

DUBLIN CITY UNIVERSITY

SAMPLE EXAM. PAPER

CA534 BIODATA ANALYSIS

M.Sc. Bio Informatics Programme

Instructions: Answer any FOUR questions. EACH question carries EQUAL marks.
Sections (only) of printout *may* be included within a selected question.
Statistical tables are provided.

Q1

- a) State the three axioms of probability and give the **special** case of the multiplication rule for the probability of the joint occurrence of two events, indicating when it is appropriate to use this.

In this context, indicate when the Hypergeometric, as opposed to the Binomial distribution, should be used.

- b) In the joint distribution $f(x,y)$ below, obtain the marginal distribution of x and y and compute $\text{cov}(x,y)$. Are x and y independent?.

(i) $f(x,y)=xy$ $0 < x < 1$ & $0 < y < 2$

(ii)

X →	0	1
Y ↓		
0	3/16	5/16
1	6/16	2/16

- c) The offspring produced by a cross between two given types of plants can be any one of three genotypes A, B or C. A simple inheritance model suggests that the offspring of types A,B,C should be in the ratio 1:2:1. An experiment was conducted in which 80 plants were bred by crossing the two parent types. The genetic classifications of the offspring are recorded below. Do these data support the simple genetic model? Justify.

Genotype	A	B	C
Observed frequency	18	35	27

[Question: 25 marks]

Q2

- a) Explain what is involved in Maximum Likelihood estimation and give brief details of two commonly-used methods of obtaining the MLE of a parameter θ .

- b) Contrast (in brief) the Bayesian and Likelihood methods for parameter estimation. For a given random variable Y with p.d.f

$$f(x) = 2e^{-2x} \quad x > 0$$

Obtain the m.g.f and the first two moments about the mean

- c) Given the prior distribution for the proportion p of people with a given condition is:

p	0.1	0.2
$f(p)$	0.6	0.4

Find the Bayes estimate for the proportion of people with the condition, if a random sample of size 2 gives 1 with the disease.

[Question: 25 marks]

Q3

- a) State three different approaches to obtaining a confidence interval and contrast them.
- b) Expected genotypic frequencies for a backcross (AaBb×aabb) model are given in the table below, where θ is the recombination fraction between A and B and f_{ij} is the observed genotypic count for the i th genotype of locus A and the j th genotype of B.

Genotype	Observed count(f_{ij})	Expected frequency(p_{ij})
AaBb	f_{11}	$0.5(1-\theta)$
Aabb	f_{12}	0.5θ
AaBb	f_{21}	0.5θ
aabb	f_{21}	$0.5(1-\theta)$

Write down the likelihood function and obtain the MLE of the recombination fraction in terms of the frequencies and total number of individuals N in the sample.

Give an expression for the average information content for an individual and, hence, the variance expression of the recombination fraction for a sample size of N .

- c) A study in the 90's was conducted on injecting drug-users. In a sample of 200 long-term regular methadone (LTM) users, 50 were female. In a sample of 250 IDUs who were not LTM, 40 were female.

Construct a 90% confidence interval for the difference between the proportions of males in the two populations. What result would you expect from the complementary test of hypothesis?

[Question: 25 marks]

Q4

- a) In each case, state the type of data and principal hypotheses being tested for
(i) ANOVA (ii) Friedman (iii) Chi-squared
- b) Describe a randomized block design and explain the terms and assumption, giving the form of the ANOVA table and expected mean squares.
In the design of an experiment, what do you understand by random, fixed and mixed effects?

- c) In a particular laboratory experiment on yield of a substance, two variables are of interest C (the catalyst used in the experiment) and T (the washing/cooling time). Two observations were available for each combination of variable values.

Source	dof	SSQ	MSQ
Model	11	76.7683	6.9789
Error	12	14.9100	1.2425
Total	23	91.6783	

T	2	14.5233	
C	3	40.0817	
T*C	6	22.1633	

Interpret the results and how you would reproduce this analysis in SPSS (minimal commands only).

[Question: 25 marks]

Q5

- a) Give three examples of the type of question that you might want to investigate, using A non-parametric (distribution-free) approach and indicate what is involved in testing for one of them.
- b) Give the basic idea and brief details of the Kolmogorov-Smirnov test for goodness-of-fit for empirical distribution functions and summarize its advantages and disadvantages when compared to the chi-squared test.
- c) The table shows the results of a mouse-infection experiment in which 12 mice in group A and 10 in group B received the same challenge dose of bacteria and were then observed daily for death or survival. Are the median death times in Groups 1 and 2 significantly different? Note: S* implies survival for the duration of the experiment = 14 days and must be treated in the same way as S = survival. (You are given that Wilcoxon-Mann-Whitney U for a 12,10 comparison =29)

Table: Results of an infection experiment in mice:

Mouse Group	Initial No. in group	Day of death (post-infection) of Individual animals
1	12	2, 3, 3, 4, 4, 5 5, S, S, S, S, S, S*
2	10	1, 1, 2, 2, 3, 3, 3, 4, 5, S*

[Hint: Check for total ranks in groups size m,n : Sum of all ranks = $1/2(m+n)(m+n+1)$ and note that the smaller group/lower ranks are quicker to calculate].

[Question: 25 marks]

Q6

- a) Distinguish between **regression** and **correlation** and briefly describe the problem of Multicollinearity, (also known as Collonearity).
- b) Using matrix notation for the basic model, explain the principle of least squares estimation, obtaining the form for the estimates of unknown parameters. Hence, give the basis for the principal hypothesis test and the form of the test statistics.
- c) The results of a multiple linear regression procedure are given in the table below. Yields (Y) of a particular substance are dependent on four properties of the original material, labelled X_1, X_2, X_3, X_4 .

Analysis of Variance (ANOVA)

Source	dof	SSQ	MSQ	F-value	p-value
Regression	4	3429.27	857.31	171.71	0.000
Error	27	134.80	4.99		
Total	31	3564.07			

Adjusted R-square = 0.957

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t-test	p-value
Intercept	-6.82	10.123	-0.674	0.506
X_1	0.227	0.100	2.274	0.031
X_2	0.554	0.370	1.498	0.146
X_3	-0.150	0.029	-5.116	0.000
X_4	0.155	0.006	23.992	0.000

Interpret the results carefully, describing the principal results of the tests and the meaning of the R^2 value.

[Question: 25 marks]