Project 2

ISyE 3770: Statistics and Application

Spring 2025

Due Date : April 29, 2025

Instructions

The project will require you to use a scientific software suite of your choice (e.g., MATLAB, Python) to perform the necessary statistical analyses. The work can be collaborative, but the writing of the report, the generation of figures, and any software code must all be completed on an individual group basis. The report must be typewritten using the software suite of your choice. The evaluation will consider the clarity of explanations, the quality of the figures, the quality of the writing, and the technical analysis performed. Some concepts may extend beyond what we have covered in class. Therefore, you may need to perform additional bibliographic research (e.g., on statistical testing) using the textbook or online resources. This additional effort aligns with the spirit of a group project. You must upload the project report on Canvas on April 29th, 2025 by midnight (Paris Time). No extensions will be granted.

Bootstrap estimation of confidence intervals for average counts of palyndromes in DNA sequences

Part 1 - Genomic Data Analysis and hypothesis testing

Scientists have found 296 palindromes[1] of 10 nucleotides or more in the DNA sequence of the human cytomegalovirus (CMV). They also found that the number of palindromes is non-uniformly distributed throughout the sequence without following any regular pattern.

Scientists have wondered whether the distribution of palindromes in the DNA sequence of CMV follows a well-established probability distribution. In this first section of the project, you will analyze the associated data and test the hypothesis of the placements of palindromes in the CMV's DNA sequence following a Poisson distribution. You will base your analysis on the data attached to the project, which gives the location of the palindromes detected in the DNA sequence of CMV as an integer in the range $[1, 229354]$.

1.1 Homogeneous Poisson Model

Question 1: (i) Compute the normalized histogram of the distance between consecutive palindromes. (ii) Determine an estimation of the average distance between palindrome (justify your choice of the estimator) and (iii) compare it to an exponential distribution $\mathcal{E}(\lambda_e)$ and justify your choice for the parameter λ_e. Comment.

We are now interested in the number of palindromes in various intervals of the DNA sequence. In the following, we will choose bin widths of 3000 for the intervals and counts the number of palindromes detected in these intervals.

Question 2: Compute from the data the table with the observed number of palindromes per interval. Display the table in the following form

Palindromes' count	Observed number O_i
0	O_0
1	O_1
$\vdots$	$\vdots$

[1]The meaning of palindrome in the context of genetics is slightly different from the definition used for words and sentences. Since a double helix is formed by two paired antiparallel strands of nucleotides that run in opposite directions, and the nucleotides always pair in the same way (adenine (A) with thymine (T) in DNA or uracil (U) in RNA; cytosine (C) with guanine (G)), a (single-stranded) nucleotide sequence is said to be a palindrome if it is equal to its reverse complement. For example, the DNA sequence ACCTAGGT is palindromic with its nucleotide-by-nucleotide complement TGGATCCA because reversing the order of the nucleotides in the complement gives the original sequence (*source: Wikipedia*)

Question 3: Using the method of moment estimation and maximum likelihood estimation, determine an estimator for parameter λ_p for a Poisson Distribution.

Question 4: (i) Using Matlab / Python, implement a χ^2-Goodness-of-Fit Test and search evidence of the number of palindrome following a Poisson distribution of parameter λ_p. Use significance level $\alpha = 0.05$ (ii) Compare your estimated Poisson parameter λ_p with the parameter λ_e. (iv) Comment and conclude this analysis.

We will be later interested in the project into generating confidence interval on the unknown parameter λ_p.

1.2 Negative binomial model

Another popular probabilistic model for random counting processes is the *Negative binomial distribution*, which counts the number of failures in a sequence of independent and identically distributed Bernoulli trials with probability of success p before a specified number of successes (denoted r) occurs. The probability mass function of $X \sim \mathcal{NB}(r, p)$ is defined as follows :

$$\mathbb{P}(X = k) = C_k^{k+r-1} p^r (1-p)^k \text{ with } k \in \mathbb{N}. \tag{1}$$

This model is also interesting because it allows for different mean and variance, contrary to the Poisson model. We admit here that if $X \sim \mathcal{NB}(r, p)$ then

$$\mathbb{E}(X) = \frac{r(1-p)}{p} \text{ and } \mathrm{Var}(X) = \frac{\mathbb{E}(X)}{p} = \frac{r(1-p)}{p^2}. \tag{2}$$

Question 5: Using the moment estimation technique, provide estimator for parameter r and p for a Negative Binomial distribution.

Question 6: Do a χ^2-Goodness-of-Fit Test with the data table found in Question 2 to search for statistical evidence of a Negative Binomial distribution. Use significance level $\alpha = 0.05$. Conclude.

Part 2 - Introduction to bootstrap estimation of confidence intervals

In the following section, We propose to present a computer-intensive technique used frequently in statistics and known as *Bootstrap* (see Appendix for more detail) to determine these confidence intervals.

2.1 Sample mean estimator and confidence intervals for the mean

Estimating a population's unknown mean relies mainly on using the sample mean estimator, whose behavior is well understood from a theoretical standpoint. It is a minimum variance unbiased estimator with an approximately normal sampling distribution (according to the Central Limit Theorem), with standard error decreasing proportionally to $1/\sqrt{n}$, with n the size of the random sample. We propose first to observe this behavior numerically through numerical simulation.

Question 7: Generate $N = 2000$ different random samples of size $n = 25$ noted as follows $\{X_i^{\#j}\}$ with $i = 1, \dots n$ and $\#j = 1, \dots, 2000$. $X_i^{\#j}$ are iid random variable following a standard normal distribution $\mathcal{N}(0, 1)$. For each sample $\#j$, compute the sample mean estimation $\overline{x}^{\#j}$ and then display the normalized histogram using the set of values $\left\{\overline{x}^{\#j}\right\}^{\#j=1,\dots 2000}$ and compare it to the theoretical distribution of the sample mean.

Question 8: Compare the sample mean's theoretical mean and standard error with your simulated values.

We are also interested in the two-sided confidence intervals on the mean, which in theory, given by $\left[\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ when the variance of the population is known. We can also compute it numerically using, for example, the `prctile` function in Matlab.

Question 9: Using the `prctile` function in Matlab (or the equivalent in Python), draw the mean and confidence interval on the sampling distribution histogram.

2.2 Non-parametric Bootstraping confidence interval from a single random sample

Now, consider having a single random sample ($N = 1$) of $n = 25$ samples. As a result, we can only do a single point estimation on the sample mean and not get an as refined numerical estimation of its standard error (or variability) from this single random experiment. The idea behind non-parametric bootstrapping is to circumvent this problem.

Question 10: Read the description of non-parametric bootstrapping in the Appendix and apply the technique to obtain a 95% confidence interval on the population mean μ using the `prctile` function in Matlab (or equivalent in Python). use $N_b = 2000$ bootstrap samples.

Question 11: Compare the bootstrap distribution of the sample mean with the sampling distribution obtained in Question 7. Comment.

2.3 - Parametric Bootstraping confidence interval from a single random sample

In this problem section, we are interested in computing a confidence interval for the mean of a standard normal distribution using the parametric bootstrap method. We will use the same initial random sample of $n = 25$ random sample used in Section 2.2 to make a quantitative comparison.

Question 12: Read the description of the parametric bootstrapping in the Appendix and apply the technique to obtain a 95% confidence interval on the population mean μ using the `prctile` function in Matlab (or equivalent in Python). use $N_B = 2000$ bootstrap samples.

Question 13: Compare the parametric and non-parametric bootstrap confidence intervals. Comment.

Part 3 - Application to the estimation of mean distribution of palindromes in DNA data

Considering the data used in part 3, we would like to determine a confidence interval in the Poisson rate λ_p and exponential rate λ_e.

Question 14: Using the parametric bootstrap method, compute 95% two-sided confidence intervals for λ_p and λ_e. Use $N_B = 1000$ bootstrap samples.

Appendix

A.1 Non-parametric boostrap for computing confidence intervals

This appendix section is adapted from the textbook "Introduction to Computational Finance and Financial Econometrics with R", E. Zivot.

The non-parametric bootstrap is amongst the simplest and most widely used type of bootstrap. It is reminiscent of Monte Carlo Simulations in how it operates. In the non-parametric bootstrap, we randomly resample the original data with replacement (*i.e.* a sample can be selected multiple times) to create synthetic random samples from the original available data. The bootstrap process is described below:

1. **Resampling.** We create N_B bootstrap samples by randomly sampling with replacement the original data $\{x_1, \ldots, x_n\}$. We obtain the following bootstrap samples :

Bootstrap Index	Bootstrap Sample
1	$\left\{x^*_{1,1}, \ldots, x^*_{n,1}\right\}$
2	$\left\{x^*_{1,2}, \ldots, x^*_{n,2}\right\}$
$\vdots$	$\vdots$
N_B	$\left\{x^*_{1}, \ldots, x^*_{n,N_B}\right\}$

2. **Estimating.** From each bootstrap sample, compute the point estimation of the parameter of interest θ and denote it $\hat{\theta}^*_i$

3. **Computing.** From $\left\{\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{N_B}\right\}$ compute an estimate of bias, standard error, and or approximate confidence interval.

There are various methods for estimating confidence intervals with the bootstrap, and the properties of the distribution of the bootstrap estimates determine which confidence interval to employ in practice. In this section, we present a purely numerical method based on the distribution of bootstrap estimations $\left\{\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{N_B}\right\}$. We consider for a two-sided $100(1-\alpha)\%$ confidence interval

$$[q^*_{\alpha/2}, q^*_{1-\alpha/2}], \tag{3}$$

with $q^*_{1-\alpha/2}$ and $q_{\alpha/2}$ the $(1-\alpha/2)\%$ and $(\alpha/2)\%$ empirical quantiles of $\left\{\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{N_B}\right\}$.

For example, searching for a 95% two-sided confidence interval would require the 2.5% and 97.5% quantiles. These quantiles partition the ordered set $\{\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(N_B)}\}$ such that 2.5% values are smaller than $q_{0.025}$ and 97.5% are smaller than $q_{0.975}$.

The above approach will work well if the bootstrap distribution is approximately normally distributed. However, the confidence intervals may have a variable coverage probability if the bootstrap distribution is asymmetric, especially if the asymmetry is significant (NB: a $QQ-$plot can

be used to assess the distribution's approximate normality visually). In the case of asymmetric distribution, there are bias and skewness confidence adjusted (BCA) intervals, which are beyond this project's scope.

A3. Parametric boostrap for computing confidence intervals

This appendix section are adapted from the textbook "Applied Statistics and Probability for Engineers", 3rd Ed. By D. C. Montgomery and G. C. Runger.

Considering an arbitrary parameter of interest θ, it can be challenging in the general case to find the proper values of the critical point to compute a confidence interval. In the general case, when sampling distributions are not symmetric (similar to the case of the χ^2 distribution), we may generalize the expression of the lower and upper bounds $[L, U]$ of a two-sided confidence interval with confidence level $100\%(1-\alpha)$ as follows:

$$L = \hat{\theta} - 100(1-\alpha/2) \text{ percentile of } (\hat{\theta} - \theta), \tag{4}$$

$$U = \hat{\theta} - 100(\alpha/2) \text{ percentile of } (\hat{\theta} - \theta). \tag{5}$$

Typically, in the case of a confidence interval on the unknown mean, the $100(1-\alpha/2)$ and $100\alpha/2$ percentile[2] of $(\hat{\mu} - \mu)$ are given by $\pm z_{\alpha/2}\sigma/\sqrt{n}$ (due to the symmetry of the sampling distribution).

Unfortunately, the percentiles of $\hat{\theta}$ may not be as easy to find as in the case of the normal distribution of the sample mean. However, they could be estimated using a computer-intensive technique called the parametric *bootstrap* that was developed in recent years.

Suppose we are sampling $f(x;\theta)$ from a distribution. The random sample results in data values $x_1, \ldots, x_n$, and we obtain $\hat{\theta}$ as the point estimate of θ. We would now use a computer to obtain bootstrap samples from the distribution $f(x;\hat{\theta})$, and for each of these samples, we calculate the bootstrap estimate. This results in the following table :

Bootstrap sample	Observations	Bootstrap Estimate
1	$x^*_{1,1}, \ldots, x^*_{n,1}$	$\hat{\theta}^*_1$
2	$x^*_{1,2}, \ldots, x^*_{n,2}$	$\hat{\theta}^*_2$
$\vdots$	$\vdots$	$\vdots$
N_B	$x^*_{1,N_B}, \ldots, x^*_{n,N_B}$	$\hat{\theta}^*_{N_B}$

Usually, $N_B = 100s$ to $1000s$ of these bootstrap samples are taken. We can then define the sample mean estimate of the bootstrap estimates as

$$\overline{\theta}^* = \frac{1}{N_B}\sum_{i=1}^{N_B} \hat{\theta}^*_i. \tag{6}$$

[2]In statistics, percentiles are found by taking a large set of numerical data, arranging it in ascending order, and then dividing it into 100 groups with an equal number of data points. Each of the 99 dividing points is called a percentile of the data set.(*Source: Wikipedia*)

In the parametric bootstrap technique, the required percentiles can be obtained directly from the differences between $\hat{\theta}_i^* - \overline{\theta}^*$. For example, if $B = 200$ and we are interested in a 95% confidence interval on θ is desired, the fifth smallest and fifth largest of the differences $\hat{\theta}_i^* - \overline{\theta}^*$ are the estimates of the necessary percentiles.

We will illustrate this procedure using an exponential distribution parameter λ. We consider a random sample of $n = 8$, and the estimation of λ obtained was $\hat{\lambda} = 0.0462$, where $\hat{\lambda} = 1/\overline{X}$ is a maximum likelihood estimator. We use $N_B = 200$ bootstrap samples generated by simulating the distribution $\mathcal{E}(\hat{\lambda} = 0.0462)$. We obtain an asymmetrical, skewed to the right, bootstrap distribution, which indicates that the sampling distribution of $\hat{\lambda}$ also has this same shape. We subtract the sample average obtained from these bootstrap estimates $\overline{\lambda}^* = 0.5013$ from each $\hat{\lambda}_i^*$. Suppose we wish to find a 90% two-sided confidence interval for λ; the fifth percentile of the bootstrap samples $\hat{\lambda}_i^* - \overline{\lambda}^*$ is found as -0.0228, and the ninety-fifth percentile is found as 0.03135. Therefore the lower and upper 90% bootstrap confidence limits are

$$L = \hat{\lambda} - 95 \text{ percentile of } (\hat{\lambda} - \overline{\lambda}^*) = 0.0462 - (0.03135) = 0.0149, \quad (7)$$

$$U = \hat{\lambda} - 5 \text{ percentile of } (\hat{\lambda} - \overline{\lambda}^*) = 0.0462 - (-0.0228) = 0.0690. \quad (8)$$

The 90% bootstrap two-sided confidence interval for λ is $[0.0149, 0.0690]$. This value can be compared to the 90% confidence interval $[0.0230, 0.0759]$. It shows the similarity between the two confidence intervals with a slightly lengthy $(+0.0012)$ bootstrap interval. This method can be improved and is known as the bias-corrected and accelerated (BCA) method, which adjusts the percentiles in more general cases. It could be applied in this example but at the cost of additional complexity.