Last update: October 4, 2019

201-DDD-05 - Statistical Methods
COMPREHENSIVE PROJECT - PART 3

Part 3 is due Wednesday December 11, before 9:00am

CONTENTS

1. PART 3: SIMULATIONS IN R AND FURTHER EXPLORATIONS

In this part of the project, we will be using R to run simulations of the model described in Part 2, that is, the model proposed by J. Chang in his paper "Recent common ancestors of all present-day individuals". **Before proceeding, make sure you fully understand that model**.

1.1 Function to simulate the TMRCA

Write an R function that receives as input the size `n` of the population, runs a simulation of that model having that population size, and returns the number of generations it took to reach a MRCA. (Make sure you test this function is free of bugs by testing it for "small" values of `n`.)

1.2 Simulating the TMRCA

Now we are going to "collect our data". That is, we will choose a "large" value of `n` and call the function defined above many times using that value of `n`. Make sure you collect the data (TMRCA) for each run of that function. Repeat this procedure for different values of `n`.

We would like `n` to be very large, as close as possible to the size of the actual human population today. As you will see, if `n` is that large, the R code will run for too long. So you will have to find a suitable value of `n` that is not too small but at the same time allows you to actually run that function in a reasonable time.

How many times should you call the function? Good question! If you read Chang's paper, Table 1 shows he called the function 25 times. You should also have at least 25 simulations (for each value of `n`).

How small could `n` be? Another good question. Again, if you read Chang's paper, you will see in the same table he chose the largest `n` to be 4000. Your largest `n` should be at least 4000.

1.3 Presenting the results of the simulation

Present the results of your simulations in a "nice" way. The bare minimum is a table like the one in Chang's paper. For full marks, though, your presentation of the simulations will have to be "nicer" both in terms of its form and of its content. This is the time to use your knowledge of descriptive statistics and data organization skills.

Include also how much time it took to run the simulations. (In R, you can get the current time by calling `Sys.time()`.)

1.4 (Bonus) Further exploration

If you want to play more with these simulations, this is your chance! Let your imagination ask interesting questions and then seek the answers! If you can come up with your own ideas, it is better but, if not, here are some ideas:

- You could find the time to "a generation in which each individual is either a CA of all present-day individuals or an ancestor of no present-day individual" (this is what Chang calls $\mathcal{U}_n$).
- Can you make the model slightly more realist and then write an R code to run simulations of those modified versions of the model?
- At the generation of the MRCA, are there usually more than one MRCA? If so, how many?
- You could do some meaningful statistics (confidence interval, hypothesis testing, ...) with the data collected.

If you have a really good question, even if you cannot find the answer, that can count too. We help Mathematics (and the sciences in general) move forward by asking the "right" questions. Welcome to the amazing world of research!

1.5 Conclusion

- Write an appropriate conclusion for your paper.
- What did you learn?
- What are the benefits of using probabilistic/statistical methods (and their implementation on computers) to address the question of the TMRCA?

1.5.1 Division of Labour

- Explain, in details, the contribution of each team member.
- Each student needs to be involved in **both** the R coding **and** the writing of the report.

1.6 Report Guidelines

Some miscellaneous comments about your final report:

- It needs to be handed in on paper (printed double-sided if you can). I will not accept electronic submissions.
- There is no requirements in terms of font size, interline spacing or margin sizes. Use your own judgment about what looks good.
- Come up with a good descriptive title for your paper.
- Organize your paper into sections, with section titles, so I can navigate around your paper with ease.
- Do not forget to include your R code.
- The number of pages is flexible.
- Part 3 is worth 5% of your final grade.
- The late submission penalty is 10% (0.5 points) for each day late.

2. GRADING SCHEME

Below is a tentative grading scheme for the project.

PART 3	
Presentation (*readability, structure, clarity, intro, effort, etc.*)	1
R code (*correctness, organization/indentation, easiness of understanding*)	3
Data presentation (*choice of items presented, relevance, discussion*)	1
TOTAL (Part 3)	**5**