IEDA 2540 Statistics for Engineers
Sampling Methods

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Spring, 2025

Descriptive and Inferential Statistics

Statistics enable us to **summarize** and **predict** more precisely than we normally would in everyday conversation.

- **Descriptive Statistics**
  Methods used to summarize or describe our observations.
  - **Example:** "On average, I cycle about 100 miles a week."
- **Inferential Statistics**
  Using observations as a basis for making estimates or predictions, i.e., inferences about situations that have not yet been investigated.
  - **Example:** "We can expect a lot of rain at this time of year."
  - **Example:** "The earlier you start preparing, the more likely you are to do well in exams."

Inferential Statistics and Population

Inferential Statistics goes beyond what has been observed (the so-called **samples**) and attempts to predict what has not been observed (the **population**).

Population

A population is the entire set of individuals to which study findings will be generalized. In other words, population is the subjects of interest of a study.

Example:

- All light-bulbs that have been or will be manufactured.
- All meteorites fallen or will fall onto Earth.
- All past and future test results of a local school.

Key Definitions

It is often unrealistic to study all members of a population.

- Prohibitive cost: light-bulbs are “tested to destruction,” none left to sell.
- Population size may be astronomical: all meteorites in the universe.
- Subject may not be available yet: future test results of a school.

Sample

A sample is a subset of the target population. From the sample, you collect the measurements (observations) known as data.

The hope is to generalize our conclusions from a sample to the population.

How Safe are Such Generalizations?

- The reliability of our generalizations depends on how well the sample mirrors the population.
- Is the sample truly **representative** of the population?

Example: A researcher finds that a group of 12-year-old Mandarin-speaking girls learn Cantonese more quickly when taught by conversation than by self-teaching materials.

She wants to infer how other people might best learn Cantonese.

Which do you think is the most/least reasonable population to generalize to?

- All 12-year-old Mandarin-speaking girls;
- All 12-year-old Mandarin-speaking children;
- All 12-year-old children;
- All girls;
- All learners.

Margin of Error in Generalizations

Even for all 12-year-old Mandarin speaking girls, there is a big margin of doubt: how typical was the group of samples?

- The group of girls may come from Guangdong Province.
- They may have a higher interest or natural ability in learning Cantonese.

Attempts to generalize to populations less and less like the sample will be more and more liable to error.

Example: Stagflation (stagnation + inflation). The 1973 oil crisis and subsequent energy shocks led to a combination of high inflation and stagnant economic growth. Traditional policies to reduce inflation might further dampen economic growth.

Paradox of Sampling

The Paradox

A sample is misleading unless it is representative of the population. But how can we tell if it is representative unless we already know what we need to know about the population?

The paradox cannot be completely resolved – uncertainty must remain.

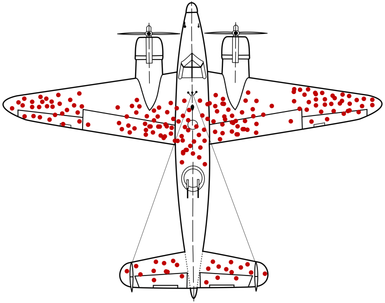
Statistical methods enable us to collect samples that are likely to be as representative as possible.

Selection Bias

Sampling usually unintentionally biases towards a sub-population.

Example: Survivorship Bias

- This hypothetical pattern of damage of surviving aircraft shows locations where they can sustain damage and still return home.
- If the aircraft was reinforced in the most commonly hit areas, this would be a result of survivorship bias because crucial data from fatally damaged planes was being ignored; those hit in other places did not survive.
- **Samples "choose themselves".**



Selection Bias: Inspection Paradox

Example: Estimating the length-of-stay of tourists at the Disneyland Resort

- **Method 1**: Survey 100 tourists randomly at the exit.
- **Method 2**: Survey 100 tourists randomly in the park.

Which method do you think better represents the population of all tourists?

Example: Estimating the effective reproduction number of Covid-19

- **Method 1**: Backward tracing.
- **Method 2**: Forward tracing.

Which method do you think provides a more accurate estimation of the effective reproduction number?

Simple Random Sampling

For samples to be representative, the subjects must be chosen at "random" from the entire population.

Simple random sampling

Each subject has an equal chance of being selected.

- This is not easy, as we saw in survivorship bias/inspection paradox examples.
- To avoid such bias, it is preferable to use mechanical or blind methods of selecting a random sample.

  Example: For a lucky draw at an alumni dinner, the organizer puts all names in a box, shuffles it, and selects a name without looking into the box.
- In general, with larger population sizes, we assign each subject an "ID" number and refer to a "table of random numbers" to pick samples.

Systematic Sampling

Systematic sampling

The first item is chosen randomly, and then every n-th element is selected from the population, where n is the sampling interval.

Example: Interview on the street:

- If you simply interview passers-by “randomly,” you may fail to obtain a representative sample – people who appear approachable or are not in a hurry are more likely to be interviewed.
- To mitigate this, you need to select subjects “blindly,” which can be done by systematic sampling – interview every tenth passer-by (no matter how unapproachable they might look!).
- Of course, they may still refuse to be interviewed...

Potential Issue with Simple Random Sampling

Say you want to sample 100 students and study their attitudes to the foods in the HKUST.

- Even if you sample uniformly at random, it is possible that the sample collected will contain only girls (even though extremely unlikely)!
- In such cases, it will be more preferable to use what is called a **stratified random sample**.

Stratified Sampling

Stratum

A **stratum** is a layer or a series of layers of rock in the ground.

Stratified sample

A stratified sample is obtained by dividing the population into **homogeneous and non-overlapping** groups called **strata**, and then obtaining a simple random sample from each stratum.

Stratified Sampling: Example

Example: Use age groups as strata: (10-25, 25-65, 65+).

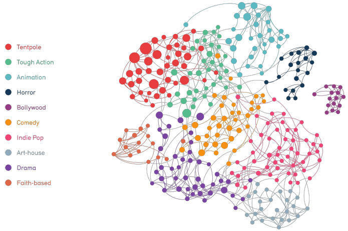
- Select subjects randomly from each stratum.
- More appropriate when individuals in the same stratum are similar, but the strata differ from each other.
- More appropriate than simple random sampling when investigating specific groups of subjects.
- Particularly useful when certain groups are deemed “uncommon”.
  Example: Investigating the COVID-19 lethal rate among elderly people (65+).

Cluster Sampling

When researchers want to collect a large amount of data, it may be impractical to use simple random sampling.

Cluster Sampling

Divide a population into smaller groups known as **clusters** and then randomly select among these clusters and collect all individuals from the selected cluster(s) to form the sample.



Example: Using pre-existing units as clusters

- By geographic location, e.g., zip codes, city blocks, schools.
- By social media, e.g., Facebook groups.

Cluster Sampling Requirements

Cluster sampling requires the following to achieve a representative sample:

- **Heterogeneity within each cluster:** Because we only select a few clusters, each cluster needs to be representative of the population.
- **Homogeneity between clusters:** If clusters are drastically different from each other, the selected sample may be biased.
  Example:
  - Studying hobbies but clustering the population by genders.
  - Estimating the Covid-19 lethal rate but clustering the population by age group.

Cluster Sampling: Cost Efficiency and Feasibility

Cluster Sampling is a practical, cost-effective method for sampling large, geographically dispersed populations.

- **Cost Efficiency:** Reduces travel and administrative costs by focusing on clusters.
- **Feasibility:** Easier to list and sample clusters (e.g., schools) than the entire population.
- **Workload Reduction:** Less resource-intensive, making it ideal for pilot studies.

Example: Survey the health habits of residents in a large metropolitan area.

- Instead of visiting scattered individual households, they divide the city into clusters by neighborhoods. They randomly select a few neighborhoods and survey all residents within those clusters.
- This method reduces travel and administrative costs by focusing data collection on a few localized areas.

Cluster Sampling versus Stratified Sampling

Cluster Sampling	**Stratified Sampling**
Population is divided into clusters	Population is divided into strata
Homogeneity between clusters	Homogeneity within strata
Heterogeneity within clusters	Heterogeneity between strata
Randomly select clusters	Select randomly from each stratum
Cost saving, large scale data collection	More representative and unbiased sample

Focus on Simple Random Sampling

We shall nonetheless focus on simple random sampling throughout this course.

- It is usually the best we can do.
- It serves as the building block for other sampling methods (e.g., stratified sampling, cluster sampling).
- It yields the simplest and richest quantitative models.