Questions

Turro Group
Icahn School of Medicine at Mount Sinai

Please return your answers to the questions and a file containing your computer code. Please make your computer code as concise as possible and please comment your code.

Question 1

In a programming language of your choice, write either an interpretable script or a compilable program, which takes a positive integer n as input and computes the nth term of the infinite sequence

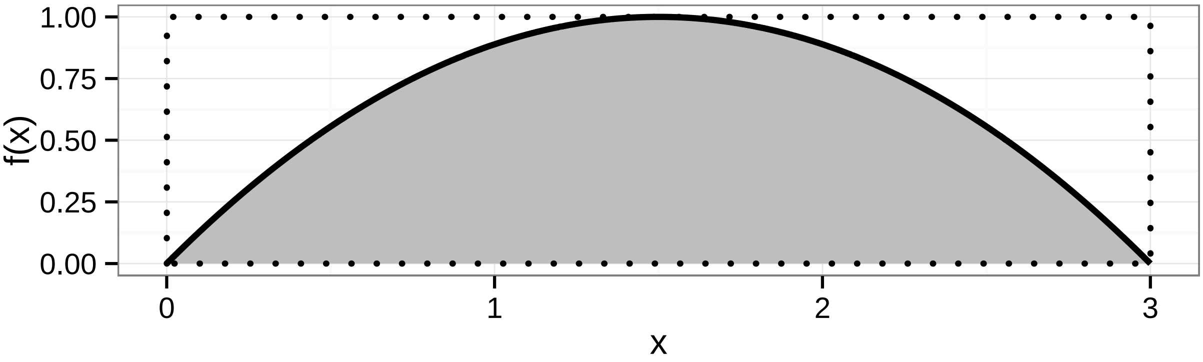
$$1, 2, 6, 24, 120, 720, 5040, 40320, 362880, 3628800, \ldots$$

efficiently. You may structure the code as you wish, but use only elementary operations ($+$,$-$, $\times$, / etc.) to perform numerical calculations.

Question 2

f is a parabolic function $f : [0, 3] \to \mathbf{R}$. The figure below shows the graph $\{(x, f(x)) : x \in [0, 3]\}$. Note that the region bounded by the dotted curve encloses the shaded region completely and that the area of the shaded region represents the definite integral

$$I = \int_0^3 f(x)dx. \tag{1}$$



(a) Give an expression for $f(x)$.

(b) Write a script in the R programming language to approximate the integral I stochastically (i.e. to generate a random estimator $\hat{I}^{(n)}$ of I), by simulating n points uniformly in the dotted region. Your script should be as short as possible. It should include the definition of a function for computing f. Your script may call the built-in R function `runif` to generate uniformly distributed random numbers. It should not rely on any additional random number generation.

(c) Give an expression for $\mathrm{se}(\hat{I}^{(n)})$, the standard error of your estimator, in terms of I and n.

(d) By substituting $\hat{I}^{(n)}$ for I in the expression for $\mathrm{se}(\hat{I}^{(n)})$ obtained in (c) one obtains an estimator $\widehat{\mathrm{se}}(\hat{I}^{(n)})$ for $\mathrm{se}(\hat{I}^{(n)})$. For each

$$n \in \{1, 9, 36, 81, 144, 225, 324, 441, 576, 729, 900, 1089\}$$

simulate $1,000$ replicate estimates $\hat{I}_1^{(n)}, \hat{I}_2^{(n)}, ..., \hat{I}_{1,000}^{(n)}$ of I using the script you wrote in (b). Plot the points $(\log \mathrm{se}(\hat{I}^{(n)}), \log \widehat{\mathrm{se}}(\hat{I}_i^{(n)}))$ across all i and n.

(e) Write a program to estimate π stochastically. Give a point estimate. Choose n so that the corresponding 95% confidence interval has width less than 0.01.

Question 3

The *lexicographic ordering* defines an ordering on sets of strings in terms of an underlying ordering on letters. When the underlying ordering is the alphabetical order, the lexicographic order is the familiar dictionary order.

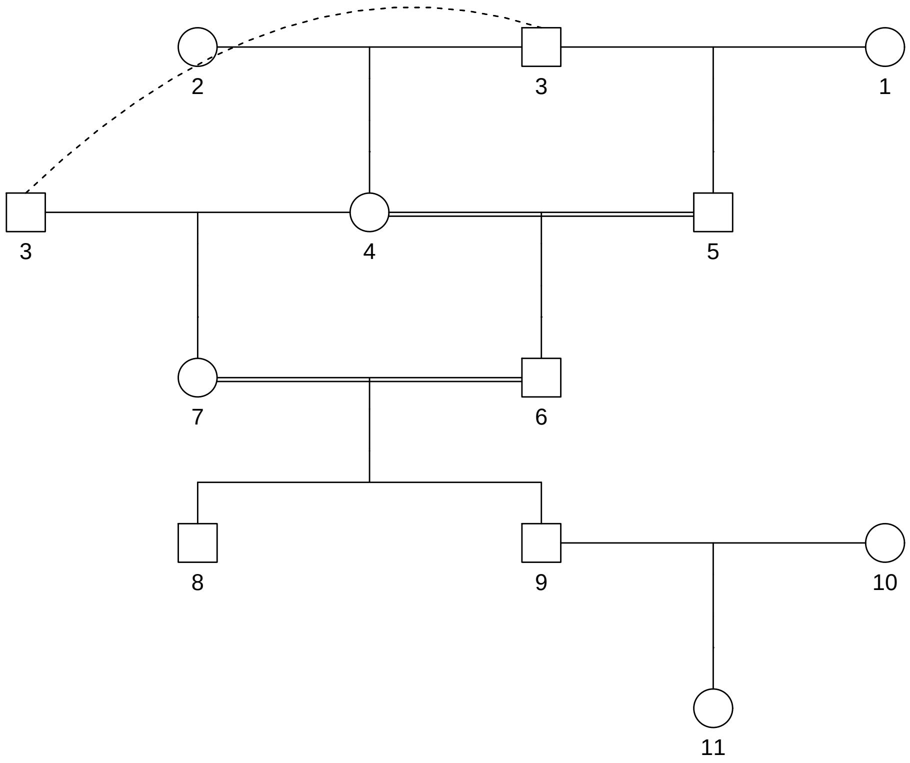
Research the lexicographic ordering as well as the sorting algorithm Bubble Sort. (Wikipedia should be a sufficient reference for all three). Using R implement Bubble Sort to sort sets of strings of length 10 on the alphabet of 26 lower case letters. Your algorithm should accept an arbitrary unsorted list of strings as input. It should output the list of strings sorted in the lexicographic ordering from smallest to greatest. However, rather than the usual alphabetical ordering, the underlying ordering on letters should be:

$$h < w < n < x < o < z < e < b < c < y < v < k < j < ...$$
$$...g < s < a < i < t < r < u < m < p < d < f < q < l.$$

Your implementation should not rely on any pre-written functions or libraries for sorting.

Question 4

The figure below is a pedigree. It illustrates the familial relationship between 11 individuals. Males are shown as squares and females as circles. Individuals 1, 2, 3 and 10 are the *founders*, their parents do not appear in the pedigree. The parents of the remaining individuals can be identified by following the lines up the diagram. For example, individuals 9 and 10 are respectively the father and mother of individual 11. For clarity individual 3 is represented in two places, the dashed line indicates the two squares represent the same individual.



Research Mendel's laws and biallelic genotypes.

(a) Write an R script to simulate biallelic genotypes at a single locus for each individual in the pedigree when arbitrary founder genotypes are given.

(b) Simulate realised genotypes $10,000$ times under each of the following scenarios for founder genotypes:

Founder	Scenario 1	Scenario 2	Scenario 3
1	aa	Aa	AA
2	AA	Aa	Aa
3	AA	Aa	AA
10	Aa	Aa	AA

(c) Generate bar plots showing the probability distribution over genotype for individual 8 and individual 11 under each scenario.

(d) Give a stochastic estimate of the probability that the genotypes of individuals 8 and 11 are identical under each scenario.

(e) Compute analytically the probability distribution over genotype for individual 8 under scenario 1.