Norwich Medical School Cancer Genetics Team Pre-Interview Tasks

Please return results to Daniel Brewer (d.brewer@uea.ac.uk) at least 48 hours prior to your interview.

Task 1 - The Sock Problem

Introduction

There is a large pile of socks that must be paired by colour. Given an array of integers representing the colour of each sock, determine how many pairs of socks with matching colours there are.

Example

```
pile = [1,2,1,2,1,3,2]
```

There is one pair of colour `1` and one of colour `2`. There are three odd socks left, one of each colour. The number of pairs is `2`.

Please provide all **code** and **answers** in a **Markdown document**. Any language or combination of languages can be used. Please do not spend more than two hours on this task (this will not be monitored or checked in anyway).

Part 1

Determine the number of pairs for the following pile

```
pile = [3,3,1,2,3,2,1,3,2,3,1,3,3,1,3,1,1,2,2,2]
```

Part 2 - Simulation

Assume that you have an infinite pool of `n` different colours of socks.

1. Produce a random pile of 20 socks if `n = 3`.
2. Repeat this 100 times

3. Determine the number of pairs of each pile

Repeat the above with different numbers of colours in your pool between `n = 3` and `n = 18` with step size of 3 i.e. `n = 3, 6, 9, 12, 15, 18`.

Part 3 - Visualisation

Create a plot to visualise the how the number of pairs varies as the pool of different colours increases using results from Part 2.

Part 4 - Interpretation

- Is there a relationship between the number of colours to select from and the average number of pairs.
- Perform a statistical test to examine this.
- Given these results train a statistical model and estimate how many pairs there will be when `n = 30` (any type of model or machine learning algorithm). Is this answer accurate?

Task 2 - Example of your pipelines/code/analysis

Please provide a real-life example of your code, pipeline (snakemake or nextflow) or an analysis script which is fully commented. Preferably within bioinformatics and relevant to proposed project. Any language or combination of languages is acceptable. Maximum 1000 lines.