Data Analytics (CS40003)

Practice Set VIII
(Topic: Analysis of Variance)

One-way ANOVA

P1:
Four treatments for fever blisters, including a placebo (A), were randomly assigned to 20 patients. The data below show, for each treatment, the numbers of days from initial appearance of the blisters until healing is complete.

Treatments	Number of days				
A	5	8	7	7	8
B	4	6	6	3	5
C	6	4	4	5	4
D	7	4	6	6	5

Test the hypothesis, at 5% significance level, that there is no difference between the four treatments with respect to mean time of healing.

P2:
The following data give the lifetimes, in hours, of three types of battery.

Type	I.	20.1	49.9	49.8	49.7	50.0
	II.	51.0	50.8	50.9	50.9	50.6
	III.	49.5	50.1	50.2	49.8	49.3

Analyze these data for a difference between mean lifetimes. (Use a 5% significance level.)

P3:
Three different brands of magnetron tubes (the key component in microwave ovens) were subjected to stress testing. The number of hours each operated before needing repair was recorded.

Brand	A	36	48	5	67	53
	B	49	33	60	2	55
	C	71	31	140	59	224

Although these times may not represent lifetimes, they do indicate how well the tubes can withstand stress.

Use a one way analysis of variance procedure to test the hypothesis that the mean lifetime under stress is the same for the three brands.

What assumptions are necessary for the validity of this test? Is there reason to doubt these assumptions for the given data?

P4:
Three special ovens in a metal working shop are used to heat metal specimens. All the ovens are supposed to operate at the same temperature. Itis known that the temperature of an oven varies, and it is suspected that there are significant mean temperature differences between ovens. The table below shows the temperatures, in degrees centigrade, of each of the three ovens on a random sample of heatings.

Oven	**Temperature (^{0}C)**					
1	494	497	481	496	487	
2	489	474	479	478		
3	489	483	487	472	472	477

Stating any necessary assumptions, test for a difference between mean oven temperatures.

Estimate the values of μ (1 value), L_i (3 values) and ε_{ij} (15 values) for the model

(Temperature) $_{ij} = x_{ij} = \mu + L_i + \varepsilon_{ij}$. Comment on what they reveal.

P5:
Eastside Health Authority has a policy whereby any patient admitted to a hospital with a suspected coronary heart attack is automatically placed in the intensive care unit. The table below gives the number of hours spent in intensive care by such patients at five hospitals in the area.

Hospital				
A	**B**	**C**	**D**	**E**
30	42	65	67	70
25	57	46	58	63
12	47	55	81	80
23	30	27		
16				

Use a one factor analysis of variance to test, at the 1% level of significance, for differences between hospitals.

P6:
An experiment was conducted to study the effects of various diets on pigs. A total of 24 similar pigs were selected and randomly allocated to one of the five groups such that the control group, which was fed a normal diet, had 8 pigs and each of the other groups, to which the new diets were given, had 4 pigs each. After a fixed time the gains in mass, in kilograms, of the pigs were measured. Unfortunately by this time two pigs had died, one which was on diet A and one which was on diet C. The gains in mass of the remaining pigs are recorded below.

Diet	**Gain in mass (kg)**			
Normal	23.1 14.6	9.8 11.2	15.5 15.7	22.6 10.5
A	21.9	13.2	19.7	
B	16.5	22.8	18.3	21.0
C	30.9	21.9	29.8	
D	21.0	25.4	21.5	21.2

Use a one factor analysis of variance to test, at the 5% significance level, for a difference between diets.

What further information would you require about the dead pigs and how might this affect the conclusions of your analysis?

Two-way ANOVA

P1:
Prior to submitting a quotation for a construction project, companies prepare a detailed analysis of the estimated labour and materials costs required to complete the project. A company which employs three project cost assessors, wished to compare the mean values of these assessors' cost estimates. This was done by requiring each assessor to estimate independently the costs of the same four construction projects. These costs, in £0000s, are shown in the next column

	Assessor		
	A	B	C
Project1	46	49	44
Project2	62	63	59
Project3	50	54	54
Project4	66	68	63

Perform a two factor analysis of variance on these data to test, at the 5% significance level, that there is no difference between the assessors' mean cost estimates.

P2:
In an experiment to investigate the warping of copper plates, the two factors studied were the temperature and the copper content of the plates. The response variable was a measure of the amount of warping. The resultant data are as follows.

Temp (^{o}C)	**Copper content (%)**			
	40	60	80	100

50	17	19	23	29
75	12	15	18	27
100	14	19	22	30
125	17	20	22	30

Stating all necessary assumptions, analyze for significant effects.

P3:
In a study to compare the body sizes of silkworms, three genotypes were of interest: heterozygous (HET), homozygous (HOM) and wild (WLD).

The length, in millimeters, of a separately reared cocoon of each genotype was measured at each of five randomly chosen sites with the following results.

		Site				
Silkworm		1	2	3	4	5
	HOM	29.87	28.24	32.27	29.85	
	WLD	35.76	34.14	36.54	34.95	36.11
	HET	32.51	30.82	34.46	34.01	32.99

Identify the blocking factor. Has it proved useful? Explain.

Test, at the 1% significance level, for a difference in mean lengths between genotypes.

P4:
Four different washing solutions were being compared to study their effectiveness in retarding bacteria growth in milk containers. The study was undertaken in a laboratory, and only four trials could be run on any one day. As days could represent a potential source of variability, the experimenter decided to use days as a blocking variable. Observations were made for five days with the following (coded) results.

Day					
Solution	1	2	3	4	5
A	12	21	17	38	29
B	15	23	16	43	35
C	6	11	7	32	28
D	18	27	23	43	35

Stating any necessary assumptions, analyze for significant differences between solutions.

Was the experimenter wise to use days as a blocking factor? Justify your answer.

P5:
The Marathon of the South West took place in Bristol in April 1982. The table below gives the times taken, in hours, by twelve competitors to complete the course, together with their type of occupation and training method used.

Types of occupation			
Training Method	**Official Worker**	**Manual Worker**	**Professional Sportsperson**
A	5.7	2.9	3.6
B	4.5	4.8	2.4
C	3.9	3.3	2.6
D	6.1	5.1	2.7

Carry out an analysis of variance and test, at the 5% level of significance, for differences between types of occupation and between training methods.

The age and sex of each of the above competitors are subsequently made available to you. Is this information likely to affect your conclusions and why?

Q6:
Information about the current state of a complex industrial process is displayed on a control panel which is monitored by a technician. In order to find the best display for the instruments on the control panel, three different arrangements were tested by simulating an emergency and observing the reaction times of five different technicians. The results, in seconds, are given below.

	Technician				
Arrangement	P	Q	R	S	T
A	2.4	3.3	1.9	3.6	2.7
B	3.7	3.2	2.7	3.9	4.4
C	4.2	4.6	3.9	3.8	4.5

Carry out an analysis of variance and test for differences between technicians and between arrangements at the 5% significance level.
Currently arrangement C is used and it is suggested that this be replaced by suggested that this be replaced by arrangement A. Comment, briefly, on this suggestion and on what further information you would find useful before coming to a definite decision.

Miscellaneous Exercises

Q1:
After completing a six month typing course with the Speedy fingers Institute, four people, A, B, C and D, had their typing speed measured, in words per minute, on each of five kinds of work. The results are given in the table below.

	Legal	**Business**	**Numeric**	**Prose I**	**Prose II**
A	40	47	42	45	53
B	34	32	14	36	44
C	33	40	31	48	44
D	24	26	25	27	45

Carry out an analysis of variance and test, at the 5% level of significance, for differences between the people and between kinds of work.

Subsequently it transpired that A and C used electric typewriters, whilst B and D used manual typewriters. Does this information affect your conclusions and why?

Q2:
A batch of bricks was randomly divided into three parts and each part was stored by a different method. After one week the percentage water content of a number of bricks stored by each method was measured.

Method of storage	% water content					
1	7.4	8.5	7.1	6.2	7.8	
2	5.5	7.1	5.6			
3	4.8	5.1	6.2	4.9	6.1	7.1

Making any necessary assumptions, use a one factor analysis of variance to test, at the 5% significance level, for differences between methods of storage.

If low water content is desirable, state which method of storage you would recommend, and calculate a 95% confidence interval for its mean percentage water content after one week. [You may assume that the estimated variance of a sample mean is given by (Within samples mean square) ÷ (sample size).]

Q3:
A textile factory produces a silicone proofed nylon cloth for making into rainwear. The chief chemist thought that a silicon solution of about 12% strength would yield a cloth with a maximum waterproofing index. It was also suspected that there might be some batch to batch variation because of slight differences in the cloth.
To test this, five different strengths of solution were tested on each of three different batches of cloth. The following values of the waterproofing index were obtained.

Strength of silicone solution (%)					
Cloth	6	9	12	15	18
A	20.8	20.6	22.0	22.6	20.9

B	19.4	21.2	21.8	23.9	22.4
C	19.9	21.1	22.7	22.7	22.1

[You may assume that the total sum of squares of the observations ($\sum x^2$) = 7022. 79.]

Carry out an analysis of variance to test, at the 5% significance level, for differences between strengths of silicone solution and between cloths.

Comment on the chief chemist's original beliefs in the light of these results and suggest what actions the chief chemist might take. (AEB)

Q4.

(a): A catering firm wishes to buy a meat tenderizer, but was concerned with the effect on the weight loss of meat during cooking. The following results were obtained for the weight loss of steaks of the same pre-cooked weight when three different tenderizers were used.

		Weight loss in grams				
Tenderizer	A	36	28	42	58	
	B	17	59	33		
	C	36	74	29	55	48

Carry out a one factor analysis of variance and test at the 5% significance level whether there is a difference in weight loss between tenderizers.

(b) Time and temperature are important factors in the weight loss during cooking. As these had not been taken account of during the first trial, a further set of results was obtained where all the steaks were cooked at the same temperature and cooking times of 20, 25 and 30 minutes were used. An analysis of these data led to the following table.

Source of variation	**Sum of squares**	**Degrees of freedom**
Between tenderizers	321	2
Between times	697	2
Error	85	4
Total	1103	8

Test at the 5% significance level for differences between tenderizers and between times.

(c) Contrast the results obtained in (a) and (b) and comment on why the two sets of data can lead to different conclusions. (AEB)

Q5:
A commuter in a large city can travel to work by car, bicycle or bus. She times four journeys by each method with the following results, in minutes.

Car	Bicycle	Bus
27	34	26
45	38	41
33	43	35
31	42	46

(a) Carry out an analysis of variance and test at the 5% significance level whether there are differences in the mean journey times for the three methods of transport.

(b) The time of day at which she travels to work varies. Bearing in mind that this is likely to affect the time taken for the journey, suggest a better design for her experiment and explain briefly why you believe it to be better.

(c) Suggest a factor other than leaving time which is likely to affect the journey time and two factors other than journey time which might be considered when choosing a method of transport. (AEB)

Q6.
(a): As part of a project to improve the steer ability of trucks, a manufacturer took three trucks of the same model and fitted them with soft, standard and hard front springs, respectively. The turning radius (the radius of the circle in which the truck could turn full circle) was measured for each truck using a variety of drivers, speeds and surface conditions. Use the following information to test for a difference between springs at the 5% significance level.

Source	Sum of squares	Degree of freedom
Between springs	37.9	2
Within springs	75.6	18
Total	113.5	20

(b) A statistician suggested that the experiment would be improved if the same truck was used all the time with the front springs changed as necessary and if the speed of the truck was controlled.

The following results for turning circle, in metres, were obtained.

	Springs		
Speed	**Soft**	**Standard**	**Hard**
15 km/h	42	43	39
25 km/h	48	50	48

Carry out a two factor analysis of variance and test at the 5% significance level for differences between springs and between speeds. [You may assume that the total sum of squares about the mean (SS_T) is 92.]

(c) Compare the two experiments and suggest a further improvement to the design. (AEB)

Q7:
A drug is produced by a fermentation process. An experiment was run to compare three similar chemical salts, X, Y and Z, in the production of the drug. Since there were only three of each of four types of fermenter A, B, C and D available for use in the production, three fermentations were started in each type of fermenter, one containing salt X, another salt Y and the third salt Z. After several days, samples were taken from each fermenter and analyzed. The results, in coded form, were as follows.

Fermenter Type			
A	**B**	**C**	**D**
X 67	Y 73	X 72	Z 70

Z 68	Z 65	Y 80	X 68
Y 78	X 69	Z 73	Y 69

State the type of experimental design used.

Test, at the 5% level of significance, the hypothesis that the type of salt does not affect the fermentation.

Comment on what assumption you have made about the interaction between type of fermenter and type of salt.

Q8:

A factory is to introduce a new product which will be assembled from a number of components. Three different designs are considered and four employees are asked to compare their speed of assembly. The trial is carried out one morning and each of the four employees assembled design A from 8.30 am to 9.30 am, design B from 10.00 am to 11.00 am and design C from 11.30 am to 12.30 pm. The number of products completed by each of the employees is shown in the following table.

	Employee			
Design	1	2	3	4
A	17	4	38	8
B	21	6	52	20
C	28	9	64	22

(a) Carry out a two factor analysis of variance and test at the 5% significance level for differences between designs and between employees. [You may assume that the total sum of squares about the mean (SS_T) is 3878.9.]

(b) Comment on the fact that all employees assembled the designs in the same order. Suggest a better way of carrying out the experiment.

(c) The two factor analysis assumes that the effects of design and employee may be added. Comment on the suitability of this model for these data and suggest a possible improvement. (AEB)

Q9:
In a hot, third world country, milk is brought to the capital city from surrounding farms in churns carried on open lorries. The keeping quality of the milk is causing concern. The lorries could be covered to provide shade for the churns relatively cheaply or refrigerated lorries could be used but these are very expensive. The different methods were tried and the keeping quality measured. (The keeping quality is measured by testing the pH at frequent intervals and recording the time taken for the pH to fall by 0.5. A high value of this time is desirable.)

Transport method	**Keeping quality (hours)**					
Open	16.5	20.0	14.5	13.0		
Covered	23.5	25.0	30.0	33.5	26.0	
Refrigerated	29.0	34.0	26.0	22.5	29.5	30.5

(a) Carry out a one factor analysis of variance and test, at the 5% level, for differences between methods of transport.

(b) Examine the method means and comment on their implications.

(c) Different farms have different breeds of cattle and different journey times to the capital, both of which could have affected the results. How could the data have been collected and analyzed to allow for these differences?

Q10:
A hospital doctor wished to compare the effectiveness of 4 brands of painkiller A, B, C and D. She arranged that when patients on a surgical ward requested painkillers they would be asked if their pain was mild, severe or very severe. The first patient who said mild would be given brand A, the second who said

mild would be given brand B, the third brand C and the fourth brand D. Painkillers would be allocated in the same way to the first four patients who said their pain was severe and to the first four patients who said their pain was very severe. The patients were then asked to record the time, in minutes, for which the painkillers were effective.

The following data were collected.

Brand	A	B	C	D
Mild	165	214	173	155
Severe	193	292	142	211
Very Severe	107	110	193	212

(a) Carry out a two factor analysis of variance and test at the 5% significance level for differences between brands and between symptoms. You may assume that the total sum of squares (SS_T) = 28590.92

(b) Criticize the experiment and suggest improvements.