Probability and statistics - Winter 2023/24 - Homework

General guidelines

Work in groups of two or three students. The homework is to be submitted either in print or in pdf latest on January 31st 2024 or five days before your final exam, whichever comes first.

The submitted report should consist of solutions of the tasks along with a **thorough explanation**, interpretation, graphs and key parts of your computer code.

Note: If you wish to forgo solving some parts of the assignment and receive no points for it, please note that in your solution.

Data and parameters

Select a teamleader among your group. Find the parameters of your assignment as follows:

- K = day in month when the teamleader was born $(1 - 31)$.
- L = number of letters in the teamleader's last name.
- $M = ((K + L) * 47) \pmod{11} + 1$.

Load a data set from the R library Sleuth2 corresponding to the following table:

M	Data set	Description
1	case0101	creativity score based on motivation type
2	case0102	salary based on gender
3	case0201	humerus length based on sparrow survival
4	case0202	hippocampus volume based on schizofrenia
5	case0301	rainfall based on cloud seeding
6	case0302	dioxin concentration based on service history
7	case0402	time to complete a problem based on geometry training type
8	ex0112	blood pressure based on diet
9	ex0211	guinea pig survival based on experimental treatments
10	ex0221	weight based on sparrow survival
11	ex0222	cholesterol levels based on living environment

Each of the data sets contains observations from an experiment, where a continuous variable was measured among two groups.

The library can be downloaded here:
https://cran.r-project.org/package=Sleuth2

The documentation for the data sets is available here:
https://cran.r-project.org/web/packages/Sleuth2/Sleuth2.pdf

Assignment

1) (1pt) Load the data set and separate the data into the two observed parts. Provide an overview of each of them by estimating the expectation, variance and median of the corresponding distribution and briefly describing the nature of the studied problem.

2) (1pt) For each group separately, estimate the density and distribution function of the data using the histogram and the empirical distribution function.

3) (3pts) For each of the observed parts separately, find the most similar distribution: Estimate the parameters of the normal, exponential and uniform distribution. Insert the corresponding densities with estimated parameters into the plot of the histogram. Discuss which of them fits the data best.

4) (1pt) For each of the groups, generate a random sample of 100 observations from the distribution you have chosen in the previous part, with parameters estimated from the data. Compare the histogram of the simulated values with the original data.

5) (1pt) For both parts separately, compute the two-sided confidence interval for the expected value with confidence level 95%.

6) (1pt) Perform a test of the hypothesis, whether the expectation of either of the parts of the data set is equal to K (assignment parameter) against the two-sided alternative, on level of significance 5%. You can use either the previous result or an in-built function.

7) (2pts) Perform a test of the hypothesis, whether the expectations of both observed parts are equal. Use level of significance 5%. Choose the type of test and the alternative hypothesis in a way which corresponds with the examined problem best.

Software

It is recommended to use the freeware statistical system R. You can also use a different software with needed functionalities (Python, Matlab, Excel, ...). Computations can be done either manually using the corresponding formulas, or with the help of statistical packages and functions.
Either way, the solutions should be thoroughly explained and interpreted.

R references

Installation	`http://www.r-project.org/`
R reference card	`http://cran.r-project.org/doc/contrib/Short-refcard.pdf`
R tutorial	`http://www.cyclismo.org/tutorial/R/index.html`
R guide	`http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf`
Simple R	`http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`
Introduction to R	`http://cran.r-project.org/doc/contrib/Lam-IntroductionToR_LHL.pdf`

Important R commands

Install library	`install.packages("Sleuth2")`
Load library	`library(Sleuth2)`
Export data	`write.table(ex0331,"D:/my_data.csv",row.names=F,sep=";",dec=",")`
Rows, columns and fields of a table	`ex0331[k,]; ex0331[,l]; ex0331[k,l];`
Other important commands	`subset();mean();ecdf();hist();t.test();`
Documentation for a given command	`?subset`