McGill University
Faculty of Science

Department of Mathematics and Statistics

April 2007

MATH 204

PRINCIPLES OF STATISTICS II

SOLUTIONS

1. (a) In an experimental study, the treatment is **assigned** to the experimental units by the experimenter; in an observational study, the experimenter merely **records or observes** the treatments received.

4 MARKS

(b) This is a **completely randomized design**.

2 MARKS

(c) The ANOVA table is as follows:

SOURCE	DF	SS	MS	F
TREATMENTS	3	11.355	3.785	3.355
ERROR	21	23.698	1.128	
TOTAL	24	35.053		

10 MARKS

(d) Null hypothesis is

$$H_0 \quad : \quad \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a \quad : \quad \text{At least one pair of } \mu\text{s different.}$$

The test statistic is $F = 3.355$ and if H_0 is true,

$$F \sim \text{Fisher-F}(k-1, n-k) \equiv \text{Fisher-F}(3, 21)$$

Using the tables, for $\alpha = 0.05$, we have

$$\mathsf{F}_\alpha(3, 21) = 3.07 < 3.355$$

so we **reject** the null hypothesis at $\alpha = 0.05$.

5 MARKS

(e) We need to assume **independence** and **normality** of the random errors, and to have **equal population variances** in the four groups.

4 MARKS

2. (a) In a balanced complete randomized block design with replication, we have r replicates for each of the $b \times k$ blocking factor level/treatment factor level combinations. For each level of the blocking factor, we randomly select kr experimental units, and then allocated r at random to each of the treatment levels. In a factorial design, no distinction is made between the two factors in terms of population substructure; we do not block according to the blocking factor.

6 MARKS

(b) (i) The model fitted is

id + **dose**

that is, a main effects for blocking factor patient and dose.

1 MARK

In total, two hypothesis are tested;

- the first concerns the differences between levels of **id** (differences between patients); this test produces a p-value of 0.001, which is significant at $\alpha = 0.05$ level. Hence we **reject** the hypothesis that the responses at different levels of **id** are equal, confirming that the blocking by **id** is necessary,

4 MARKS

- the second concerns the differences between levels of **dose**; this test produces a p-value of 0.625, which is not significant at $\alpha = 0.05$ level. Hence we **do not reject** the hypothesis that the responses at different levels of **dose** are equal.

4 MARKS

(ii) In this analysis, the model with interaction

id + **dose** + **id . dose**

However, with no replication, we cannot test this hypothesis, as the number of parameters equals the number of data points, leading to the result $SSE = EDF = 0$. Thus the F and p-values cannot be computed.

6 MARKS

(c) You could use the **non-parametric** procedure, Friedman's test, which takes into account the blocking structure. Alternatively, you could use a randomization/permutation procedure, taking care to preserve block structure when randomizing.

4 MARKS

3. (a) This is a balanced complete factorial design.

3 MARKS

(b) The five models are (in the order analyzed)

M_1	$C + T + C.T$
M_2	$C + T$
M_3	T
M_4	C
M_0	Null

5 MARKS

(c) The analysis fits the models in the above sequence; using inspection of the p-values in this **balanced design**, it appears that the backward selection sequence should be

$$M_1 \longrightarrow M_2 \longrightarrow M_3 \longrightarrow M_0$$

The sums of squares in each comparison is listed below

Complete	SSE_C	k	Reduced	SSE_R	g	F	$k-g$	$n-k-1$	F_α	Conclusion
M_1	10.929	14	M_2	11.248	6	0.055	8	15	2.64	Do not reject
M_2	11.248	6	M_3	13.907	2	1.359	4	23	2.80	Do not reject
M_3	13.907	2	M_0	20.284	0	6.190	2	27	3.35	Reject

Hence

(i) The interaction should be omitted

(ii) The most appropriate model is model M_3 that contains the main effect T only, that is, there is a different response for the different temperature levels, but no effect of concentration.

(iii) For this model, the R^2 and adjusted R^2 quantities are 0.314 and 0.264 respectively. Hence the explanatory power overall is not very high. We cannot comment on the appropriateness of the normality assumptions, or the presence of outliers, which also need to be checked.

Note that most of these conclusions can be deduced from the original ANOVA tables, as the design is balanced and complete.

12 MARKS

(d) The model

$$C + T$$

contains **six** non-intercept parameters in total if C is fitted as a factor predictor. Of these, four parameters correspond to the $5 - 1 = 4$ contrasts from baseline due to concentration. If concentration is fitted as a covariate, then there is only a single parameter for concentration, and the number of non-intercept parameters is $2 + 1 = 3$. Hence the difference is **three**.

5 MARKS

4. (a) By inspection, it seems that **Height** is the only variable that should be in the model; from Analysis 1, it seems that **Age** is not significant in the model, and despite the results of Analysis 2, **Sex** is also not significant in the presence of **Height** .

We compare the models

Height and **Height** + **Sex** + **Height** . **Sex**

using the F-test. The test statistic is

$$F = \frac{(42.163 - 38.894)/(3-1)}{38.894/28} = 1.389$$

which we compare against the Fisher-F$(2, 28)$ distribution. From tables, $\mathsf{F}_\alpha(2, 28) = 3.34$, so we do not reject this hypothesis that the reduced model is an adequate simplification of the complete model.

We cannot assess whether **Height** would not be significant in the presence of **Age** as this output is not recorded, but this is unlikely as **Age** and **Height** are unlikely to be dependent in adults. Also Analyses 1, 5 and 6 indicate that **Age** is not significant.

The R^2 and Adjusted R^2 for the preferred model are 0.484 and 0.467, so the explanatory power is only moderate.

12 MARKS

(b) There is dependence between **Sex** and **Height** and hence their apparent effects are **confounded**. Simply, we might suspect that height predicts sex well, that is, taller individuals are likely to be men.

4 MARKS

(c) Using the results from Analysis 3, the prediction is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_{\textsf{Height}} \times 165 = -9.740 + 0.095 \times 165 = 5.935$$

6 MARKS

(d) This note is given as the corresponding parameter is estimated as the *Intercept*, as is always the case for factor predictors.

3 MARKS

5. (a) We have for the fitted values

Swim ≥ 6 hrs.	Erosion of Enamel Yes	No	Total
Yes	25	125	150
No	24	120	144
Total	49	245	294

6 MARKS

(b)

$$X^2 = \frac{(32-25)^2}{25} + \frac{(17-24)^2}{24} + \frac{(118-125)^2}{125} + \frac{(127-120)^2}{120} = 4.802$$

4 MARKS

(c) The $\alpha = 0.05$ quantile for Chisquared(1) distribution is 3.841. Thus we **reject** the hypothesis of independence.

4 MARKS

(d) The assumptions behind the chi-squared test may in general be violated for a case control study, as we do not have independent multinomial sampling overall; actually in this case the assumptions are met. Also, all expected cell entries are greater than five, so the chi-squared approximation seems to be valid.

3 MARKS

(e) Here we have

$$\log \widehat{\psi} = \log\left(\frac{n_{11}\,n_{22}}{n_{12}\,n_{21}}\right) = \log\left(\frac{32 \times 127}{118 \times 17}\right) = 0.706$$

and

$$\text{s.e.}(\log \widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{32} + \frac{1}{118} + \frac{1}{17} + \frac{1}{127}} = 0.326$$

so that

$$Z = \frac{\log \widehat{\psi}}{\text{s.e.}(\log \widehat{\psi})} = \frac{0.706}{0.326} = 2.166$$

Given the critical values ± 1.96, we conclude that the log odds ratio is significantly different from zero.

8 MARKS

6. (a) (i) Levene's Test: Testing the equality of variances, for example in a one-way or two-way layout, as a precursor for an ANOVA test.

(ii) Friedman's Test: Non-parametric equivalent to ANOVA for the randomized block design.

(iii) Fisher's Exact Test: Exact test for independence in a 2×2 table.

9 MARKS

(b) We have

$$R_1 = 55 \qquad R_2 = 36 \qquad R_3 = 80$$

and hence

$$H = 5.696$$

We compare this with the Chisquared$(k-1)$ = Chisquared(2) distribution; Chisq$_{0.05}(2) = 5.991$. Thus the test result suggests that **we do not reject** the null hypothesis of different locations in the three groups.

12 MARKS

(c) With a normality assumption, we may use one-way ANOVA, provided that the variances in the three groups could be proved to be equal using Levene's Test.

4 MARKS