STAT 513/413, Midterm Examination, March 5, 2021

Show all your reasoning: if it is mathematics, the steps taken, if it is computing in R, then all relevant code. Your computer work must be reproducible – that is, anybody else must be able to take your code and produce your results. To this end, you must initialize random generators in an appropriate and visible manner (use you personal number for that) and give the transcript of all code that you used – not only the code for an implemented function, say, but also how it ran and what results it produced. No handwritten transcriptions of computer work will be accepted.

You may include scans of handwritten mathematics, if necessary, but there may be not that much need for that in these problems. For the justification of certains steps, it is sufficient to include the graph of a function in question, together with some explanation; no formal proofs are required if the graph shows the claimed property in a reasonably transparent manner.

Therefore, it is advised that before you do anything with the problem, you graph the function in question (in R, or merely by hand, if you are good at it). The text of this examination contains three problems, and has two pages. It is advised that you read carefully the entire problems before attempting any partial solutions. Good luck!

1. As a part of a review process, you would like to reproduce certain computation done ("by them") elsewhere, within your R environment. The computation involves random numbers; you have their code at your disposal, and you also know what value they set the random generator seed to before running the computation (like you know they started with `set.seed(007)` or something similar). The computation was then done in *one interrupted session* of R on their computer.

So, checking on your computer to see whether you will obtain the same results, should not pose any problem. Your computer can also run R (the same version as they did), and you know the value of seed they started with. The only problem is that their computation was somewhat lengthy: your computer is not that fast, so the uninterrupted run may well not end in one day, and running computer overnight is not possible in you situation (regulations, power outage at midnight, whatever).

What you can do? Fortuitously for you, the computation can be divided into several sessions, each producing a part of the whole result; and each of these sessions can be safely completed in a day. (For instance, one of those sessions computes the estimate, another one its precision, and so on.)

The question thus is: describe – also in terms of the appropriate R commands – how would you reproduce their uninterrupted session in your two or more sessions, when between these sessions you have to quit R and switch off the computer.

Be reminded that you do not and will not have any additional information (there is no possibility to ask them anything, for instance); you are entirely on your own regarding this.

2. The density of a probability distribution is

$$f(x) = 20x(1-x)^3 \qquad \text{when } 0 \le x \le 1 \qquad \qquad \text{(and 0 otherwise)}$$

(a) Construct and implement the function generating random numbers with this distribution.

Note: as here is no method explicitly prescribed for this, you have to think carefully which method would allow you to carry the task to the succesful end – before committing your efforts to a particular one. No piecemeal, partial attempts (like: "...and I would continue like this if I knew how") will earn any credit here; only methods carried to the stage that they can handle part (b), which says:

(b) By repeatedly evaluating your function N=10000 times, estimate the mean and variance of the distribution given above.

3. This problem is about integration via simple Monte Carlo method, in the typical setting using uniform distribution on the integration domain. Two integrals are sought, of the same function, but on different domains,

$$(A) \quad \int_1^4 (\log(x))^2 \, dx \qquad \text{and} \qquad (B) \quad \int_0^1 (\log(x))^2 \, dx.$$

Be reminded that "log" means in R and also here, as well as in every area significantly touched by mathematical education, the logarithm with the basis $e = 2.718282$ (which is in engineering sometimes written as "ln").

(a) Compute the integral (A) in R using, as outlined above, simple Monte Carlo. Use $N = 50000$ random numbers, and repeat the computation 3-4 times, to observe the variability of the result.

(b) What is the best perfomance guarantee you can derive for this case, that is computing integral (A) with $N = 50000$ as you did in (a), aiming at the precision of one (1) decimal digit? (One digit after the decimal point, to be clear.) The performance guarantee is in terms of the probability that the result does not meet this precision: can you show that this probability is less than 0.05? Or can you show that this probability is less than 0.01?

If the answer to one of the latter questions is "no", you do not have to do anything, even not writing it down. If it is "yes", it will be credited only if you give a correct and reasonably complete justification, as well as show necessary R calculations regarding the numerical aspects of your answer. Of course, if you can provide a well-justified "yes" on the second question, you do not have to answer or justify the first one.

(c) And now the integral (B): we would like to know which N would guarantee that the simple Monte Carlo method ran to compute integral (B) with N repetitions would achieve the same precision as in (b) (that is, one decimal digit) with probability 0.99. Can you give some computation and/or estimation showing what would have this value of N be? Once you obtain the desired N, try it on 3-4 runs of simple Monte Carlo computing the integral (B).

Remember, the due time for this is Sunday, March 7, noon, Edmonton time