Probability and Statistics
Lab assignment 3:
Parameter estimation and Unbiasedness of Estimators

General comments:

- This is a team assignment. Complete solution will give you **4** points (out of 100 total). Submission deadline is **22:00 of 22 November 2025**.
- You have to submit both the **Rmd** source file and the **html** output (links to GitHub repositories are NOT accepted!)
- **At the beginning of the notebook, provide a work breakdown structure estimating the efforts of each team member.**
- 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.
- Take into account that not complying with these instructions may result in point deduction regardless of whether or not your implementation is correct.

Part I: Parameter estimation

Aim: In problems 1 and 2, 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 1 (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 four 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 2 (1 pt)**.** Repeat parts (2)–(4) of Problem 1 (with corresponding amendments) for a Poisson distribution $\mathscr{P}(\theta)$.
Task and **Directions** remain the 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

```
set.seed(000)
M <- 1000
N <- 100
mu = 5
## sample N rvs; then replicate M times and write the results as 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 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", sep=" ")
  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
```

Part II: Unbiasedness of Estimators

Problem 3 (1 pt)**.** In this task, you need to analyze the estimators of sample variance and their properties: we'll prove the unbiasedness of variance estimator theoretically and check it on practice!

Example 2. Here is code to create a dataset (with known population variance; for simplicity we simulate the normal distribution)

```
set.seed(42)
n <- 100
mu <- 10
sigma_squared <- 4
sigma <- sqrt(sigma_squared)
dataset <- rnorm(n, mean = mu, sd = sigma)
head(dataset)

## [1] 12.741917  8.870604 10.726257 11.265725 10.808537  9.787751

cat("Population Mean (mu):", mu, "\n")

## Population Mean (mu): 10

cat("Population Variance (sigma_squared):", sigma_squared, "\n")

## Population Variance (sigma_squared): 4

sample_mean <- mean(dataset)
sample_variance <- var(dataset)
cat("Sample Mean:", sample_mean, "\n")

## Sample Mean: 10.06503

cat("Sample Variance:", sample_variance, "\n")

## Sample Variance: 4.337697
```

There are two common estimators of sample variance:

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\sigma_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Task:

(a) write the code to find the variance for the dataset;

(b) find σ_n^2 and σ_{n-1}^2 for n = 10, n = 50, n = 100, n = 1000;

(c) find the biases $Bias(\sigma_n^2) = E(\sigma_n^2) - \sigma^2$ and $Bias(\sigma_{n-1}^2) = E(\sigma_{n-1}^2) - \sigma^2$;

(d) comment on the results for the different values of n;

(e) derive analytically the expected value of each estimator $E(\sigma_n^2)$ and $E(\sigma_{n-1}^2)$;

(f) using the expected values found above, show mathematically what of the above two estimators are unbiased;

(g) comment on the results behind theoretical and practical tasks.