Last updated on August 12, 2025

BM5033 Statistical Inference Methods in Bioengineering

Problem Set 1

Instructions

1. The dataset for this assignment can be downloaded from here and here.
2. These problems are to make you comfortable with the topics covered in the class. Some of the problems require you to analyze data and make plots. Use these problems as examples to learn these methods on your own.
3. You are supposed to work on these problems independently and do not have to submit the answers.

Questions

1. Consider the following data obtained from 4 repetitions of an experiment measuring the rate of cell death due to an anti-cancer drug

Starting cell count	Time taken	Rate of cell death
10^6	2.0hrs	5.0×10^5 cells/hr
5×10^6	8.0hrs	6.3×10^5 cells/hr
2×10^6	4.0hrs	5.0×10^5 cells/hr
2×10^6	5.0hrs	4.0×10^5 cells/hr

   What is the average rate of cell death in this experiment?

2. Consider the following three data sets
   (a) Volume (in mm^3) of AV malformation in the brain (A1Q1a.dat)
   (b) Elastic modulus (in GPa) of femoral cortical bone (A1Q1b.dat)
   (c) Hematocrit levels (in %) in dengue patients (A1Q1c.dat)

   and answer the following
   (i) Make box plots for these data.
   (ii) Identify and calculate the appropriate measure of central tendency and dispersion you will use for each. Justify your answer in 30 words for each case.

3. Consider the data
   (a) Degree of anisotropy in compressive strength and age for male and female donors (A1Q2a.dat)
   (b) Hemoglobin (in g/dl) and hematocrit levels (in %) of dengue patients (A1Q2b.dat)

   and answer the following
   (i) Make scatter plots of these datasets.
   (ii) Calculate the mean, median, and mode for three datasets. Plot histograms using a tool of your choice.
   (iii) Calculate variance, standard deviation, and IQR for the datasets.

4. In a study to identify the role of air pollution on pulmonary fibrosis (PF), a survey was performed in Patancheru (where several pharmaceutical plants are located), and in Zaheerabad. The data are summarized in the following table

	Patancheru	Zaheerabad
With PF	546	183
No PF	1753	1202

   (a) What is the fraction of the sample suffering from PF and exposed to air pollution?

   (b) What is the fraction of the sample suffering from PF and not exposed to air pollution?

   (c) What can you conclude from here about a possible association of air pollution with PF?

5. Check if Python, MS Excel, Google Sheets, and/or any other tool of your choice use $(n-1)$ or n in the denominator for variance calculation.

6. At the end of a course last year (not necessarily BM5033), the data from the student feedback for the course was the following.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
3	2	3	7	30

Table 1: Ratings showing the course of effective in achieving its target

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
2	3	1	15	24

Table 2: Ratings of the effectiveness of the instructor

   Identify the correct metrics to summarize (central tendency and dispersion) the data and calculate them.

7. For the following data sets make histogram plots using R.

   (a) Sheet Q1A in this file

   (b) Sheet Q1B in this file

   (c) Sheet Q1C in this file

   (d) Sheet Q1D in this file

8. Calculate Pearson, Spearman's and Kendal's correlation coefficients using R for the following data.

   (a) Sheet Q2A in this file

   (b) Sheet Q2B in this file

9. Also explain the appropriate choice of the correlation coefficient for each data in the last question.

10. Construct a data-set containing 10 pairs of values (these values are not supposed to represent anything of biological or physical relevance) such that the Pearson correlation, Spearman's correlation and Kendall's correlation give values which differ with each other by a magnitude of 0.1 or more.