Statistical Method

Midterm Exam

9:00-12:00, November 21, 2023

Write down your answers to Questions 1, 2, 3, and 4 on the paper.
Prepare your analysis procedure of Question 5 in a pdf file, and submit it to Moodle.

1. (**15**%) Choose one of the following comparisons to introduce the mentioned methods, including the types of variables and the purpose of the tests. After the introduction, provide similarities and differences between two given methods. You can use examples to show the similarities and differences.

   (A) A Chi-square test and a McNemar test.

   (B) A paired t-test and a Wilcoxon signed-rank test.

   (C) A Welch's test and Mann-Whitney U-test.

   (D) A paired t-test and an independent t-test.

2. (**20**%) 60 people are randomly assigned to undertake one type of diets. There are two types of diets, and the aim of the study is to assess which diet has better performance on the weight loss. In the dataset, there are four columns, including participant IDs, types of diets, the weights before the diet, and the weights after 10-week diet. The dataset is at "Q2.csv

   (a) What are the types for the types of diets and the weights, respectively?

   (b) For the purpose, what would you choose for the response for the analysis?

   (c) Based on (b), what is the method you are going to use?

   (d) According to the purpose and (c), set up the null hypothesis and the alternative hypothesis.

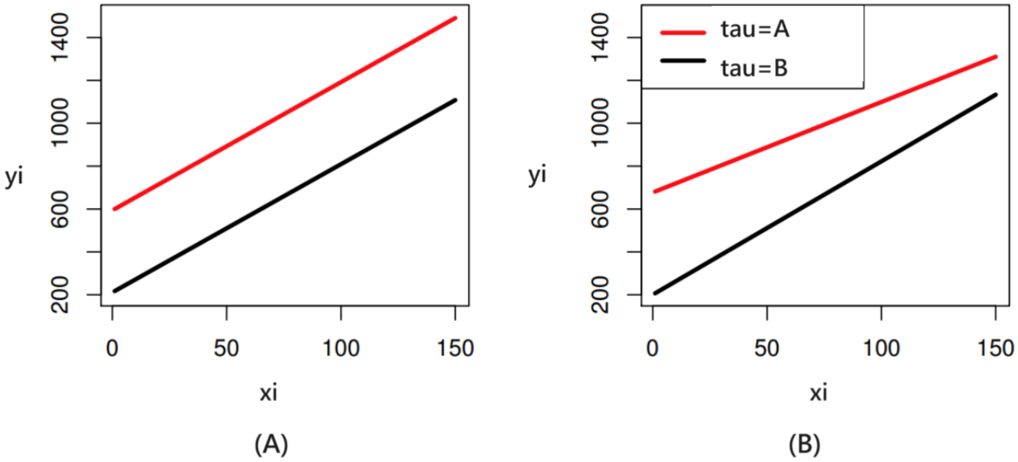
   (e) Conducting the analysis based on (c) and (d), what is the conclusion? Please provide the statistical evidence.

3. (**30**%) The analysis of covariance (ANCOVA) can be used for testing the main effect of categorical variable (τ_i) on a continuous dependent variable (y_i). The ANCOVA allow us to have a controlled continuous variable (x_i), which also correlated with the dependent variable. The control variable is called the "covariates." For a two-level categorical variable, a general model can be expressed as follows:

$$y_i = \mu + \beta_1 x_i + \beta_2 \tau_i + \epsilon_i,\ \epsilon_i \sim^{iid} N(0, \sigma^2). \tag{1}$$

(a) According to the description of Question 2, define the variables to be y_i, x_i, and τ_i.

(b) An important assumption for the ANCOVA is the homogeneity of covariate regression coefficients, which means the covariate coefficients (the slopes of the regression lines) are the same for each group formed by the categorical variable (i.e. the parallel lines). Based on the assumption, specify which figure satisfies the assumption.



(c) Try to re-express the model in Equation (1) if the assumption does not satisfy.

(d) Based on (b) and (c), set up a test if the homogeneity of covariate regression coefficient is satisfied using the dataset in Question 2.

(e) When the assumption of ANCOVA is satisfied, which coefficient is the most important in this analysis?

(f) Set up the null hypothesis and the alternative hypothesis based on (e).

(g) Conducting the ANCOVA, what is the conclusion? Please provide the statistical evidence.

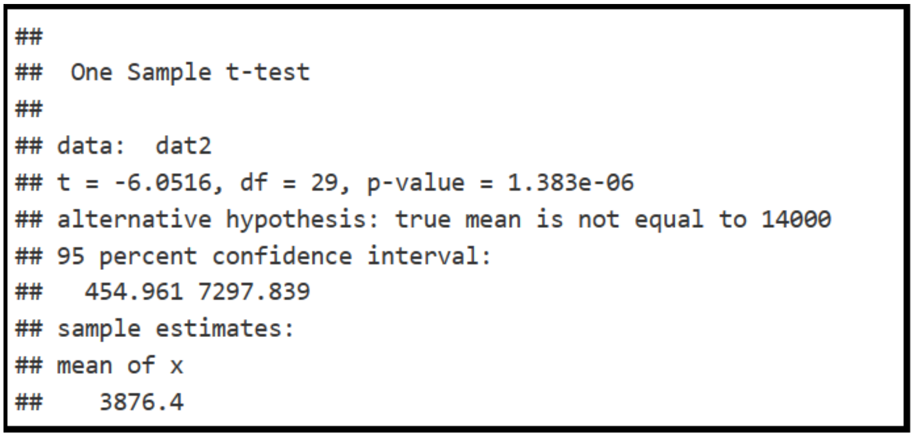
(h) Based on your results, what are the meanings of the coefficient β_1 and β_2?

(i) Is there any agreement on the conclusions between Question 2 and Question 3?

4. (**40**%) An engineer wants to know if the mean lifetime (expected lifetime) of the product is 14000 hours. Assume the random variable to be the lifetime, T_i, where $i = 1, \ldots, 30$. That is, the mean lifetime is $E(T) = \mu$. The dataset is at "Q4.csv".

(a) Before analyzing the data, what are the average and the standard deviation of the lifetime? Make a guess if the the mean lifetime is equal to 14000 or not.

(b) Let's use some possible methods. First, use the one sample t-test to see if the mean lifetime is equal to 14000 and the null hypothesis is $H_0 : \mu = 14000$. The result is shown in Figure 1. What can you conclude?



```
##
##  One Sample t-test
##
## data:  dat2
## t = -6.0516, df = 29, p-value = 1.383e-06
## alternative hypothesis: true mean is not equal to 14000
## 95 percent confidence interval:
##   454.961 7297.839
## sample estimates:
## mean of x
##    3876.4
```

Figure 1: Results of the one sample t-test.

(c) Let's check the assumption of the one sample t-test, indicating that the data are from a normal distribution. After obtaining the maximum likelihood estimates, $\hat{\mu} = 3876.4$ and $\hat{\sigma} = 9008.775$. According to the figure, what is your conclusion?

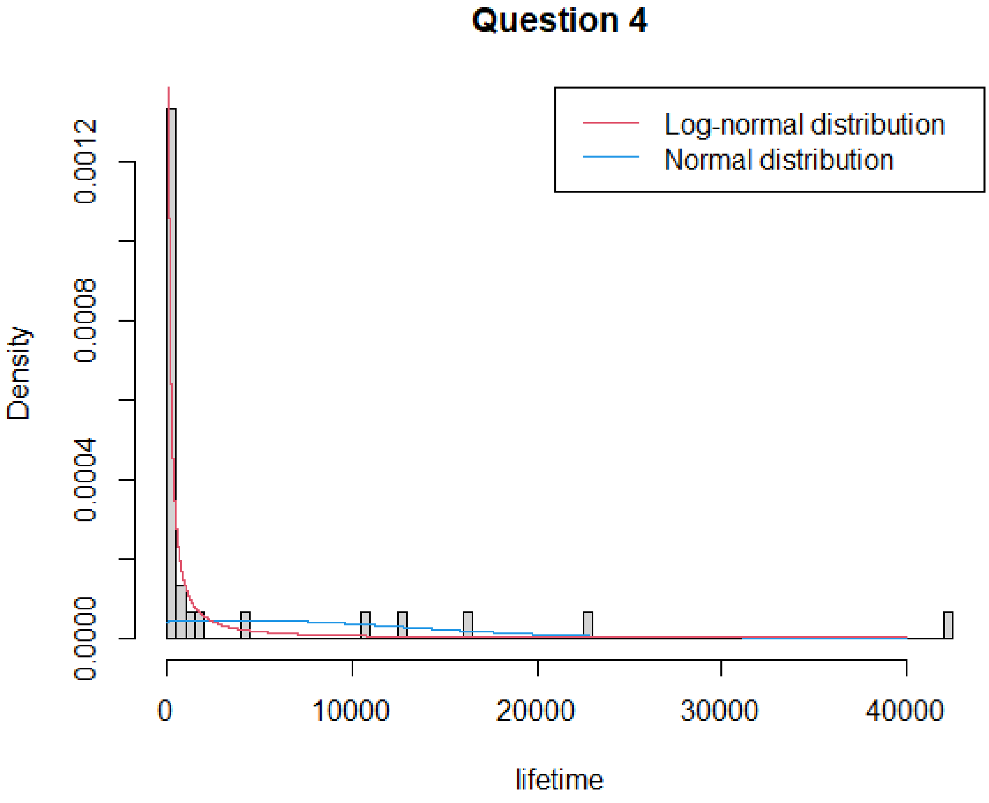


Figure 2: Histogram with the probability density functions (pdf).

(d) Let's try the other method. Based on (c), the log-normal distribution with parameters (μ, σ) can be assumed to model the lifetime data. What is the probability density function and the expected value of the log-normal distribution?

(e) Use the maximum likelihood estimates to estimate the model parameters (μ, σ). What is the estimated values of (μ, σ)?

(f) Provide the p-value of the Kolmogorov-Smirnov test if the log-normal distribution with the estimated values of (μ, σ) in (e) for the lifetime data is good enough. Give the conclusion from its p-value.

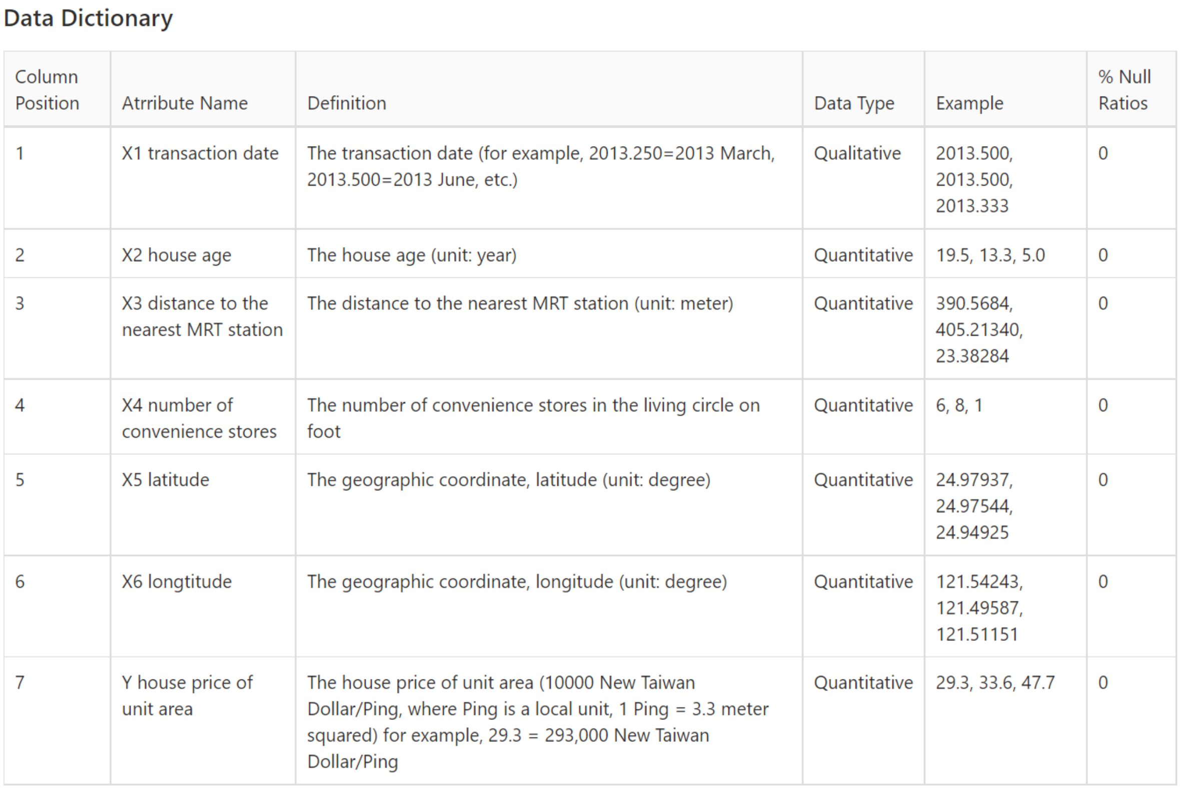
(g) Use the 95% confidence interval to answer if the expected value of lifetime is 14000. If the confidence interval of $E(T)$ is [4000, 20000], what can you conclude?

(h) Provide the values of the 95% confidence interval for the expected value of lifetime, $E(T)$ by using the bootstrap confidence interval. What can you conclude if the mean lifetime of the product is 14000 hours? Submit your bootstrap code to Moodle.

5. **(15%)**

For the Estate dataset, "Q5.csv", please use the predictors in the dataset to model the house price of unit area (Y) or the log of price $(\log(Y))$. The descriptions of the variables are show in the following:

`https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Real%20Estate%20Valuation`



Data Dictionary

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
1	X1 transaction date	The transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)	Qualitative	2013.500, 2013.500, 2013.333	0
2	X2 house age	The house age (unit: year)	Quantitative	19.5, 13.3, 5.0	0
3	X3 distance to the nearest MRT station	The distance to the nearest MRT station (unit: meter)	Quantitative	390.5684, 405.21340, 23.38284	0
4	X4 number of convenience stores	The number of convenience stores in the living circle on foot	Quantitative	6, 8, 1	0
5	X5 latitude	The geographic coordinate, latitude (unit: degree)	Quantitative	24.97937, 24.97544, 24.94925	0
6	X6 longtitude	The geographic coordinate, longitude (unit: degree)	Quantitative	121.54243, 121.49587, 121.51151	0
7	Y house price of unit area	The house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) for example, 29.3 = 293,000 New Taiwan Dollar/Ping	Quantitative	29.3, 33.6, 47.7	0

The file also includes an index, ”Year”, for the year of transaction. You can use either variable X_1 or variable ”Year” as the predictor. The response could be the original scale Y or transformation term $\log(Y)$. The goals are:

(a) What's your settings for the response? Do you treat some variables as the categorical variables?

(b) To fit a regression model, provide some useful figures to show the relationship.

(c) Try to fit a better regression model using some useful predictors. What is your fitted regression model?

(d) What are the adjusted R^2 and the significant predictors?