Statistics

I. (30 points) Miscellaneous problems. Briefly answer the following questions.

(1). Starting at a fixed time, each car entering (independently) the intersection of Victory St. and University Rd. is observed to see whether it turns left, right or goes straight ahead. The experiment terminates as soon as a car is observed to turn left. Let X be the number of cars observed. What is the distribution of X. Why?

(2).**Choose the most appropriate one.** A new teaching method is tried on a large class. At the end of the course, a final is given. The average score is *73* out of *100* and only *3%* of the class failed.

(a). This shows the new method was successful.
(b). This shows nothing because there was no control group.
(c). This shows nothing because there was no randomization.
(d). This shows nothing because it was not double-blind.

(3). Fill in the missing values in the table below and calculate the correlation coefficient.

X	*Y*	*X in standard units*	*Y in standard units*	*Product*
1	*8*	*-1.414*	*1.265*	*-1.789*
2	*7*	•	•	•
3	*4*	*0*	*0*	*0*
4	*1*	*0.707*	*-0.949*	*-0.671*
5	*0*	*1.414*	*-1.265*	*-1.789*

(4). A time-distance data collected by the technicians at Gait Laboratory of National Cheng Kung University Hospital is to be analyzed. The researcher thought that no intercept simple linear regression model is appropriate to the data. After a SAS run, she found that the sum of the residuals $\Sigma_i e_i$ is not zero, but $\Sigma_i x_i e_i = 0$, where x is the independent variable. She was confused. Suppose you are a consultant, explain to her why we must have the result.

(5). It is suspicious that the value $Y = 50$ (evaluated at $x = 3$) may be an outlier for a data set. Suppose a simple linear regression model is used to fit the data without *(50,3)*. The fitted line is $\hat{Y} = 2 + 10x$, and $MSE = 4$. Can you claim that the point *(50,3)* is an outlier if $\alpha = 0.01$ is used? Why?

II. (40 points) A manufacture suspects that the batches of raw material furnished by her supplier differ significantly in calcium content. There is a large number of batches currently in the warehouse. Five of these are randomly selected for study. A chemist makes five determinations on each batch and obtains the following data:

Batch 1	*Batch 2*	*Batch 3*	*Batch 4*	*Batch 5*
23.46	*23.59*	*23.51*	*23.28*	*23.29*
23.48	*23.46*	*23.64*	*23.40*	*23.46*
23.56	*23.42*	*23.46*	*23.37*	*23.37*
23.39	*23.49*	*23.52*	*23.46*	*23.32*
23.40	*23.50*	*23.49*	*23.39*	*23.38*

(1). (25 points) Suppose the chemist is interesting in whether or not there is a significant variation in calcium from batch to batch and, from the above data, she got the following summary statistics:

$\Sigma\Sigma Y_{ij}^2 = 13740.2441$, $\Sigma\Sigma Y_{ij} = 586.09$, $\Sigma_j Y_{1j} = 117.29$,
$\Sigma_j Y_{2j} = 117.46$, $\Sigma_j Y_{3j} = 117.62$, $\Sigma_j Y_{4j} = 116.90$,
$\Sigma_j Y_{5j} = 116.82$, $\Sigma_i(\Sigma_j Y_{ij})^2/5 = 13740.1565$

Perform the test for the chemist at level $\alpha = 0.05$. You should state (a) the model with suitable assumptions; (b) why the model is chosen; (c) the null and alternative hypothesis; (d) ANOVA table; and (e) the test statistic and your conclusion. (Note: $F(0.05;4,20) = 2.87$, $F(0.05;5,20) = 2.71$, $F(0.05;4,21) = 2.84$, $F(0.05;5,21) = 2.68$, $F(0.05;4,16) = 3.01$), $F(0.025;4,20) = 3.51$, $F(0.025;20,4) = 8.56$.

(2). (15 points) Estimate the components of variance and give a 95% confidence interval for the ratio = variance of treatment/(variance of treatment + variance of pure error). Comment on the interval you obtained.

III. (50 points) A researcher in AVRDC (Asian Vegetable Research and Development Center) wants to know the relationship between DEF (DEFoliation in %) and DAP (Days After Planting) of soybean in Fall season. He collects data on DEF and DAP and performs preliminary analysis by using simple linear regression model. The following results are obtained from the output of a SAS run:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F value	P-value
Model	1	( )	( )	( )	0.0001
Error	( )	( )	( )		
C. Total	57	4529.89603			

Root MSE	3.92827	R-square	0.8092
Dep Mean	11.94172		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H_o: Parameter = 0	Prob > |T|
Intercept	1	-3.29981	1.14133	( )	0.0018
DAP	1	3.34161	0.24605	( )	0.0001

(1). (5 points) From the above results, can we say that Days After Planting really has something to do with the Defoliation?

(2). (30 points) From the scatterplot (not shown here), the researcher notices that the deliforation at $DAP = 7$ seems to depart from the trend of the previous six DAPs. The researcher wants to know whether the slope between DAPs 5 and 6 and the slope between DAPs 6 and 7 are the same. One way to see it is to test the hypothesis H_o: $\mu_5 - \mu_6 = \mu_6 - \mu_7$, where μ_i is the expected value of the defoliation observed at $DAP = i$, $i = 5,6,7$. The following are the defoliation data obtained at $DAP = 5,6,7$, respectively.

DAP	Defoliation (in %)
5	12.10, 11.10, 13.30, 11.10, 12.10, 12.10, 13.10, 13.30
6	12.10, 14.10, 13.10, 13.30, 15.20, 12.10, 14.10, 14.40
7	28.30, 24.20, 27.80, 27.80, 31.30, 30.30, 31.30, 31.30

Assuming that the variances in defoliation are the same for $DAP = 5,6$ and 7, perform the above test at level $\alpha = 0.05$. You should list the appropriate assumptions so that your test is suitable. Note: $F(0.05;1,21) = 4.32$, $F(0.05;1,22) = 4.30$, $F(0.05;1,23) = 4.28$, $F(0.05;1,24) = 4.26$, $F(0.05;3,21) = 3.07$, $F(0.05;3,24) = 3.01$.

(3). (15 points) As a consequence of the test in (2), suppose, from the scatterplot; the following piecewise linear regression model is appropriate for the data

$$Y = \beta_o + \beta_1 DAP + \beta_2 (DAP - 6)_+ + \varepsilon, \qquad (*)$$

where $(DAP - 6)_+ = 0$ if $DAP \leq 6$, and $= DAP - 6$ if $DAP > 6$. After fitting the data to model in (*), we further, besides those in (1), have $SSE(DAP, (DAP - 6)_+) = 309.048813$. If the researcher wants to know whether the piecewise linear regression model is suitable as compared to the simple linear regression model, perform a test for her at $\alpha = 0.05$. Note: $F(0.05,1,40) = 4.08$, $F(0.05,1,60) = 4.00$, $F(0.05,2,40) = 3.23$, $F(0.05,2,60) = 3.15$, $F(0.05,3,40) = 2.84$, $F(0.05,3,60) = 2.76$.

IV. (30 points) Mr. Chen owns a fleet of 12 taxicabs. He plans to buy 6 Firestone tires and 6 Goodyear tires to put on the rear wheels of his cabs. The tires will be checked at 500-mile intervals and will be considered to be worn out when the tread wear bars show. He can either (A) put one new tire on each of the 12 cabs or (B) put one of each brand on the rear tires of 6 cabs.

(1). (15 points) From a statistical viewpoint, which would be the preferred procedure? Why? (Hint: This discussion should be based on the nature of the problem, the number of degrees of freedom under both plans, the magnitude of σ^2.) For this part only we assume $\sigma_1^2 = \sigma_2^2 = (3500)^2$.

(2). (5 points) If Mr. Chen followed plan A and observed the following data, can he conclude that one brand is better than the other? (See note after (3)). You can directly make necessry asumption(s) to perform the test.

Mileage on Goodyear tires:	32,000	31,500	38,500	37,000	40,000	37,000
Mileage on Firestone tires:	34,500	31,000	39,000	41,000	38,000	38,500

(3). (5 points) If Mr. Chen followed Plan (B) and if he observed the following mileage data, can he conclude that one brand is better than the other?

	Cabs					
	1	*2*	*3*	*4*	*5*	*6*
Goodyear	32,000	31,500	38,500	37,000	40,000	37,000
Firestone	34,500	31,000	39,000	38,500	41,000	38,000

Note: Make sure in parts (2) and (3) that you state the necessary steps for hypothesis testing. Use $\alpha = 0.10$ in both parts. $F(0.10;1,5) = 16.26$, $F(0.10;1,6) = 13.75$, $F(0.10;1,10) = 10.04$, $F(0.10;1,12) = 9.33$.

(4). (5 points) Compare briefly your conclusions in (2) and (3).