UNIVERSITY OF ABERDEEN **SESSION 2003–04**

Degree Examination

ST2504 Data Analysis with a Statistical Package (SPSS)

Thursday 27 May 2004 (9am to 11am)

Only calculators approved by the Department of Mathematical Sciences may be used in this examination. Calculator memories must be clear at the start of the examination.

Marks may be deducted for answers that do not show clearly how the solution is reached.

Answer THREE questions. All questions carry equal weight.

1. The table below gives data on the lean body mass (the weight without fat) and resting metabolic rate for twelve men who were the subjects in a study of obesity.

Man number	Lean body mass (kg) x_i	Resting metabolic rate (in suitable units) y_i
1	36.1	995
2	54.6	1425
3	48.5	1396
4	42.0	1418
5	50.6	1502
6	42.0	1256
7	40.3	1189
8	33.1	913
9	42.4	1124
10	34.5	1052
11	51.1	1347
12	41.2	1204

$$[\sum x_i = 516.4, \quad \sum x_i^2 = 22,741.34, \quad \sum y_i = 14,821, \quad \sum y_i^2 = 18,695,125, \quad \sum x_i y_i = 650,264.8]$$

Regard resting metabolic rate (Y) as the dependent variable and lean body mass (X) as the independent variable. You should assume that $Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$, where β_0, β_1 and σ^2 are unknown constants.

(a) Calculate the least squares fit regression line (in which resting metabolic rate is modelled as the response and the lean body mass as the explanatory variable.)

(b) Find a 95% confidence interval for the slope coefficient (β_1) of the model.

(c) Use the fitted model to construct 95% confidence intervals for the mean resting metabolic rate when

(i) the lean body mass is 50kg;

(ii) the lean body mass is 75kg.

(continued on next page)

(d) Comment briefly on the appropriateness of each of the confidence intervals given in **(c)**.

(e) Consider man number 4.

(i) Calculate his "predicted" resting metabolic rate ($\hat{y}_4$), and find the residual $e_4 = y_4 - \hat{y}_4$.

(ii) Find the Studentised residual, e'_4, and state whether the observation y_4 may be regarded as an "outlier" (which is taken to mean an observation with Studentised residual exceeding 3 in absolute value).

2. **(a)** Describe the following experimental designs briefly:

(i) a Completely Randomised design;

(ii) a Randomised Block design;

(iii) a Latin Square design.

(b) In an agricultural experiment with $r = 6$ treatments and $s = 5$ blocks, the following figures were obtained for the yield of grain in kilograms per 40 square metres.

Block number	Treatment number						Total
	1	2	3	4	5	6	
1	12.0	11.5	11.5	11.0	9.5	9.3	64.8
2	10.8	11.4	12.0	11.1	9.6	9.7	64.6
3	13.2	13.1	12.5	11.4	12.4	10.4	73.0
4	14.0	14.0	14.0	12.3	11.5	9.5	75.3
5	14.6	13.2	14.2	14.3	13.7	12.0	82.0
Total	64.6	63.2	64.2	60.1	56.7	50.9	359.7

(i) Construct an appropriate ANOVA table.

(ii) Test the hypotheses that

(1) there is no "block effect",

(2) there is no "treatment effect",

using a 0.1% significance level in each case.

(iii) Express your conclusions in **(b)**(ii) in non-technical language.

3. **(a)** The following table refers to a random sample of 12 European ski resorts:

	Height (above sea level), x metres	Early morning temperature (on a given day), y°C
Arosa	1742	7
Davos	1543	4
Wengen	1277	11
Andermatt	1439	7
Brunwald	1254	10
Champery	1049	13
Goschenen	1109	11
Leysin	1398	13
St Moritz	1778	8
Zermatt	1609	6
Villars	1256	10
Gstaad	1049	10

$$[\sum x_i = 16,503, \quad \sum x_i^2 = 23,402,167, \quad \sum y_i = 110, \quad \sum y_i^2 = 1094, \quad \sum x_i y_i = 145,964]$$

It may be assumed that the random variables X (height) and Y (temperature) are normally distributed.

(i) Calculate the sample correlation coefficient, r.

(ii) Find an approximate 95% confidence interval for the true correlation coefficient, ρ.

(iii) Using a 1% significance level, test the hypothesis that $\rho = 0$.

(b) The following table refers to the coat colour of 1000 mother and daughter pairs of racehorses:

		Coat colour of mother					
		Black	Brown	Bay	Chestnut	Grey	Total
Coat colour of daughter	Black	7	8	11	11	5	42
	Brown	7	40	75	20	9	151
	Bay	13	95	230	101	42	481
	Chestnut	6	23	113	82	17	241
	Grey	5	7	18	16	39	85
	Total	38	173	447	230	112	1000

(i) Describe how you would calculate the χ^2-statistic for conducting a test of the hypothesis H_0 that there is no association between the coat colour of mother and daughter racehorses. (You are not required to calculate χ^2 numerically.)

(ii) You are given that $\chi^2 = 181.55$. Conduct a test, at the 0.1% significance level, of the hypothesis H_0, and state your conclusion in non-technical language.

4. **(a)** Consider a clinical trial in which n_1 patients receive a certain treatment and n_2 are controls. Suppose that m_1 of the "treatment" patients and m_2 of the "control" patients suffer adverse effects within one year. Let p_1 denote the true (but unknown) probability that a "treatment" patient suffers adverse effects within one year, and let p_2 denote the corresponding quantity for a "control" patient.

(i) Define the (true) odds ratio, OR, in respect of the incidence of adverse events of "treatment" patients relative to that of "control" patients.

(ii) Let $\alpha = \log(OR)$ be the (true) log-odds ratio. Give formulae for

(1) $\hat{\alpha}$, an estimator of α,

(2) its estimated variance,

in terms of n_1, n_2, m_1 and m_2.

(iii) Suppose that you have obtained an estimate $\hat{\alpha}$ of the log-odds ratio, α, with approximate standard deviation s.e.$(\hat{\alpha})$. Give a formula for an approximate 95% confidence interval for the (true) odds ratio OR.

(b) Consider a clinical trial involving a treatment group (group 1) and a control group (group 0). Let $S_1(t)$ and $S_0(t)$ denote their respective survival functions.

(i) State the proportional hazards assumption.

(ii) Show that, under this assumption,

$$g(t) = \log\{-\log[S_1(t)]\} - \log\{-\log[S_0(t)]\} = \text{constant} \quad (t > 0).$$

(iii) Describe how the result of (ii) may be used in practice to check the proportional hazards assumption.