香港公開大學
THE OPEN UNIVERSITY OF HONG KONG

STAT S251F

2020 Spring Take-home Assignment (Final Examination) (UG)

Statistical Data Analysis (2020 Autumn Term)

Date to Release Take-home Assignment	Submission Deadline
11 January 2021, 14:40	13 January 2021, 14:39

Name								
Student Number								

Admissible/Inadmissible materials in this take-home assignment:

1. Calculators are allowed. (Please refer to the OUHK Approved List of Calculators.)
2. Dictionaries are NOT allowed.
3. Formula Handbook is allowed. Circles, underlines or corrections to errors in the texts of the handbook are allowed, but no marginal notes are permitted.

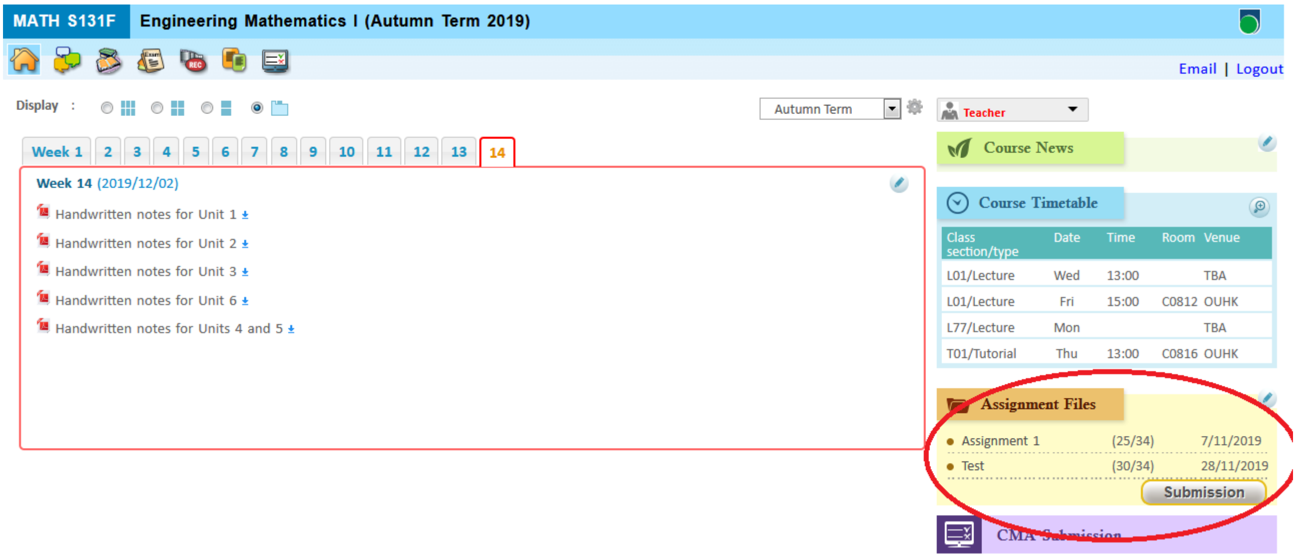
Violation of the above may lead to disqualification.

Instructions:

1. Answer this take-home assignment in English. ***All the answers must be handwritten***.
2. Read the rubric(s) in the assignment paper carefully and write your answers using ***A4 papers*** or the spaces provided in the assignment. Answers recorded elsewhere will not be marked. ***Begin each question on a new page and write the question number at the top of each page you have worked on.***
3. Rough work will not be marked.
4. Write clearly. It is not possible to award marks where the writing is very difficult to read.
5. ***Write your name and student number on the cover of the assignment paper***. ***Failure to do so will mean that your work cannot be identified***.
6. ***The take-home assignment must be finished completely by yourself***. Showing your assignment to or discussing with anyone else, including your classmate(s), is totally forbidden. In addition, ***you are not allowed to consult any external resources***, such as internet searches, materials from other classes, books or any teaching notes you have taken in other classes. ***We may check your answers to see if they are copied from the internet or other sources.*** Certainly, checking your assignment answers with any other person(s) is prohibited. ***Students involved in cheating / plagiarism will definitely receive ZERO MARKS without notice!!!***

7. After completion of your assignment, ***scan your answer script AND assignment cover page with your name and student number in PDF format and submit it to the OLE no later than the deadline specified above. Late submission will not be accepted.***

For Full-time F2F courses, you can download and submit the take-home assignment via the OLE.



8. Same as last time, the General Office has a mobile phone for receiving students' submissions of take-home assignments / exam scripts. Such an arrangement is mainly for emergency cases of computer / system crashes making students unable to submit their assignments / exam scripts through the Online Learning Environment (OLE) or by e-mail, students can still send their submissions to a mobile phone using WhatsApp or WeChat to get the submissions time-stamped. Please note the following information for communication:

Mobile phone no.	5933 7390
WeChat ID	OUHKST2019
This mobile phone is held by	Ms. Susie Liu

Such an arrangement is not a regular method of submission and should be used ***for emergency only***. Should students send in their work to the above mobile phone using WhatsApp/WeChat, the following information must be stated clearly in the message:

(1) Student number
(2) Course code
(3) Name of Course Coordinator (to whom the submission should be sent)

9. ***If a course coordinator finds that his student(s) get(s) extremely high scores which are not consistent with his/her Overall Continuous Assessment Score, then the course coordinator may invite the student(s) to return to school after the exam period to have a short viva exam to verify his/her performance.***

[END OF INSTRUCTIONS]

Question 1 (12 marks)

[Parts (a), (b) and (c) are NOT related to each other.]

(a) The 95% confidence interval for the mean height (cm) of 100 students chosen from different schools is found to be (172.3, 174.1).

(i) Find the mean $\bar{x}$ of the sample and the standard deviation σ of the population. [5]

(ii) Calculate the 99% confidence interval for the mean height. [2]

(b) The result of a test is known to be a normally distributed random variable with mean μ and standard deviation 12. It is required to have a 90% confidence interval for μ with total width less than 15. Find the least number of tests that should be carried out to achieve this. [2]

(c) A random sample of 110 measurements taken from a normal distribution with the following results: $\sum x_i = 973.5$, $\sum x_i^2 = 8780.775$. Find a 99% confidence interval for the population mean μ. [3]

Question 2 (4 marks)

(a) One hundred alumni of OUHK joined the gathering for the 2021 New Year Celebration Lucky Draw. Twenty gifts will be given to those alumni randomly chosen from the lucky-draw box. Explain briefly if this is a random sampling method. [2]

(b) One day when you go out for shopping, you are stopped by an interviewer near the entrance of a MTR station for a questionnaire opinion poll about the Hong Kong government's performance on monitoring COVID-19. Do you think that this is a random sampling method? Explain your answer briefly. [2]

Question 3 (5 marks)

The Table 3.1 below shows the ratings (5-point Likert scale) of 5 brands of brandy in a winery judged by 2 different sommeliers.

Table 3.1

Brandy Brand	1	2	3	4	5
1st Sommelier Rating	2	3	1	5	3
2nd Sommelier Rating	4	3	5	5	1

(a) Calculate the Spearman's rank correlation coefficient, r_s for the data above. Your answers must include the ranks of data. [4]

(b) Interpret the value of r_s. [1]

Question 4 (18 marks)

An auditing firm was hired to determine if a particular TV plant was overstating the value of their inventory items. It was decided that 15 items would be randomly selected. For each item, the difference between the recorded amount and the audited amount was determined. The manager of the firm interests to see if the average difference is greater than $25, in which case the plant would be subject to a loss of contract and financial penalties. The following 15 differences were obtained:

17, 35, 31, 22, 50, 42, 56, 23, 27, 38, 20, 25, 43, 45, 21

Assume that the recorded and audited amounts of inventory items are both normally distributed.

(a) Compute the mean and standard deviation of the sample data of differences. [1]

(b) Define the population mean for this problem. [1]

(c) State the null and alternative hypotheses, H_0 and H_1, for the above problem. [1]

(d) Which one of the tests: t test and z test, is appropriate for testing the above hypotheses? Explain your answer briefly. [2]

(e) Assume that α = 0.05, determine the rejection region. [1]

(f) Compute the value of the test statistic. [2]

(g) Test at 5% level of significance to see if there is any evidence to reject H_0. State your decision and conclusion clearly. [3]

(h) Construct a 95% confidence interval for the population mean time, μ. [3]

(i) Is it appropriate to employ the formula below to calculate the sample size, n, required for having 99% confidence in estimating μ to within an error, $e = \pm\$10$? Explain your answer briefly. [2]

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

(j) What conclusion can be drawn from the interval constructed in part (h) related to the rejection or non-rejection of the null hypothesis. State your reason briefly. [2]

Question 5 (12 marks)

Kong Lung Airlines recently received 30 complaints on its services. The following data represent the number of days between the receipt of the complaints and the resolution of the complaints:

14, 5, 22, 37, 31, 27, 12, 2, 13, 8,

34, 31, 26, 5, 12, 3, 5, 32, 29, 28,

29, 26, 25, 10, 14, 13, 13, 10, 14, 16

(a) Estimate the sample mean and sample standard deviation of the above data. [1]

(b) State the sampling distribution of the means. Explain your answers briefly [2]

(c) (i) Construct 94%, 96% and 98% large-sample confidence intervals for the population mean number of days between the receipt of the complaint and the resolution of the complaints, μ. [3]

(ii) What conclusion can be drawn for the reliability of confidence interval based on the lengths of the confidence intervals obtained in part (c)(i)? [2]

(d) What sample size is needed to have 90% confidence in estimating the population mean to within ± 3 days? [4]

Question 6 (20 marks)

The relationship between production cost (y) and the number of units produced (x) of a certain commodity is modelled by the following simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon, \text{ where } \varepsilon \text{ is normally distributed.}$$

A sample of data of 10 observations is listed in the Table 6.1 below:

Table 6.1

Units Produced (x)	30	20	60	80	40	50	60	30	70	60
Production Cost (y) (Million \$)	73	50	128	170	87	108	135	69	148	132

(a) Find the parameter estimates for β_0 and β_1. Hence, write down the fitted regression model. [2]

(b) Construct the ANOVA table. Show your calculations for *MSR*, *MSE* and *F*-statistic. [6]

(c) Test the linearity of the model at the 5% level of significance. State the hypotheses to be tested, critical value, decision and conclusion. [4]

(d) Compute the coefficient of determination, R^2. Interpret your result. [2]

(e) (i) Let $e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1, 2, \cdots, 10$. Find the values of $e_1, e_2, \cdots, e_{10}$. [2]

(ii) Hence, compute $\sum_{i=1}^{10} e_i$. [1]

(f) Justify if ANY fitted linear regression line, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, must pass through the point $(\bar{x}, \bar{y})$. Numerical proof is not accepted. [3]

Question 7 (13 marks)

A researcher, Dr. Brown, of the medical school, Hong Kong Island University asserts that a new drug can reduce blood cholesterol level for patients with heart disease. A total of 19 patients having heart disease were randomized to take the new drug or a placebo for one month. At the end of the period, the patients had their blood cholesterol levels (measured in suitable units), as follows:

New drug: 280, 233, 274, 283, 235, 230, 258, 241, 226
Placebo: 249, 238, 280, 247, 273, 289, 284, 256, 267, 250

Let

μ_1 = the population mean cholesterol level of the patients receiving the new drug.
μ_2 = the population mean cholesterol level of the patients receiving the placebo.

Assume that the two groups of patients are independent simple random samples and their variances are equal and unknown.

(a) Work out the mean cholesterol levels, $\overline{x_1}$ and $\overline{x_2}$, and standard deviations, s_1 and s_2 for the two groups of patients (new drug and placebo), respectively. [2]

(b) A statistical test is applied to test the equality of μ_1 and μ_2. Formulate the hypotheses, H_0 and H_1 in terms of μ_1 and μ_2 for testing Dr. Brown's assertion. [1]

(c) What is the name of the statistical test? [1]

(d) What distributional assumption is needed for the statistical test to be valid in part (c)? [1]

(e) Take α = 5%. Find the critical value of the assumed distribution. [1]

(f) Compute the value of the test statistic. [4]

(g) By using the results of parts (e) and (f), test whether the mean cholesterol level after receiving the new drug is significantly smaller than that of using placebo treatment. State your decision and conclusion clearly. [3]

Question 8 (16 marks)

Case Study

The body weights (pounds) and exercise level (0 = "none", 1 = "light", 2 = "moderate") of 15 secondary students are shown in Table 8.1 below:

Table 8.1

Student ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Exercise Level	2	1	0	1	0	0	2	0	1	1	2	2	0	1	2
Body Weight	120	100	178	162	90	163	180	165	147	142	210	139	194	200	234

Assume that the body weights are normally distributed.

Choose an appropriate statistical test from the various statistical tests of the course, STAT S251F to test if different exercise levels have an effect on body weight. Take 5% level of significance.

Your testing procedures must include the following:

(i) Name of the statistical test [3]

(ii) Hypotheses to be tested [2]

(iii) Critical value [1]

(iv) Rejection region [1]

(v) Value of the test statistic [6]

(vi) Decision to be made [2]

(vii) Conclusion to be drawn [1]

[END OF TAKE-HOME ASSIGNMENT (FINAL EXAMINATION)]