Lecture 02a – Sampling designs: simple random sampling

ENVX2001 Applied Statistical Methods

Januar Harianto

The University of Sydney

Feb 2025

THE UNIVERSITY OF SYDNEY

Outline

- Last week
- Observational studies
- Interpreting sampled data
- Confidence intervals
- Data story: soil carbon
- Simple random sampling
- Tomorrow: stratified random sampling
- Thanks!

Last week

- Population vs. sample
- Parameters (pop) and statistics (sample)
  - ➟ central tendency (mean, median, mode)
  - ➟ spread/dispersion (variance, standard deviation)
- Confidence intervals – a brief introduction

Observational studies

Overview

Aspect	Observational study	Controlled experiment
Control	No control over the variables of interest: **mensurative** and **absolute**	Control over the variables of interest: **comparative** and **manipulative**
Causation	Cannot establish causation, but perhaps **association**	Can establish **causation**
Feasibility	Can be done in many cases	May be destructive and thus cannot always be done
Examples	Surveys, monitoring studies, correlational studies, case-control studies, cohort studies	Clinical trials, A/B testing, laboratory experiments, field experiments
Statistical Tests	Correlation, regression, chi-squared tests, **t-tests**, **one-way ANOVA**, time series analysis	**T-tests**, **one-way ANOVA**, factorial ANOVA, regression

We will focus on the fundamentals behind **observational studies** this week.

Two common types

Surveys

- Estimate a statistic (e.g. mean, variance), but
- **no temporal change** during estimate.
- *E.g. measuring species richness in a forest*.

Monitoring studies

- Estimate a ***change*** in statistic (same as above), and
- **temporal change** across observations, i.e. before and after.
- *E.g. measuring species richness in a forest* ***before and after a fire***.

Sampling designs

Simple random sampling:

- Each unit has an equal chance of being selected.
- **Randomly sample units from the entire population.**

Simple random sample of 20 numbers from population of 100 numbers

Stratified random sampling

- The population is first divided into *strata* (more on this later).
- **Randomly sample units within each strata by simple random sampling, *standardised* by the inclusion probability (or weight) of each strata.**

What is “random” sampling?

Random selection of **finite** or **infinite** population units.

> What does *random* mean?

Within a population, **all** units have a > 0 probability of being selected *i.e. everything has a chance to be selected*.

- This *chance* is called the **inclusion probability (π_i)**:
    - ➟ π_i is **equal** *within* a population unit – *i.e. all units have the same chance of being selected*.
    - ➟ π_i **not** necessarily equal *between different* population units – *i.e. a unit from one population unit may have a different chance of being selected than a unit from another population unit* - more on this later.

How do we perform random sampling in real life?

- **Random number generator** (RNG) – e.g. R’s `sample()` function.
- **Random number table** – e.g. Random number table by the National Institute of Standards and Technology (NIST).

Interpreting sampled data

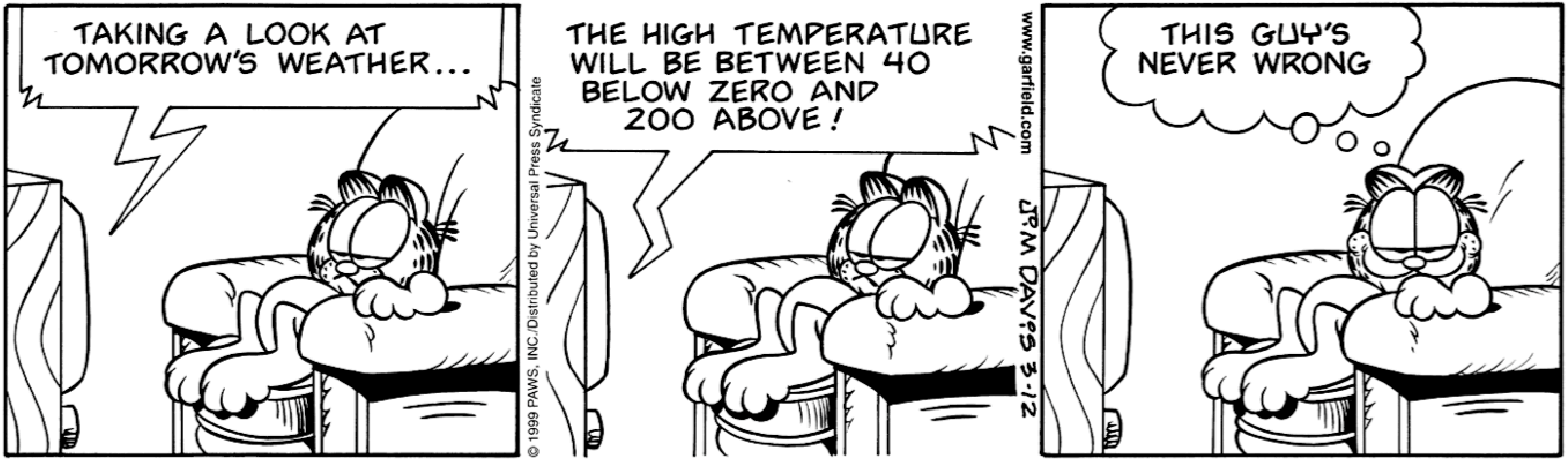
We know that…

From the previous lecture

- **Sample mean** is a good measure of central tendency.
- **Sample variance** is a good measure of dispersion.
- **Sample size** affects the precision of the sample mean.

Can we combine all of the above in a single statistic?

Confidence intervals



Combining an estimate with its precision

A **confidence interval (CI)** is:

- A range of values, derived from a sample of data, that is used to estimate the range of values for a population **parameter**
- Crucial for **hypothesis testing** and **estimation**, the basis of statistical inference

You will often see CIs in scientific papers, reports, and news articles.

Calculating confidence intervals

What we need

1. **Estimate** of the population parameter, i.e. the **sample mean** $(\bar{x})$
2. **Critical value** (t_{n-1}) from the sampling distribution of the estimate, which depends on the **number of samples** and the **confidence level**. This is usually based on the **t-distribution**
3. **Standard error** of the estimate, standardised by the number of samples i.e. **SE of the mean** $(SE_{\bar{x}})$

All these (mean, standard error and critical value) are *combined* to form the confidence interval.

Breakdown

In general, a CI has the form:

$$\text{estimate} \pm \text{margin of error}$$

where the margin of error is a **function of the standard error** of the estimate:

$$\text{estimate} \pm (\text{critical value} \times \text{standard error}\ (\text{estimate}))$$

where the critical value is based on the **sampling distribution** of the estimate i.e. the ***t*-distribution**.

Formula for 95% Confidence Interval (CI)

$$\bar{x} \pm \left(t_{n-1} \times \frac{s}{\sqrt{n}}\right)$$

Step-by-step calculation by hand

1. Calculate the sample mean, $\bar{x}$
2. Calculate the sample standard deviation, s
3. Determine the standard error of the mean, $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$
4. Look up the t-value, t_{n-1}, from the t-distribution table for the 95% confidence level and $n-1$ degrees of freedom (manual table or function in R).
5. Compute the margin of error:

$$\text{Margin of Error} = t_{n-1} \times SE_{\bar{x}}$$

6. Finally, the 95% CI is:

$$\bar{x} \pm (t_{n-1} \times SE_{\bar{x}})$$

You need to be able to calculate this by hand/calculator.

More definitions

Degrees of freedom (df)

The number of independent values in a sample. It is calculated as:

$$\mathrm{df} = n - 1$$

where n is the number of samples. We subtract 1 because of Bessel's correction.

Example

- Four numbers have a mean value of 5. The first three numbers can be any value, so they are 3, 10 and 7
- **The fourth number must be 5 to make the mean 5**
- Therefore, the group of four numbers have 3 i.e. $n-1$, degrees of freedom

More definitions

t-critical value

Given a **confidence level** (e.g. 95%) and **degrees of freedom** (df), the t-critical value is a value that is used to determine the **margin of error** in a confidence interval.

It assumes: - A **t-distribution** of the data - A **symmetric distribution** around the mean

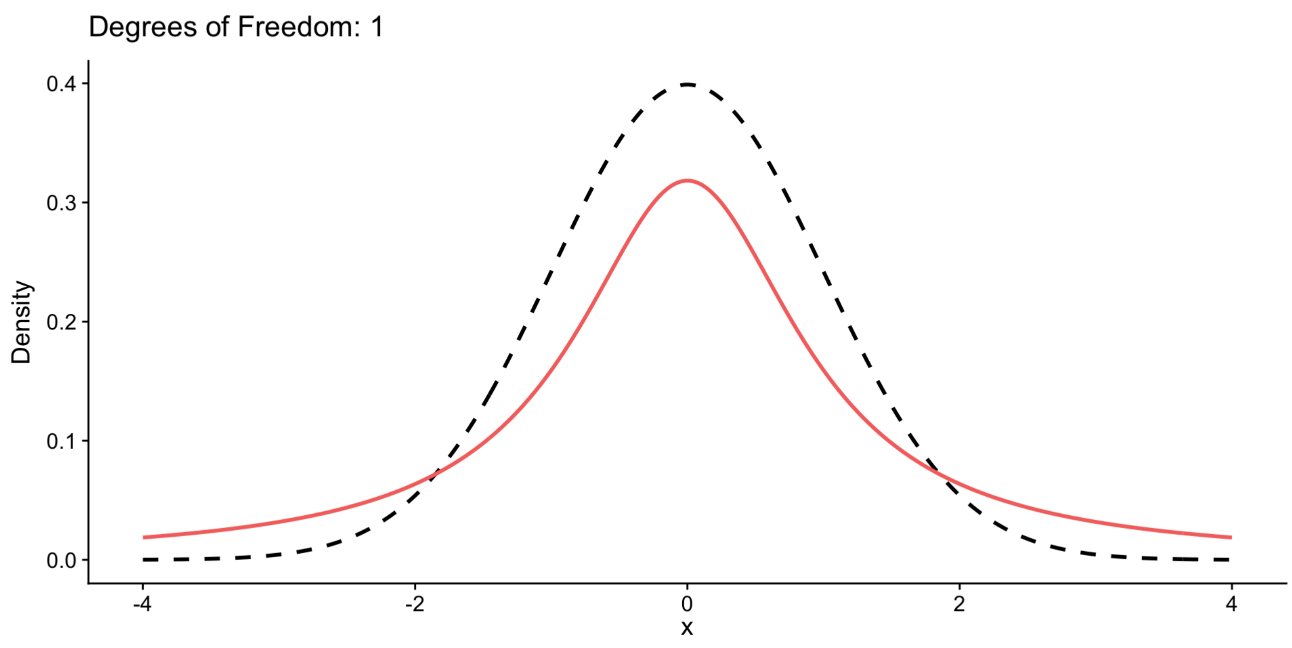
More definitions

t-distribution

A probability distribution that is used to estimate the population. Using the correct distribution gives better estimates.

- Similar to the normal distribution, but with **heavier tails**.
- As the sample size increases, the t-distribution approaches the normal distribution, so we do not need to worry about it for large sample sizes.

▶ Code



Interpreting confidence intervals

- Confidence intervals depend on a specified **confidence level** (e.g. 95%, 99%) with higher confidence levels producing wider intervals (i.e. more conservative).
- Another way to think of it: a **range of values** that we are fairly sure contains the true value of the population parameter.

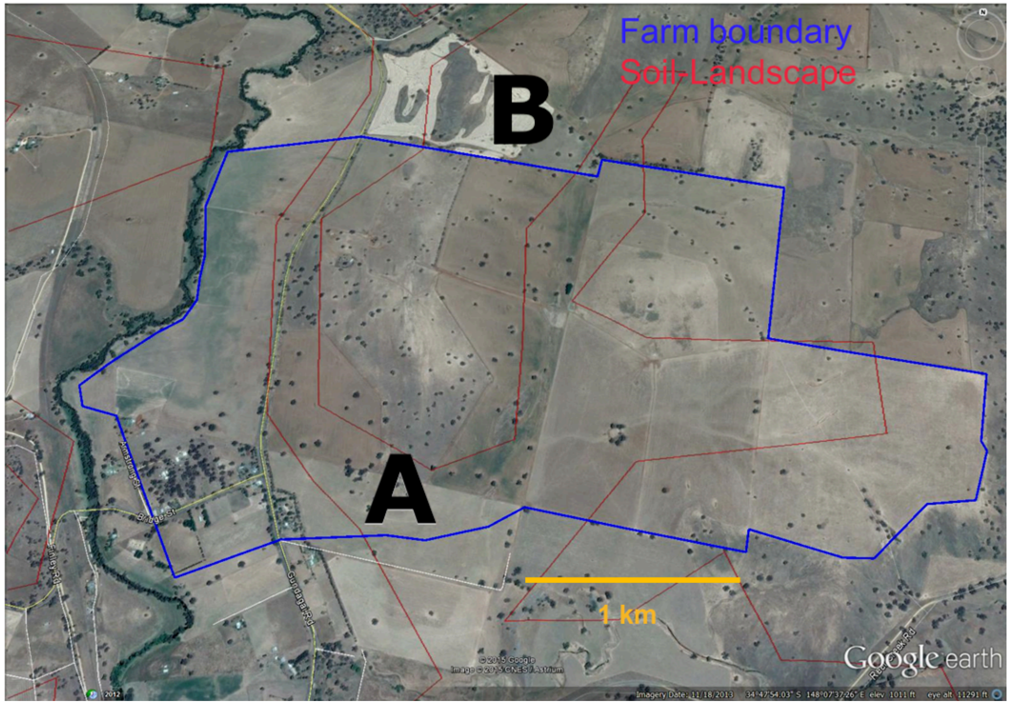
Fishing analogy

A confidence interval is like a fishing net:

- A wider net (interval) is more likely to catch the fish (true value)
- A spear is *less* likely to catch the fish
- The net width represents our uncertainty about the true value

Data story: soil carbon

Soil carbon



Soil carbon content was measured at 7 locations across the area. The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha).

▶ Code

```
[1] 48 56 90 78 86 71 42
```

What is the mean soil carbon content and how confident are we in this estimate? How this is calculated depends on whether we used simple random sampling or stratified random sampling.

Simple random sampling

Assuming that the soil carbon content is a simple random sample from the population, let's calculate the 95% confidence interval.

Mean and 95% CI

Step-by-step calculation

1. Mean: $\bar{x} = \frac{48+56+90+78+86+71+42}{7} \approx 67.3$
2. Standard deviation: $s \approx 18.84$
3. Standard error: $SE = \frac{s}{\sqrt{7}} \approx 7.12$
4. t-value (95% CI, df = 6): $t_{0.975,6} \approx 2.447$
5. Margin of error: $t_{0.975,6} \times SE \approx 17.43$
6. Which gives: $(67.3 - 17.43, 67.3 + 17.43) = (49.87, 84.73)$

And so we report the mean soil carbon content as 67.3 t/ha with a 95% CI of (49.87, 84.73) t/ha or 67.3 ± 17.43 t/ha.

Implementation in R

Manual calculation

```r
# Step 1: Calculate the sample mean of soil carbon content
mean_soil <- mean(soil)

# Step 2: Calculate the sample standard deviation of soil carbon content
sd_soil <- sd(soil)

# Step 3: Calculate the standard error of the mean (SE)
se_soil <- sd_soil / sqrt(length(soil))

# Step 4: Calculate the t-critical value for a 95% confidence interval
t_crit <- qt(0.975, df = length(soil) - 1)

# Step 5: Calculate the margin of error (t_crit * SE)
# and then determine the lower and upper bounds of the confidence interval
ci <- mean_soil + c(-1, 1) * (t_crit * se_soil)

# Step 6: View the calculated 95% confidence interval
ci
```

```
[1] 49.84627 84.72516
```

There are ways to calculate this in R quickly, but it is important to understand the manual calculation.

Questions

- How precise is our estimate?
- How big a change must there be to estimate a statistically significant change?
- **Can we sample more efficiently?**

To answer these questions, we need to compare simple random sampling with a hypothetical stratified random sampling design (i.e. what if we had considered stratification before sampling?)

Tomorrow: stratified random sampling

Thanks!

Questions?