Statistical Inference

HW Author: Parham Sazdar, Reyhane Vahedi
Instructor: Mohammadreza A. Dehaqani

Fall 2024

Homework 5

- This assignment entails **five mandatory questions**, which are in a similar format to those you might encounter in your final exam. These will directly contribute to your grade for this assignment. Additionally, there are **eleven optional questions**, which can be submitted at a later date to earn extra points and help compensate for any missed points from previous assignments.
- For the 5 mandatory questions, we recommend reviewing the relevant lecture materials thoroughly before attempting the questions. Approach each one as if it were an exam question. After writing down what you can, revisit the lecture content to refine and complete your responses. Full credit will be awarded if at least 80% of each solution is completed.
- Note that the deadline for turning in **the mandatory questions is Bahman 9th**, and the solutions for this section will be available on the course page before the final exam. **The optional section of this assignment is due on Bahman 15th**.
- Feel free to use the class group to ask questions — our TA team will do their best to help out!
- For any simulations or computations, focus on explaining your results clearly with proper graphs and analyses — these carry the most weight for scoring. Codes without a report will not be given any credit.
- Quick reminder: A short ($\sim$ 5-minute) hand-in session will be held to discuss your work. The date and time will be announced in the class group.

Question 1: Analysis of Statements Regarding ANOVA

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 the 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.

Question 2: Comparison of Flashlight Battery Lifetimes (Mandatory)

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
1	20	24	28	42
2	32	36	36	30
3	38	28	31	39
4	28	28	32	28
5	25	33	27	29

Table 1: Lifetimes of flashlight batteries (in hours)

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 the null and alternative hypotheses.
3. Construct the ANOVA table.

Question 3: Effects of Medication on Reducing Blood Pressure

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:

Source	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 2: ANOVA Table

Answer the following questions:

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:

	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 3: Summary Statistics

4. State one advantage and one disadvantage of the Bonferroni correction method.

Question 4: Analyzing a Randomized Controlled Study on Chia Seeds

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.

Question 5: Controlling False Discovery Rate in Multiple Comparisons (Mandatory)

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

$$P\text{-values} : 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019$$

Questions

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.

Question 6: Effect of Aromas on Anagram Solving Efficiency (Mandatory)

Recent research studies suggest that having certain aromas or fragrances present in a work environment can enhance the productivity levels of workers. In one such study, subjects were placed in environments with different aromas and asked to solve as many anagrams (word jumbles) as possible within 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
- No aroma (control group)

Further, suppose that 12 individuals of similar intelligence participated in the study, with three subjects assigned at random to each of the four aroma treatments. The subjects were exposed to the given aroma for half an hour of anagram solving. The table below shows the number of anagrams each person solved:

Table 4: Anagrams Solved by Aroma

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

Questions

1. Write the Null and Alternative hypotheses and conduct an 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 Tukey's method. (Use a 5% significance level).
3. State one limitation of Tukey's procedure.

Question 7: Bootsrtapping for a Problem of Ants' Population (Mandatory)

"How many ants would gather on a piece of sandwich on the ground, near an ant nest?"

In this question, we will be exploring the answer to this question via an experiment. For this experiment, we put a piece of sandwich near the opening of an ant colony for a few minutes and then cover it with a glass, and count the number of ants trapped beneath the glass. We have repeated this experiment 8 times in total, and we've listed the results in the table below:

Num. of Experiment	1	2	3	4	5	6	7	8
Num. of Ants	43	5	77	70	36	42	15	71

(i) Find the mean and standard deviation of the sample.

(ii) Explain how you can use 8 pieces of paper to create a Bootstrap Statistics.

(iii) What do we expect the shape, distribution, and the center of distribution of this Bootstrap be?

(iv) What would be a good parameter for representing this population of interest? What would be the best estimate for that parameter?

(v) One Bootstrap distribution has a standard error of 4.85. Use this standard error to find and interpret the 95% confidence interval for the parameter defined in part (iv).

Question 8: Permutation Tests for Paired Samples

Let $D_1, D_2, \ldots, D_n \overset{\text{IID}}{\sim}$ f for a probability density function f on $\mathbb{R}$, and consider a test of the null hypothesis

$$H_0 : f \text{ is symmetric about 0.}$$

(Against some alternative, say $H_1 : f$ is symmetric about a value $\mu > 0$) that rejects for large values of a test statistic $T = T(D_1, D_2, \ldots, D_n)$.

(i) Describe the distribution of T conditional on $|D_1|, \ldots, |D_n|$, under H_0. (What values can T take conditional on $|D_1|, \ldots, |D_n|$, and with what probabilities? You may assume no value of D_i is exactly equal to 0.)

(ii) Explain how computer simulation can be used to approximate the conditional distribution of T in part (i) (even if n is very large), and hence to perform a $level - \alpha$ test of H_0 based on T.

(iii) If each D_i is the difference $D_i = X_i - Y_i$ of values from two paired samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, explain how your test in part (ii)may be interpreted as a "permutation" test. Generalize your procedure to the following setting: Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be random paired samples of "objects" represented in some data space χ, and consider the null hypothesis $H - 0$ that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are IID pairs such that (X_i, Y_i) has the same (joint) distribution as (Y_i, X_i). For a test statistic $T = T(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$, how can you use simulation to determine the rejection threshold of a test of H_0 based on T?

Question 9: Simple Linear Regression (Mandatory)

Considering the table below, which pertains to a simple linear regression problem where the values of Y correspond to a specific attribute of several types of alloys at temperatures X, answer the following questions.

i	x_i	y_i
1	0.5	40
2	1.0	41
3	1.5	43
4	2.0	42
5	2.5	44
6	3.0	42
7	3.5	43
8	4.0	42

Table 5: Alloy Attribute vs Temperature

(i) Using maximum likelihood estimation, calculate the values of the parameters $\beta_0, \beta_1, \sigma^2$ and the variance related to β_0, β_1.

(ii) Determine the correlation related to the two parameters β_0, β_1.

(iii) Suppose we want to estimate the value of $\theta = 5 + \beta_0 - 4\beta_1$. Find an unbiased estimator for θ and then calculate its value and MSE.

(iv) Now, considering $\theta = 5 + \beta_0 - c\beta_1$, find an unbiased estimator for θ. For what value of c will the MSE of the corresponding estimator reach its minimum?

(v) For a new sample with $x = 3.25$, predict the output value. What is the MSE of this prediction?

(vi) For what value of x will a predicted sample of alloy reach the minimum MSE?

Question 10: Intercept/Slope Covariance in Linear Regression

Considering the following two equations for β_0 and β_1, determine the covariance of these two parameters. Using the yielded covariance, explain how and when these parameters can be independent of each other.

$$\beta_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{x}$$

Question 11: An Investigation in the Statistical Attributes of Linear Regression Equations

Suppose $\epsilon_1, \ldots, \epsilon_n \sim$ i.i.d. $N(0, \sigma^2)$, and for $i = 1, \ldots, n$ we have $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. (Capital "Y" and lower-case "x"; the former is random; the latter is constant.) This model can be written as

$$Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

or as

$$Y \sim N_n(X\beta, \sigma^2 I_n).$$

(i) Find the entries in the matrix X.

(ii) Fill in the blanks:

$$(X'X)^{-1}X'Y \sim N_?(?, ?).$$

What is $(X'X)^{-1}X'Y$ an unbiased estimator of? (X and Y are "observable.")

(iii) Use the result of (ii) to find an unbiased estimator $\hat{\beta}_1$ of the slope β_1. Write this estimator in the form

$$\hat{\beta}_1 = [\text{some row vector (specify and simplify!)}]Y.$$

Fill in the blanks:

$$\beta_1 \sim N_?(?, ?).$$

(iv) Use the result of (iii) to find a 90% confidence interval for the slope β_1, assuming (unrealistically) that σ is known. I.e., we want

$$\Pr(\text{some statistic} < \beta_1 < \text{some statistic}) = 0.9.$$

Question 12: An Exercise in Practical Regression Models

A chain of small convenience food stores performs a regression analysis to explain variation in sales volume among 16 stores. The variables in the study are as follows:

- **Sales:** Average daily sales volume of a store, in thousands of dollars
- **Size:** Floor space in thousands of square feet
- **Parking:** Number of free parking spaces adjacent to the store
- **Income:** Estimated per household income of the zip code area of the store

DF	Sum Sq	Mean Sq	F-value	Pr(>F)	
Model	3	??	9.04320188	15.16	< 0.001
Residuals	12	7.15923792	0.59660316		
Total	15	??	2.2859229		

(i) Fill in the blanks in the above table.

(ii) Compute the goodness of fit measure or the R^2.

(iii) Is there clear evidence that there is predictive value in the model using $\alpha = 0.01$?

The following model was fit to predict sales:

(iv) Write the regression equation.

(v) Interpret the coefficient of the variable **Parking**.

(vi) Find a 95% confidence interval for the intercept.

	Estimate	Std. Error	t-value	Pr(> |t|)
(Intercept)	0.872716	1.945615	0.449	0.662
Size	2.547936	1.200827	2.122	0.055
Parking	0.2202793	0.1553877	1.418	0.182
Income	0.5893221	0.1780576	3.310	0.006

Question 13: Fisher Information in the Normal Model

Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$. We know that the MLEs for μ and σ^2 are given by $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

(i) By computing the Fisher information matrix $I(\mu, \sigma^2)$, derive the approximate joint distribution of $\hat{\mu}$ and $\hat{\sigma}^2$ for large n. (Hint: Substitute $v = \sigma^2$ and treat v as the parameter rather than σ.)

(ii) Suppose it is known that $\mu = 0$. Compute the MLE $\tilde{\sigma}^2$ in the one-parameter sub-model $N(0, \sigma^2)$. The Fisher information matrix in part (i) has off-diagonal entries equal to 0 when $\mu = 0$ and n is large. What does this tell you about the standard error of $\tilde{\sigma}^2$ as compared to that of $\hat{\sigma}^2$?

Question 14: A Discrete Model - Based on Rice 8.4

Suppose that X is a discrete random variable with

$$P[X = 0] = \frac{2}{3}\theta, \quad P[X = 1] = \frac{1}{3}\theta, \quad P[X = 2] = \frac{2}{3}(1 - \theta), \quad P[X = 3] = \frac{1}{3}(1 - \theta)$$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: $3, 0, 2, 1, 3, 2, 1, 0, 2, 1$. (For parts (i) and (ii), feel free to use any asymptotic approximations you wish, even though $n = 10$ here is rather small.)

(i) Find the method of moments estimate of θ, and compute an approximate standard error of your estimate using asymptotic theory.

(ii) Find the maximum likelihood estimate of θ, and compute an approximate standard error of your estimate using asymptotic theory. (Hint: Your formula for the log-likelihood based on n observations $X_1, \ldots, X_n$ should depend on the numbers of 0's, 1's, 2's, and 3's in this sample.)

(iii) Compute, instead, an approximate standard error for your MLE in part (ii) using the nonparametric bootstrap and $B = 10000$ bootstrap simulations. (Provide both your code and the standard error estimate.)

Question 15: Step-by-step Simulation of a Non-parametric Bootstrap

At last, we've moved on from the realm of probability to statistics where we have unknown population i.e. all we know is that we have data and statistics computed from the data. Our objective is to estimate the sampling distribution of statistic.

Bootstrap is a technique that uses the given sample as the population (from which this sample is drawn) and creates a new distribution called Bootstrap Distribution. This bootstrap distribution is then used for approximating the sampling distribution of mean (or other statistics).

We draw bootstrap samples i.e. resamples of size n from the original sample with replacement and compute the desired statistic for each of the resample. The idea behind the bootstrap is that if the given sample is a representative of the population then the bootstrap distribution of the mean (or any other statistic) will resemble approximately to the sampling distribution of mean (or the statistic under consideration). This means, that "bootstrap distribution of mean will have roughly the same shape and spread as that of the sampling distribution of the mean". However, the mean of the bootstrap distribution will have the mean of the sample and not necessarily that of the population from which this sample came.

(i) **Creating a Population** For completing this section, we will first be creating a population that has a gamma distribution with parameters r=5 and $\lambda = 1/4$. Run the script below in Python to create the population we need. (Note that you can similarly create a population distribution in R by using "set.seed(1234) & rgamma(5000, 5, 1/4)"). The shape of the population distribution has been given in Fig 1.

```
import numpy as np
# here we are generating random samples from a gamma distribution
# while setting a specific seed for reproducibility purposes
np.random.seed(1234)
dist = np.random.gamma(5, 1/4, 5000)
```

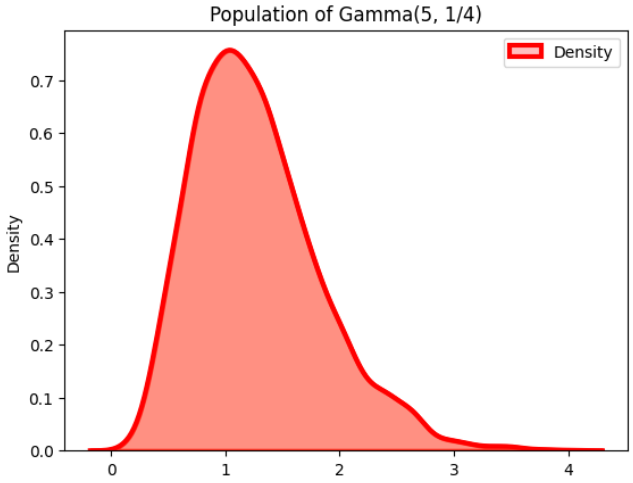

Figure 1: Gamma Population

(ii) **Simulating the Sample Distribution** First, report the mean and standard deviation of the previously created population. For a population distributed as Gamma(r,λ), the sampling distribution of sample means of that population are distributed as Gamma(nr, n*lambda), where n is the population size under consideration. For this part, you need to first compute the statistics of this sampling distribution such as its minimum and maximum, its mean, SD, value of 1st and 3rd quartiles, and its median, and thereafter, plot a histogram of the Sampling Distribution to visualize your results along with fitting a density function to it.

(iii) **Preparations for Bootstrapping** create a random sample of size 200 from the population and report a histogram of this sample population. This sample will be the base for simulating our bootstrap distribution. Just like part (ii), report all statistics of this sample population.

(iv) **Bootstrap - Resampling with Replacement** Finally, write a short script and use the method of resampling i.e. sampling with replacement, for creating a bootstrap distribution of our means. Repeat this step and create bootstrap distributions for sample sizes of 200, 50, and 10. Report all histograms and make a table comparison using the 10 and 90 percent quantiles of each bootstrap distribution, alongside the same quantiles for the original 200 size sample population of part (iii).