ADS2 Practical 2: Probabilities and Simulation

Melanie Stefan, Duncan MacGregor, and Dmytro Shytikov

2022-08-26

Work through this guide alone or in groups. Facilitators are here to help. The time it takes to complete this practical can vary between individuals - this is OK. Do not worry if you do not finish within the session.

Learning Objectives

- Estimate a number or probability
- Draw random numbers in R
- Write loops in R
- Visualise data in R

RemembeR?

Create a script in RStudio, so that you can run several commands at once. For instance, create a script for this practical (or one for every exercise in this practical).

Open a file and save the data to a dataframe. We provide the file `Chicago2013.csv`, which contains results from 85 randomly drawn male finishers of the 2013 Chicago Marathon. Import that file and save the data to a dataframe called `chicago`.

- Examine the dataset. It contains names, country of origin, age, and finishing time (in hours).
- Can you tell what countries the athletes are from and how many there are from each country?
- Can you draw a histogram of finishing times? What does it look like?
- Select 10 people at random from the dataset and draw a histogram of their finishing time. You can use the `sample()` command. You can check the exact syntax of this command using the help in RStudio. Try repeated runs of this, does the histogram change?

Heights of students in a virtual classroom

Imagine an undergraduate biomedical sciences class with 100 students, of which 45 are men and 55 are women. The average height of the men is 172 cm (standard deviation: 7 cm), and the average height of the women is 158.5 cm (standard deviation: 6 cm).

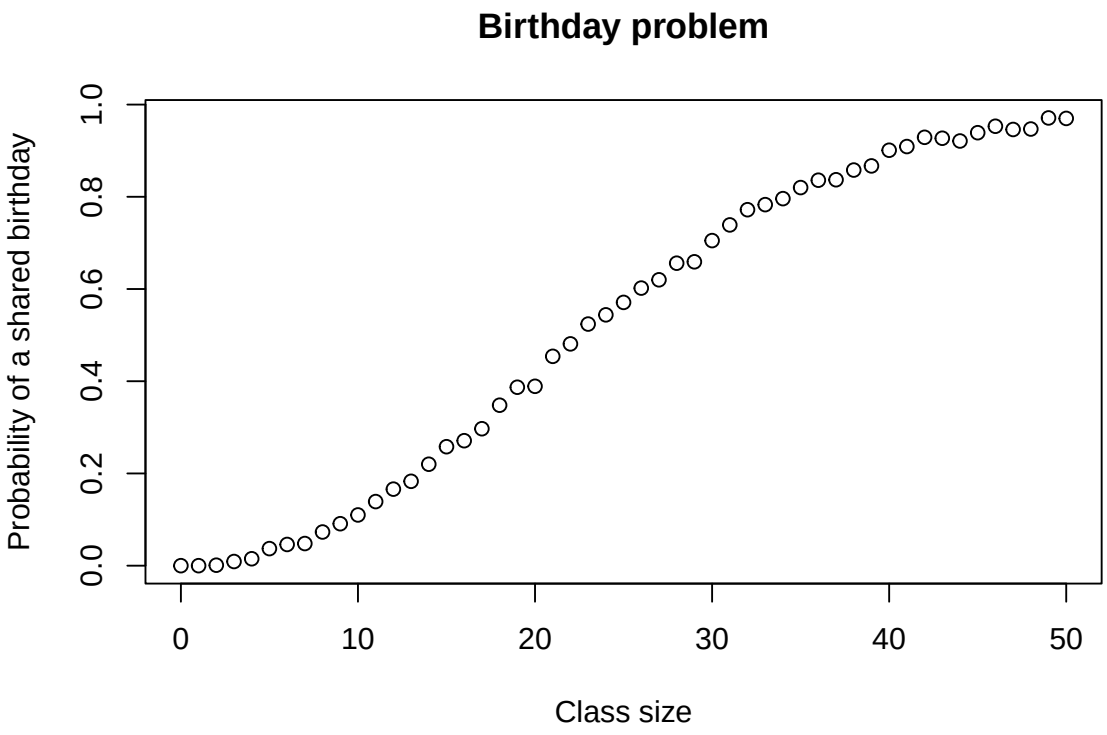
- Can you "create" a group of 100 people with this property in R?
  Hint: Maybe this will help: What does the following command do - `rnorm()`?

- How can you verify whether you did this (approximately) correctly?
- Draw box plots of
  – the men
  – the women
- How tall is the tallest student in your simulated class? How tall is the shortest student? How many students are taller than you are?
- Give a short summary about the mean, median, and quantile distribution of the values in both fictional samples.

Revisiting the birthday problem

In IBI1, we talked about the "Birthday Problem": In a group with n people, how likely is it that at least two of them share a birthday? In IBI1, you learned to tackle this problem mathematically, but you could also instead just run a computer simulation.

- Write a command that creates a group of 26 students, and assigns a day of the year to each of them as their birthday. (For the purpose of this exercise, you can ignore leap years).
- Are there shared birthdays in this group? How could you find out?
  Hint: You may find a combination of the functions `length()` and `unique()` useful.
- This gives you an answer for your particular group of students, but how would you go about computing the overall probability of a shared birthday for n=26?
- How about computing and plotting the probabilities of shared birthdays from n=1 to n=50? We are after a plot that looks similar to this:



Bonus (if you have time)

Let's try to work with probability distributions. Recently, your colleagues measured the activity of serum alanine aminotransferase (ALT) in young healthy mice. Here are the values: 33.45, 24.67, 24.16, 21.27, 26.86, 27.38, 27.91, 26.15, 31.63, 28.12 IU/L.
Considering that this sample represents the whole population, calculate its mean value and standard deviation (SD) and identify the probability of getting the following readings derived from the same population randomly:

1. 40.2 IU/L;
2. higher than 33 IU/L;
3. 22-25 IU/L;
4. 27-31 IU/L;
5. which values are within 40-65% highest values;
6. which value is less than 99.995% of all the other values derived from this population.
7. Back to the bonus question 4, identify the expected probability of getting values within the specified range and confirm these probabilities simulating random sampling.