

INTRODUCTION TO STATISTICS AND DATA SCIENCE (4ECON006C_n) PORTFOLIO

Semester: Spring 2025

Weight: 20%

Date available: February 19th, 9:00 am

Deadline of submission: by 11:59 pm on February 20th, 2025





Link of submission for zipped R file:

<https://intranet.wiut.uz/Coursework/UploadsDT?courseworkID=4337>

Link of submission of the Report (.doc or .pdf):

<https://intranet.wiut.uz/Coursework/UploadsDT?courseworkID=4399>

(You can also find the submission pages by visiting Module Intranet -> Course Work -> Portfolio and Report)

ID	Assignment title	Turnitin enabled	Deadline date	Created date	Created user	Edited date	Edited user	Action
4337	Portfolio (.R file in zipped format)	No	2/20/2025 11:59:59 PM	1/24/2025 9:48:33 AM	Olmas Isakov	2/13/2025 2:43:19 PM	Olmas Isakov	 
4399	Report (.doc or .pdf)	Yes	2/20/2025 11:59:59 PM	2/13/2025 2:44:41 PM	Olmas Isakov	2/13/2025 2:44:41 PM	Olmas Isakov	 

Instructions to students. Please read them carefully before you start.

1. This is an individual work and you are solely responsible to submit your own work. Any type of collaboration with others is not allowed and will be subject to academic misconduct policies.
2. Do not change the variable names (such as task_1a, task_1b, etc.) in the R file. **This is very important.** Your solutions will be checked by those variable names.
3. Most of the exercises require the student to insert his/her last 5 digits of student ID in place of \overline{abcde} values. Check several times to ensure you are using the correct values.
For example, a student with an ID of 00014040 has the following values:
 $\overline{a} = 1, \overline{b} = 4, \overline{c} = 0, \overline{d} = 4, \overline{e} = 0.$
A student with an ID of 00009540 has the following values:
 $\overline{a} = 0, \overline{b} = 9, \overline{c} = 5, \overline{d} = 4, \overline{e} = 0.$
4. You **must** replace the \overline{abcde} values with the correct values from your own ID in each exercise. Do not provide the solutions with a reference to a, b, c, d, e values (i.e. your code should not have these letters,

replace them with your own ID values). For example, if you declare these values in the beginning of your R file and continue to refer to these letters in the exercises, then **this work is not acceptable**.

5. **Do not round your answers unless it is stated in the question.** Your answers should be provided in R codes. Watch the recorded Portfolio instructions before you start.

6. **Submission instructions:**

- R file (zipped): You must use the R Template file to write your codes, rename it with your student ID and upload it in compressed (zipped) format. Ensure that your file contains the correct file.
- Report: This file must contain your own explanations for each task. Plagiarism and use of AI to generate explanations are not allowed and will be subject to Academic Misconduct regulations.

7. Late submissions are not acceptable. Please upload your work at least 30 minutes before the deadline. The Intranet might become unresponsive when many students start uploading at the same time.

TASKS

Task 1. [10 marks] Oriyat FM, a radio station, finds that the distribution of the lengths of time listeners are tuned to the station follows the normal distribution. The mean of the distribution is $(30+\bar{a})$ minutes and the standard deviation is $(8+\bar{b})$ minutes.

- a. What is the probability that a particular listener will tune in between $(30+\bar{c})$ and $(45+\bar{d})$ minutes? [5 marks]
- b. What **percent** of the listeners tune in for more than $(28+\bar{a})$ minutes? **Do not include the percentage sign (%) in your solution.** [5 marks]

Task 2. [10 marks] Light bulbs are tested for their life-span. It is found that $(\bar{c}+\bar{d}+5)\%$ of the light bulbs are rejected. A random sample of $(20+\bar{a})$ bulbs is taken from the stock and tested. The random variable X is the number of bulbs that are rejected.

- a. What is the probability that the value of X is at least $(2+\bar{a})$? [5 marks]
- b. What is the $(80+\bar{e})^{\text{th}}$ percentile value for X? [5 marks]

Task 3. [10 marks] Suppose the closing stock price (X, in USD) of Bank of America Corp follows the following continuous distribution in one year:

$$p(X) = \begin{cases} \frac{1}{35} & \text{for } 20 + \bar{c} < X < 55 + \bar{c} \\ 0 & \text{for other values of X.} \end{cases}$$

- a. What is the probability that the stock price will close above $$(30 + \bar{b})$ in a randomly chosen trading day? [5 marks]
- b. Suppose there are 252 trading days in 2025. How many days should we expect the stock price to close below $$(40 + \bar{d})$? **Round your answer to the nearest integer value.** [5 marks]

Task 4. [10 marks] In a cafe, the customers arrive at a mean rate of $(\bar{c}+4)$ per every 12 minutes. The variance of customer arrivals is equal to $(5*\bar{c}+20)$ per hour.

- a. Find the probability of arrival of at most $(\bar{a}+1)$ customers in the **next minute**. [5 marks]
- b. Let x denote the number of customer arrivals per 12 minutes.
Find $P(\bar{c}+2 \leq x < \bar{c}+6)$. [5 marks]

Task 5. [20 marks: 2 marks for each correct answer]

- Create a numeric vector with a sequence of only **even** numbers from $(\bar{a}+10)$ to $(\bar{b}+25)$.
- Create a vector of consecutive integer numbers which starts from 1 and has the same length as the vector from part 5a.
- Add the vectors from part 5a and 5b and store the new vector under the given name in R file.
- Sum up the elements of the vector from part 5c.
- Find the median value of the vector created in part 5c.
- Find the mean of the vector created in part 5c.
- Find the sample standard deviation of the vector created in part 5c.
- Find the $(40+\bar{a})^{\text{th}}$ percentile of the vector created in part 5c.
- Find the IQR of the vector created in part 5c. Use built-in R function(s) to find the IQR.
- Calculate the absolute difference between $(80 + \bar{c})^{\text{th}}$ and $(20+\bar{d})^{\text{th}}$ percentiles of the vector from part 5c.

Task 6. [10 marks]

A man has two bags. Bag A contains $(2+\bar{d})$ keys and bag B contains $(12+\bar{e})$ keys. Only one of those keys fits the lock which he is trying to open. The man selects a bag at random, picks out a key from the bag at random and tries that key in the lock. What is the probability that the key he has chosen fits the lock?

Task 7. [10 marks]

A company wants to investigate the quality of its current customer service with the one it had two years ago. To do so, the company research team randomly selected customer enquiries and looked at the summary statistics of waiting times (in minutes) in those enquiries. The following table provides the data:

Year	Average waiting time	Sample standard deviation	Sample size
2024	$(7 + \bar{b})$	$(3 + \bar{c})$	$(35 + \bar{d})$
2022	$(7.5 + \bar{b})$	$(4 + \bar{c})$	$(40 + \bar{e})$

Compute the p-value of the hypothesis test whether the average waiting times have decreased in 2024 compared to 2022.

Task 8. [20 marks]

Read the following college data into your RStudio using the url option:

Link: "https://s3.amazonaws.com/itao-30230/college.csv"

Remove the following rows from the dataset all at once: $(\bar{a}+5)$, $(\bar{a}+15)$, $(\bar{b}+105)$, $(\bar{b}+255)$, $(\bar{c}+405)$, $(\bar{c}+455)$, $(\bar{d}+600)$, $(\bar{d}+700)$, $(\bar{e}+1001)$.

This is your unique college data and you will be working with this dataset for the following questions:

- Find the average tuition for private universities with more than $\$(12000 + 300*d + 400*e)$ tuition rate in the state of New York. **[3 marks]**
- How many universities from the South region have an acceptance rate higher than 40% and SAT average above 1050? **[3 marks]**
- How many university names in the data contain the string "Virginia" (Consider all letter cases, such as "virginia", "VIRGINIA" if any)? **[3 marks]**
- Group the universities by *state* and compute the average of *tuition* by each *state*.
Which state has the minimum value? Your answer code should ONLY produce the two-letter abbreviation from the state column. **[3 marks]**
- Create a set of boxplots of tuition rates based on 4 regions in a 2 by 2 frame.
Make sure to have 4 different colors for your boxplots using only hex color codes.
Your student ID must be shown in the main title of the plot. **[4 marks]**
- Create the histograms of tuition rates for private and public universities in overlapped plots in two different transparent colors (blue for private and green for public). Your ID must be given in the title of the plot (see the graph on the next page). **[4 marks]**

Overlapped Histograms of Tuition Fees (Student ID: _____)

