Introduction to Statistical Inference

Erfan Rasouli, Mahyar Maleki
erfanrasouli121@gmail.com, mahyar.7667@gmail.com
Instructor: Mohammad-Reza A. Dehaqani
Deadline:
12 Bahman 1402

I. Problems

Problem 1

In this question, we are going to examine the bootstrapping method. For this question, please utilize the diabetes data set that has been provided to you.

Consider the population to be equal to the BMI, Body Mass Index, of individuals who do not have diabetes.

1) Choose a sample size of 100 from population.
2) Construct a 90% confidence interval for the mean using bootstrapping on this sample. (use the percentile method and set the number of repetitions equal to 1000)
3) Report bootstrap mean and population mean.
4) Plot the histogram of the bootstrap distribution and show the vertical lines of the confidence interval on it.
5) Repeat steps 1-3 for a smaller sample size (10) and compare with the previous results.
6) Are the confidence intervals symmetric around the point estimate? If not, what might be the reason?

Problem 2

Four brands of flashlight batteries are to be compared by testing each brand in five flashlights. Twenty flashlights are randomly selected and divided randomly into four groups of five flashlights each. Then each group of flashlights uses a different brand of battery. The lifetimes of the batteries, to the nearest hour, are as follows.

Brand D	Brand C	Brand B	Brand A
20	24	28	42
32	36	36	30
38	28	31	39
28	28	32	28
25	33	27	29

TABLE I
Problem 2 Table

Preliminary data analyses indicate that the independent samples come from normal populations with equal standard deviations. At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

In your solution, please take into consideration the following items :

1) Basic conditions of using this test and checking them.
2) State null and alternative hypothesis
3) ANOVA table

Problem 3

An experiment was conducted to investigate the effects of three types of medication on reducing blood pressure. The ANOVA table for this experiment is as follows.

	Df	Sum Sq	Mean Sq	F value	$P(>F)$
Treatment	2	639.48	319.74	3.33	0.0461
Residuals	39	3740.43	95.91		

TABLE II
ANOVA Table

1) What are the hypotheses?
2) What is the conclusion of the test? Use a 5% significance level.
3) Conduct pairwise tests (bonferroni – t test) to determine which groups are different from each other. Summary statistics for each group are provided below
4) State one advantage and one disadvantage of the bonferroni correction method.

	Tr 1	Tr 2	Tr 3
Mean	6.21	2.86	-3.21
SD	12.3	7.94	8.57
n	14	14	14

TABLE III
Summary Statistics

Problem 4

Suppose that in the stage of multiple comparisons in an experiment, the p-values are as follows.

$$P-values : 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019$$

1) Use the Benjamini-Hochberg method, which is a method to control the FDR (false discovery rate), and determine the significant p-values. (Consider a control level of 5%)
2) Plot the p-value chart according to their rank and show the cut off line.
3) Briefly explain the difference between FDR control methods (such as Benjamini-Hochberg) and FWER (family wise error rate) control methods such as Bonferroni.

Problem 5

Determine whether the following statements are true or false and correct the false statements.

1) If the number of groups increases, then the type 1 error increases in multiple comparisons tests, so the corrected significance level should increase.
2) If the number of samples increases, the degree of freedom for the residuals also increases.
3) The F distribution is a symmetric distribution around the zero mean.
4) Using ANOVA test, we can conclude that all means are different from each other
5) If the initial hypothesis is rejected in the ANOVA test, the standardized variability between groups is higher than the standardized variability within groups.

Problem 6

Recent research studies suggest that having certain aromas or fragrances present in a work environment will enhance the productivity levels of the workers. In one such study, subjects were put in environments with different aromas present and asked to try to solve as many anagrams (word jumbles) as possible in a given amount of time. Suppose that four different aromas were compared in one such study. These aroma treatments were: Lemon fragrance, Floral fragrance, Fried food aroma, and No aroma (the control group). Further suppose that 12 persons of similar intelligence participated in such a study, with three being assigned at random to each of four aroma treatments. The subjects were put in a room with the given aroma for a half hour of anagram solving. The table below shows the number of anagrams each person solved.

Aroma	# Anagrams Solved		
Lemon	11	10	12
Floral	11	14	11
Fried Food	5	5	8
None	8	7	6

TABLE IV
ANAGRAMS SOLVED BY AROMA

1) Write the Null and Alternative hypotheses and conduct analysis using one-way ANOVA.(use the R or Python programming language to solve this part).
2) Determine the significantly different pairs of means using the Tukey's method.(Use a 5% significance level).
3) State one limitation of Tukey's procedure.

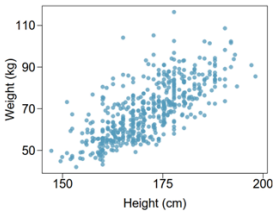
Problem 7

In one study, a team of researchers recruited 38 men and evenly divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.

1) What type of study is this?
2) What are the experimental and control treatments in this study?
3) Has blocking been used in this study? If so, what is the blocking variable?
4) Has blinding been used in this study?
5) Has double-blinding been used in this study?
6) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

Problem 8

he scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.



	Estimate	Std. Error	T value	Pr ($>\|t\|$)
(Intercept)	-105.0113	7.5394	-13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

1) Describe the relationship between height and weight.
2) Write the equation of the regression line. Interpret the slope and intercept in context.
3) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.
4) The correlation coefficient for height and weight is 0.72. Calculate R2 and interpret it in context.

Problem 9

Examine the data presented in the provided table for a simple linear regression scenario.

Xi	2.5	8.7	1.2	7.9	0.8	5.3	4.1	7.4	9.6	0.4
Yi	1.3	3.9	0.6	3.9	0.5	2.4	2.1	3.0	4.4	0.2

1) Using maximum likelihood estimator, calculate $\beta_1, \beta_0, \sigma^2, var(\beta_1), var(\beta_0)$.
2) Test the following hypotheses at the level of significance 0.05:
   a) $H_0 : \beta_0 = 0.5$
      $H_1 : \beta_0 \neq 0.5$
   b) The regression line passes through the origin in the XY plane.

Problem 10

Suppose that in a problem of simple linear regression, a confidence interval with confidence coefficient $1 - \alpha_0 (0 < \alpha_0 < 1)$ is constructed for the height of the regression line at a given value of x. Show that the length of this confidence interval is shortest when $x = \bar{x}_n$.

Problem 11

Suppose that $X \sim bin(n, p)$.

1) Show that the MLE (Maximum Likelihood Estimate) of p is $\bar{p} = X/n$.
2) Show that MLE of part (a) attains the Cramer-Rao lower bound.

Problem 12

(**Programing**) Consider the prostate dataset for programing parts:

https://hastie.su.domains/ElemStatLearn/data.html

Examine the validity of the theoretical Cramer-Rao lower bound for the linear regression model using ("lcavol", "lpsa") columns. Calculate the Cramer-Rao lower bound, and then compare it with the empirical results obtained from the dataset.

Problem 13

(**Programing**) Answer the questions using "lweight" variable as response variable and ("age", "lpsa") as explanatory variables.

1) Which explanatory variable do you guess is the more significant predictor and why?
2) For each explanatory variable:
   a) Investigate the linearity of data points using scatter plot of residuals.
   b) Compute the least squares regression.
   c) Write the predictive equation for the response variable and interpret its parameters.
   d) Draw a scatter plot of the relation between these two variables overlaid with the least-squares fit as a dashed line.
3) Using the previous part results, try to explain which variable is the more significant predictor.
4) Choose a random sample of 100 data points from the dataset.
   a) By 90 percent of data, build two Linear Regression models and design hypothesis tests to see if these explanatory variables are a significant predictor of the response variable or not.
   b) Calculate the 95% confidence interval for the slope of the relationship between response variable and explanatory variables. Interpret these CIs.
   c) Use your models to predict the values of the response variable for the remaining percent of samples.
   d) Compare the predicted values with actuals. Report the success rate.

II. Submission

For the programming section, each student is required to submit a well-structured, typed PDF report that presents a concise summary of their analysis. The report should include the figures mentioned in the problem description and offer a detailed discussion of each. Please avoid uploading theoritical problem in .jpg format and upload them in a single .pdf file.

For each section of the report, a separate script is expected, which can be written in MATLAB (.m), Python 3 (.py or .py3), or R (.r). Avoid submitting scripts in formats like MATLAB live scripts, Python notebooks, or R Markdown. It is crucial that the submitted code is compatible with the grader's system. Be sure to include all relevant functions and any non-standard libraries used in your code.

The report should be treated as an academic piece of writing, and it should not contain any code snippets or explanations of coding logic. Instead, it should provide the author's insights about the results and demonstrate a strong grasp of the reference article. Academic reports typically maintain a concise and highly formal tone.

Each section of the report should briefly outline the hypothesis being tested. The responsibility for designing and implementing the tests lies with the students, as does explaining the results. Interpretations should be comprehensive without unnecessary verbosity.

The report can be written in either Persian or English, with no preference for either. In Persian reports, use B Nazanin with a font size of 14 for the text body and B Titr with a font size of 18 for titles. English reports should use Times New Roman 12 for the body text and Times New Roman 16 for titles. Sentences should be written in the passive tense. In Persian reports, the correct usage of the zero-width non-joiner is mandatory. In all reports, equations, figures, and tables must be labeled with unique numbers and referenced accordingly. Referring to figures as "the following figure," "the figure above," and similar expressions is considered incorrect.

Every figure in the report should be accompanied by a descriptive caption below it, while tables should have captions above them. Feel free to use footnotes and citations as necessary for clarity and proper attribution.