MAT220 Lecture 1: Concepts in Statistics

Contents

1 What is Statistics? 2

1.1 The workflow of a statistical study 2

1.2 Why statistics can feel boring 3

1.3 Descriptive and inferential statistics 3

2 Types of Data 3

2.1 Data acquisition 3

2.2 Quantitative and categorical data 4

2.3 Levels of measurement 4

3 Sampling as a Concept 7

3.1 Core definitions 7

3.2 Worked example: Average Salaries in Tokyo 8

4 Sampling Methods 9

4.1 Simple random sample 10

4.2 Systematic sample 10

4.3 Stratified sample 11

4.4 Cluster sample 11

4.5 A common confusion: stratified versus cluster sampling 12

4.6 Other sampling methods 12

4.7 Returning to Jenny's Tokyo study 13

5 When Data Misleads Us 13

5.1 Sampling error and sampling bias 13

5.2 Non-sampling errors 14

5.3 Example: a biased survey question 15

5.4 Common ways conclusions become biased 15

5.5 Inferential errors 16

5.6 Correlation and causation 16

5.7 A checklist for trusting a statistical claim 16

This is the first lecture of MAT220, so our goal is to establish the basic language of statistics. We begin by asking what statistics *is*, and why it is not merely a collection of formulas. We then discuss types of data, including the four levels of measurement: nominal, ordinal, interval and ratio. After that, we introduce sampling, compare several sampling methods, and explain why different methods can lead to different conclusions. Finally, we discuss the main ways in which data can mislead us: sampling problems, non-sampling errors and mistakes in inference.

---

1 What is Statistics?

Statistics is the mathematical study of *data*. We can think of data as information collected from something in the real world. The purpose of statistics is to collect, organise, analyse, interpret and present this information so that we can make better decisions.

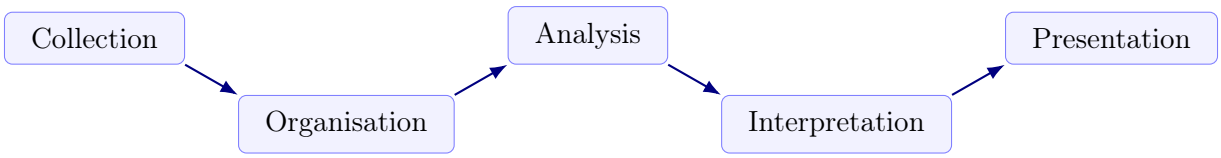
Statistics is used throughout science, business, government, medicine, sport, education and technology. Any time somebody tries to understand a complicated real-world system using data, they are using statistical thinking.

Definition

Statistics is the study of data, especially the collection, organisation, analysis, interpretation and presentation of data.

1.1 The workflow of a statistical study

A good statistical study mixes real-world context with mathematical techniques in order to discover new information about a system. One basic format for a statistical study is the following workflow:



Here:

- **Collection** involves measuring some system in the real world and recording data from it.
- **Organisation** involves rearranging the data into a useful format.
- **Analysis** involves using informed mathematics to describe useful features of the data set.
- **Interpretation** involves explaining what the results mean within a real-world context.
- **Presentation** involves communicating the results clearly, concisely and honestly.

Notice that the above workflow is not simply a collection of mathematical tools. Instead, a statistical study is an informed blend of mathematical techniques and real-world, practical considerations.

1.2 Why statistics can feel boring

Statistics can often feel boring and dry because it is formulaic. In fact, lots of the formulas used in statistics have a deep mathematical theory behind them, but practical statistics rarely uses that theory directly. Evaluated from the perspective of mathematical sophistication, statistics can often feel boring because it is so simple, and many of the most common formulas use only basic arithmetic.

In fact, statistics is also very interesting for exactly the same reason: even from the most basic mathematical operations, statistics creates a bunch of clever formulas that can be used to describe overall structures in data. Statistics is interesting in that it is such a powerful and useful tool that comes from such humble beginnings. The practical use of statistics cannot be overstated: the world is made up of information, and many properties about you and society can be converted into data and analysed.

1.3 Descriptive and inferential statistics

There are two main pillars of statistics: *descriptive* statistics and *inferential* statistics.

- **Descriptive statistics** is the art of describing the data that we already have. This includes tables, charts, averages, percentages, variation and other summaries.
- **Inferential statistics** is the science of making conclusions about a larger group based on a small amount of data. This includes confidence intervals, hypothesis testing and statistical modelling.

Example. Suppose we measure the commute times of 50 students.

- If we calculate the average commute time of those 50 students, we are doing **descriptive statistics**.
- If we use those 50 students to estimate the average commute time of all students at the university, we are doing **inferential statistics**.

2 Types of Data

2.1 Data acquisition

In reality, we do not always have access to a complete data set, and instead, usually, we only have access to the data that we are able to collect. Inferential statistics builds tools that help us figure out things about a larger group without perfectly knowing everything about it. These are extremely powerful and insightful methods; however, *our data is only as good as our method of acquiring it.* If we happen to have collected bad data in the first place, then no amount of fancy mathematics will be helpful in the analysis. Unwanted effects can accidentally be built into a statistical analysis, and this can cause incorrect conclusions to be made. It is our job to correctly identify these potential effects and to minimise them.

Important point

Bad data cannot usually be saved by good mathematics. If the data was collected in a confused, biased or careless way, then the final statistical conclusion may be unreliable even if the mathematics is technically correct.

2.2 Quantitative and categorical data

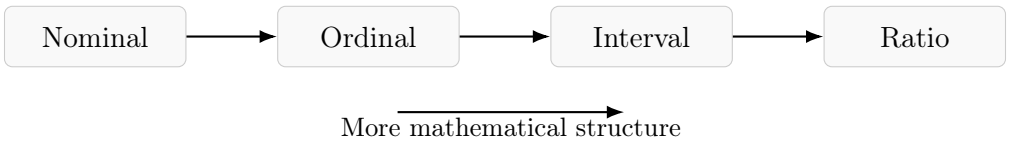
Data can roughly be divided into two broad types:

- **Quantitative data** is numerical data.
- **Categorical data** is non-numerical data.

Quantitative data has numerical structure, so it can often be analysed using arithmetic operations. Categorical data is more like a list of labels, names or groups, and therefore has much less mathematical structure to play around with.

2.3 Levels of measurement

There are many different types of data that we may collect in the real world. So, depending on the type of data we have access to, we may or may not be able to perform certain mathematical operations on it in a meaningful way. One way to navigate this potentially messy subject is to use *levels of measurement*, introduced by Stevens in 1946. There are four basic levels of measurement:



Roughly speaking, the more mathematical structure we have, the more operations that we can perform on our data. We will now review these levels of measurement one-by-one.

What is the point?

Levels of measurement help us be direct and explicit about our data. By classifying the type of data we have, we tell the audience what kinds of mathematical operations are meaningful and what kinds of claims would be inappropriate.

2.3.1 Nominal scales

Nominal data consists of labels or categories with no intrinsic order; it is *categorical*. There is no numerical structure to this level of measurement, so all we can do is test whether two observations are the same or different, and count their frequency of occurrence. But, there is no prescribed order or arithmetic that we have access to.

Nominal scale

Nominal data consists of qualitative categories with no natural order. Examples include:

- favourite colours,
- blood type,
- nationality,
- movie genre,
- train line used for commuting.

For nominal data, it makes sense to test equality and count frequencies.

2.3.2 Ordinal scales

The word "ordinal" means something similar to the word "order". As such, ordinal data is data that has a meaningful order or ranking. However, the gaps between ranks are not assumed to be equal or measurable in any sense.

Ordinal scale

Ordinal data consists of categories or values with a meaningful order, but unclear distances between neighbouring values.
Examples:

- ranking of competitors,
- pain scale,
- satisfaction responses such as "very dissatisfied", "dissatisfied", "neutral", "satisfied", "very satisfied",
- class standing such as first year, second year, third year and fourth year.

For ordinal data, medians and percentiles are often meaningful. Means can sometimes be used in practice, but one should be careful because the distance between neighbouring categories may not be equal.

2.3.3 Interval scales

Interval data has a meaningful order and equal differences between values. However, zero does not represent absolute absence of the quantity. Because of this, ratio statements such as "twice as much" are not generally meaningful.

Interval scale

Interval data has meaningful order and meaningful differences, but no genuine zero.
Examples:

- temperature measured in Celsius,
- year of birth.

For example, 20°C is not "twice as hot" as 10°C in a physically meaningful sense, because 0°C does not mean the absence of temperature. However, the difference between 20°C and 30°C is the same

size as the difference between 30°C and 40°C.

2.3.4 Ratio scales

Ratio data has order, equal intervals and a genuine zero representing absence of the quantity. Because of that, both differences and ratios are meaningful.

Ratio scale

Ratio data has meaningful order, meaningful differences and a genuine zero.
Examples:

- weight,
- height,
- length,
- time duration,
- number of students in a classroom,
- salary.

For example, a commute time of 60 minutes is genuinely twice as long as a commute time of 30 minutes, because there is an absolute zero of 0 minutes that can be used as a baseline. This kind of comparison makes sense because 0 minutes means no time has passed.

2.3.5 Levels of measurement: summary

Level	Main structure	Allowed comparisons	Example
Nominal	Categories only	same or different	blood type
Ordinal	Categories with order	greater or less than	satisfaction rating
Interval	Equal differences, no true zero	differences	Celsius temperature
Ratio	Equal differences and true zero	differences and ratios	height

Exercise

Classify each variable as nominal, ordinal, interval or ratio.

(a) A student's favourite convenience store.
(b) A satisfaction rating from 1 to 5.
(c) A temperature measured in Celsius.
(d) The time taken to commute to campus.

Solution

(a) Nominal, because the responses are category labels.

(b) Ordinal, because the responses have an order, but the gaps between ratings may not be equal.

(c) Interval, because differences are meaningful but zero Celsius is not a genuine absence of temperature.

(d) Ratio, because time duration has a genuine zero and ratios are meaningful.

3 Sampling as a Concept

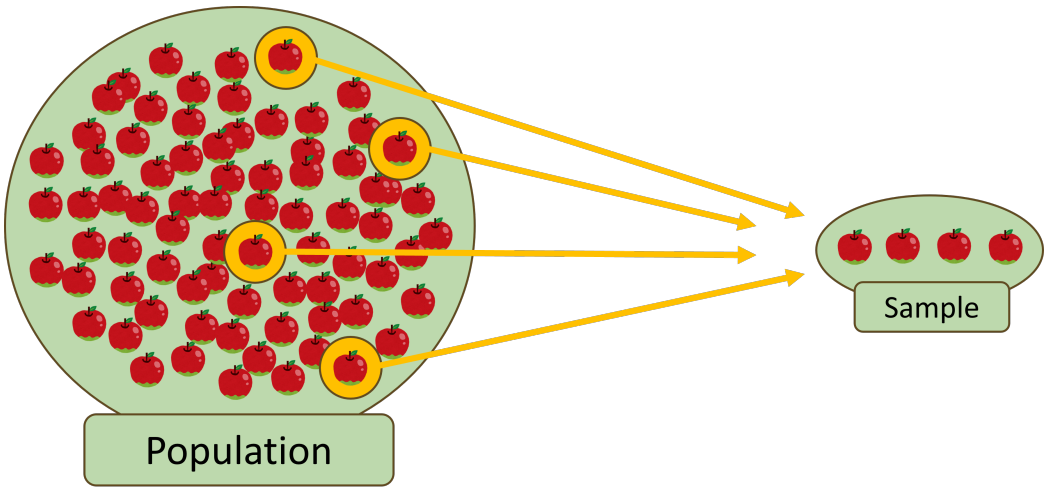
3.1 Core definitions

As previously mentioned, when studying the real world it is very often the case that the collection of things we want to study is simply too large to work with. So, in practice we will often just take a smaller portion and study that instead. This realistic, practical consideration motivates the following terminology.

Definitions

- **Population:** the complete group of people or things you want to learn about.
- **Sample:** the smaller group you actually collect data from, chosen from the population.
- **Population parameter:** a true number that describes the whole population, even if you do not know it exactly.
- **Sample statistic:** a number calculated from the sample, often used to estimate a population parameter.

In other words, a population is the collection we actually want to study. The facts that we want to discover are called population parameters, but often we can't discover these directly because the population is too big. So, instead we take a subset, which is just a smaller group selected from the overall population. We can then perform easier calculations on this smaller group, and this gives us some partial information known as a sample statistic. Here is a visual example of a sample:



3.2 Worked example: Average Salaries in Tokyo

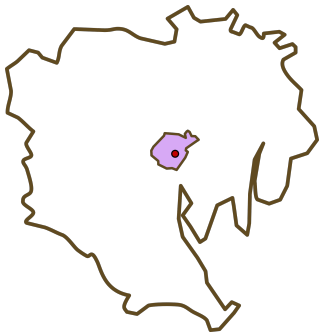
The process of obtaining a sample is known as *sampling* from the population. Sampling is an extremely important part of statistics, and large parts of our course will be dedicated to the relationship between samples and the overall population. However, we have to be very careful when sampling, because we may accidentally introduce some effects that are not representative of the true population. We will now illustrate this idea with an example.

Suppose that Jenny is an intern at a financial firm in Chiyoda, central Tokyo. She has been tasked by her boss to find out what the average salary of residents of Tokyo is. So, Jenny needs to perform a statistical study in order to find this out. Since Jenny does not work for the municipal government, she does not have access to complete salary data about the population of Tokyo. Instead, she decides to take a sample of 30 residents of Tokyo, and then take the average salary of these people.

When performing her statistical study, Jenny realises that her results will depend heavily on the people that she asks. For instance, the sample will be different based on:

- the time of day that she takes the sample,
- the location where she takes her survey,
- the age of the people involved,

and so on. Below are four different possible ways that Jenny might take a sample of 30 people, together with the problems inherent to each.

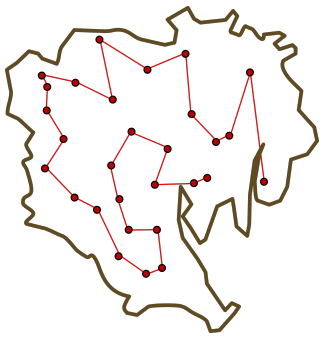


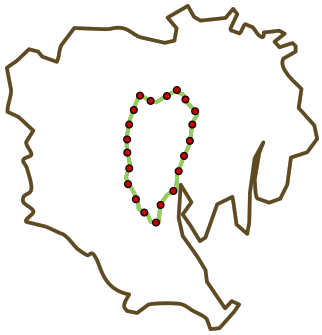
Method 1: Selecting 30 participants from her office in Chiyoda.

Lack of diversity: it is certainly convenient to take a sample of people from her local neighbourhood. However, her numbers may not represent Tokyo overall. Some neighbourhoods are richer or poorer, older or younger, or have different types of jobs. So even if she asks 30 people, she may accidentally measure the character of that neighbourhood rather than Tokyo as a whole. In the case of Chiyoda, her sample is likely to overestimate average salary.

Method 2: Randomly selecting 30 participants from anywhere in Tokyo.

Logistical difficulty: although it may feel unbiased to sample people randomly from across the city, actually doing so would be expensive and time-consuming. Not only would it be difficult to travel to the various neighbourhoods she has randomly selected, she would also run into other potential biases depending on how she selects a representative to ask. Nevertheless, a plan that looks unbiased on paper may fail in practice because of resource constraints.

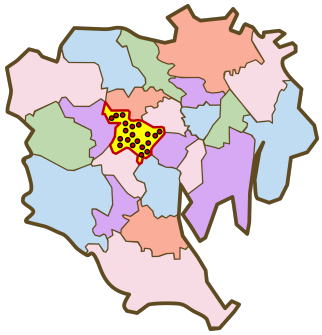




Method 3: Travelling the Yamanote line, stopping at each of the 30 stations and asking one person at each station.

Lack of accessibility: sampling "one person per station" may include tourists and outsiders who do not live in Tokyo. It can also accidentally introduce a bias towards tourist stations, commuter hubs or business districts, which can skew the sample toward certain kinds of workers. Put differently, this sample may have a bias towards a particular type of Tokyo resident: those who work or travel around the Yamanote line.

Method 4: Randomly selecting one of the 23 wards of Tokyo, and then randomly selecting 30 participants from within that ward.



Lack of diversity: although it is more convenient than travelling throughout all of Tokyo, and less biased than Jenny's local neighbourhood of Chiyoda, there will *still* be a lack of diversity in this type of sample. Randomness does not guarantee that the diversity of a population is fully represented in small samples. Instead, it only makes the first stage of selection random.

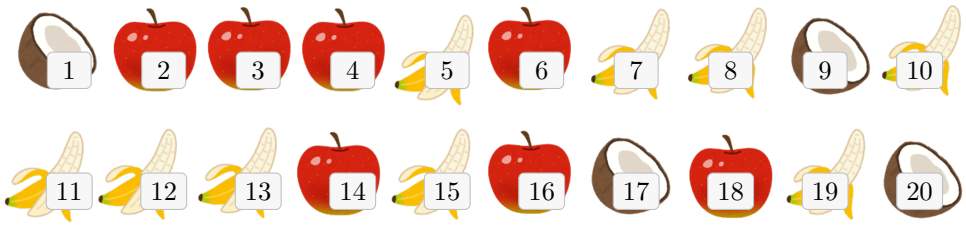
So, what is the best sampling method here? In fact, it depends on what Jenny is trying to do. In practice, there is often no single "best method". Instead, sampling is often a balancing act between resources and error, and sometimes we must compromise. Ideally, these biases ought to be understood by the researcher running the study, and this should be made as explicit as possible when communicating results.

Sampling in practice

Sampling is often a balancing act between resources and error. The most accurate method of data collection is usually more expensive, more difficult or more time-consuming. Sometimes we need to make a compromise, but we should make the compromise explicit when presenting the results.

4 Sampling Methods

Motivated by Jenny in the previous example, we will now review some common sampling methods. Throughout this section, we will suppose that we have a population consisting of 20 pieces of fruit: 7 apples, 9 bananas and 4 coconuts.



Our goal is to choose a sample of size 5 from this population using different sampling methods.

4.1 Simple random sample

Definition: Simple Random Sample

A **simple random sample** is obtained so that every possible sample of the given size is equally likely.

For example, from our fruit population, one possible simple random sample of size 5 is:

Sample 5 9 10 12 18

Another possible simple random sample of size 5 is:

Sample 2 4 8 14 17

In a simple random sample, we do not deliberately control how many apples, bananas and coconuts appear in the sample. The sample is fair in its selection rule, but it may still be unbalanced by chance.

4.2 Systematic sample

Definition: Systematic Sample

A **systematic sample** is obtained by choosing a random starting point and then selecting every kth member of the population after that.

For example, if we start at 1 and choose every 4th member, we obtain:

Sample 1 5 9 13 17

If we start at 3 and choose every 4th member, we obtain:

Sample 3 7 11 15 19

Systematic sampling can be convenient, especially when the population is already arranged in a list. However, it can be dangerous if the list itself has a repeating pattern. For example, if every 4th item in a factory list comes from the same machine, then choosing every 4th item may accidentally sample only one machine.

4.3 Stratified sample

Definition: Stratified Sample

A **stratified sample** is obtained by first dividing the population into groups, called **strata**, based on a relevant characteristic, and then randomly sampling from each stratum.

In our fruit example, the natural strata are the 3 types of fruit:

$$\text{apples}, \qquad \text{bananas}, \qquad \text{coconuts}.$$

A stratified sample might deliberately choose some apples, some bananas and some coconuts, so that the sample represents all three fruit types.

Example stratified sample of size 5:

Sample 2 6 9 14 19

Here we have sampled from each fruit type. The goal is not just randomness, but representation across important groups.

Stratified sampling is useful when the population has important subgroups and we want to make sure that each subgroup is represented.

4.4 Cluster sample

Definition: Cluster Sample

A **cluster sample** is obtained by dividing the population into clusters, usually based on natural or convenient groupings, and then randomly selecting one or more whole clusters.

For example, suppose our list of fruit is divided into clusters of size 5:

$$\{1,2,3,4,5\}, \quad \{6,7,8,9,10\}, \quad \{11,12,13,14,15\}, \quad \{16,17,18,19,20\}.$$

If we then randomly choose the second cluster, our sample is:

Sample 6 7 8 9 10

If we randomly choose the third cluster, then our sample is:

Sample 11 12 13 14 15

Cluster sampling is useful when it is expensive or difficult to sample individuals scattered across the whole population. For example, a researcher might randomly choose several classrooms and survey every student in those classrooms.

4.5 A common confusion: stratified versus cluster sampling

Stratified sampling and cluster sampling can feel similar because they both divide the population into groups. The difference is the role that the groups play.

Method	How groups are formed	How the sample is chosen
Stratified sampling	Groups are formed using important characteristics, such as fruit type, gender, year level or ward.	Randomly sample from every stratum.
Cluster sampling	Groups are often natural or convenient clusters, such as classrooms, neighbourhoods or list blocks.	Randomly choose whole clusters.

In short: in **stratified sampling**, we sample *from every group*, but in **cluster sampling**, we sample *whole groups themselves*.

4.6 Other sampling methods

There are many other sampling methods. Some are useful in practice, but they often come with serious limitations.

Method	Basic idea
Convenience sampling	Sample whoever is easiest to reach.
Voluntary response sampling	People choose whether to participate.
Panel sampling	Follow the same group of people over time.
Haphazard sampling	Sample in an informal or careless way without a clear random rule.
Snowball sampling	Existing participants help recruit future participants.
Extreme-case sampling	Deliberately sample extreme cases, such as the smallest and largest values.

These methods are not automatically wrong, but they must be interpreted carefully. A convenience sample, for example, may be cheap and fast, but it may tell us more about who was easy to reach than about the full population.

4.7 Returning to Jenny's Tokyo study

We can now describe Jenny's methods using sampling language:

- Method 1: **Asking 30 people near her office in Chiyoda.** This is closest to a convenience sample. The main concern is that Chiyoda may not represent Tokyo.
- Method 2: **Randomly sampling 30 people from Tokyo.** This is closest to a simple random sample. The main concern is that it is accurate in principle, but difficult in practice.
- Method 3: **Asking one person at each Yamanote station.** This is closest to a route-based convenience sample. The main concern is that this may sample commuters and tourists rather than residents.
- Method 4: **Randomly picking one ward, then sampling inside it.** This is closest to a cluster-style sample. The main concern is that one ward may not represent all wards.

The vocabulary does not solve the practical problem by itself, but it helps us describe the strengths and weaknesses of the study clearly.

5 When Data Misleads Us

There are three main reasons why a statistical conclusion might fail.

Three ways a statistical conclusion can fail

1. **Sampling problems:** the sample does not honestly represent the population.
2. **Non-sampling errors:** the data itself is distorted, even if the sample was chosen well.
3. **Inferential errors:** the conclusion goes beyond what the data actually supports.

A good statistical study must deal with all three. It is not enough to have a large sample, and it is not enough to use the correct formula. The whole chain of reasoning must be trustworthy.

5.1 Sampling error and sampling bias

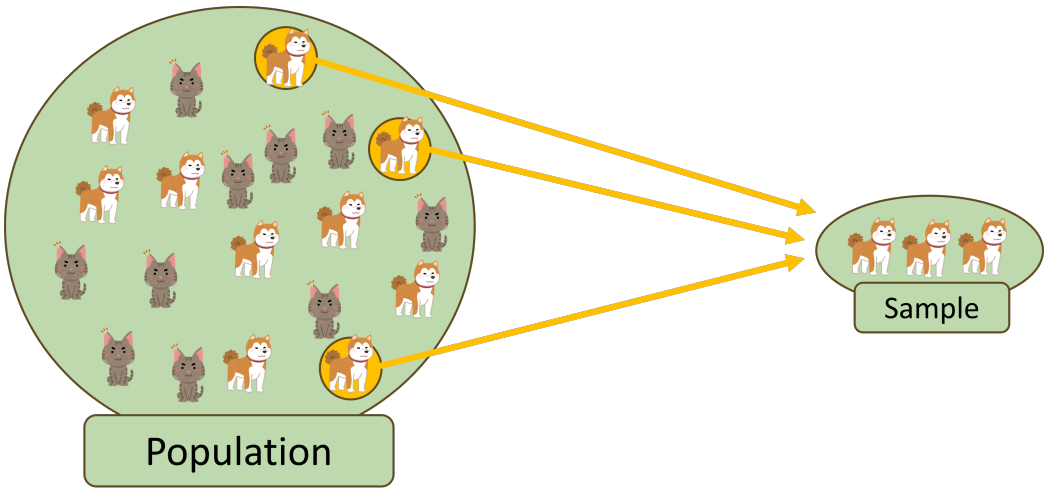
There are two important types of sampling problems.

Sampling error versus sampling bias

- **Sampling error** is a random mismatch between the sample and the population. This can happen even when the sampling method is fair.
- **Sampling bias** is systematic unfairness in the sampling method. The method itself makes some parts of the population more or less likely to be included.

5.1.1 Example: sampling error

Suppose the population contains cats and dogs in equal numbers. We take a simple random sample of size 3, and by chance all 3 selected animals are dogs.



This is a sampling error, because the method was fair but the particular sample was unlucky.

5.1.2 Example: sampling bias

Suppose Jenny only asks people in a wealthy neighbourhood such as Chiyoda about salary. This is not merely bad luck. The sampling method itself is systematically unfair because it overrepresents wealthy people.

This is a sampling bias. If Jenny repeats the same method many times, she will keep getting samples that tend to overestimate the average salary of Tokyo.

5.2 Non-sampling errors

Even if the sample is chosen well, the data itself can still be distorted. These distortions are called **non-sampling errors**.

Definition: Non-sampling Error

A **non-sampling error** is an error in a statistical study that is not caused by the random act of choosing a sample.

Common non-sampling errors include:

- unclear or leading questions,
- inaccurate measurements,
- dishonest answers,
- missing responses,
- recording mistakes,
- data entry mistakes,
- faulty equipment,

and so on. Importantly, these errors may occur even if the sampling method itself is reasonable.

5.3 Example: a biased survey question

Suppose you survey commuters in Tokyo and ask:

"How awful and terrible was your train ride this morning?"

This is a bad question because it assumes the train ride was awful and/or terrible, and therefore pushes the respondent toward a negative answer. A better version would be:

"How was your train ride this morning?"

Very bad	Bad	Fine	Good	Very good

Here the wording is much more neutral, and does not encourage a particular value judgement from the respondent.

5.4 Common ways conclusions become biased

Some biases are not just technical errors. Instead, they are sometimes failures in how we select, remember or interpret evidence. Here are three common examples of this phenomenon.

Survivorship bias

Survivorship bias happens when we only look at the examples that survived a selection process.
Example: "Architecture in the past was so beautiful, and modern buildings are so ugly."
This may be misleading because many ugly old buildings were destroyed, while beautiful old buildings were preserved. The surviving buildings are not a random sample of past buildings.

Selection bias

Selection bias happens when the data is selected in a way that favours a certain conclusion.
Example: "I asked people in Kabukicho whether Tokyo is noisy and chaotic. They said yes. Therefore Tokyo is noisy and chaotic."
Kabukicho is not a representative sample of Tokyo. The conclusion may tell us more about Kabukicho than about Tokyo as a whole.

Funding bias or conflict of interest

Funding bias can occur when a study is funded by a group that benefits from a particular conclusion.
For example, if an industry funds research about whether its own product is harmful, we should pay careful attention to the study design, the data, the analysis and the interpretation.

5.5 Inferential errors

Even when the sample is reasonable and the data is recorded accurately, the final conclusion can still be wrong. This happens when the inference goes beyond what the evidence supports.

Common inferential errors include:

- generalising from a small sample to a huge population,
- treating correlation as causation,
- ignoring lurking variables,
- ignoring confounding variables,
- using a statistic correctly but interpreting it incorrectly,
- making a conclusion stronger than the evidence justifies.

5.6 Correlation and causation

Two variables may be related without one causing the other. If two things happen together, this does not automatically mean that one is responsible for the other.

Example. Suppose that on very hot days in Tokyo, convenience stores sell more bottled water and more people report heat exhaustion.
It would be silly to conclude that buying bottled water causes heat exhaustion. A more reasonable explanation is that hot weather affects both variables.

In this example, bottled water sales and heat exhaustion are correlated, but the causal explanation involves another variable: temperature.

5.7 A checklist for trusting a statistical claim

Before trusting a statistical conclusion, ask the following questions.

1. **Population:** What population is the claim really about?
2. **Sample:** Who or what was actually sampled?
3. **Sampling method:** How were the sample members selected?
4. **Data quality:** Were the measurements, questions and records reliable?
5. **Missing data:** Who did not respond, and could that matter?
6. **Variables:** Are there lurking or confounding variables?
7. **Conclusion:** Does the conclusion follow from the data, or does it overreach?

Summary

A statistical claim is only as strong as the full chain behind it:

$$\text{collection} \longrightarrow \text{organisation} \longrightarrow \text{analysis} \longrightarrow \text{interpretation} \longrightarrow \text{presentation}.$$

If one part of the chain is weak, the final conclusion may be weak as well.