Probability and Statistics
Lab assignment 3: Markov chains and Parameter estimation

General comments:

- This is a team assignment. Complete solution will give you **3** points (out of 100 total). Submission deadline is **23:59 of 29 November 2022**.
- The assignment must be completed in **R** language for statistical computing (`https://www.r-project.org/`). It can be installed from the official site. RStudio (`https://www.rstudio.com/`) is a convenient GUI.
- You will need just a few basic **R** commands to complete the task. As a quick reference guide, use the official manual `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf` or help section of RStudio.
- The report must be prepared as an R notebook; you must submit to `cms` (github submission will not be accepted) both the code and html-generated notebook.
- For each task, include
  - problem formulation and discussion (what is a reasonable question to discuss);
  - the corresponding **R** code with comments (usually it is just a couple of lines long);
  - the statistics obtained (like sample mean or anything else you use to complete the task) as well as histograms etc to illustrate your findings;
  - justification of your solution (e.g. refer to the corresponding theorems from probability theory);
  - conclusion (e.g. how reliable your answer is, if it agrees with common sense expectations).
- The `team id number` referred to in tasks is the **two-digit** ordinal number of your team in the R random selection. Observe that the answers **do** depend on this `team id number`! You **must** include the line `set.seed(**)` at the beginning of your code (with `**` being the `team id number`) to make your calculations reproducible.

Part I: Markov Chain

Determine the `TLN` (that stands for the `team lucky number`) as a three-digit number which is the `team id number` with an extra zero added from the left; e.g., `TLN` is 028 for `team id number` 28. In this part, you will study the questions about chances to see the `TLN` in random sequences of digits.

Problem 1 (1 pt.). In the first part, we will estimate the probability that a random digit sequence of length n contains the `TLN` (consider the cases $n = 100$, $n = 200$, $n = 1000$).

1. Estimate numerically the probability $\hat{p}_n$ of the event that your `TLN` occurs in a random digit sequence $d_1 d_2 d_3 \dots d_n$.

   Hint: Such a sequence can be generated with R command `sample(0:9, n, replace=T)`; you will need to generate a sample of such sequences of sufficiently large size N

2. Identify the the Markov chain structure with four states S_0, S_1, S_2, S_3 in this sequence with S_k denoting the number of correct last digits (eg., for the `team id number` 028 these states will be $S_0 =$ "*", $S_1 =$ "0", $S_2 =$ "02", $S_3 =$ "028"). Determine the transition probabilities matrix P and find the limiting probability p_n for the state "028". Compare with the result obtained in part 1.

   Hint: you can find the limiting probabilities by either solving the corresponding system, or by calculating P^k in **R** for large enough k, or by finding the left eigenvector of P using **R**

3. Determine approximately the sample size N which guarantees that the absolute error $|\hat{p}_n - p_n|$ of the estimate $\hat{p}_n$ is below 0.03 with confidence level of at least 95 percent. Rerun the experiments for $n = 1000$ with the determined size N to illustrate the confidence interval and confidence level.

   Hint: estimate the standard deviation of the corresponding random variable by the standard error of the sample

Problem 2 (1 pt.). In the setting of Problem 1, assume that the random digit generation stops at the first occurrence of the `TLN` (i.e., that the state S_4 of the Markov chain is now absorbing). In this problem, you will estimate the average length of such sequences (i.e., the average time till absorption in the Markov chain).

1. Make necessary amendments to the transition probabilities matrix P above and solve the corresponding system to find the expected time $\mathsf{E}(T)$ till absorption
2. Estimate numerically the expected length $\mathsf{E}(T)$ till the first occurrence of the `TLN` by running a sufficiently large number N of experiments.

   Hint: Clearly, the unbiased estimator for $\theta := \mathsf{E}(T)$ is the sample mean $\hat{\theta} = \overline{T} = \frac{1}{N}(T_1 + \dots T_N)$
3. Find the sample size N which guarantees that the absolute error $|\hat{\theta} - \theta|$ of the estimate does not exceed 10 with confidence level of at least 95 percent.

   Hint: use Chebyshev inequality and estimate the standard deviation of T by the standard error of the sample $T_1, T_2, \dots, T_N$

Part II: Parameter estimation

Aim: In problems 3 and 4, you will have to verify that the interval estimates produced by the known rules indeed contain the parameter with probability equal to the confidence level.

Problem 3 (1 pt.). The expected value of the exponential distribution $\mathscr{E}(\lambda)$ is $1/\lambda$, so that a good point estimate of the parameter $\theta := 1/\lambda$ is the sample mean $\overline{\mathbf{x}}$. Confidence interval for θ can be formed in several different ways:

(1) Using the exact distribution of the statistics $2\lambda n\overline{\mathbf{X}}$ (show it is χ^2_{2n} and then use quantiles of the latter to get the interval endpoints)

(2) Using the normal approximation $\mathscr{N}(\mu, \sigma^2)$ for $\overline{\mathbf{X}}$; the parameters are $\mu = \theta$ and $\sigma^2 = s^2/n$, where $s^2 = \theta^2$ is the population variance (i.e., variance of the original distribution $\mathscr{E}(\lambda)$). In other words, we form the Z-statistics $Z := \sqrt{n}(\overline{\mathbf{X}} - \theta)/\theta$ and use the fact that it is approximately standard normal $\mathscr{N}(0,1)$ to find that

$$\mathsf{P}(|\theta - \overline{\mathbf{X}}| \le z_\beta \theta/\sqrt{n}) = \mathsf{P}(|Z| \le z_\beta) = 2\beta - 1.$$

in other words, θ is with probability $2\beta - 1$ within $\overline{\mathbf{X}} \pm z_\beta\theta/\sqrt{n}$.

(3) The confidence interval constructed above uses the unknown variance $s^2 = \theta^2$ and is of little use in practice. Instead, we can solve the double inequality

$$|\theta - \overline{\mathbf{X}}| \le z_\beta \theta/\sqrt{n}$$

for θ and get another confidence interval of confidence level $2\beta - 1$ that is independent of the unknown parameter.

(4) Another (and a more universal approach) to get rid of the dependence on θ in (2) is to estimate s via the sample standard error and use approximation of $\overline{\mathbf{X}}$ via Student t-distribution; see details in Ross textbook on statistics or in the lecture notes

Task:

(a) verify that the confidence intervals of level $1 - \alpha$ constructed via 1.–4. above contain the parameter $\theta = 1/\lambda$ approx. $100(1 - \alpha)\%$ of times

(b) compare their precision (lengths)

(c) give your recommendation as to which of the three methods is the best one and explain your decision

Directions:

- use $\theta = \texttt{id\_num}/10$ and $\alpha = 0.1; 0.05; 0.01$;
- vary the sample sizes n and the number m of repetitions to estimate the probability and comment on the results.

Problem 4 (1 pt). Repeat parts (2)–(4) of Problem 3 (with corresponding amendments) for a Poisson distribution $\mathscr{P}(\theta)$.
Task and **Directions** remain he same; in other words, you have to check that confidence intervals constructed there contain the parameter θ with prescribed probability.

Example 1. Assume we need to test how good a Student-type confidence intervals are for samples from two combined normal distributions $\mathscr{N}(\mu, \sigma^2)$ with alternating $\mu = \mu_0 - 1$ and $\mu_0 + 1$ and $\sigma = 1$ and 4 and are too lazy to calculate the resulting variance

Lab assignment 2: confidence intervals

We illustrate the notion of the confidence level on the following example:

```
set.seed(000)
M <- 1000
N <- 100
mu = 5
## sample N rv; then replicate M times and write the results as an N*M matrix
x <- matrix(rnorm(N*M,mean = c(mu-1,mu+1), sd = c(1,4)), nrow = N)
## calculate sample mean in each column
sample_mean <- colMeans(x)
## calculate sample sd of each column; 2 in `apply' indicates the coordinate to keep as output
sample_sd <- apply(x, 2, sd)
## check how good the CI are:
for (alpha in c(.01, .05, .1)){
  cat("For confidence level", 1-alpha, "\n")
  cat("    the fraction of CI's containing the parameter is",
      mean(abs(sample_mean-mu) < qt(1-alpha/2, N-1)*sample_sd/sqrt(N)), "\n", sep = " ")
## The maximal and mean CI length:
  cat("    maximal CI length is", 2*qt(1-alpha/2, N-1)*max(sample_sd)/sqrt(N), "\n", sep = " ")
  cat("    mean CI length is", 2*qt(1-alpha/2, N-1)*mean(sample_sd)/sqrt(N), "\n", sep = " ")
}
```

```
## For confidence level 0.99
##     the fraction of CI's containing the parameter is 0.992
##     maximal CI length is 2.141762
##     mean CI length is 1.618653
## For confidence level 0.95
##     the fraction of CI's containing the parameter is 0.973
##     maximal CI length is 1.618075
##     mean CI length is 1.222873
## For confidence level 0.9
##     the fraction of CI's containing the parameter is 0.935
##     maximal CI length is 1.354004
##     mean CI length is 1.023299
```