Chapter 1: Introduction to Statistics

STS3401: Probability and Statistics 1

Dr. Ahmad **Hakiim** Jamaluddin

Room 1.34
Department of Mathematics & Statistics
Universiti Putra Malaysia

"Statistics is the grammar of science." — Karl Pearson

Last modified: October 11, 2025

Tickler!

"Excellent health statistics - smokers are less likely to die of age related illness"

1

[1] Source: https://www.pinterest.com/pin/489907265711896401/

Outline

What is Statistics?

Definition

Statistics is the science of **collecting**, **organising**, **analysing**, and **interpreting** data in order to make decisions.

Applications:

- Politics (election polling)
- Industry (product testing)
- Medicine and healthcare
- Engineering and quality control
- Social sciences research

Uses of Statistics:

- Theoretical discipline
- Tool for researchers
- Drawing general conclusions
- Making predictions
- Decision making under uncertainty
- Uses probability to model uncertainty and make inferences

Common Problem in Statistics

The Central Challenge

Making decisions or predictions about a **large body of measurements** which cannot be totally enumerated.

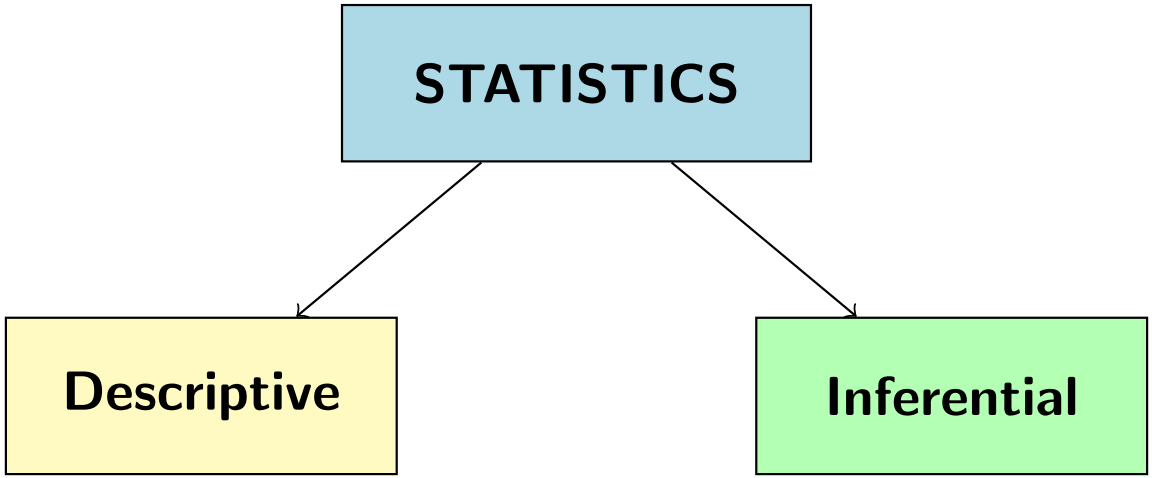
Examples of This Problem:

- **Light bulbs**: Cannot test all bulbs (testing is destructive)
- **Elections**: Population too big; people change their minds
- **Quality control**: Cannot inspect every product
- **Medical research**: Cannot test treatment on entire population

Solution

Collect a **smaller set of measurements** that will (hopefully) be **representative** of the larger set.

Two Branches of Statistics



- Organise data
- Summarise data
- Present data
- Numerical & graphical

- Generalisations
- Test hypotheses
- Make predictions
- Estimate parameters

Descriptive Statistics

Definition

Methods for **organising**, **displaying**, and **describing** data using tables, graphs, and summary measures.

Key Functions:

- Organise raw data into meaningful formats
- Calculate summary measures (mean, median, standard deviation)
- Create visual representations (charts, graphs, histograms)
- Identify patterns and trends
- Provide numerical and graphical descriptions

Example

A pet shop sells cats, dogs, birds and fish. If 100 pets are sold, and 40 out of the 100 were dogs, then one description of the data would be that 40% were dogs. This summarises the observed data without making broader claims.

Inferential Statistics

Definition

Methods that use **sample data** to make generalisations, estimates, decisions, predictions, or other generalisations about a **larger set of data** (population).

Two Main Areas:

1. **Estimating parameters**: Using sample statistics to estimate population parameters
2. **Hypothesis testing**: Using sample data to answer research questions

Key Functions:

- Make generalisations from samples to populations **using probability models**
- Test hypotheses about population parameters **based on probability**
- Estimate population parameters with confidence intervals **using probabilistic methods**
- Predict future outcomes **with probability-based models**
- Assess reliability and significance of results **using probability theory**

Example

A pharmaceutical company tests a drug on 1000 patients and finds 75% improve. They use inferential statistics and **probability models** to estimate that 70-80% of **all patients** would improve (with 95% confidence).

Descriptive vs. Inferential: Comparison

Aspect	**Descriptive**	**Inferential**
Purpose	Summarise and describe	Generalise and predict
Scope	Sample or population	Sample to population
Tools	Tables, graphs, measures	Hypothesis tests, confidence intervals
Uncertainty	No uncertainty	Involves uncertainty
Example	Average height is 170cm	Population mean is 170 ± 3cm

Population

Definition

A **population** is the collection of all outcomes, responses, measurements, or counts that are of interest in a particular study.

Characteristics:

- Contains **all** possible observations of interest
- Often very large or infinite in size
- May be difficult or impossible to study completely
- **Parameters** describe population characteristics
- Denoted by Greek letters: μ (mean), σ (standard deviation), π (proportion)

Examples:

- All likely voters in the next election
- All parts produced by a factory
- All sales receipts for November

Population = Everyone/Everything of Interest

Sample

Definition

A **sample** is a subset of the population that is selected for study.

Characteristics:

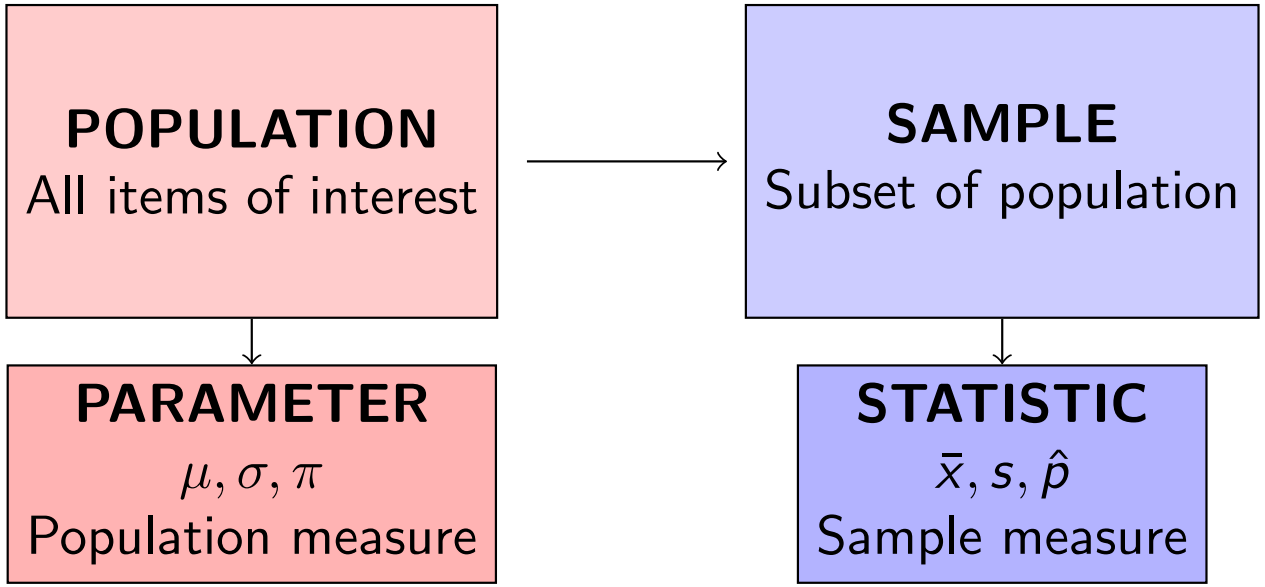
- Smaller and more manageable than the population
- Should be **representative** of the population
- **Statistics** describe sample characteristics
- Denoted by Roman letters: $\bar{x}$ (mean), s (standard deviation), $\hat{p}$ (proportion)
- Used to make inferences about the population

Examples:

- 1000 voters selected at random for interview
- A few parts selected for destructive testing
- Every 100th receipt selected for audit

Sample = A Representative Subset

Parameters vs. Statistics



Key Distinction

Parameter: Numerical description of a **population** characteristic
Statistic: Numerical description of a **sample** characteristic

Population vs. Sample: Examples

Example 1: Election Polling

- **Population:** All eligible voters in Malaysia
- **Sample:** 1,500 randomly selected voters surveyed

Example 2: Quality Control

- **Population:** All light bulbs manufactured
- **Sample:** Every 100th bulb tested

Example 3: Medical Research

- **Population:** All patients with the disease
- **Sample:** 500 patients in clinical trial

Example 4: Student Survey

- **Population:** All university students
- **Sample:** 200 students surveyed

Key Statistical Terms

Essential Definitions

Data: Information coming from observations, counts, measurements, or responses.

Variable A characteristic that changes or varies over time and/or for different individuals or objects under consideration.

Experimental Unit The individual or object on which a variable is measured.

Measurement Results when a variable is actually measured on an experimental unit.

Observation The value of a variable at a particular period for a particular experimental unit.

Example

Variable: Hair color
Experimental unit: Person
Measurements: Brown, black, blonde, etc.
Observation: "John has brown hair"

Data Dimensionality

How many variables have you measured?

Univariate data One variable is measured on a single experimental unit.

- Example: Heights of students

Bivariate data Two variables are measured on a single experimental unit.

- Example: Height and weight of students

Multivariate data More than two variables are measured on a single experimental unit.

- Example: Height, weight, age, and GPA of students

This classification helps determine appropriate statistical methods

Levels of Measurement

Understanding the scale of measurement is crucial for choosing appropriate statistical methods

Nominal Mutually exclusive categories with no natural ordering

- Examples: Gender (male, female), religion, blood type
- Operations: Count frequencies, find mode

Ordinal Categories that can be ordered but differences aren't precise

- Examples: Health quality (excellent, good, adequate, bad, terrible)
- Operations: Median, percentiles, rank correlation

Interval Numerical measurements with no meaningful zero point

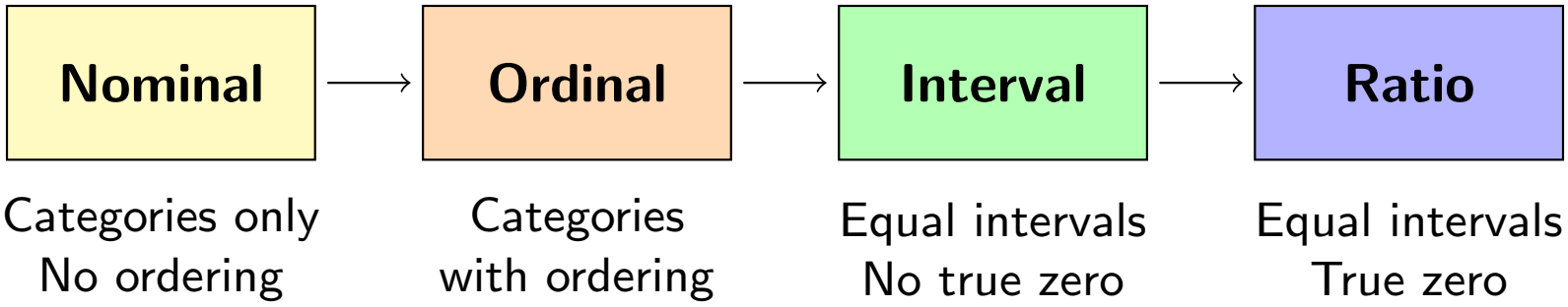
- Examples: Temperature (°C, °F), calendar years
- Operations: Mean, standard deviation (but no ratios)

Ratio Numerical measurements with meaningful zero and precise differences

- Examples: Height, weight, time, income
- Operations: All statistical measures, including ratios

Measurement Scales: Hierarchy

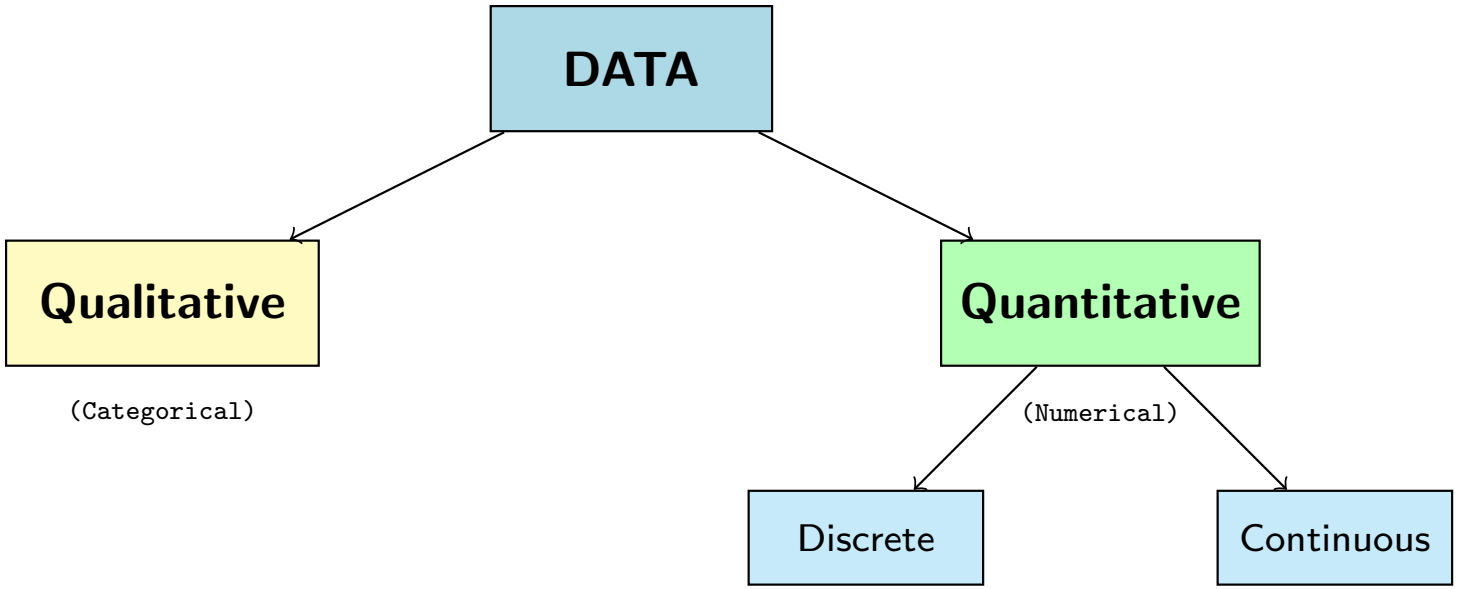
Increasing Statistical Power



Important

Higher levels of measurement allow for more sophisticated statistical analyses.
Always identify the measurement scale before choosing analytical methods.

Classification of Data



Qualitative (Categorical) Data

Definition

Qualitative variables measure a quality or characteristic on each experimental unit.

Characteristics:

- Represents categories or attributes
- Cannot be subjected to arithmetic operations
- May or may not have natural ordering
- Analysed using frequencies and proportions

Nominal (No order):

- Hair color (black, brown, blonde)
- Make of car (Honda, Toyota, Ford)
- Gender (male, female)
- State of birth

Ordinal (Has order):

- Grades: A, B, C, D, F
- Size: Small, Medium, Large
- Satisfaction: Poor, Fair, Good, Excellent
- Health quality ratings

Quantitative (Numerical) Data

Definition

Quantitative variables measure a numerical quantity on each experimental unit.

Characteristics:

- Represented by numbers
- Can be subjected to arithmetic operations
- Can be ordered from smallest to largest
- Allows for meaningful mathematical analysis

Examples:

- Height: 172 cm, 165 cm, 180 cm
- Age: 20, 21, 19, 22
- Income: \$45,000, \$52,000
- Test scores: 85, 92, 78

Operations possible:

- Addition and subtraction
- Multiplication and division
- Calculate means and variances
- Perform statistical tests

What is Probability?

Definition

Probability is the mathematical study of uncertainty, quantifying the likelihood that a specific event will occur, expressed as a number between 0 and 1.

Key Concepts:

- **Event**: A specific outcome or set of outcomes (e.g., rolling a 6 on a die).
- **Probability of an event**: $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total possible outcomes}}$ (for equally likely outcomes).
- **Random experiment**: A process with uncertain outcomes but a defined set of possibilities (e.g., flipping a coin).
- **Probability models**: Describe the likelihood of outcomes for random variables.

Example

When rolling a fair six-sided die, the probability of rolling a 6 is $P(6) = \frac{1}{6} \approx 0.1667$, assuming all outcomes are equally likely.

How Probability and Statistics Complement Each Other

Complementary Roles

Statistics uses data to describe and infer characteristics of a population, while **probability** provides the mathematical framework to model uncertainty and make predictions about future events.

Statistics:

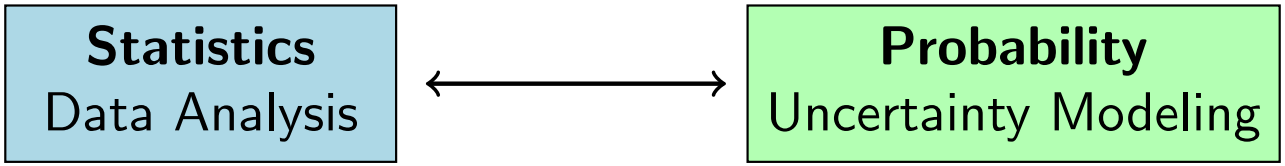
- Observes and analyses **real data** (samples).
- Descriptive: Summarises observed data (e.g., mean, variance).
- Inferential: Makes conclusions about populations (e.g., confidence intervals).
- Relies on probability to quantify uncertainty in inferences.

Probability:

- Models **theoretical outcomes** of random processes.
- Provides rules to calculate likelihoods (e.g., probability distributions).
- Enables inference by describing expected behaviour of samples.
- Underpins random sampling and hypothesis testing.

Probability and Statistics: Visual and Example

Complementary Relationship



Example

In a clinical trial, statistics summarises recovery rates from a sample of patients, while probability models the likelihood that a new patient will recover based on the sample data.

Probability provides the theory; statistics applies it to data.

Quantitative Variables: Discrete vs. Continuous

Discrete Variables

Can assume only a **finite or countable number** of values.

Continuous Variables

Can assume the **infinitely many values** corresponding to the points on a line interval.

Discrete Examples:

- Number of oranges on a tree
- Number of cars entering campus
- Number of students in class
- Number of defective items

Result of COUNTING

Continuous Examples:

- Time until light bulb burns out
- Height of students
- Weight of products
- Temperature measurements

Result of MEASURING

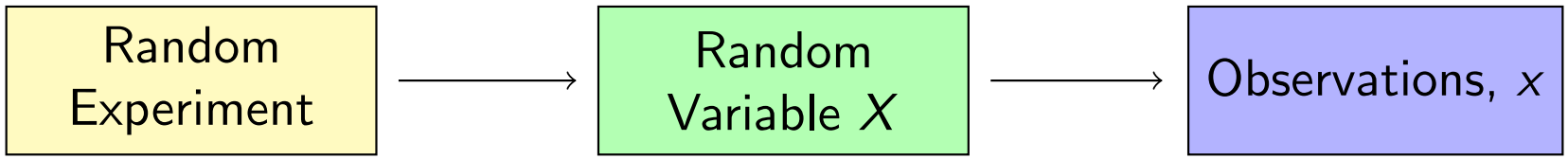
Random Variables

Definition

A **random variable** is a function that assigns numerical values to the outcomes of a random experiment, with probabilities described by probability distributions.

Notation:

- Random variables denoted by capital letters: X, Y, Z
- Specific values (observations) denoted by lowercase letters: x, y, z



Discrete Random Variables

Definition

A **discrete random variable** can take on only a countable number of distinct values.

Characteristics:

- Values can be counted (finite or countably infinite)
- Often result from **counting** processes
- Gaps exist between possible values
- Probability described using **probability mass functions (PMFs)**, which assign probabilities to each possible value

Examples

- Number of oranges on trees in a grove: $X = \{0, 1, 2, 3, \dots\}$
- Number of cars entering campus: $X = \{0, 1, 2, 3, \dots, n\}$
- Score on a die roll: $X = \{1, 2, 3, 4, 5, 6\}$
- Number of defective items in a batch: $X = \{0, 1, 2, 3, \dots, n\}$

Continuous Random Variables

Definition

A **continuous random variable** can take on any value within a given interval or range of values.

Characteristics:

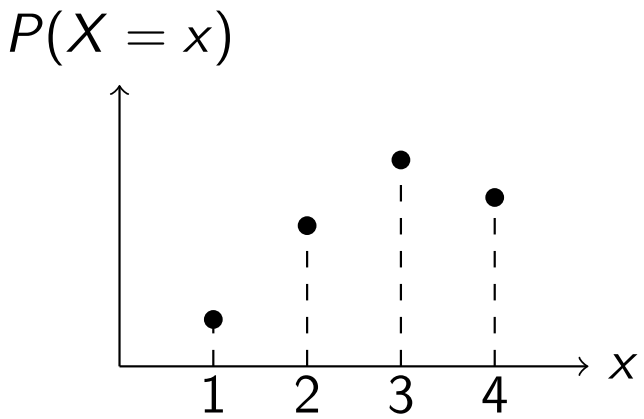
- Values form a continuum (uncountably infinite)
- Often result from **measuring** processes
- No gaps between possible values
- Probability described using **probability density functions (PDFs)**, where probabilities are areas under the curve
- Probability of any single specific value is zero

Examples

- Time until light bulb burns out: $X \in [0, \infty)$ hours
- Height of students: $X \in [100, 220]$ cm
- Temperature measurements: $X \in (-\infty, \infty)$ °C
- Weight of products: $X \in [0, \infty)$ kg

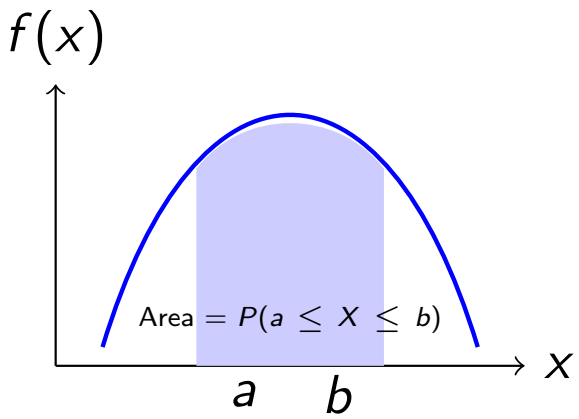
Discrete vs. Continuous: Visual Comparison

Discrete Random Variable



Probability at specific points

Continuous Random Variable



Probability as area under curve

Aspect	**Discrete**	**Continuous**
Values	Countable	Uncountable
Process	Counting	Measuring
$P(X = \text{specific value})$	> 0	$= 0$

Identifying Variable Types: Practice

Classify each variable as qualitative/quantitative and discrete/continuous:

1. Number of cars in a car park
2. Time taken to run a marathon
3. Blood type (A, B, AB, O)
4. Amount of rainfall in millimetres
5. Student satisfaction (Poor, Fair, Good, Excellent)
6. Speed of a car
7. Number of students in a classroom
8. Weight of a baby at birth

Think: What type of measurement? Counting or measuring?

Why Representative Sampling Matters

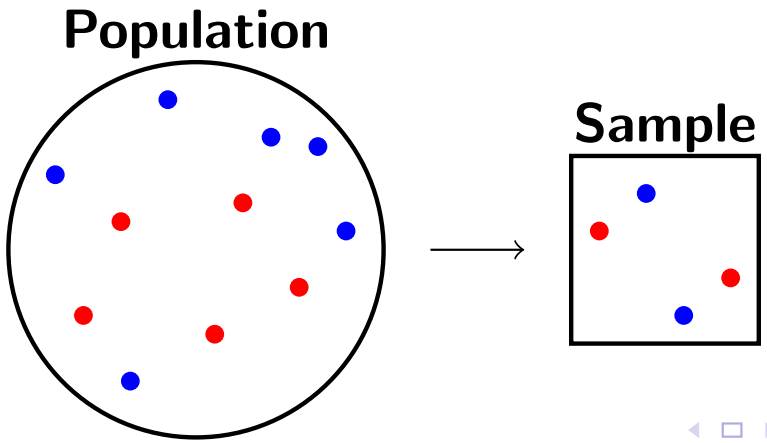
The sampling plan determines the amount of information you can extract

Key Principle

Good samples should be **representative** of the population to allow valid inferences.

Simple Random Sampling

A method of sampling that allows each possible sample of size n an **equal probability** of being selected, ensuring unbiased representation based on probability theory.



Types of Sampling Situations

Sampling can occur in two practical situations:

1. **Observational Studies**: The data existed before you decided to study it.
   - Watch out for **nonresponse bias**: Are responses biased because only opinionated people responded?
   - Watch out for **undercoverage**: Are certain segments systematically excluded?
   - Watch out for **wording bias**: Is the question too complicated or poorly worded?

2. **Experimentation**: The data are generated by imposing an experimental condition or treatment.
   - Hypothetical populations can make random sampling difficult
   - Samples must be chosen to be representative of the whole population
   - Samples must behave like random samples!

Random Sampling Methods

These methods involve randomization and ARE appropriate for statistical inference:

Simple Random Each possible sample of size n has equal probability of selection

Stratified Random Divide population into subpopulations (strata) and select a simple random sample from each stratum

Cluster Sample Divide population into subgroups (clusters); select a simple random sample of clusters and take a census of every element in the selected clusters

1-in-k Systematic Randomly select one of the first k elements in an ordered population, then select every k-th element thereafter

Important

All these methods use randomization and allow for valid statistical inference.

Random Sampling Examples

Stratified Random Sample:

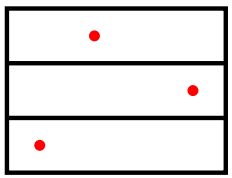
- Divide California into counties
- Take simple random sample within each county

1-in-50 Systematic Sample:

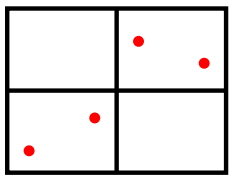
- Choose an entry at random from phone book
- Select every 50th number thereafter

Cluster Sample:

- Divide California into counties and take simple random sample of 10 counties
- OR: Divide city into blocks, choose 10 blocks randomly, interview all residents



Stratified



Cluster

Non-Random Sampling Methods

These methods do NOT involve randomization and should NOT be used for statistical inference:

Convenience Sample A sample that can be taken easily without random selection

- Example: People walking by on the street
- Problem: Not representative of population

Judgment Sample The sampler decides who will and won't be included

- Example: Selecting "typical" cases based on judgment
- Problem: Introduces personal bias

Quota Sample Sample makeup must reflect population makeup on selected characteristics

- Example: Matching race, ethnic origin, gender proportions
- Problem: May miss other important characteristics

Warning

These methods can lead to biased results and invalid conclusions about the population!

Pop Quiz: Population, Sample, and Sampling

Test your understanding!

1. What is the main difference between a parameter and a statistic?
   A. A parameter is smaller than a statistic
   B. A parameter describes a population, a statistic describes a sample
   C. A parameter uses Roman letters, a statistic uses Greek letters
   D. There is no difference
2. Which sampling method should NOT be used for statistical inference?
   A. Simple random sampling
   B. Stratified random sampling
   C. Convenience sampling
   D. Cluster sampling

Real-World Applications

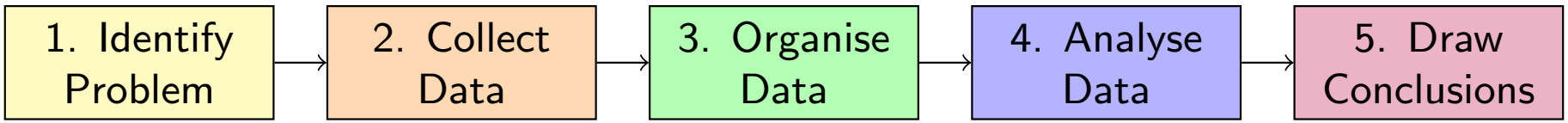
Understanding data types and measurement scales is crucial for choosing appropriate statistical methods:

Data Type	Examples	Statistical Methods
Nominal	Gender, blood type, car make	Frequency tables, bar charts, chi-square tests, mode
Ordinal	Satisfaction ratings, grades	Median, percentiles, rank correlation, non-parametric tests
Interval	Temperature (°C), IQ scores	Mean, standard deviation, correlation (no ratios)
Ratio	Height, weight, income, time	All statistical measures including ratios and geometric means
Discrete	Number of events, counts	Mean, variance, Poisson distribution, binomial tests
Continuous	Measurements like height	Mean, standard deviation, normal distribution, t-tests

Important

Different data types and measurement scales require different analytical approaches!

The Statistical Process



Key Steps:

1. Identify the research question or problem
2. Collect data using appropriate sampling methods
3. Organise data according to type and measurement scale
4. Analyse using methods appropriate for the data type
5. Draw valid conclusions and inferences

Chapter 1 Summary

Key Concepts We've Learnt

1. **Statistics** is the science of collecting, organising, analysing, and interpreting data
2. **Probability** quantifies uncertainty and underpins statistical inference
3. **Two branches**: Descriptive (summarise) and Inferential (generalise)
4. **Population vs. Sample**: Complete set vs. representative subset
5. **Parameters vs. Statistics**: Population measures vs. sample measures
6. **Measurement scales**: Nominal, ordinal, interval, ratio
7. **Data types**: Qualitative vs. quantitative (discrete vs. continuous)
8. **Sampling methods**: Random (valid for inference) vs. non-random (invalid)
9. **Key terms**: Variable, experimental unit, measurement, observation

Foundation for the Course

These concepts form the foundation for all subsequent topics in probability and statistics, including descriptive statistics, probability theory, and statistical inference.

Next Steps

What's Coming Next:

- **Chapter 2**: Descriptive Statistics
  - Graphical methods for data visualisation (histograms, box plots)
  - Measures of central tendency (mean, median, mode)
  - Measures of dispersion (variance, standard deviation, range)
  - Measures of relative standing (percentiles, quartiles)
  - Coefficient of variation and other summary measures

Preparation for Next Class:

- Review today's concepts, especially data types, probability, and sampling methods
- Practice identifying measurement scales and variable types
- Read assigned sections from textbook (Walpole et al.)
- Complete homework exercises on populations, samples, and data classification

Questions?

Thank you for your attention!

Next Class: Chapter 2 - Descriptive Statistics

Consultation Hours:

- Mon: 14:00 – 17:00
- Wed: 14:00 – 17:00