commonly used examples, and evaluate their advantages and limitations in health and environmental research.
- **SLO 5.2**: **Demonstrate basic operations** in statistical packages, including data entry, coding, descriptive statistics, and visualisation.
- **SLO 5.3**: **Calculate and interpret** crude, specific, and standardised rates in health statistics, with examples relevant to environmental health.
- **SLO 5.4**: **Apply statistical packages and rate calculations** to environmental health research, using case examples to analyse exposures and outcomes, and assess challenges in low-resource settings.

5.1 Introduction to Statistical Packages

5.1.1 Definition and Purpose of Statistical Packages

Statistical packages are software tools designed to manage, analyse, and interpret data. They simplify complex calculations, generate descriptive and inferential statistics, and produce charts and graphs for data presentation.

Fill in the blank: Software tools designed to manage and analyse data are called _____ packages.

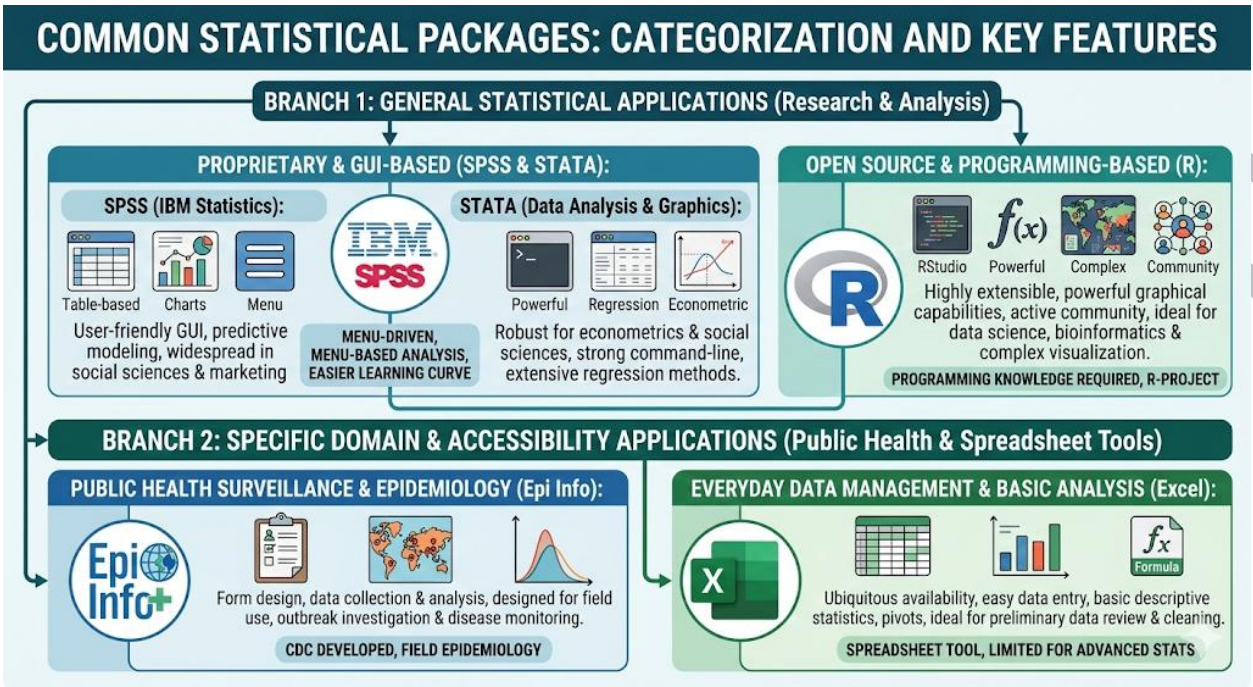


Figure 5.1 Placeholder: Diagram showing examples of statistical packages. Caption: Scope of statistical packages.

5.1.2 Commonly Used Packages

- **SPSS**: Widely used in social sciences and health research.
- **STATA**: Known for econometrics and epidemiology applications.
- **R**: Open-source, powerful for advanced statistical modelling.
- **Epi Info**: Free software developed by CDC for epidemiology.
- **Excel**: Common spreadsheet tool with basic statistical functions.

Fill in the blank: A free statistical package developed by CDC for epidemiology is called _____ Info.

Software	Primary Features	Key Applications	User Level
SPSS (IBM)	**Menu-driven interface, robust data management, "Point-and-click" analysis.**	**Social sciences, market research, and clinical trials.**	**Beginner to Intermediate**

Software	Primary Features	Key Applications	User Level
STATA	**Command-line driven, excellent for longitudinal/panel data, publication-quality graphics.**	**Econometrics, biostatistics, and large-scale public health surveys.**	**Intermediate to Advanced**
R (Open Source)	**Programming-based, highly extensible (thousands of packages), superior data visualization.**	**Bioinformatics, advanced data modeling, and academic research.**	**Advanced (Coding required)**
Epi Info (CDC)	**Free, specialized for outbreak investigation, customizable data entry forms, mapping.**	**Field epidemiology, disaster response, and local health department surveillance.**	**Beginner to Intermediate**
MS Excel	**Spreadsheet-based, ubiquitous, excellent for data cleaning, basic pivots, and charts.**	**Preliminary data review, budgeting, and simple descriptive reporting.**	**Beginner**

Table 5.1 Placeholder: Table listing statistical packages with features and applications. Caption: Commonly used statistical packages.

5.1.3 Advantages and Limitations of Statistical Packages

Advantages

- Handle large datasets efficiently.
- Provide user-friendly interfaces.
- Offer advanced statistical functions.
- Generate visual outputs (charts, graphs).

Limitations

- Cost (some packages are expensive).
- Require training and technical skills.

- Dependence on computer access.
- Limited functions in basic tools like Excel.

Fill in the blank: One limitation of statistical packages is that some are _____ to purchase.

Feature	Excel	SPSS/STATA	R / Python
Data Integrity	**Low (easy to break)**	**High (structured)**	**High (scripted)**
Visualizations	**Basic**	**Standard**	**Exceptional**
Automation	**Difficult**	**Moderate**	**High**
Cost	**Subscription**	**Expensive**	**0 (Free)**

Box 5.1 Placeholder: Text box listing advantages and limitations of statistical packages.
Caption: Advantages and limitations.

5.1.4 Applications in Health and Environmental Research

- Analysing survey data on disease prevalence.
- Evaluating effectiveness of health interventions.
- Studying associations between environmental exposures and health outcomes.
- Producing reports for policy and decision-making.

Fill in the blank: Statistical packages are used to study associations between environmental _____ and health outcomes.

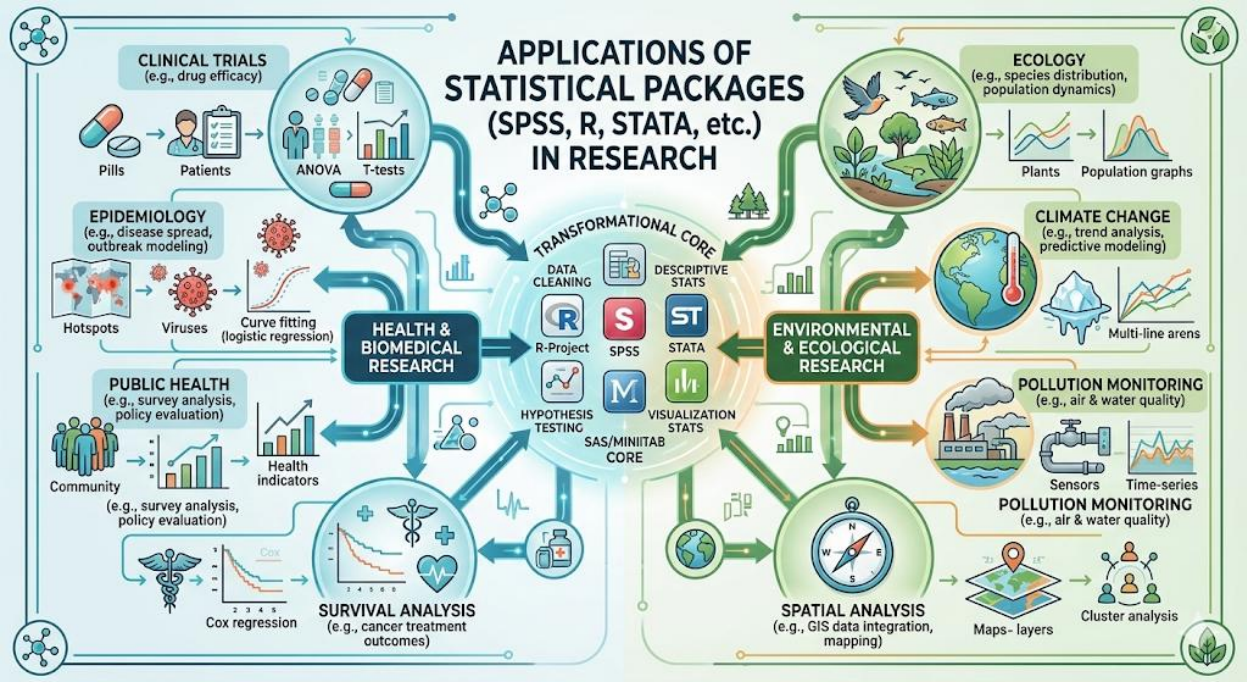


Figure 5.2 Placeholder: Diagram showing applications of statistical packages in health and environmental research. Caption: Applications of statistical packages.

Pilot questions 5.1

1. Software tools designed to manage and analyse data are called _____ packages.
2. A free statistical package developed by CDC for epidemiology is called _____ Info.
3. One limitation of statistical packages is that some are _____ to purchase.
4. Name one commonly used statistical package.
5. Statistical packages are used to study associations between environmental _____ and health outcomes.

5.2 Basic Operations in Statistical Packages

5.2.1 Data Entry and Management

- Importing datasets from spreadsheets or databases.
- Creating variables and defining data types (numeric, string, categorical).
- Cleaning data: handling missing values, duplicates, and errors.

Fill in the blank: Removing duplicate entries from a dataset is part of data _____.

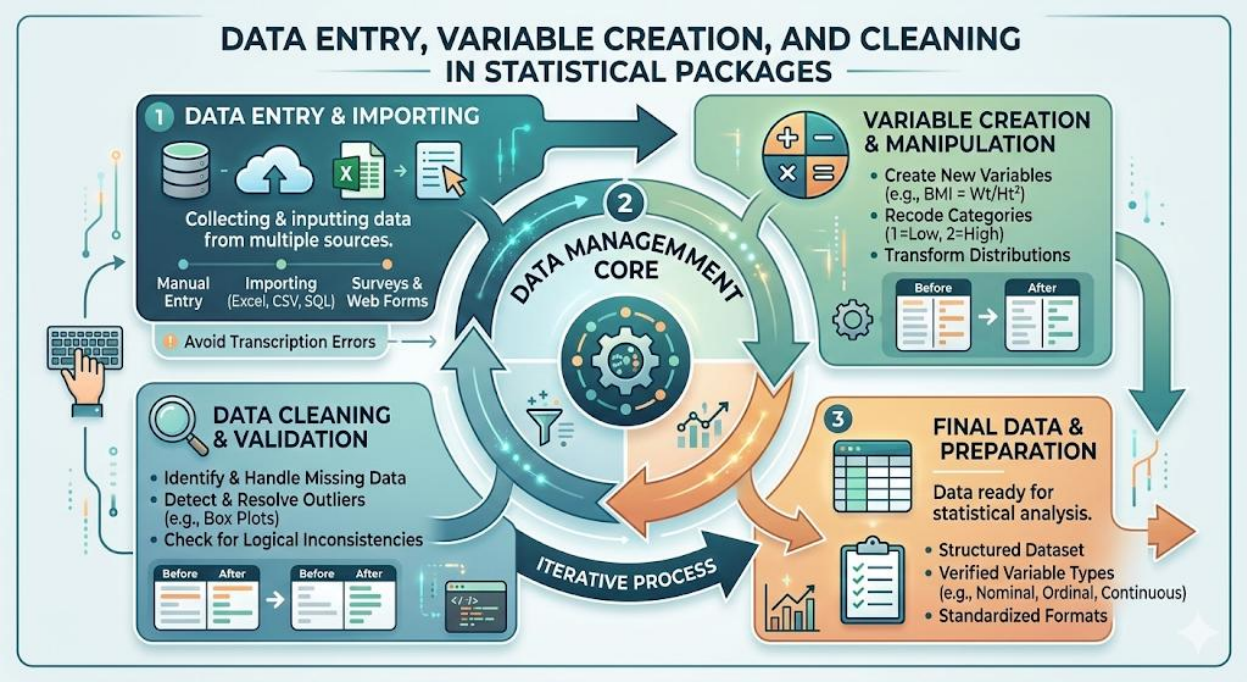


Figure 5.3 Placeholder: Diagram showing data entry and management steps. Caption: Data entry and management in statistical packages.

5.2.2 Coding and Recoding Variables

- **Coding**: Assigning numerical values to categorical data (e.g., Male = 1, Female = 2).
- **Recoding**: Transforming variables into new categories (e.g., age groups: 0–14, 15–49, 50+).
- Useful for simplifying analysis and creating meaningful groups.

Fill in the blank: Assigning numerical values to categorical data is called _____.

Operation	Original Variable (Raw Data)	Coded/Recoded Variable	Purpose of Transformation
Initial Coding	**Gender: "Male", "Female", "Non-binary"**	**$0 = \text{Male}$, $1 = \text{Female}$, $2 = \text{Non-binary}$**	**Converts text (string) data into numeric values for calculation.**

Operation	Original Variable (Raw Data)	Coded/Recoded Variable	Purpose of Transformation
Categorization	Age (Continuous): $14, 22, 45, 67, 82$	$1 = \text{Youth (<18)}$, $2 = \text{Adult (18-64)}$, $3 = \text{Senior (65+)}$	Simplifies analysis by grouping continuous data into age brackets.
Dichotomization	Smoking Status: "Never", "Former", "Current"	$0 = \text{Non-smoker}$, $1 = \text{Smoker (Former/Current)}$	Creates a "Binary" or "Dummy" variable for simple comparisons.
Reverse Coding	Survey Likert Scale: $1 = \text{Strongly Agree}$ to $5 = \text{Strongly Disagree}$	$1 = \text{Strongly Disagree}$ to $5 = \text{Strongly Agree}$	Ensures that a higher score always represents a "higher" level of the trait being measured.
Variable Creation	Weight (kg) and Height (m)	$BMI = \frac{\text{Weight}}{\text{Height}^2}$	Combines multiple raw data points into a single clinical index.

Table 5.2 Placeholder: Table showing examples of coding and recoding variables. Caption: Coding and recoding in statistical packages.

5.2.3 Generating Descriptive Statistics

- Calculating measures of central tendency (mean, median, mode).
- Calculating measures of variability (range, variance, standard deviation).
- Producing frequency tables for categorical data.

Fill in the blank: The most frequently occurring value in a dataset is called the _____.

Data Type	Example	Central Tendency	Dispersion
Nominal	**Blood Type, Gender**	**Mode**	**None**
Ordinal	**Pain Scale (1–10)**	**Median**	**Range / IQR**
Interval/Ratio	**Age, Heart Rate**	**Mean (if symmetrical)**	**Standard Deviation**

Box 5.2 Placeholder: Text box listing descriptive statistics functions. Caption: Descriptive statistics functions.

5.2.4 Producing Charts and Graphs

- Bar charts for categorical data.
- Histograms for continuous data.
- Pie charts for proportions.
- Scatter plots for relationships between variables.

Fill in the blank: A chart used to show frequency distribution of continuous data is called a _____.

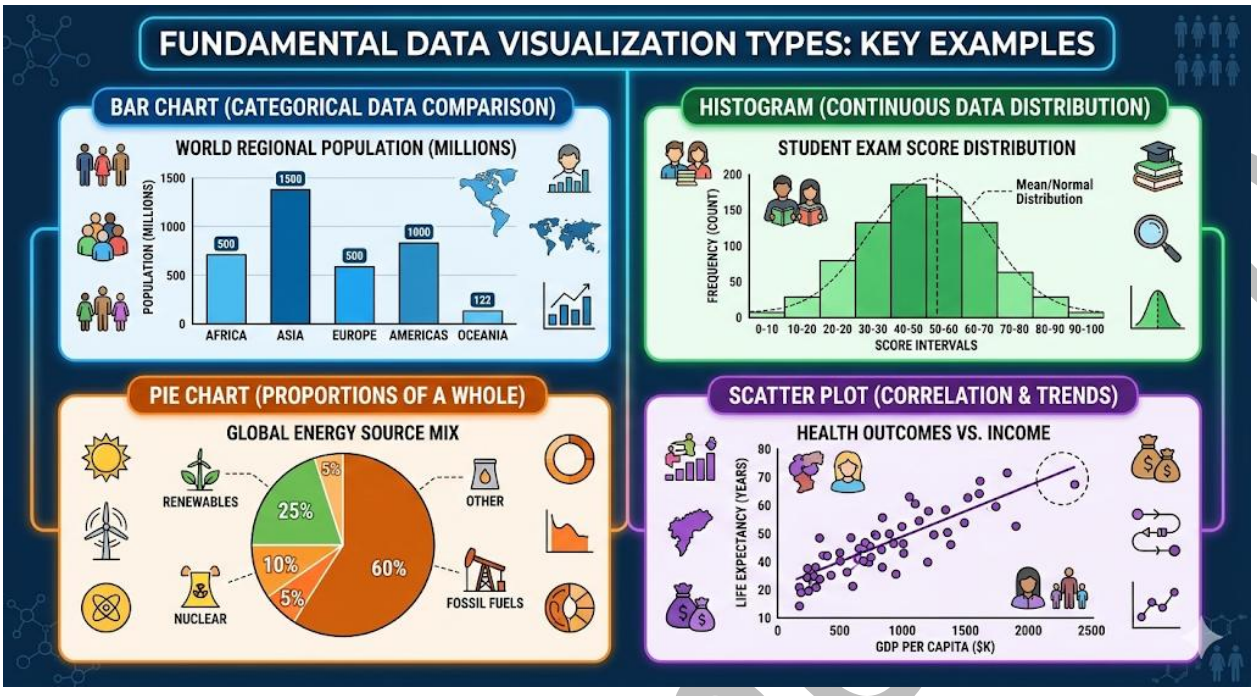


Figure 5.4 Placeholder: Diagram showing examples of charts and graphs. Caption: Charts and graphs in statistical packages

Pilot questions 5.2

1. Removing duplicate entries from a dataset is part of data _____.
2. Assigning numerical values to categorical data is called _____.
3. The most frequently occurring value in a dataset is called the _____.
4. Name one measure of variability.
5. A chart used to show frequency distribution of continuous data is called a _____.

5.3 Calculation of Rates in Health Statistics

5.3.1 Crude Rates

- **Crude birth rate (CBR)**: Number of live births per 1,000 population in a given year.
- **Crude death rate (CDR)**: Number of deaths per 1,000 population in a given year.
- Simple to calculate but does not account for differences in age or other factors.

Fill in the blank: The number of live births per 1,000 population in a given year is called the crude _____ rate.

Vital Statistic	Mathematical Formula	Description of Variables
Crude Birth Rate (CBR)	**$\text{CBR} = \left(\frac{B}{P} \right) \times 1,000$**	**$B = \text{Total live births during a specific year}$** **$P = \text{Total mid-year population for that year}$**
Crude Death Rate (CDR)	**$\text{CDR} = \left(\frac{D}{P} \right) \times 1,000$**	**$D = \text{Total deaths during a specific year}$** **$P = \text{Total mid-year population for that year}$**

Table 5.3 Placeholder: Table showing formulas for crude birth rate and crude death rate. Caption: Crude rates in health statistics.

5.3.2 Specific Rates

- **Age-specific death rate**: Deaths in a specific age group per 1,000 population of that age group.
- **Cause-specific death rate**: Deaths from a specific cause per 100,000 population.
- Provide more detailed insights than crude rates.

Fill in the blank: Deaths from a specific cause per 100,000 population are called _____-specific death rates.

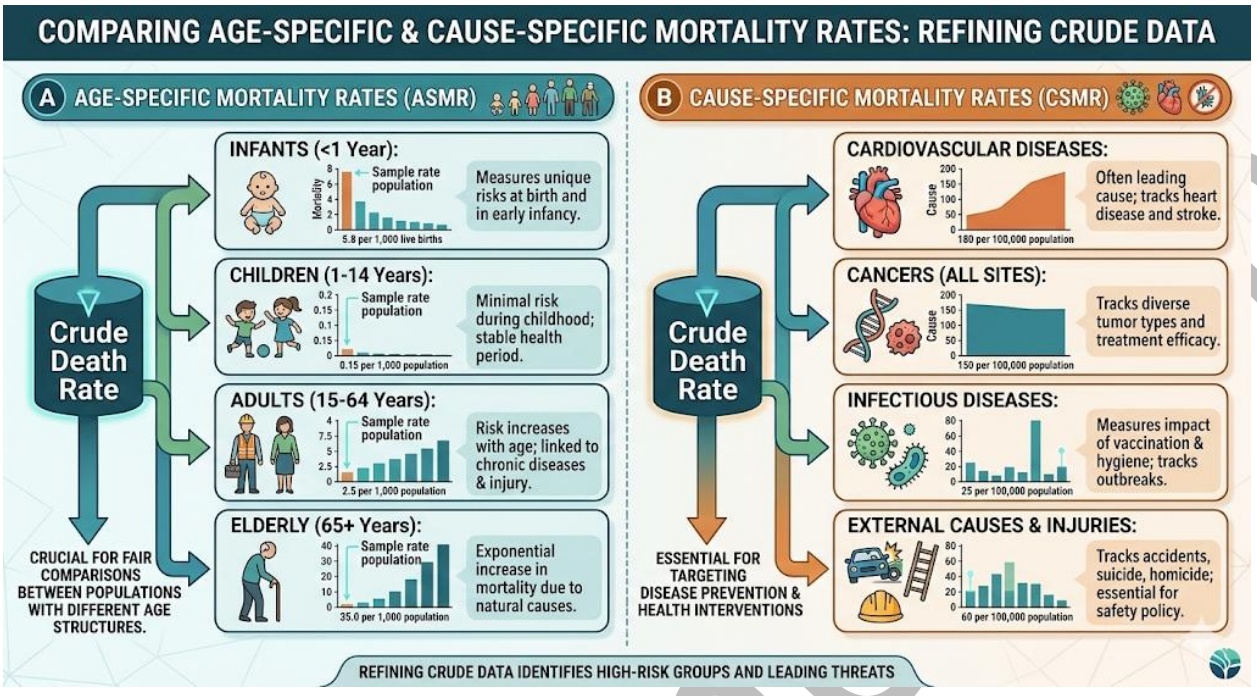


Figure 5.5 Placeholder: Diagram showing age-specific and cause-specific rates. Caption: Specific rates in health statistics.

5.3.3 Standardised Rates

- **Direct standardisation**: Adjusts rates using a standard population to allow comparison between groups.
- **Indirect standardisation**: Uses expected deaths based on standard rates to calculate Standardised Mortality Ratio (SMR).
- Useful for comparing populations with different age structures.

Fill in the blank: Adjusting rates using a standard population is called _____ standardisation.

Feature	Direct Standardization	Indirect Standardization
Requires	**Age-specific rates of the study groups.**	**Age-specific rates of a standard (large) population.**
Common Output	**Age-Adjusted Rate.**	**Standardized Mortality Ratio (SMR).**

Feature	Direct Standardization	Indirect Standardization
Best For	**Comparing two large regions/countries.**	**Comparing a small sub-group (e.g., a factory) to the general public.**
Key Question	**"What would the rate be if this group had a standard age structure?"**	**"How many deaths would we expect if this group were like the general public?"**

Box 5.3 Placeholder: Text box explaining direct and indirect standardisation. Caption: Standardised rates.

5.3.4 Examples in Environmental Health Contexts

- Calculating diarrhoeal disease incidence rate in communities with unsafe water.
- Measuring respiratory disease prevalence in areas with high air pollution.
- Comparing mortality rates in populations exposed to poor sanitation versus improved sanitation.

Fill in the blank: Measuring respiratory disease prevalence in areas with high _____ pollution is an example of applying rates in environmental health.

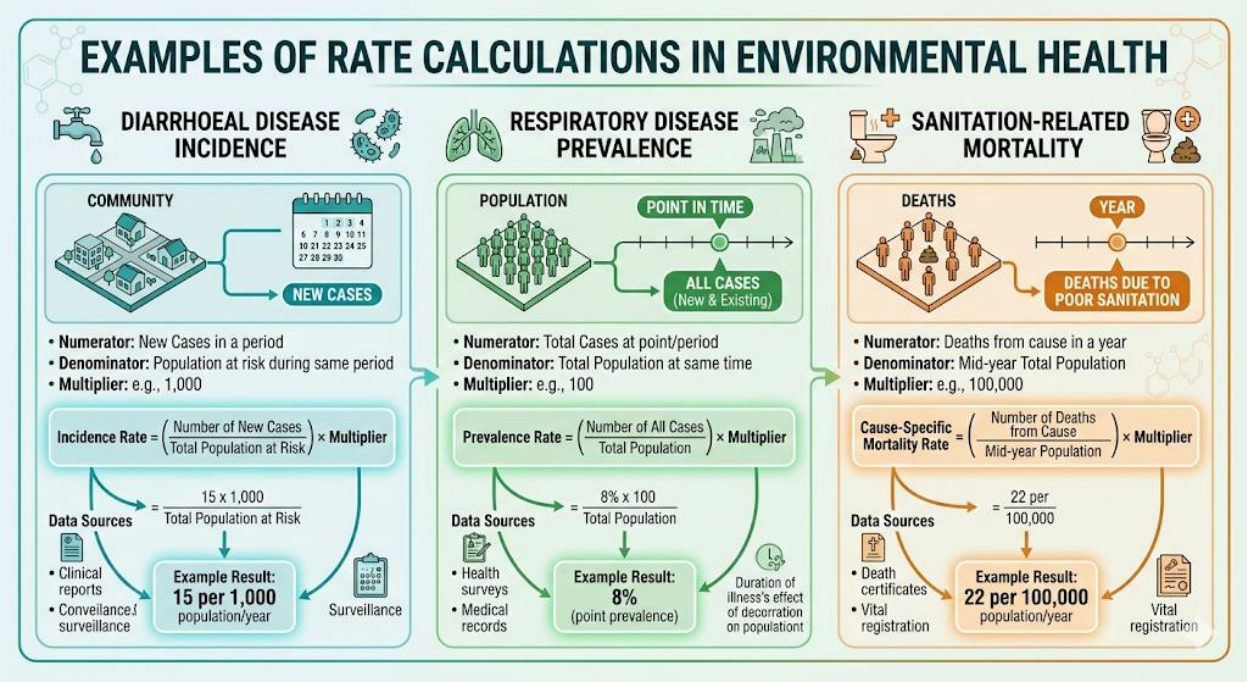


Figure 5.6 Placeholder: Diagram showing examples of rate calculations in environmental health. Caption: Rate calculations in environmental health contexts.

Pilot questions 5.3

1. The number of live births per 1,000 population in a given year is called the crude _____ rate.
2. Deaths from a specific cause per 100,000 population are called _____-specific death rates.
3. Adjusting rates using a standard population is called _____ standardisation.
4. Name one example of a specific rate.
5. Measuring respiratory disease prevalence in areas with high _____ pollution is an example of applying rates in environmental health.

5.4 Applications in Environmental Health Research

5.4.1 Using Statistical Packages to Analyse Environmental Health Data

- Statistical packages help manage large datasets from surveys and surveillance systems.
- They allow researchers to run descriptive and inferential analyses on environmental exposures.
- Examples: analysing water quality data, air pollution levels, or sanitation coverage.

Fill in the blank: Statistical packages help manage large _____ from surveys and surveillance systems.

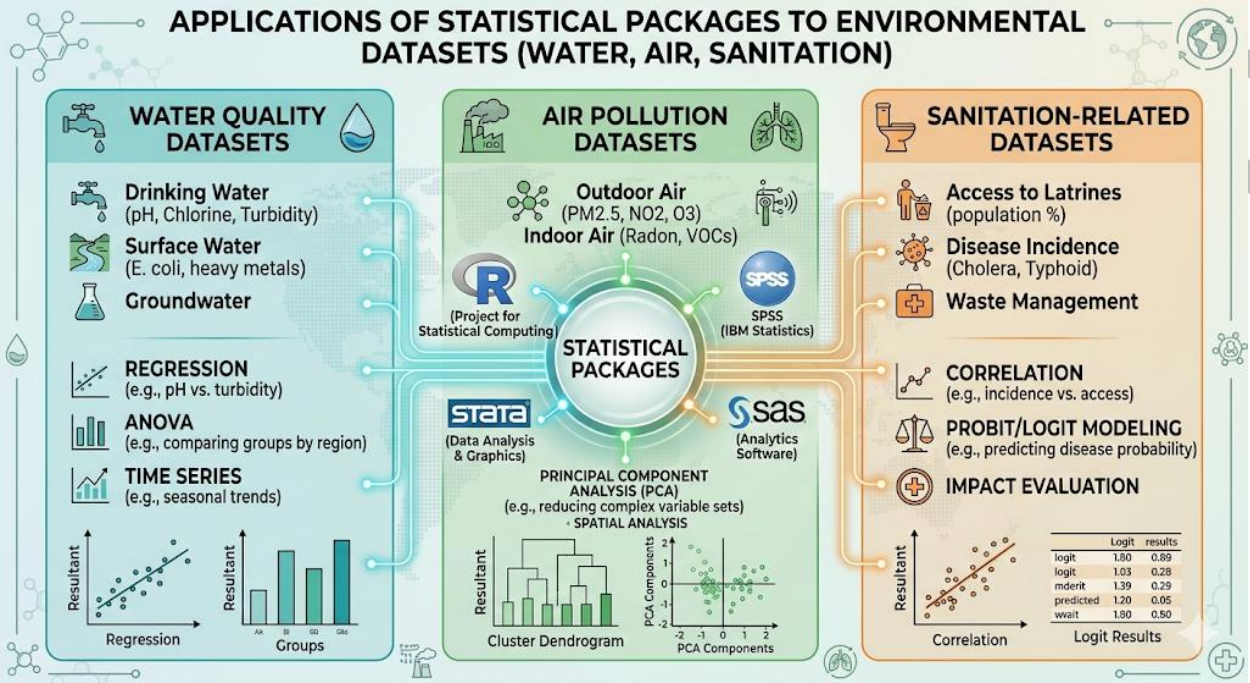


Figure 5.7 Placeholder: Diagram showing statistical packages applied to environmental health datasets. Caption: Using statistical packages in environmental health.

5.4.2 Calculating Morbidity and Mortality Rates Linked to Environmental Exposures

- Morbidity rates: incidence of diarrhoeal diseases in communities with unsafe water.
- Mortality rates: deaths due to respiratory illnesses in polluted areas.
- These rates help quantify the health burden of environmental hazards.

Fill in the blank: Incidence of diarrhoeal diseases in communities with unsafe water is an example of calculating _____ rates.

Environmental Risk Factor	Associated Morbidity (Illness/Condition)	Leading Causes of Mortality (Death)	Key Statistical Indicator
Unsafe Water Sources	**Cholera, Typhoid, Giardiasis, Dysentery, Dracunculiasis (Guinea Worm).**	**Severe Dehydration (from diarrheal diseases), Organ Failure, Sepsis.**	**Incidence Rate of waterborne pathogens per 1,000.**

Environmental Risk Factor	Associated Morbidity (Illness/Condition)	Leading Causes of Mortality (Death)	Key Statistical Indicator
Poor Sanitation & Hygiene	**Chronic Diarrhea, Intestinal Worms (Helminths), Trachoma (blindness), Stunting.**	**Diarrheal Disease, Pneumonia (due to inadequate hand hygiene), Malnutrition.**	**Prevalence of intestinal parasite infection in children under 5.**
Polluted Air (Ambient & Household)	**Asthma, Chronic Bronchitis, Reduced Lung Function, Cataracts (from indoor smoke).**	**Ischemic Heart Disease, Stroke, Lung Cancer, Chronic Obstructive Pulmonary Disease (COPD).**	**Age-Standardized Mortality Rate for cardiovascular & respiratory disease.**

Table 5.4 Placeholder: Table showing morbidity and mortality rates linked to environmental exposures. Caption: Rates linked to environmental exposures.

5.4.3 Case Examples in Environmental Health Research

- **Waterborne diseases**: Statistical packages used to calculate incidence and prevalence rates.
- **Air pollution**: Mortality rates compared across regions with different pollution levels.
- **Sanitation coverage**: Standardised rates used to compare child mortality in areas with improved vs. poor sanitation.

Fill in the blank: Standardised rates are used to compare child _____ in areas with improved vs. poor sanitation.

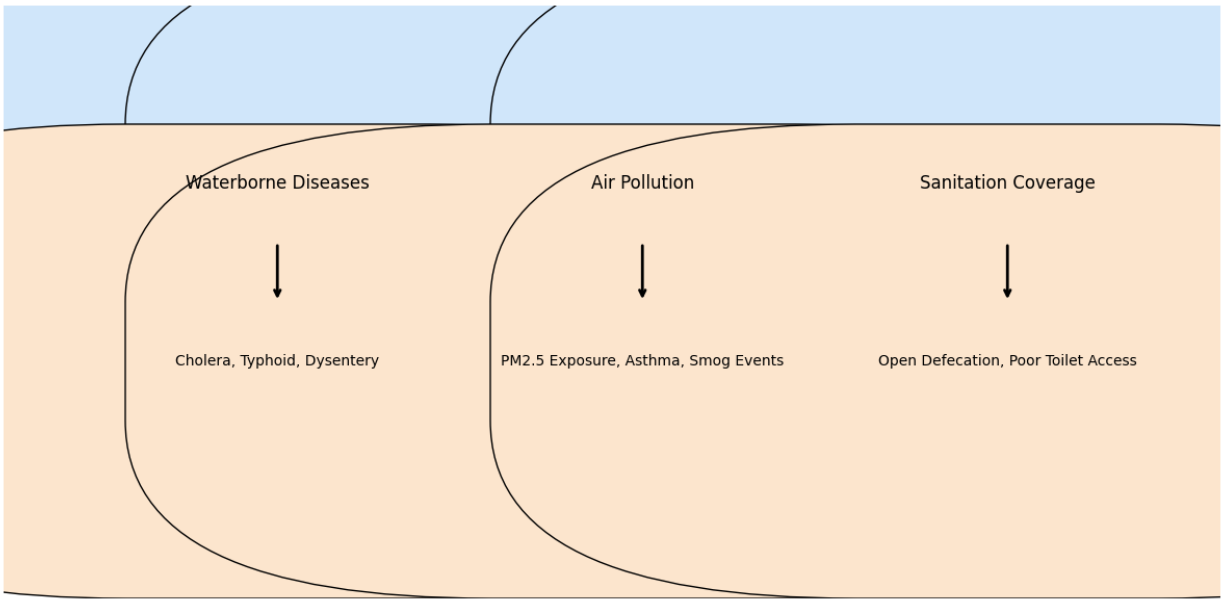


Figure 5.8 Placeholder: Diagram showing case examples of environmental health research. Caption: Case examples in environmental health research.

5.4.4 Challenges and Opportunities in Applying Statistical Packages

Challenges

- Limited access to software in low-resource settings.
- Need for trained personnel.
- Incomplete or poor-quality data.

Opportunities

- Open-source tools like R and Epi Info.
- Integration of digital health records.
- Improved surveillance systems.

Fill in the blank: A major challenge in applying statistical packages in low-resource settings is limited access to _____.

Feature	Challenge	Opportunity
Data Complexity	**Multicollinearity (pollutants often move together).**	**Principal Component Analysis (PCA) to simplify variables.**

Feature	Challenge	Opportunity
Privacy	**Sensitive health records are hard to access.**	**De-identified "Big Data" sets and secure cloud processing.**
Policy Impact	**Communicating complex stats to lawmakers.**	**Dynamic data visualizations and interactive charts.**

Box 5.4 Placeholder: Text box listing challenges and opportunities. Caption: Challenges and opportunities in applying statistical packages.

Pilot questions 5.4

1. Statistical packages help manage large _____ from surveys and surveillance systems.
2. Incidence of diarrhoeal diseases in communities with unsafe water is an example of calculating _____ rates.
3. Standardised rates are used to compare child _____ in areas with improved vs. poor sanitation.
4. Name one case example of environmental health research using statistical packages.
5. A major challenge in applying statistical packages in low-resource settings is limited access to _____.

Series 5 Summary – Statistical Packages and Calculation of Rates

In this Series, you learned how **statistical packages** and **rate calculations** are applied in health and environmental research:

- **Part 5.1** introduced statistical packages, defining them as software tools for managing and analysing data, with examples like SPSS, STATA, R, Epi Info, and Excel. Their advantages (efficiency, advanced functions, visual outputs) and limitations (cost, training needs, incomplete functions) were highlighted.
- **Part 5.2** explored basic operations in statistical packages, including data entry and management, coding and recoding variables, generating descriptive statistics, and producing charts and graphs.

- **Part 5.3** focused on calculating rates in health statistics, covering crude rates (birth and death), specific rates (age-specific, cause-specific), and standardised rates (direct and indirect). Examples were provided in environmental health contexts such as waterborne diseases and air pollution.
- **Part 5.4** applied statistical packages and rate calculations to environmental health research, showing how they are used to analyse exposures, calculate morbidity and mortality rates, and address challenges like limited access to software in low-resource settings.

By completing this Series, you should now be able to:

- **Define and explain** statistical packages and evaluate their role in health and environmental research (SLO 5.1).
- **Demonstrate basic operations** in statistical packages (SLO 5.2).
- **Calculate and interpret** crude, specific, and standardised rates (SLO 5.3).
- **Apply statistical packages and rate calculations** to environmental health research (SLO 5.4).

This Series bridges **technical proficiency** with **practical application**, preparing learners to use statistical tools and rate calculations for evidence-based decision-making in health and environmental contexts.

Glossary of Terms – Series 5: Statistical Packages and Calculation of Rates

- **Statistical packages**: Software tools designed to manage, analyse, and interpret data (e.g., SPSS, STATA, R, Epi Info, Excel).
- **SPSS**: A widely used statistical package in social sciences and health research.
- **STATA**: A statistical package popular in econometrics and epidemiology.
- **R**: An open-source statistical software known for advanced modelling and flexibility.
- **Epi Info**: Free software developed by the CDC for epidemiological analysis.
- **Excel**: A spreadsheet tool with basic statistical functions.
- **Data entry and management**: The process of importing, cleaning, and organising datasets for analysis.
- **Coding**: Assigning numerical values to categorical data (e.g., Male = 1, Female = 2).
- **Recoding**: Transforming variables into new categories (e.g., age groups).
- **Descriptive statistics**: Measures that summarise data, including mean, median, mode, variance, and frequency tables.

- **Charts and graphs**: Visual representations of data, such as bar charts, histograms, pie charts, and scatter plots.
- **Crude birth rate (CBR)**: Number of live births per 1,000 population in a given year.
- **Crude death rate (CDR)**: Number of deaths per 1,000 population in a given year.
- **Age-specific death rate**: Deaths in a specific age group per 1,000 population of that age group.
- **Cause-specific death rate**: Deaths from a specific cause per 100,000 population.
- **Direct standardisation**: Adjusting rates using a standard population to allow comparison between groups.
- **Indirect standardisation**: Using expected deaths based on standard rates to calculate Standardised Mortality Ratio (SMR).
- **Morbidity rate**: Measure of illness in a population, such as incidence or prevalence.
- **Mortality rate**: Measure of death in a population, often linked to specific causes or exposures.

Concept Check Questions – Series 5: Statistical Packages and Calculation of Rates

Objective questions (linked to SLOs 5.1–5.4)

1. Software tools designed to manage and analyse data are called _____ packages. (SLO 5.1)
2. A free statistical package developed by CDC for epidemiology is called _____ Info. (SLO 5.1)
3. The most frequently occurring value in a dataset is called the _____. (SLO 5.2)
4. The number of live births per 1,000 population in a given year is called the crude _____ rate. (SLO 5.3)
5. Adjusting rates using a standard population is called _____ standardisation. (SLO 5.3)

Short-answer questions (linked to SLOs 5.1–5.4)

6. Distinguish between coding and recoding variables with one example each. (SLO 5.2)
7. Name two advantages and two limitations of statistical packages. (SLO 5.1)
8. Give one example of a crude rate and one example of a specific rate. (SLO 5.3)
9. Explain the difference between direct and indirect standardisation. (SLO 5.3)

10. State one challenge and one opportunity in applying statistical packages in environmental health research. (SLO 5.4)

Long-answer questions (linked to SLOs 5.1–5.4)

11. Analyse the role of statistical packages in health and environmental research, highlighting their applications and limitations. (SLO 5.1, 5.4)

12. Discuss basic operations in statistical packages, explaining how data entry, coding, descriptive statistics, and visualisation support research. (SLO 5.2)

13. Describe crude, specific, and standardised rates, and assess their usefulness in health statistics with examples from environmental health. (SLO 5.3)

14. Evaluate how statistical packages and rate calculations can be applied to environmental health research, using case examples such as waterborne diseases, air pollution, and sanitation coverage. (SLO 5.4)

Answers to Pilot Questions – Series 5: Statistical Packages and Calculation of Rates

Pilot questions 5.1 (Introduction to Statistical Packages)

1. Statistical
2. Epi
3. Expensive
4. SPSS / STATA / R / Epi Info / Excel (any one)
5. Exposures

Pilot questions 5.2 (Basic Operations in Statistical Packages)

1. Cleaning
2. Coding
3. Mode
4. Range / Variance / Standard deviation (any one)
5. Histogram

Pilot questions 5.3 (Calculation of Rates in Health Statistics)

1. Birth
2. Cause
3. Direct
4. Age-specific death rate / Cause-specific death rate (any one)

5. Air

Pilot questions 5.4 (Applications in Environmental Health Research)

1. Datasets
2. Morbidity
3. Mortality
4. Waterborne diseases / Air pollution / Sanitation coverage (any one)
5. Software

References and Attributions – Series 5: Statistical Packages and Calculation of Rates

General references

- Course content adapted from *EHS 212: Biostatistics (2 Units)*.
- Concepts based on standard biostatistics, epidemiology, demography, and environmental health literature, as well as open educational resources.

Figures

- *Figure 5.1: Scope of statistical packages* – AI prompt: "Generate a diagram showing statistical packages such as SPSS, STATA, R, Epi Info, and Excel."
- *Figure 5.2: Applications of statistical packages* – AI prompt: "Generate a diagram showing applications of statistical packages in health and environmental research."
- *Figure 5.3: Data entry and management steps* – AI prompt: "Generate a diagram showing data entry, variable creation, and cleaning in statistical packages."
- *Figure 5.4: Charts and graphs in statistical packages* – AI prompt: "Generate a diagram showing bar chart, histogram, pie chart, and scatter plot examples."
- *Figure 5.5: Specific rates in health statistics* – AI prompt: "Generate a diagram showing age-specific and cause-specific rates."
- *Figure 5.6: Rate calculations in environmental health contexts* – AI prompt: "Generate a diagram showing examples of rate calculations in environmental health such as diarrhoeal disease incidence, respiratory disease prevalence, and sanitation-related mortality."
- *Figure 5.7: Using statistical packages in environmental health* – AI prompt: "Generate a diagram showing statistical packages applied to water, air, and sanitation datasets."

- *Figure 5.8: Case examples in environmental health research* – AI prompt: "Generate a diagram showing case examples of waterborne diseases, air pollution, and sanitation coverage analysed with statistical packages."

Tables

- *Table 5.1: Commonly used statistical packages* – AI prompt: "Generate a table listing SPSS, STATA, R, Epi Info, and Excel with their features and applications."
- *Table 5.2: Coding and recoding variables* – AI prompt: "Generate a table showing examples of coding and recoding variables in statistical packages."
- *Table 5.3: Crude rates in health statistics* – AI prompt: "Generate a table showing formulas for crude birth rate and crude death rate."
- *Table 5.4: Rates linked to environmental exposures* – AI prompt: "Generate a table showing morbidity and mortality rates linked to unsafe water, polluted air, and poor sanitation."

Boxes

- *Box 5.1: Advantages and limitations of statistical packages* – AI prompt: "Create a text box listing advantages and limitations of statistical packages."
- *Box 5.2: Descriptive statistics functions* – AI prompt: "Create a text box listing descriptive statistics functions such as mean, median, mode, variance, and frequency tables."
- *Box 5.3: Standardised rates* – AI prompt: "Create a text box explaining direct and indirect standardisation with examples."
- *Box 5.4: Challenges and opportunities in applying statistical packages* – AI prompt: "Create a text box listing challenges and opportunities in applying statistical packages in environmental health research."