STAT S251F
Statistical Data Analysis
2020-21
Take-home Assignment 2 (Mid-term Test)

Student Name: _____

Student ID:_____Mark: _____

Points to Note:

(1) Submission Deadline: ≤ 11:59pm, 25 December 2020.

(2) Important Note: LATE submission is NOT accepted and no marks will be given.

(3) Students involved in PLAGARISM / CHEATING will receive ZERO marks WITHOUT NOTICE!!!

(4) This is a mid-term test in the form of a take-home assignment. Thus, the course coordinator is NOT ALLOWED to give assistance or hints to any students for fairness.

(5) If you think that some question wordings are NOT clear, you can seek for help from your course coordinator.

(6) Use A4 paper to complete your assignment.

(7) Your solutions to the questions MUST be HAND-WRITTEN.

(8) Scan your completed assignment in a SINGLE pdf file.

(9) Submit your pdf file via the OLE. Submission of assignment through email is NOT accepted.

(10) Give 4 decimal places when your answers cannot be expressed as integers or fractions.

Question 1 (24 marks)

[The two parts (a) and (b) are NOT related to each other.]

Part (a) (15 marks)

An athlete finds that his times for running the 100m race follow a normal distribution with mean 10.6 seconds. He trains intensively for a week and then runs 100m on each of 5 consecutive days. His times (measured in seconds) were 10.70, 10.65, 10.75, 10.80, and 10.60. After training, he wants to know if his mean running time is greater than his past mean value.

(i) Compute the sample mean and sample standard deviation of the data. [2]

(ii) State the null and alternative hypotheses, H_0 and H_1, for the above problem. [2]

(iii) Suggest an appropriate hypothesis test (Z test or t test) to test the above hypotheses. Give reason(s) to support your choice. [3]

(iv) Assume that $\alpha = 0.05$, determine the critical value and, hence, the rejection region(s). [2]

(v) Compute the value of the test statistic. [3]

(vi) Test at 5% level of significance to see if there is any evidence that the training has improved his times. State the decision and draw conclusion clearly. [3]

Part (b) (9 marks)

A social study enrolled 15 sets of ***identical twins*** to measure the effect of home environment on some social attitudes. There are 2 different environments: minority and home. Every twin from each set of twins was randomly assigned to one of them. The twin assigned to the minority environment had to stay with an African American family for one year. An attitude survey was conducted after the end of one year. The researcher of the study wants to know if living in the minority environment can lead to higher mean score of social attitudes. The data are shown in the table below:

Twin ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Home Environment	65	67	75	77	69	65	73	78	70	72	73	79	68	73	71
Minor Environment	83	75	72	76	78	80	72	81	70	78	77	71	87	70	75

(i) Suggest an appropriate test of hypothesis. [1]

(ii) State the null and alternative hypotheses. [1]

(iii) Determine the critical value. [1]

(iv) Test if living in the minority environment can lead to higher mean score of social attitudes at the 1% level of significance. [6]

Question 2 (33 marks)

The Census & Statistics Department, HKSAR uses sample surveys to produce important economic estimates. One pilot study estimated rice prices in July and September using independent samples from rice retailers in the two months. The summary statistics, in HK$/kg, are shown below:

Month	Sample Size, n	Sample Mean, $\bar{x}$ (HK$)	Sample SD, s (HK$)	Population Variance
July	27	7.43	0.24	σ_1^2
September	25	7.16	0.27	σ_2^2

The researcher intends to test if the mean prices in July and September are not the same.

(a) Set up the null and alternative hypotheses, H_0 and H_1. [2]

(b) Suggest an appropriate statistical test to test the hypotheses set in part (a). Give reason(s) to support your choice. [3]

(c) Take $\alpha = 5\%$ and assume that both σ_1^2 and σ_2^2 are <u>equal</u>.

(i) Determine the critical value and the rejection rejection(s). [4]

(ii) Compute the value of the test statistic. [8]

(iii) Test at 5% level if the mean prices in July and September are not the same. State the decision and draw the conclusion clearly. [3]

(d) Take $\alpha = 5\%$ and assume that both σ_1^2 and σ_2^2 are <u>unequal</u>.

(i) Determine the critical value and the rejection rejection(s). [4]

(ii) Compute the value of the test statistic. [6]

(iii) Test at 5% level if the mean prices in July and September are not the same. State the decision and draw conclusion clearly. [3]

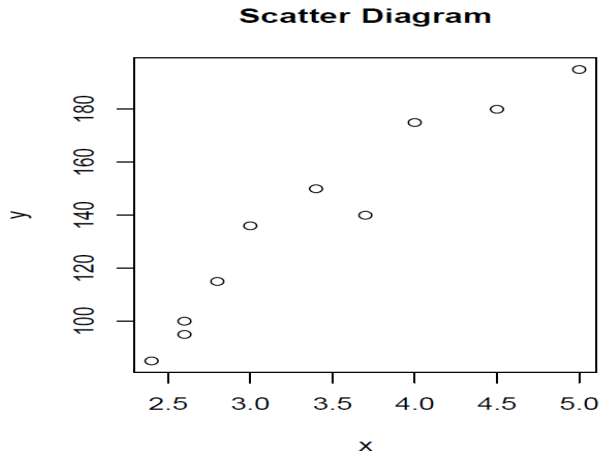
Question 3 (43 marks)

A firm administers a test to sales trainees before they go into the field. The management of the firm is interested in determining the relationship between the test scores and the sales made by the trainees at the end of one year in the field. The following data were collected for 10 sales who have been in the field for one year.

Test Score	2.6	3.7	2.4	4.5	2.6	5.0	2.8	3.0	4.0	3.4
No. of Units Sold	95	140	85	180	100	195	115	136	175	150

(a) State the independent (x) and dependent (y) variables of this question. [2]

(b) The scatter diagram of the above data is plotted below. State what can be seen from the diagram ***with reason(s).*** [3]



(c) Calculate the Pearson sample correlation coefficient, r. Comment on your computed value. [3]

(d) (i) Find the sample estimates, a and b, for the population regression coefficients β_0 and β_1. [2]

(ii) Hence, write down the fitted regression equation. [1]

(e) (i) Use the fitted equation obtained in part (dii) to predict the numbers of units that would be sold by two trainees who received test scores of 3.2 and 6.0, respectively. ***Give your answer in whole number.*** [2]

(ii) Are your fitted/predicted values obtained in part (ei) reliable? Explain your answers briefly. [5]

(f) Construct a ANOVA table for the above data by filling in the 9 numbers of an ANOVA table. ***Show your calculations for MSR, MSE and F_0 clearly.*** [9]

SV	*DF*	*SS*	*MS*	*F Test Statistic*
Regression	*df(R)*	*SSR*	*MSR*	F_0
Error	*df(E)*	*SSE*	*MSE*	
Total	*df(T)*	*SST*		

(Go onto the next page)

(g) Based on the ANOVA table obtained part (f), test for the linearity of the regression model at the 5% level of significance. State the following for your testing procedures:

(i)	H_0 and H_1	[2]
(ii)	Critical value of F distribution	[1]
(iii)	Decision rule	[1]
(iv)	Value of test statistic	[1]
(v)	Decision	[2]
(vi)	Conclusion	[1]

(h) Calculate the coefficient of determination, R^2. Interpret your result briefly. [4]

(j) (i) In accordance with the R computer output below, justify if the fitted regression model obtained in part (d)(ii) above is appropriate ***with reason(s)***. [2]

(ii) How to adjust the model if it is really inappropriate. [2]

```
Coefficients:
            Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)     ?          14.486    -0.319       0.758
x               ?           4.137    10.076    8.02e-06
```

The 2 question marks "?" have already been asked in part (d)(ii).

Note: This question does not have part (i).