MAS6002/6024/468, Assignment 2019/20

You should use R Markdown for this assignment. The submission deadline is noon on Tues 12th November 2019. Use R Markdown to produce a pdf file and submit the pdf file on the `TurnitinUK Assignments` tab on the MAS6002/6024/468 MOLE page.

Description of the problem

A leaflet printing company has invested in a new printing machine. The managing director wants to estimate the probability, per leaflet printed, that the printing machine produces a defective leaflet. Assume that the managing director can tell, without error, whether a leaflet is defective by looking at it. There are some complicating factors:

- the printing machine itself has two internal print devices and it is possible that the per-leaflet probability of printing a defective leaflet may differ between the internal print devices;
- the leaflets produced by the two internal print devices are placed on top of each other as they are printed so it is not possible to tell which internal print device has printed any given leaflet;
- the managing director only has large batches of printed leaflets available and as a result takes a random sample of the batch to test for defects.

Questions

Part 1

Write a function that will return the proportion of defective leaflets in the random sample. Assume that internal print device i ($i \in \{1, 2\}$) prints N_i of the leaflets in a batch. Assume further that the per-leaflet probability that internal print device i prints a defective leaflet is p_i where p_1 and p_2 are allowed to be different. Assume that p_i depends only on which print device has printed the leaflet, not on whether any other leaflets are defective (i.e. independent realisations). Assume that the managing director samples n leaflets at random from the batch of $N_1 + N_2$ leaflets.

Part 2

Assuming that $p_1 = 0.05$, $p_2 = 0.02$, $N_1 = 10000$ and $N_2 = 20000$, use your function from Part 1 to describe the probability distribution of the proportion of defective leaflets that the managing director will observe in a random sample of size 200. Note that just calculating the expected number of defective leaflets in the sample is **not** what is expected.

Part 3

The printer manufacturer claims that $p_1 = p_2 = p$. Assume that the manufacturer's claims are true. The manager director needs an accurate estimate of p but wants to use as small a sample as possible. They ask an intern to solve the problem. The intern is smart and knows that they should **not** use the Central Limit Theorem to answer this question since the samples from the $N_1 + N_2$ leaflets are not independent. As a result they choose to solve the problem using Monte Carlo simulation instead. Let $\hat{p}$ represent the estimate of p based on the random sample.

Write a function that will allow the intern to use trial and error to find the smallest value of n possible to ensure that $\hat{p}$ is within ϵ of p with probability exceeding 0.95 for given values of N_1, N_2, ϵ and p. I.e. find the smallest n such that $Pr(\mid p - \hat{p} \mid < \epsilon) > 0.95$ for given values of N_1, N_2, ϵ and p.

Part 4

Assume that the manufacturer is correct that $p_1 = p_2 = p$. Use your function in Part 3 to find the smallest sample size (n) needed to ensure that $\hat{p}$ is within 0.05 of p with probability exceeding 0.95 for $N_1 = 10000$ and $N_2 = 20000$ and for $p = 0.1$.

Part 5

In Part 4 you found the smallest value of n (where $1 \leq n \leq N_1 + N_2$) that estimated $\hat{p}$ to be within $\epsilon = 0.05$ of $p = 0.1$ with probability exceeding 0.95 ($Pr(\mid p - \hat{p} \mid < \epsilon) > 0.95$).

- Describe any values, or combinations of values, of p and ϵ that allow **all** values of n in the set $1 \leq n \leq N_1 + N_2$ to satisfy $Pr(\mid p - \hat{p} \mid < \epsilon) = 1$;
- For a given value of $0 < p < 1$ and $\epsilon > 0$ describe, as succinctly as possible, the values of n in the set $1 \leq n \leq N_1 + N_2$ that **cannot** satisfy $Pr(\mid p - \hat{p} \mid < \epsilon) > 0$.

You should answer the above **without** running any R code.

What else to include in your answers

As well as answering the questions directly, where appropriate, you need to justify your choice of the simulation parameters under your control. You should also try to quantify the uncertainty whenever there is any.

Administrative Information

Marking is anonymous, so **do not** write your name anywhere in the report. Your registration number should be shown on the first page of the report. You should submit a single pdf file for this assignment. The report should be named `studentnumber-modulenumber-A2.html`.

Report Structure and Content

There is no page limit for the report but as a guide 5 pages or fewer should be sufficient. Think about how to communicate your answers concisely and clearly whilst providing sufficient detail. Unnecessarily long reports will be penalized. Just answer the questions directly and justify your answers. Use R Markdown to produce your pdf file. Make sure all the code used is visible in the report (use `echo = TRUE`). There is an optional R Markdown assignment template on MOLE that you can use use if you prefer. Describe what the R code does where you think another MSc student might not easily understand it (for example if you are using functions **not** in the R notes for this module).

Assessment and Feedback

The Assignments should be **your own work** and should be undertaken in accordance with the University's rules for non-invigilated assessment. Marks will be given on the scale described in Chapter 1 of the notes. You will receive individual feedback and some general comments will be posted on MOLE.