Lecture 02a – Sampling designs I

ENVX2001 Applied Statistical Methods

Dr. Januar Harianto
The University of Sydney

Feb 2024

Outline

- Last week
- Today: observational studies
- Interpreting sampled data
- Confidence intervals
- Calculating confidence intervals
- Data story: soil carbon
- Simple random sampling: estimates
- Tomorrow: stratified random sampling

Last week

Observational study vs. controlled experiment

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 cannot always be done

Today: 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*.

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

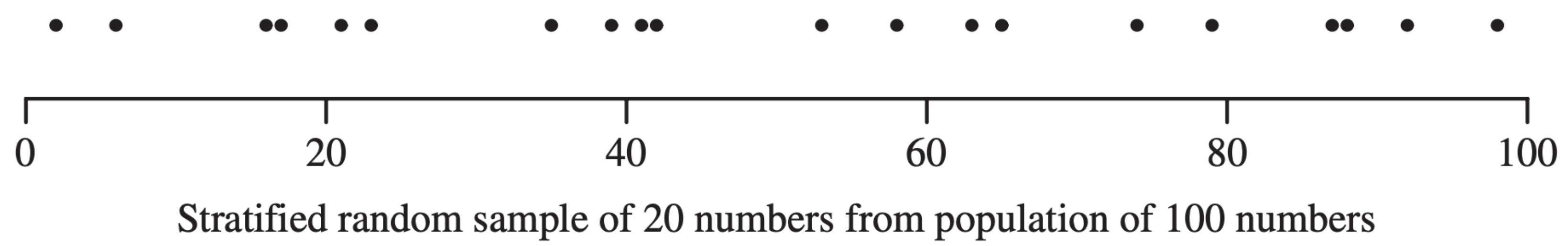
Simple random sampling:

- 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.**


0
20
40
60
80
100
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. 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

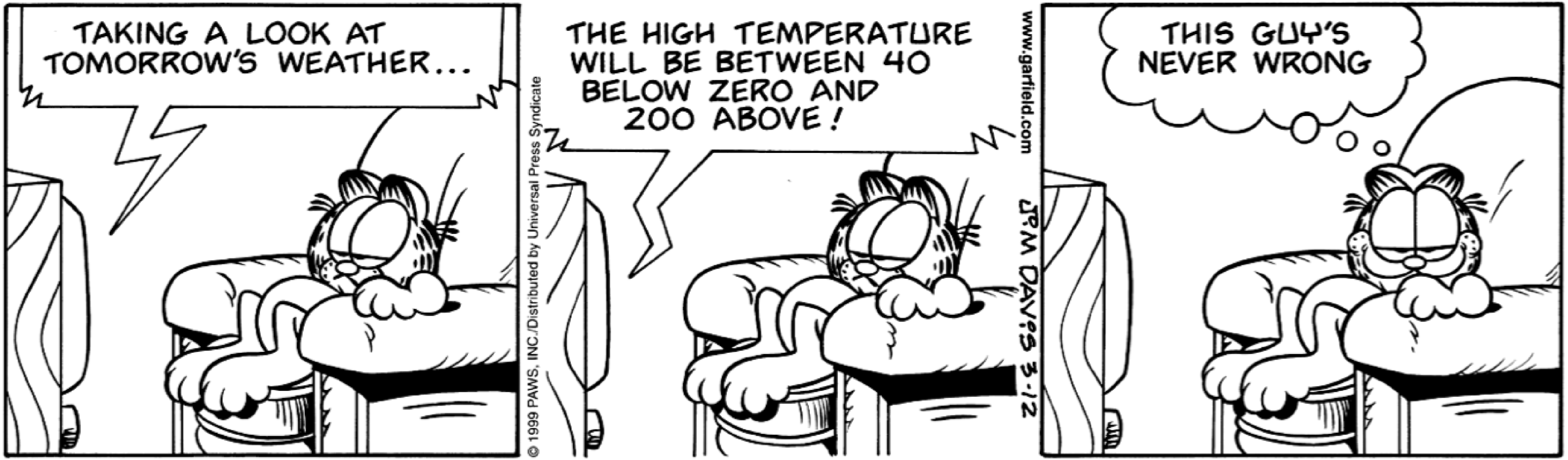
We know that…

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.
- Will be frequently mentioned throughout this unit!

General form

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**.

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).
- Think of it as a **range of values** that we are fairly sure contains the true value of the population parameter.

Fishing net analogy

Imagine that we are fishing in a river and we want to catch a fish that we saw.

- If we use a *spear* and throw it at a fish, we might miss it.
- If we use a **net**, we have a better chance of catching the fish.
- The *bigger* the net, the *more likely* we are to catch the fish.

Analogy: **The net is the confidence interval, and the fish is the true population parameter.**

Calculating confidence intervals

What we need

1. **Estimate** of the population parameter, e.g. the **sample mean**.
2. **Critical value** 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**.

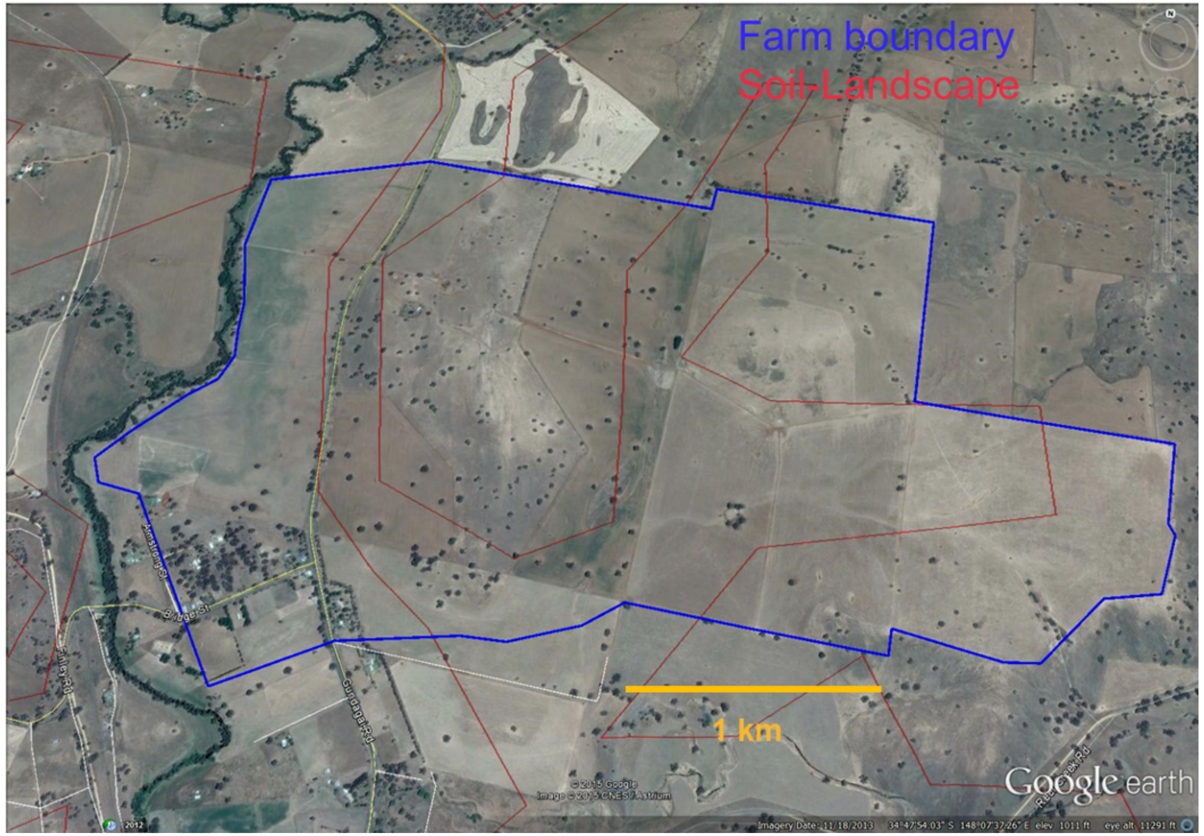
Why the t-distribution?

- The t-distribution results from **standardising the sample variance by the number of samples**.
  - Used when the true population variance is unknown.
- It resembles the normal distribution, but with heavier tails for small sample sizes.
- As sample size increases, the t-distribution converges to the normal distribution.

Data story: soil carbon

Soil carbon

Data story



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).

We start with the sampled data:

```
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?

Simple random sampling: estimates

95 % Confidence interval

The formula

$$95\%\ CI = \bar{y} \pm t_{n-1}^{0.025} \times SE(\bar{y})$$

Recall:

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

So:

$$95\%\ CI = \text{sample mean} \pm \text{t-critical value} \times \text{standard error of the mean}$$

We need to calculate each of these components:

① Sample mean $\bar{y}$; ② Critical value $t_{n-1}^{0.025}$; and ③ Standard error of the mean $SE(\bar{y})$

Sample mean

$$\bar{y} = \frac{1}{n} \times \sum_{i=1}^{n} y_i$$

> The **sum** of all sampled **values**, divided by the number of **samples**.

Relatively straightforward to calculate.

```
mean_soil <- mean(soil)
mean_soil
```

```
[1] 67.28571
```

t-critical value

What is the t-distribution?

The t-distribution is a **family of distributions** indexed by a **parameter** called **degrees of freedom**.

Understanding degrees of freedom for a mean estimate

- Degrees of freedom (df) represent the count of independent data points used to estimate a parameter.
- For the mean, df equals `n - 1`. For a sample size `n`, the last sample isn't independent – it **must** satisfy the mean.
- For instance, in a 3-value data set with a mean of 6, if two values are 7 and 3, the final value **must** be 8 and df = 2.

Calculating the t-critical value

We refer to the t-distribution table to find the critical value for a given confidence level and degrees of freedom. These days, we can use the `qt()` function in R. For a 95% confidence level, we use the 0.975 quantile since the t-distribution is symmetric.

```r
t_critical <- qt(0.975, df = length(soil) - 1)
t_critical
```

Standard error of the mean

The variance of the mean, $var(\bar{y})$, is:

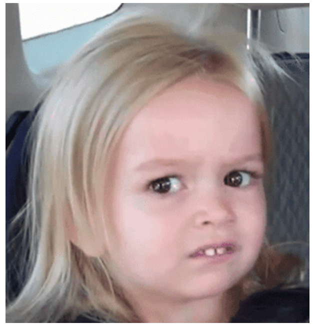
$$var(\bar{y}) = \frac{var(y)}{n}$$

Variance is standard deviation squared (s^2), so the formula is:

$$var(\bar{y}) = \frac{s^2(y)}{n}$$

Since $SE = \frac{s}{\sqrt{n}}$, then the standard error of the mean, $SE(\bar{y})$, is:

$$SE(\bar{y}) = \frac{s(y)}{\sqrt{n}} = \frac{\sqrt{s^2(y)}}{\sqrt{n}} = \sqrt{var(\bar{y})}$$



In words

> **step 1** calculate the variance $var(y)$ of the sampled values.
> **step 2** divide $var(y)$ by the number of samples (n) to obtain variance of the mean $var(\bar{y})$.
> **step 3** take the square root of $var(\bar{y})$ to obtain the standard error of the mean $\sqrt{var(\bar{y})} = SE(\bar{y})$.

In R, we can calculate the standard error of the mean using the `var()` or `sd()` function and the number of samples.

```r
se_mean <- sd(soil) / sqrt(length(soil))
# also ok:
# se_mean <- sqrt(var(soil) / length(soil))
```

Putting it all together

So far we have:

```r
mean_soil <- mean(soil)
t_critical <- qt(0.975, df = length(soil) - 1)
se_mean <- sd(soil) / sqrt(length(soil))
```

Now we can calculate the confidence interval:

```r
margin_error <- t_critical * se_mean
ci95 <- c(mean = mean_soil,
  L95 = mean_soil - margin_error,
  U95 = mean_soil + margin_error)

ci95
```

```
    mean      L95      U95 
67.28571 49.84627 84.72516 
```

Questions

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

Tomorrow: stratified random sampling

Thanks!

Questions?