1. Listed below are the coded values of serum total cholesterol concentration for a random sample of 23 baboons each being 24 months old, raised on a high cholesterol, saturated fat diet.

141	135	127	200	184	122	219	114	136	253
243	188	239	135	165	140	186	134	110	103
144	252	169							

(Minimum = 103, maximum = 253.)

(a) Construct a stem-and-leaf diagram of these data and comment on the shape of the distribution.

[7 marks]

(b) Obtain the average serum total cholesterol concentration, $\bar{x}$, for the data.

[1 mark]

(c) For a random sample of n observation from a Normal distribution with mean μ and variance σ^2, derive the mean and variance of the sample mean, $\bar{X}$.

[5 marks]

(d) It is known from previous studies that the coded serum total cholesterol concentration in the population of all 24-month-old baboon may be modelled by a Normal distribution with mean 165 and standard deviation 20.

(i) Give a 95% confidence interval for mean serum total cholesterol concentration. [3 marks]

(ii) In the light of your result, comment on whether there appears to be evidence to suppose that the high cholesterol diet raises the serum total cholesterol concentration in the 24-month-old baboons. [2 marks]

(iii) it is desired to reduce the width of the confidence interval for the mean serum total cholesterol concentration. How many 24-month-old baboons would need to be in the sample if the 95% confidence interval was required to have a width of 4? [2 marks]

2. Investigators tested the effect of a drug with antiarrhythmic properties on patients with frequent premature ventricular contractions (PVCs). For each of 12 patients, the researchers recorded the number of PVCs during a 1-minute electrocardiograph, both before and after treatment with the drug.

Patient	1	2	3	4	5	6	7	8	9	10	11	12
Pre-treatment	6	9	17	22	7	5	5	14	9	7	9	51
Post-treatment	5	2	0	0	2	1	0	0	0	0	13	0

(a) Obtain a comparative Box plot of the data and comment on your results.
[8 marks]

(b) To test whether the drug has an effect on PVCs, the differences between treatments, Pre – Post, and their mean and standard deviation are given below for the 12 patients:

Patient	1	2	3	4	5	6	7	8	9	10	11	12
Pre-Post in PVCs	1	7	17	22	5	4	5	14	9	7	- 4	51

Mean = 11.50 s.d = 14.29

Obtain a 95% confidence interval for the mean difference between pre and post treatment PVCs. Is the hypothesis of zero mean difference rejected at the 5% significance level? Does the drug seem to be effective in reducing the PVCs?
[8 marks]

(c) If the standard deviation of the differences in (b) was known (rather than estimated) to be 14, how many patients would be needed for a one-sided test of the hypothesis of zero mean difference at the 5% significance level to have power 0.95 to detect a true mean difference of 10?
[4 marks]

3.

(a) Define Type I and Type II errors and discuss how they relate to the significance level and power of a hypothesis test.

[4 marks]

(b) An educator believes that the new directed reading activities in the classroom will help elementary school pupils improve their reading ability. For testing this hypothesis, a random sample of 21 pupils was chosen to undertake the new directed reading activities, over an eight-week period, and a control group of 23 pupils was also randomly selected and asked to follow the same curriculum but without the new directed reading activities. At the end of eight weeks, pupils were given a reading test, which measure aspects of reading ability that the activities were designed to improve. The scores in the reading test for the 'Activities' and 'Control' groups of pupils are given below together with some summary statistics:

Control	19, 17, 10, 28, 26, 20, 37, 37, 33, 85, 48, 46, 43, 42, 41, 42, 42, 62, 60, 53, 54, 55, 55
Activities	61, 62, 67, 24, 33, 71, 49, 49, 46, 44, 43, 43, 43, 59, 52, 58, 53, 57, 56, 57, 54

Group	n	$\bar{x}$	s
Control	23	41.52	17.15
Activities	21	51.48	11.01

(i) Display the data in a back-to-back stem plot. In the light of your plot describe briefly the similarities and differences between the distributions of scores in the Control and Activities groups. Do the scores in the Control group appear more variable than those in Activities group?

[7 marks]

(ii) Calculate a 90% confidence interval for the ratio of the true population variances of the Control and Activities group. [6 marks]

(iii) Comment on whether the hypothesis that the two groups have equal variances is supported by the confidence interval you have constructed. [2 marks]

(iv) Would you recommend that a two-sample t-test could be used for testing the hypothesis that the directed reading activities help to improve the reading ability of the pupils. [1 mark]

4.

(a) In a one-way analysis of variance with m treatment groups and n observations in each group let y_{ij} be the jth observation in the ith group, $j = 1, 2, ..., n$ and $i = 1,2,...,m$.

The model assumed has the form

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad \text{where } \sum_i \alpha_i = 0 .$$

Stating appropriate assumptions about the error term, e_{ij}, derive the least squares estimates of the unknown parameters μ and α_i.

[8 marks]

(b) In a study of a synthetic vaccine for malaria, scientists divided 45 18-21 year old male volunteers into three groups. They assigned fifteen volunteers to a saline control group, while the other thirty men were divided equally among two different vaccine dose/treatment regimens. The allocations of volunteers to treatment were at random. After vaccination, the researchers recorded a stimulation index for each volunteer, determined from proliferation assays of peripheral blood mononuclear cells. The results are shown below.

Group (i)	Stimulation index (y_{ij})	Sum
Saline control	1.4, 1.0, 4.0, 2.1, 2.4, 1.5, 2.5, 3.1, 3.4, 3.9, 1.6, 1.1, 2.2, 2.9, 3.8	36.9
Regimen 1	1.5, 5.6, 12.4, 1.8, 2.5, 3.0, 3.5, 4.1, 5.0, 5.9, 7.6, 9.8, 10.5, 11.4, 6.0	90.6
Regimen 2	6.6, 9.1, 6.9, 6.8, 8.0, 7.1, 7.5, 8.4, 8.9, 7.3, 7.4, 9.0, 6.8, 7.9, 7.0	114.7
		242.2

Note that $\sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}..\right)^2$ = 413.99.

Carry out an analysis of variance to test the null hypothesis that the mean stimulation index is the same for each treatment group and comment on your results.

[7 marks]

Prior to the start of the study, the scientists had the following specific comparisons in mind, namely

(i) Saline control versus Regimen 1.

(ii) Regimen 1 versus regimen2

Carry out a t-test to compare the mean stimulation index for each of the comparisons.

[5 marks]

5. The observations below were taken on 10 incoming shipments of chemicals arriving at a warehouse and show the total weight of shipment (x , in hundred kilograms), and number of man-minutes required to handle the shipment (y):

x:	20.3	22.3	24.0	25.0	26.4	28.4	30.9	34.1	36.6	39.9
y:	26.6	23.5	27.0	26.3	32.6	35.7	36.5	39.8	42.2	43.2

$\sum y = 333.4$ $\quad\sum x = 287.9$ $\quad\sum x^2 = 8663.1$ $\quad\sum xy = 9997.3$

$SSE = 41.06$

(a) It has been suggested that y and x are linearly related and the relationship between y and x can be modelled as

$$y_i = \alpha + \beta x_i + e_i \qquad (i = 1,2,...,10)$$

where the errors, e_i , are independent Normal random variables each with mean zero and variance σ^2 . Derive the normal equations and show that the estimates of α and β , $\hat{\alpha}$ and $\hat{\beta}$ say, are given by

$$\hat{\beta} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

where $\bar{x}$ and $\bar{y}$ denote mean of x and y respectively and n is the number of shipments.

[7 marks]

(b) Find the line of regression of y on x and test the hypothesis that $\beta = 0$ against the alternative $\beta > 0$.

[8 marks]

Question 5 contd overleaf

Q5 contd

(c) Calculate a 95% Prediction interval for the number of man-minutes, y_0, when the total weight of shipment, x, take the value $x_0 = 32.0$.

[5 marks]

[Hint: a 95% prediction interval for a new response y_0 when $x = x_0$ is given by

$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2}(0.025)s\sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}}$ where $s^2 = \frac{SSE}{n-2}$ and $S_{xx} = \sum\left(x_i - \bar{x}\right)^2$.]

6. Foresters studied 20 strands of pine trees. For each strand, they recorded the age of the strand (units not given), the average height in feet of dominant trees, the number of pine trees per acre, and the average diameter 4.5 feet above the ground (units not given). The foresters wish to model y (average diameter) as a function of one or more of the variables x_1(age), x_2(height), x_3(number), x_4 (age X number), and x5 (height / number).The data were read into Minitab, and a multiple regression performed, resulting in the following output being produced.

```
Predictor          Coef        StDev          T        P
Constant          1.233        1.619       0.76    0.459
x1               0.0526       0.1683       0.31    0.759
x2              0.08246      0.04035       2.04    0.060
x3             0.003224     0.002532       1.27    0.224
x4           -0.0002817    0.0002300      -1.22    0.241
x5                16.03        27.89       0.57    0.575

S = 0.2952      R-Sq = 88.2%     R-Sq(adj) = 84.1%


Analysis of Variance

Source            DF           SS          MS         F        P
Regression         5       9.1651           ?         ?    0.000
Residual Error    14            ?           ?
Total             19      10.3855
```

(a) Write down the fitted model and the estimated regression equation.

[3 marks]

(b) Explain how you would interpret each of the following values appearing in the given Minitab output: P values, S, R-Sq, and R-Sq(adj).

[13 marks]

(c) Complete the Analysis of variance table appearing above Minitab output. Suggest what further analysis you might carry out.

[4 marks]

7. The data given below were collected in the US in 1989, 1990 and 1991 and are based on responses given by a simple random sample of men and women from the US population 18 years of age or older to a self administered questionnaire designed to obtain information on current patterens of adult sexual behaviour in the general population. Respondents were asked to report the number of male and female sexual partners they had had since they were 18 years of age (referred to here as 'lifetime partners')

(a) In the Table given below, the data for 1989, 1990 and 1991 are aggregated and the respondents are classified according to gender and according to whether they reported less than 10 opposite-sex lifetime partners or 10 or more such partners.

Table
Number of opposite-sex lifetime sexual partners classified according to gender and sexual activity.

Sexual activity

Gender	**Fewer than 10 partners**	**Ten or more partners**	**Total**
Male	3240	16471	**19711**
Female	3429	2717	**6146**
Total partners	**6669**	**19188**	**25857**

A puzzling anomaly is that whereas in the entire population the number of lifetime female sexual partners reported by men may be expected to equal the number of lifetime male partners reported by females, in the sample the former far exceeds the latter. To investigate this anomaly further, test the hypothesis that the variables 'Gender' and ' Sexual activity' are independent.

[8 marks]

Question 7 is continued overleaf

Q7 contd

(b) It may be gleaned from the Table above that much of the gender discrepancy in the reported sexual activity may originate in the highly active group with 10 or more lifetime partners. To investigate this observation further, consider only the data for the sexually less active group with fewer than 10 lifetime partners. For this group only, test the hypothesis that there are no gender differences in the reported sexual activity, that is, in the entire population the number of female lifetime sexual partners reported by sexually less active males and the number of male lifetime partners reported by sexually less active females are equal.

[7 marks]

(c) Relate the result obtained in (b) above to that in (a) and comment on what additional information may be collected to resolve the anomaly described in (a) above.

[3 marks]

(d) What is Cochran’s rule? Is this rule satisfied in the tests used in parts (a) and (b)?

[2 marks]