Lecture 02a – Sampling designs

ENVX2001 Applied Statistical Methods

Januar Harianto

Feb 2026

Last week

Key concepts we covered

- **Population vs. sample**
  - Population: The complete set of all items we're studying
  - Sample: A subset of the population we actually measure
- **Parameters (population) and statistics (sample)**
  - Central tendency: how data clusters around a middle value
    - mean (average), median (middle value), mode (most common)
  - Spread/dispersion: how data points vary from each other
    - variance (average squared deviation), standard deviation (square root of variance)
- **Confidence intervals** – we'll explore this further today!

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.

Tip

Mensurative studies involve measuring without manipulating variables.
Absolute studies measure actual values rather than comparing between treatments.

Observational studies: 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*.
- *Think of it as a snapshot at one point in time.*

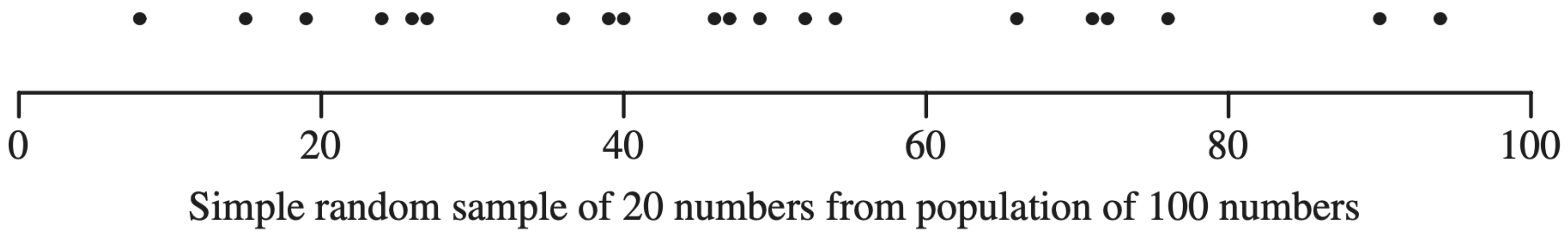
Monitoring studies

- 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*.
- *Think of it as taking multiple snapshots over time to see changes.*

Sampling designs

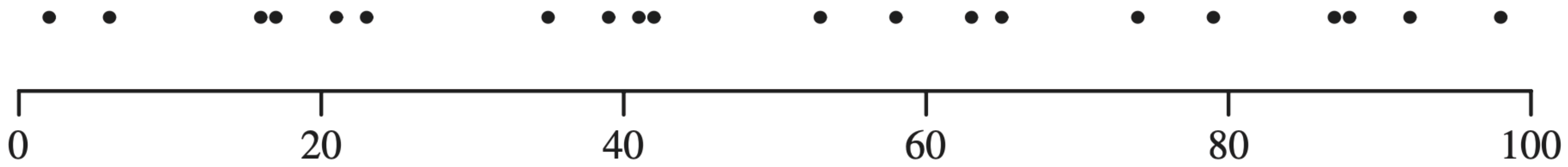
Simple random sampling:

- Each unit has an equal chance of being selected.
- **Randomly sample units from the entire population.**
- *Like putting all names in a hat and drawing some out randomly.*



Stratified random sampling

- The population is first divided into *strata* (groups with similar characteristics).
- **Randomly sample units within each strata by simple random sampling.**
- *Like separating students by year level, then randomly selecting some from each year.*



Stratified random sample of 20 numbers from population of 100 numbers

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. units from different groups may have different chances of being selected* - 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).
- *Think of it like rolling dice or drawing names from a hat, but using mathematics to ensure true randomness.*

Interpreting sampled data

We know that...

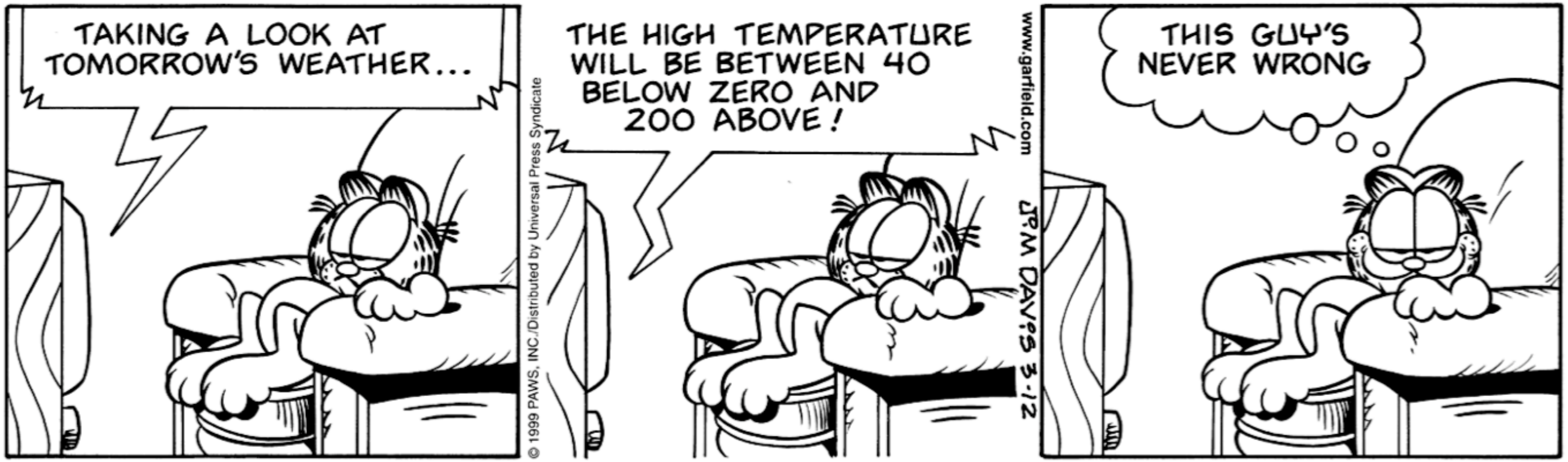
From the previous lecture

- **Sample mean** is a good measure of central tendency (the “middle” of our data).
- **Sample variance** is a good measure of dispersion (how spread out our data is).
- **Sample size** affects the precision of the sample mean (more samples = more precise).

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

Yes! That’s where confidence intervals come in.

Confidence intervals


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

Combining an estimate with its precision

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

- A range of values that likely contains the true population value
- Like saying “We’re 95% confident that out of 100 samples, 95 of them will be within this range”
- 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** ($\bar{x}$) - our best guess of the true value
2. **Critical value** (t_{n-1}) - a number that represents our chosen confidence level (like 95%)
3. **Standard error** of the estimate, **SE of the mean** ($SE_{\bar{x}}$) - tells us how precise our mean is

> **i** Note
>
> Think of standard error as "how much our sample means would vary if we took many different samples and calculated the mean each time"

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:

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

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

> **Tip**
>
> Think of margin of error like the "plus or minus" value you often see in survey results (±3%)

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, $\bar{x}$ (add all values and divide by number of samples)
2. Calculate the sample standard deviation, s (measure of spread in your data)
3. Determine the standard error of the mean, $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$ (how precise your mean estimate is)
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.
   - This is a specific number based on how many samples you have
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 values in a sample that are free to vary while still maintaining the same statistic.

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

where n is the number of samples.

Example

Imagine you have 4 numbers with a mean of 5:

- You can choose the first three numbers freely: 3, 10, and 7
- But the fourth number MUST be 0 to make the mean = 5
  - (3 + 10 + 7 + 0) ÷ 4 = 5
- So only 3 numbers (n-1) can be freely chosen = 3 degrees of freedom

More definitions

t-critical value

A number that helps determine how wide to make your confidence interval.

- Based on your **confidence level** (e.g. 95%) and **sample size**
- Larger t-values = wider intervals = more confidence but less precision
- Smaller sample sizes = larger t-values (because we're less certain)

More definitions

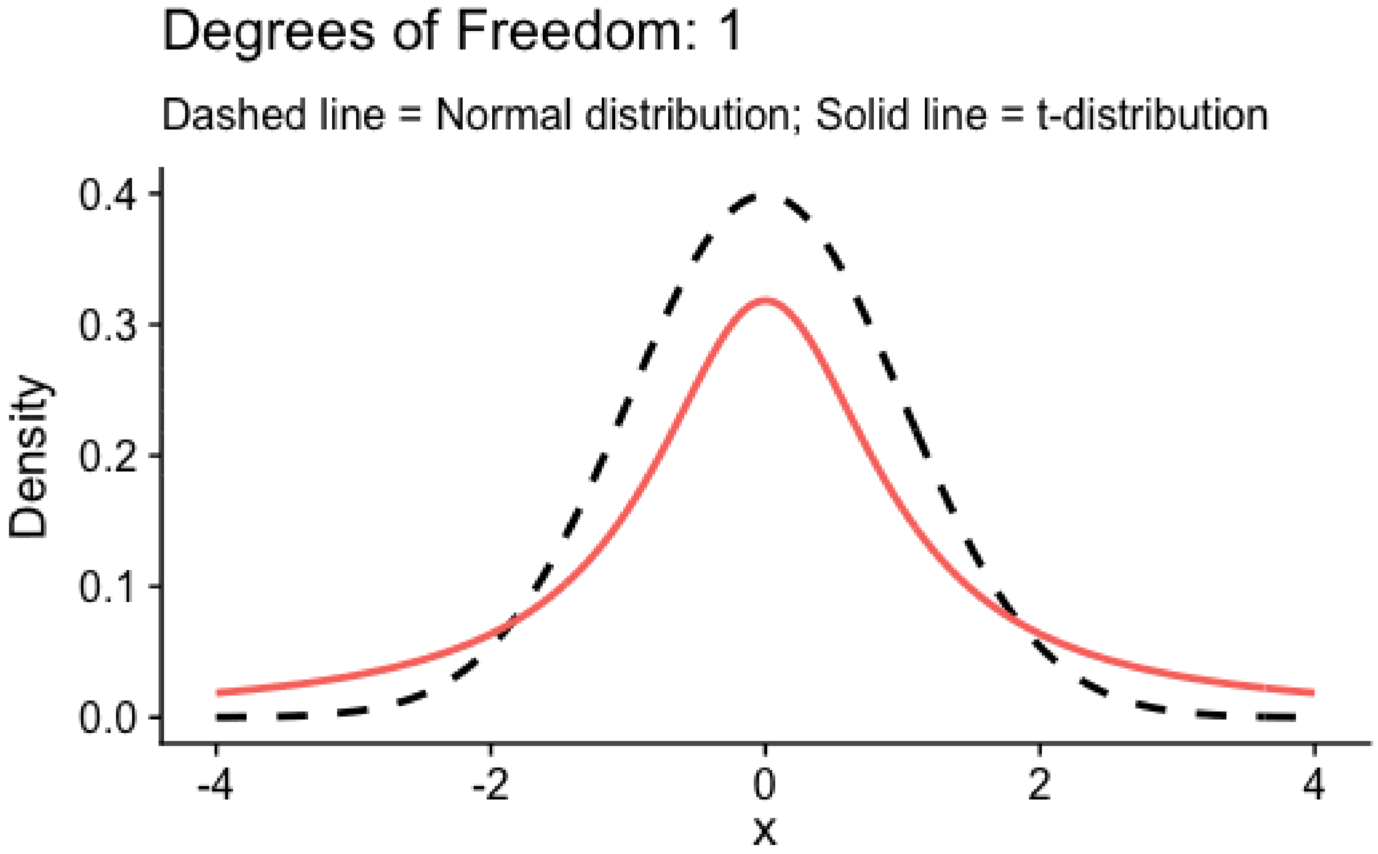
t-distribution

A probability distribution that accounts for the uncertainty when estimating from small samples.

- Similar to the normal "bell curve" distribution, but with **heavier tails**.
- With few samples (small n), the t-distribution is wider, reflecting greater uncertainty.
- As sample size increases, the t-distribution gets closer to the normal distribution.

```r
library(ggplot2)
library(gganimate)

# Variable to control animation speed (higher values = slower transitions)
anim_speed <- 1

# Create a sequence of x values for the plot
x <- seq(-4, 4, length.out = 400)

# Degrees of freedom: 1 through 5 then even numbers from 6 to 30
dfs <- c(1:5, seq(6, 30, by = 2))
```

```r
# Prepare data for t-distribution curves with a new "df" column
t_curves <- do.call(rbind, lapply(dfs, function(df) {
    data.frame(
        x = x,
        density = dt(x, df),
        df = df
    )
}))

# Create data for the standard normal distribution (static)
normal_curve <- data.frame(
    x = x,
    density = dnorm(x)
)

# Plot with gganimate: animate t-distribution curves (each frame shows one df)
p <- ggplot() +
    # Static normal distribution curve
    geom_line(
        data = normal_curve, aes(x = x, y = density),
        color = "black", linetype = "dashed", size = 1
    ) +
    # t-distribution curve that will animate
    geom_line(
        data = t_curves, aes(x = x, y = density, color = factor(df)),
        size = 1
    ) +
    labs(
        title = "Degrees of Freedom: {closest_state}",
        x = "x",
        y = "Density",
        subtitle = "Dashed line = Normal distribution; Solid line = t-distribution"
    ) +
    theme(legend.position = "none") +
    transition_states(states = df, transition_length = anim_speed, state_length = anim_speed)

p
```



Degrees of Freedom: 1

Dashed line = Normal distribution; Solid line = t-distribution ℹ Note

Notice how the t-distribution (solid line) gets closer to the normal distribution (dashed line) as the degrees of freedom increase!

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:

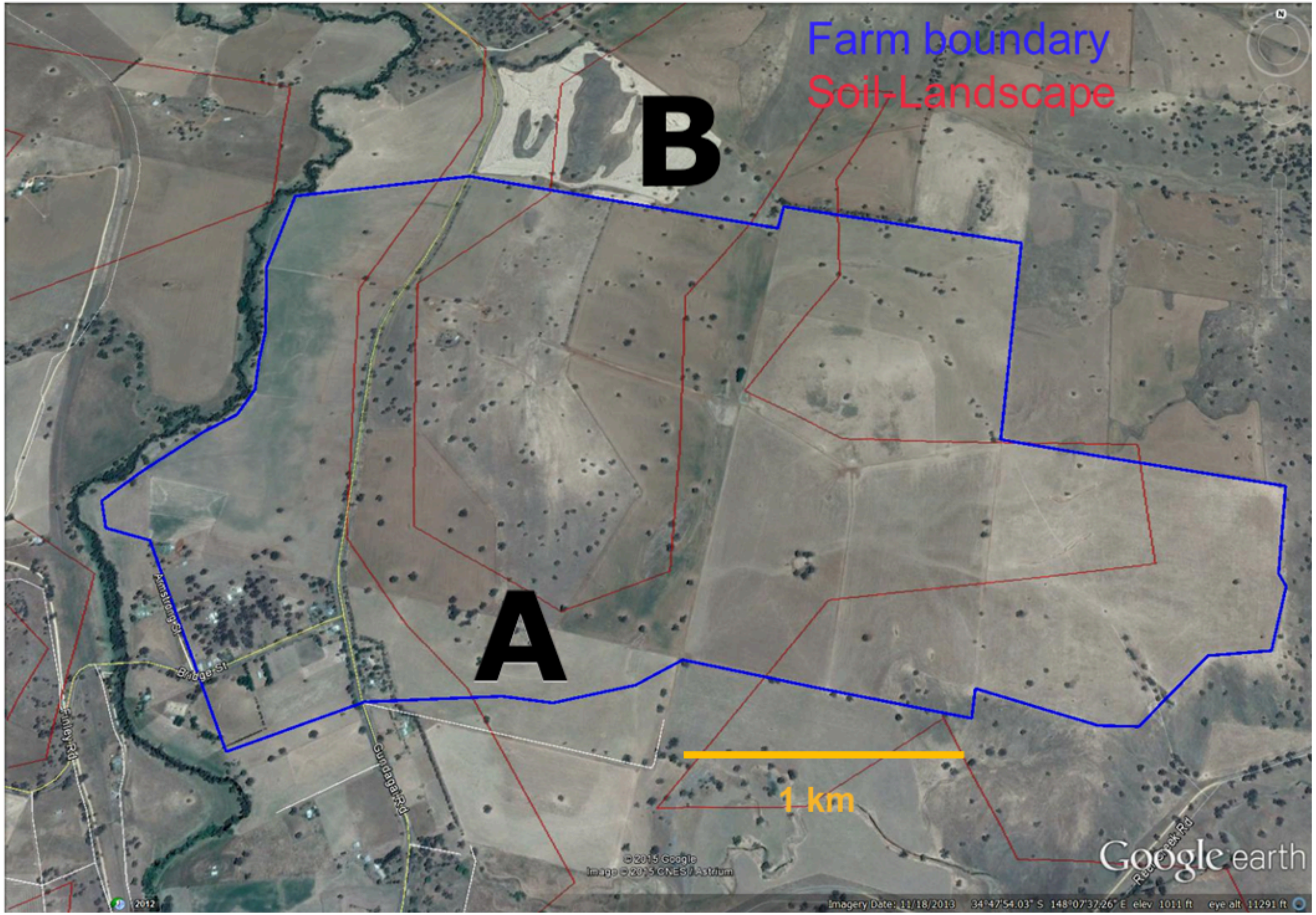
- A wider net (interval) is more likely to catch the fish (true value)
- A spear (single point estimate) is *less* likely to catch the fish
- The net width represents our uncertainty about the true value

! Important

Common misunderstanding: A 95% CI does NOT mean there's a 95% chance the true value is inside the interval. It means if you took 100 different samples and made 100 different CIs, about 95 of them would contain the true value.

Data story: soil carbon

Soil carbon


Farm boundary
Soil-Landscape
B
A
1 km
Armstrong St
Bridge St
Finley Rd
Gundagai Rd
Redcreek Rd
Google earth
© 2015 Google
Image © 2015 CNES / Astrium
2012
Imagery Date: 11/18/2013 34°47'54.03" S 148°07'37.26" E elev 1011 ft eye alt 11291 ft
N

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).

```r
soil ← c(48, 56, 90, 78, 86, 71, 42)
soil
```

```
[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: $\bar{x} = \frac{48+56+90+78+86+71+42}{7} \approx 67.3$ t/ha
2. Standard deviation: $s \approx 18.84$ t/ha
   - This tells us how much individual measurements vary from the mean
3. Standard error: $SE = \frac{s}{\sqrt{7}} \approx 7.12$ t/ha
   - This tells us how precise our estimate of the mean is
4. t-value (95% CI, df = 6): $t_{0.975,6} \approx 2.447$
   - This is the critical value for 95% confidence with 6 degrees of freedom
5. Margin of error: $t_{0.975,6} \times SE \approx 17.43$ t/ha
   - This is the “plus or minus” value for our interval
6. Which gives: $(67.3 - 17.43, 67.3 + 17.43) = (49.87, 84.73)$ t/ha

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.

Note: Our confidence interval is quite wide relative to our mean (about ±26% of the mean value), suggesting moderate uncertainty in our estimate.

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)
```

```r
# 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.

Understanding the R code

- `mean()` calculates the average value
- `sd()` calculates the standard deviation
- `qt(0.975, df=6)` gives the t-critical value (we use 0.975 because we want 95% in the middle, with 2.5% in each tail)
- `c(-1, 1)` creates a vector to subtract and add the margin of error

Questions

- How precise is our estimate?
  - *Our margin of error is about 17.43 t/ha, which is roughly 26% of our mean value.*
- How big a change must there be to estimate a statistically significant change?
  - *Changes larger than about 17.43 t/ha would likely be statistically significant.*
- **Can we sample more efficiently?**
  - *Yes! This is where stratified sampling becomes valuable.*

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?)

Break (10-15 minutes)

When we return, we'll explore how dividing our sampling area into meaningful groups can improve our precision.