Assignment 1: Resampling

Bayesian Inference and Computation — Bootstrap sampling

What you are learning (big picture)

- **Bootstrap sampling** = resampling data *with replacement* to approximate the sampling distribution of a statistic.
- In this assignment, the statistic is the **median of paired time differences** between two advert-display methods.
- You will practice:
  - turning a mathematical idea into working code,
  - writing a clean function and using replication,
  - summarising results visually and numerically.

Bootstrap sampling (definition)

Setup

You have N observations $x = \{x_1, \ldots, x_N\}$.

Bootstrap sample

A bootstrap sample x^B is formed by sampling N values from x **with replacement**.

Why?

Repeating this many times gives an empirical distribution of your statistic (median, mean, etc.), which you can summarise with a histogram, mean, and variance.

Experiment story (paired design)

- A social media company is testing a new advert-display method.
- **30 users** are chosen uniformly at random.
- Each user sees the same advert twice:
  - first via current method **X**,
  - second via new method **Y**.
- For each user, you observe watch time (seconds) under both methods.

Key point: paired structure

Each user provides a *pair* (X_i, Y_i). Your bootstrap must resample **users/pairs**, not individual X's and Y's separately.

Task 1 (core function)

Goal

Write an R function that:

1. generates **one** bootstrap sample that preserves pairing,
2. computes the **median** of the **time differences** (Y – X) in that bootstrap sample,
3. returns that median.

What you will likely compute

$$d_i = Y_i - X_i, \quad i = 1, \ldots, 30, \qquad \text{then return } \mathrm{median}(d_1^B, \ldots, d_{30}^B).$$

Task 2 (replication)

Goal

Set the seed to 7, then use `replicate` to generate **5000** bootstrap medians.

Reminder

Your submitted script should run on any computer using only:

- functions you wrote,
- **built-in** R functions (no extra packages).

Task 3 (plot + summaries)

Goal

Using the 5000 bootstrapped medians:

1. plot a **histogram**,
2. compute their **mean** and **variance**.

Built-in R functions you may use

- Histogram: `hist(meds, ...)`
- Mean: `mean(meds)`
- Variance: `var(meds)`

Plot quality

Label axes clearly (what is the statistic? what are the units?).

Why do we care about the bootstrap? (the real problem)

- You computed a statistic from your sample (here: the **median** of $Y - X$).
- But a statistic from one dataset is just **one noisy draw**.
- The question we actually care about:

If we repeated the experiment with new users, how much would that median change?

- That variability is the **sampling uncertainty** of the statistic.

Why not just use a formula?

- For some statistics (e.g. the mean), we have neat textbook formulas for uncertainty.
- For others, especially:
  - the **median**,
  - complicated estimators,
  - small/medium sample sizes,

  closed-form sampling distributions are hard or unavailable.

Key idea: use the data as a proxy for the population

The leap of faith (assumption)

Your observed sample is a reasonable stand-in for the population you sampled from.

- We do *not* know the true population distribution of (X, Y).
- We approximate it with the **empirical distribution** of the N observed users.
- Then we mimic "re-running the study" by re-sampling users **with replacement**.

Translation

Bootstrap is a way to simulate new datasets when the only thing you have is the dataset you already observed.

What bootstrap gives you in this assignment

Let

$$T = \mathrm{median}(Y_1 - X_1, \dots, Y_N - X_N)$$

be the statistic computed from the original sample.

- Each bootstrap resample gives a new statistic:

$$T^{(b)} = \mathrm{median}(Y_1^B - X_1^B, \dots, Y_N^B - X_N^B).$$

- The collection $\{T^{(1)}, \dots, T^{(B)}\}$ approximates the sampling distribution of T.
- From that distribution you can estimate:
  - **typical value** (mean of bootstrap medians),
  - **uncertainty** (variance of bootstrap medians),
  - **shape** (histogram: symmetry, skew, outliers).

When bootstrap works well (and when to be cautious)

Works well when

- the sample is representative of the population,
- observations (here: users) are approximately independent,
- N is not tiny.

Be cautious when

- data are highly dependent (e.g. time series, network effects),
- the sample is biased / unrepresentative,
- extreme outliers dominate and N is very small.

Submission requirements

- Submit **one R script** to Canvas.
- Use only:
  - functions written by you,
  - built-in R functions (no packages).
- Code should be **commented**.
- Maximum length: **100 lines**.

Assessment overview (10 marks total)

Task completion (7 marks)

- 7: completes the task in full with no errors
- 5–6: completes tasks in full but minor errors
- 3–4: some progress but incomplete / serious errors
- 0–2: little or no progress

Coding style & presentation (3 marks)

- 3: fully commented, good variable names, labelled plots
- 1–2: mostly commented / mostly consistent
- 0: little/no comments; incoherent style