1. Given n pairs of observations, (x_i, y_i) $(i = 1, \ldots, n)$, following the linear regression model

$$y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $x_1, \ldots, x_n$ are some fixed numbers and $\epsilon_1, \ldots, \epsilon_n$ are independent Normal random variables, each with mean 0 and variance σ^2, it is known that

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned}$$

where $\bar{y} = n^{-1}(y_1 + \cdots y_n)$, $\bar{x} = n^{-1}(x_1 + \cdots + x_n)$, provide the maximum likelihood estimators of β_1 and β_0, respectively.

Show that

(a) $E\left(\hat{\beta}_1\right) = \beta_1,$ [4 marks]

(b) $E\left(\hat{\beta}_0\right) = \beta_0,$ [2 marks]

(c) $\mathrm{Var}\left(\hat{\beta}_1\right) = \sigma^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2.$ [3 marks]

A widely-held belief in business and industry is that younger managers tend to be harder-driving and their appointments could help to increase the profitability of a company. It is also possible, however, that the greater experience of the older managers may lead to better decisions. For testing whether the mean age (x) of executives of a firm influences the change in annual earnings per share (y) of the firm, the mean ages (x_i) of executives of fifteen randomly chosen firms in a certain industry were recorded together with last year's percentage increase in earnings per share (y_i) of the firms. The following summary statistics were then calculated:

$$\sum x_i = 683.8, \qquad \sum y_i = 167.3, \qquad \sum x_i y_i = 7741.74,$$

$$\sum x_i^2 = 31358.58, \qquad \sum y_i^2 = 2349.61.$$

Verify that

$$\hat{\beta}_0 = -16.99, \qquad \hat{\beta}_1 = 0.62$$

provide the maximum likelihood estimates of the coefficients, β_0 and β_1, in a linear regression of y on x. [4 marks]

Construct a 95% confidence interval for the unknown coefficient, β_1, and hence test the hypothesis that y and x are linearly related. [5 marks]

Provide an interpretation of your results. [2 marks]

2. Consider the standard multiple regression model:

$$Y \;=\; X\beta + u,$$

where X is a known $n \times (k+1)$ matrix of rank $k+1$, β is a vector of $(k+1)$ unknown parameters and u is a random vector whose components are independent Normal random variables, each with mean 0 and variance σ^2. Given that

$$\hat{\beta} \;=\; \left(X^T X\right)^{-1} X^T Y$$

provides the maximum likelihood estimator of β, and $e = Y - X\hat{\beta}$ denotes an $n \times 1$ vector of errors of the fitted regression model, show that the error sum of squares may be written as

$$e^T e \;=\; Y^T Y - Y^T X \hat{\beta}.$$

[3 marks]

In a certain experiment, the following multiple regression model is envisaged:

$$Y_i \;=\; \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, \qquad i = 1, \ldots, n,$$

where the u_i are independent Normal random variables, each with mean 0 and variance σ^2, β_0, β_1 and β_2 are unknown coefficients and x_1 and x_2 may assume values ±1.

Let $\beta = [\beta_0, \beta_1, \beta_2]^T$ denote the vector of unknown parameters. For estimating β, two different values of Y were recorded at each of the four possible values of (x_1, x_2), namely $(1,1)$, $(1,-1)$, $(-1,1)$, $(-1,-1)$, and the following $n = 8$ observations were obtained:

i	x_{i1}	x_{i2}	Y_i
1	1	1	20.2
2	1	1	19.8
3	1	-1	8.0
4	1	-1	8.2
5	-1	1	9.7
6	-1	1	10.0
7	-1	-1	1.9
8	-1	-1	0.6

Find the maximum likelihood estimate, $\hat{\beta}$, of β. [7 marks]

Find also an estimate of the residual error variance, σ^2. [4 marks]

Test, at the 1% level of significance, the hypothesis that $\beta_1 = \beta_2 = 0$ against the alternative that at least one of these two coefficients does not vanish. [6 marks]

3. A drug company is testing the effectiveness of two new drugs, called Drug 1 and Drug 2, relative to an existing product, called Drug 3. It is concerned, however, about their possible side effects, and, in particular, whether the new drugs increase the blood pressures of patients taking them. For investigating this last possibility, an experiment was conducted in which n patients were allocated at random to 3 groups of possibly unequal sizes and the patients in Group i were prescribed Drug i, $i = 1, 2, 3$. The blood pressures of all patients were recorded initially at the start of the relevant treatment and once again one hour after taking a single dose of the prescribed drug. The values of blood pressure change, y_{ij}, $j = 1, \ldots, n_i$, $i = 1, 2, 3$, were recorded for all patients, where n_i denotes the number of patients receiving Drug i, with $n = n_1 + n_2 + n_3$.

   The following statistical model is suggested for analysing the collected data:

$$Y_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij} \qquad (j = 1, 2, \ldots, n_i, \;\; i = 1, 2, 3)$$

   where Y_{ij} denotes the random variable corresponding to the observed value, y_{ij}, the ϵ_{ij} are independent Normal random variables each with mean 0 and variance σ^2, μ denotes the overall mean of the Y_{ij}, and α_i denotes the possible specific effect of Drug i.

   (a) Explain why the α_i satisfy the following constraint:

$$\sum_{i=1}^{3} n_i \alpha_i \;=\; 0.$$

   [3 marks]

   (b) Prove the following sum of squares identity.

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \;=\; \sum_{i=1}^{3}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{3} n_i \left(\bar{y}_{i\cdot} - \bar{y}\right)^2,$$

   where $\bar{y}_{i\cdot} = (n_i)^{-1} \sum_{j=1}^{n_i} y_{ij}$, $\bar{y} = n^{-1} \sum_{i=1}^{3} \sum_{j=1}^{n_i} y_{ij}$. [3 marks]

(c) Hence give a heuristic justification of the standard F-test for testing the hypothesis that the mean blood pressure changes after taking each of these drugs do not differ. [N.B. You may use without proof, but should state clearly, any standard result on the sampling properties of the F-statistic that you invoke.] [4 marks]

The actual number of patients equalled $n = 36$. A numerical summary of the data that were collected is given below:

Drug (i)		1	2	3
Number of patients	n_i	12	12	12
Mean change in blood pressure	$\bar{y}_{i.}$	2.25	14.583	8.083
Sample variance*	s_i^2	76.75	62.265	68.2652

* The sample variances were calculated by using a divisor of $n_i - 1$.

Carry out an analysis of variance of the data and test whether or not the mean changes in blood pressure as a result of taking the different drugs are the same.

[8 marks]

Coment on why it is important to randomly allocate the patients to different drugs.

[2 marks]

4. (a) Explain the principal features of a randomised block design and describe generically, and/or by a specific example, the potential situations in which a use of this design may be appropriate. [5 marks]

The table below shows the time taken, in minutes, for four different city councils to reach a decision on the following issues:

I_1: whether to change the name of a street;

I_2: whether to increase the parking charges;

I_3: whether to allocate money for building a community hall.

	Issues	I_1	I_2	I_3	Means
Councils					
1		36	55	63	51.33
2		44	50	62	52.00
3		26	44	45	38.33
4		21	29	39	29.67
	Means	31.75	44.5	52.25	42.83

(b) Complete the analysis of variance table given below (copy into your submitted answers) and identify which effects are significant.

Source	Sum of squares
Councils	1049.7
Issues	
Residual	
Total	1993.7

[8 marks]

(c) Before the data were recorded, it was believed by many analysts that

i) more time is taken on average on issues involving money than on those not involving money;

ii) more time is taken on average on issues invoving spending money than on those involving the receipt of money.

Calculate the appropriate Fisher significant difference at the 6% level for a pairwise comparison of the average time taken for reaching a decision on the three issues, I_1, I_2, I_3, described above, and hence construct approximate 94% simultaneous confidence intervals for the differences between each pair of these means. [5 marks]

In the light of your results, comment on whether there is sufficient evidence to support the beliefs i) and ii) stated above. [2 marks]

5. (a) Outline some of the limitations of a Latin Square design as compared with a complete factorial experiment. [2 marks]

(b) Write down an appropriate model for analysing data from a Latin Square experiment, stating clearly all assumptions made. [3 marks]

(c) A Latin Square experiment was carried out over a six week period and using six different grocery stores for investigating the effect of shelf space on sales of powdered coffee whitener. The data, y_{ijk}, which were collected are summarised below in suitable units, and with shelf-space index shown in brackets:

	Weeks						Means
Stores	1	2	3	4	5	6	$\bar{y}_{i\cdot\cdot}$
1	27 (55)	14 (54)	18 (53)	35 (51)	28 (56)	22 (52)	24.00
2	34 (56)	31 (55)	34 (54)	46 (53)	37 (52)	23 (51)	34.17
3	39 (52)	67 (56)	31 (55)	49 (54)	38 (51)	48 (53)	45.33
4	40 (53)	57 (51)	39 (52)	70 (56)	37 (54)	50 (55)	48.83
5	15 (54)	15 (53)	11 (51)	9 (52)	18 (55)	17 (56)	14.17
6	16 (51)	15 (52)	14 (56)	12 (55)	19 (53)	22 (54)	16.33
Means $\bar{y}_{\cdot j\cdot}$	28.5	33.17	24.5	36.83	29.5	30.33	30.47

In addition, we also have

$$\bar{y}_{\cdot\cdot 1} = 30.0, \bar{y}_{\cdot\cdot 2} = 26.83, \bar{y}_{\cdot\cdot 3} = 31.0, \bar{y}_{\cdot\cdot 4} = 28.5, \bar{y}_{\cdot\cdot 5} = 28.17, \bar{y}_{\cdot\cdot 6} = 38.33,$$

$$\sum y_{ijk}^2 = 42219, \sum y_{ijk} = 1097.$$

Complete the following analysis of variance table (copy into your submitted answers):

Source	Sum of squares
Stores	6475.80
Weeks	529.47
Shelf space	
Residual	
Total	

Identify which effects are significant at the 5% level. [8 marks]

(d) Stores 1, 5 and 6 are in relatively poor areas whose residents prefer to add fresh milk to their drinks of coffee/tea. By contrast, Stores 2, 3 and 4 are in areas where many young professionals live. Before the data were collected, it was suggested that the average sales of coffee whiteners for Stores 2, 3 and 4 could be higher than those for Stores 1, 5, 6. Test whether the data provide evidence to support this suggestion. [4 marks]

(e) Comment on the extent to which the generic assumptions stated in (b) above for analysing the data from a Latin Square design are likely to hold for the particular application given above in (c) and whether there are features which suggest that caution should be exercised before accepting the results of the statistical tests carried out in (c) and (d) above. [3 marks]

6. Let x be the observed value of a random variable, X, following a Binomial distribution with parameters n and p, that is, its probability function is given by

$$f(x|p) \quad = \quad \binom{n}{x} p^x (1-p)^{n-x} (x = 0, 1, \ldots, n).$$

(a) Show that, given x and n, the maximum likelihood estimate of p is

$$\hat{p} \quad = \quad \frac{x}{n}.$$

[4 marks]

(b) The number of seeds per pod, X, say, of an honesty plant may be modelled by a Binomial distribution with parameters m and p, where p is unknown and it may vary from plant to plant. For estimating p, a random sample of k pods was taken from an honesty plant and the number, j, say, of seeds per pod was counted, $j = 0, 1, \ldots, m$. Let f_j denote the number of pods containing exactly j seeds, where $f_0 + f_1 + \cdots + f_m = k$. Write down the likelihood function for the experiment described above by treating m as known, and show that the maximum likelihood estimate of p based on the collected data is

$$\hat{p} \quad = \quad (mk)^{-1} \sum_{j=0}^{m} j f_j.$$

[8 marks]

(c) Random samples of pods were taken from two plants and yielded the following data:

Number (x)	Number of pods with x seeds	
of seeds in pods	Plant 1	Plant 2
0	0	0
1	0	0
2	0	1
3	2	2
4	5	5
5	12	17
6	37	51
7	39	45
8	19	19

Calculate $\hat{p}$ separately for the two plants. [2 marks]

(d) It is suggested that the values, p_1 and p_2, of p should be the same for the two plants, i.e. $p_1 = p_2 = p$, say. Give the maximum likelihood estimate of p, and provide a brief justification for your estimate. [2 marks]

(e) Without doing any further calculations, discuss how you would judge whether the simplified model with $p_1 = p_2 = p$, or the model in which p_1 may differ from p_2, should be adopted for the data given above. [4 marks]

7. Suppose that $Y_1, \ldots, Y_n$, $n > 1$, are independent Poisson random variables with means $\lambda_1, \ldots, \lambda_n$, respectively, so that the probability function of Y_i for each $i = 1, \ldots, n$, is given by

$$f(y_i|\lambda_i) \quad = \quad \lambda_i^{y_i} \exp(-\lambda_i) / (y_i!), \qquad (y_i = 0, 1, 2, \ldots),$$

where $\lambda_i > 0$. Also, let y_i denote the observed value of Y_i, $i = 1, \ldots, n$.

(a) Derive the maximum likelihood estimate of λ_i, $i = 1, \ldots, n$, corresponding to this maximal model. [4 marks]

(b) Also derive the maximum likelihood estimate of λ for the simplified model, which specifies that $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$. [2 marks]

(c) Show that the deviance statistic, D, for comparing the maximal and simplified models is

$$D \quad = \quad 2 \sum_{i=1}^{n} Y_i \ln\left(\frac{Y_i}{\bar{Y}}\right),$$

where $\bar{Y} = n^{-1}(Y_1 + \cdots + Y_n)$. [5 marks]

(d) By using a standard analytical result that, to terms of second order, a (truncated) Taylor expansion of the function

$$h(y) \quad = \quad y \ln\left(\frac{y}{y_0}\right)$$

around a constant, y_0, yields

$$h(y) \quad \approx \quad (y - y_0) + (2y_0)^{-1}(y - y_0)^2,$$

deduce that the deviance statistic may be expressed as

$$D \quad \approx \quad \left(\bar{Y}\right)^{-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2.$$

[3 marks]

(e) Hence explain why, for large values of n, $n^{-1}D$ may be expected to be close to 1, if the simplified model with $\lambda_1 = \cdots = \lambda_n = \lambda$ is true and comment on the possible implications of this result for a comparison of the simplified and maximal models. [6 marks]