Statistical Method

Final Exam

9:00-12:00, November 30, 2022

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 an independent t-test.

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

   (D) An one-way ANOVA and the chi-square test.

2. (**25**%) A study was performed to test if cars get better mileage on Gas A than on Gas B. Each of 16 cars was filled with Gas A or Gas B first, decided randomly. The mileage will be recorded. Then, the mileage was recorded again for the same cars using other kind of gasoline. For example, Car 1 was filled with Gas A and then filled with Gas B. Car 2 was filled with Gas B and then filled with Gas A. The dataset of each car are shown in the following:

Car No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gas A	16	20	21	22	23	22	27	25	27	28	20	19	25	18	21	20
Gas B	19	22	24	24	25	25	26	26	28	32	26	25	23	20	25	23

Table 1: Mileage of Gas A and Gas B.

   (a) Set up the null and alternate hypotheses.

   (b) Choose an appropriate method to test (a) and provide the reason.

   (c) If the p-value of (b) is smaller than 0.05, then can we conclude that the cars get better mileage on Gas A than on Gas B? If no, how to do next?

   (d) The statement "Each of 16 cars was filled with Gas A or Gas B first, decided randomly." What is "decided randomly" used for?

   (e) If we let the study to be that "The first 8 cars was filled with Gas A first and then Gas B, and the last 8 cars was filled with Gas B first and then Gas A," should we change the statistical method? Provide the reason. If yes, please give the idea about the procedure.

3. (**15**%) A publisher is interested in determining which book cover is most popular. The manager interviews 100 volunteers in each of the three regions (West, North and East), and asks each volunteer to select his or her preference on Cover A, Cover B, and Cover C. The manager wants to know if there are regional differences in people's preferences concerning these covers. The number of preference for each cover is as follows:

Cover	West	North	East	Total
A	21	15	46	82
B	19	23	24	66
C	60	62	30	152
Total	100	100	100	300

Table 2: The count of the cover among different locations

   (a) What is the type for the response of each volunteer?

   (b) According to the description, which variable are most concerned, Region or Cover?

   (c) According to the settings, the Chi-Square test is selected, then what are the null and alternate hypotheses?

   (d) If the p-value of the Chi-square test is smaller than 0.05, what could we conclude? Can you find the evidence from Table 2?

4. (**40**%) An engineer wants to know if the mean lifetime (expected lifetime) of the product is 5. Assume the random variable to be the lifetime, T_i, where $i = 1, \ldots, 34$. 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 5 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 5 and the null hypothesis is $H_0 : \mu = 5$. The result is shown in Figure 1. What can you conclude?

```
> t.test(lifetime, mu = 5)

        One Sample t-test

data:  lifetime
t = 2.0004, df = 33, p-value = 0.05374
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.961703 9.535197
sample estimates:
mean of x
  7.24845
```

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

(c) Maybe the result in (b) contradict the guess in (a). Let's check the assumption of the one sample t-test. What is the most important assumption?

[A] the lifetime should satisfy the independent assumption,

[B] the lifetime should satisfy the normality assumption,

[C] the variance of the normal distribution should be given (known),

[D] the mean of the normal distribution should be set to 0 (not 5).

(d) Select the most suitable evidence in Figure 2 (from (A)-(D)) to show the support or the violation of the assumption in (c). Is it appropriate to use the t-test for this question?

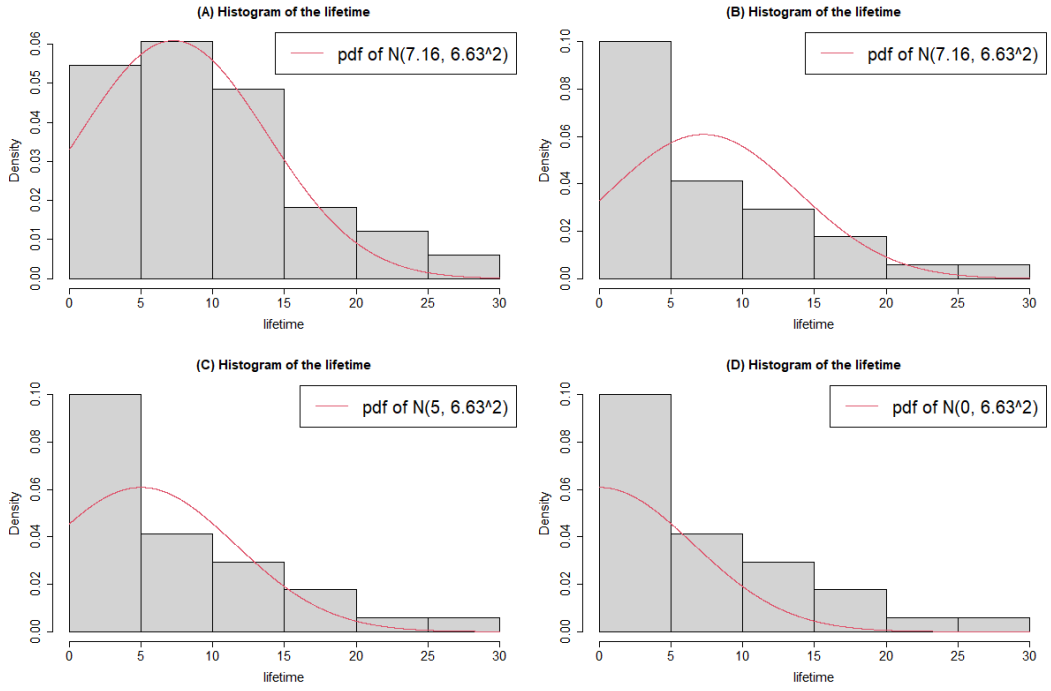

Figure 2: Histogram with the probability density functions (pdf).

(e) Let's try the other method. Assume the lifetime, T_i, is a random variable from an exponential distribution with the scale parameter θ, where $i = 1, \ldots, 34$, and the pdf is

$$f(t) = \frac{1}{\theta} \exp\left\{-t/\theta\right\}, t > 0.$$

Use Figure 3 to conclude if the expected value of lifetime is 5. That is, $E(T) = \theta = 5$. Provide your reason. (The dash line is the confidence intervals of the empirical cdf of lifetime.)

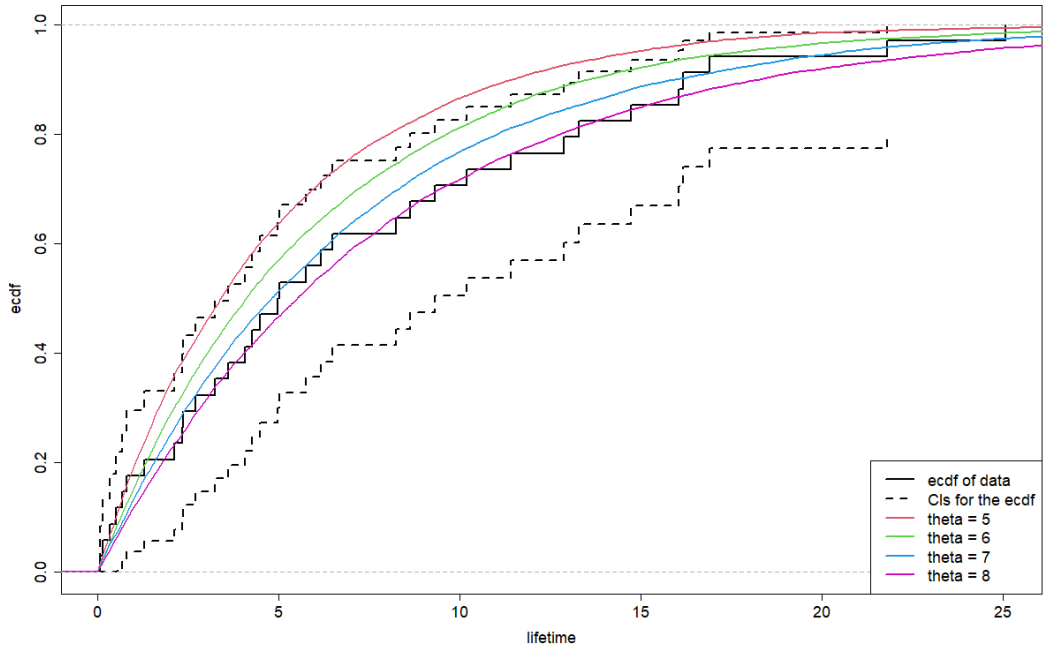

Figure 3: The empirical cdf of lifetime with the given distribution.

(f) Use the maximum likelihood estimates to estimate the model parameter θ. What is the estimated value of θ?

(g) Provide the p-value of the Kolmogorov-Smirnov test if the exponential distribution with the estimated value of θ in (f) for the lifetime data is good enough. Give the conclusion from its p-value.

(h) Use the 95% confidence interval to answer if the expected value of lifetime is 5. If the confidence interval of θ is [4.8, 7.2], what can you conclude?

(i) Provide the values of the 95% confidence interval for the expected value of lifetime, $E(T)$. (You can write down the steps of theoretical confidence interval or the bootstrap confidence interval, and then give the values.)

5. (**25**%) Choose either (A) or (B) to analyze the data by the technique of regression model.

(A) For a simple regression, there are two datasets from different groups, denoted by Group1 and Group2. The relationship of the dataset, "Q5(A).csv", is shown as Figure 4. Please use the technique of regression and dummy variables to test if the intercept or the slope is different between two groups. Give the final best model with the statistical evidence. The analysis should include the criteria of selecting models or the hypothesis testing.

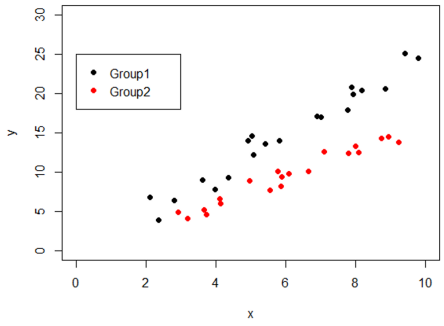

Figure 4: A simple regression

(B) For the Estate dataset, "Q5(B).csv", please use the predictors in the dataset to model the house price of unit area (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 $X1$ or variable "Year" as the predictor. The response could be the original scale Y or transformation term $\log(Y)$. The goals are:

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

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

(c) After the fitting, is there any "unusual observation" or "unusual pattern"?

(d) Provide the diagnostic figures of residuals to give the further comments and suggestions for improving the modeling in the future.