ADVANCED WORKBOOK

Task for: DLMDSAS01– Advanced Statistics

The workbook bases on the output of the parameter generator. The produced numbers are ξ_1 to ξ_{20} and are used in the assignment. This gives each student a personal set of numbers. In the tasks below, several parts start with sentences such as "If ξ_9 is 5": **Do not** perform tasks if the corresponding value is different in your parameters (e.g., if you have $\xi_9 = 3$).

Task 1: Basic Probabilities and Visualizations (1)

Please provide the requested visualization as well as the numerical results. In both cases, please either prove or cite any computation of the proof steps (calculations, code, steps, etc.), and justify why you trust the tools you used. Do not forget to include the scale of each graphics so that the reader can interpret the numbers represented. As indicated in the remark above, only perform one of the following tasks based on your personal value of ξ_1:

If ξ_1 is 0: A vote with the outcome for or $against$ follows a Bernoulli distribution where $P(\text{vote} = "for") = \xi_2$. Represent the proportion of "for" and "against" in this single Bernoulli trial using a graphic and a percentage. Can an expectation be calculated? Justify your answer with all necessary hypotheses.

If ξ_1 is between 1 and 3:
The number of meteorites falling into an ocean in a given year can be modeled by one of the following distributions. Provide a graphic showing the probability of one, two, three, etc., meteorites falling (until the probability remains less than 0.5% for any higher number of meteorites; you should also prove it). Calculate the expectation and median and present them in this graphic:

- If ξ_1 is 1: a Poisson distribution with an expectation of $\lambda = \xi_2$
- If ξ_1 is 2: a negative binomial distribution with an expectation of $k = \xi_2$ and $p = \xi_3$
- If ξ_1 is 3: a geometric distribution counting the number of Bernoulli trials with $p = \xi_2$ until it succeeds.

Task 2: Basic Probabilities and Visualizations (2)

Let Y be the random variable with the time it takes to hear an owl from your room's open window (in hours). Assume that the probability that you still need to wait to hear the owl after y hours is one of the following functions:

- If ξ_4 is 0: the probability is given by $\xi_5 \mathrm{e}^{-\xi_6 y} + \xi_7 \mathrm{e}^{-\xi_8 y}$
- If ξ_4 is 1: the probability is given by $\xi_5 \mathrm{e}^{-\xi_6 y^2} + \xi_7 \mathrm{e}^{-\xi_8 y^8}$
- If ξ_4 is 2: the probability is given by $\xi_5 \mathrm{e}^{-\xi_6 \sqrt{y}} + \xi_7 \mathrm{e}^{-\xi_8 \sqrt[3]{y}}$
- If ξ_4 is 3: the probability is given by $\xi_5 \mathrm{e}^{-\xi_6 y^2} + \xi_7 \mathrm{e}^{-\xi_8 y^2}$

Find the probability that you need to wait between two and four hours to hear the owl, and compute and display the probability density function graph, as well as a histogram where each bar represents the probability of the hearing the owl at any particular minute. Compute and display in the graphics the mean, variance, and quartiles of the waiting times.
Please pay attention to the various units of time and make sure that the parameters you receive are such that $\xi_5 + \xi_7 = 1$.

Task 3: Transformed Random Variables

A type of network router has a bandwidth total to first hardware failure called S expressed in terabytes. The random variable S is modeled by an exponential distribution whose density is given by one of the following functions:

- If $\xi_9 = 0$: $f_S(s) = \frac{1}{\theta} e^{-\frac{s}{\theta}}$
- If $\xi_9 = 1$: $f_S(s) = \frac{1}{24\theta^5} s^4 e^{-\frac{s}{\theta}}$
- If $\xi_9 = 2$: $f_S(s) = \frac{1}{120\theta^7} s^6 e^{-\frac{s}{\theta}}$

with a single parameter θ. Consider the bandwidth total to failure T of the sequence of the two routers of the same type (one being brought up automatically when the first is broken).
Express T in terms of the bandwidth total to failure of single routers S_1 and S_2. Formulate realistic assumptions about these random variables. Calculate the density function of the variable T.
Given an experiment with the dual-router-system yielding a sample $T_1, T_2, \ldots, T_n$, calculate the likelihood function for θ. Propose a transformation of this likelihood function whose maximum is the same and can be computed easily.

An actual experiment is performed, and the infrastructure team has obtained the bandwidth totals to failure given by the sequence ξ_{10} of numbers. Estimate the model-parameter with the maximum likelihood and compute the expectation of the bandwidth total to failure of the dual-router-system.

Task 4: Hypothesis Test

Over a long period of time, the production of 1,000 high-quality hammers in a factory seems to have reached a weight with an average of ξ_{11} (in g) and standard deviation of ξ_{12} (in g). Propose a model for the weight of the hammers including a probability distribution for the weight. Provide all the assumptions needed for this model to hold (even the uncertain ones). What parameters does this model have?

One aims to answer one of the following questions about a new production system:

- (if $\xi_{13} = 0$): Does the new system make *more constant* weights?
- (if $\xi_{13} = 1$): Does the new system make *lower* weights?
- (if $\xi_{13} = 2$): Does the new system make *higher* weights?
- (if $\xi_{13} = 3$): Does the new system make *less constant* weights?

To answer this question, a random sample of newly produced hammers is evaluated yielding the weights in ξ_{14}.

What hypotheses can you propose to test the question? What test and decision rule can you make to estimate whether the new system answers the given question? Express the decision rules as logical statements involving critical values. What error probabilities can you suggest and calculate? Perform the test and draw the conclusion to answer the question.

Task 5: Regularized Regression

Given the values of an unknown function $f: \mathbb{R} \to \mathbb{R}$ at some selected points, we try to calculate the parameters of a model function using OLS as a distance and a ridge regularization:

- (if $\xi_{15} = 0$): a polynomial model function of twelve α_i parameters: $f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{12} x^{12}$
- (if $\xi_{15} = 2$): a polynomial model function of ten α_i parameters: $f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{10} x^{10}$

Calculate the OLS estimate, and the OLS ridge-regularized estimates for the parameters given the sample points of the graph of f given that the values are (x, y) each of the elements of ξ_{16}. What weight do you give to the penalties? What are the qualities of each of the solutions?

Remember to include the steps of your computation, which are more important than the actual computations. If you calculate the solution with a program, make sure that you trust and cite the core functions used and that you sketch the mathematical path in a way that is coherent with the program.

Task 6: Bayesian Estimates

Following Hogg et al. (2020), exercise 11.2.2:
Let $x_1, x_2 \ldots, x_{10}$ be a random sample from a gamma distribution with $\alpha = 3$ and $\beta = 1/\theta$. Suppose we believe that θ follows a gamma-distribution with $\alpha = \xi_{17}$ and $\beta = \xi_{18}$ and suppose we have a trial $(x_1, \ldots, x_n)$ with an observed $\overline{x} = \xi_{19}$.

a) Find the posterior distribution of θ.
b) What is the Bayes point estimate of θ associated with the square-error loss function?
c) What is the Bayes point estimate of θ using the mode of the posterior distribution?