UNIVERSITY OF LONDON

IMPERIAL COLLEGE LONDON

Examination for the Master of Science in Bioinformatics 2005

Probability And Statistics

Students in Master of Science in Bioinformatics

Tuesday 1st March 2005 (9.30 am to 11.30 am)

Answer any two questions. Each question carries 25 marks.

A Formula Sheet and Statistical Tables are provided.
Calculators may be used

1. Give a description of each of the following statistical methods, quantities or terms, and in each case give an example of how the method, quantity or term is used in the analysis of data in bioinformatics or statistical genetics.

(a) p-value

(b) Family-wise error rate (FWER)

(c) permutation or randomization test

(d) hidden Markov model

(e) principal components analysis

2. The association between a single nucleotide polymorphism (SNP) at a given genomic location and a genetic disease is to be investigated in two genetically distinct populations A and B. Data on a number of individuals across the two populations typed at the locus are presented in the table below. The wild-type individuals are coded 0, and those with the polymorphism are coded 1.

Population A

	Genotype	
Disease status	0	1
Unaffected	72	12
Affected	30	15

Population B

	Genotype	
Disease status	0	1
Unaffected	108	67
Affected	22	78

(a) An individual is selected at random from all people in the study. Evaluate

(i) the probability that the person is Affected

(ii) the probability that the person is from Population B

(iii) the conditional probability that the person is Affected, given that they are from Population A

(iv) the conditional probability that the person is Affected, given that they are from Population A and have genotype code 1.

(b) To test for association between genotype and disease status in a sample **pooled across the two populations**, a Chi-squared test can be used. The chi-squared statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}$$

can be used to test the null hypothesis H_0 that there is no association (in fact, **independence**) between the SNP genotype and the disease, where $\widehat{n}_{ij}$ is the **fitted** value for cell (i, j) $(i = 1, 2$ and $j = 1, 2)$ defined by

$$\widehat{n}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

where $(n_{1.}, n_{2.})$ are the row totals, $(n_{.1}, n_{.2})$ are the column totals, and $n_{..}$ is the total number of individuals in the study. If H_0 is true, χ^2 has an approximate Chi-squared distribution with 1 degree of freedom.

Carry out a test of the hypothesis of independence using the chi-squared statistic.

(c) The test in (b) is only valid if the pooling of data from the two populations is valid, that is, if the pattern of association is the same in both populations. One way to test this is to compare the **odds-ratios** for the two samples; the **sample odds-ratio statistic** for a 2×2 table is denoted $\hat{\psi}$ and is defined by

$$\hat{\psi} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

and, for large samples, this statistic, **on the natural log scale**, has an approximate Normal distribution

$$\log \widehat{\psi} \sim Normal(\log \psi, s^2)$$

where ψ is the true value of the odds-ratio for the table, and

$$s^2 = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

(i) Using these facts, construct a two-sample Z-test of the hypothesis

$$H_0 : \log \psi_A = \log \psi_B$$

against a general alternative $H_1 : \log \psi_A \neq \log \psi_B$, where ψ_A and ψ_B are the true odds-ratios for population A and B data respectively. A suitable test statistic is

$$z = \frac{\log \hat{\psi}_A - \log \hat{\psi}_B}{\sqrt{s_A^2 + s_B^2}}$$

which has a standard normal distribution if H_0 is true.

(ii) If this null hypothesis is rejected, the odds-ratios, and hence the patterns of associations for the two populations' data, are significantly different. Comment on the result in (b), in light of the odds-ratio comparison.

NOTE: In (c), $\log$ denotes natural logarithm, sometimes written ln.

3. The differential expression of five genes in two tumour types is to be investigated. For each tumour type, 16 microarray experiments are carried out and summary statistics (sample mean and sample variance) of the gene expression measurement, relative to expression in healthy tissue, are recorded below:

	TUMOUR TYPE 1		TUMOUR TYPE 2	
	MEAN	VARIANCE	MEAN	VARIANCE
GENE 1	0.940	0.923	-1.766	2.108
GENE 2	0.127	0.366	0.171	0.797
GENE 3	0.249	0.456	1.988	1.229
GENE 4	-0.185	0.736	0.004	1.107
GENE 5	0.232	0.402	-3.040	0.681

Assume that the expression samples are independently normally distributed.

(a) Carry out **appropriate** tests to assess whether the genes are differentially expressed between the two tumour types. Justify the test used in each case, and if necessary suggest other tests that may be more appropriate. Explain your choice of significance level.

(b) In the analysis above, we assume that any observed differences between expression measurements are genuinely due to biological causes. Discuss some of the statistical problems that arise for microarray measurements when a number of replicate arrays are used, and summarize simple methods that can be used to check for and rectify these problems.

(c) Describe in detail one method that allows the normality assumption in (a) to be relaxed.

4. In the analysis of gene-expression data, partitioning a large number of genes or gene profiles (where expression is recorded across a number of time points) into subsets that are similar in terms of expression or regulatory pattern is often of considerable interest, as it gives insight into possible functional similarity or co-regulation.

(a) Discuss how **hierarchical clustering** can be used to cluster the expression data. In your discussion, explain (with examples where appropriate) the following aspects of the procedure

(i) the **distance** or **similarity** measure between individual objects

(ii) the difference between **agglomerative** and **divisive** hierarchical clustering

(iii) different forms of **linkage** method that determine how the clustering method proceeds.

(b) Describe how the **number** of clusters present in the data may be determined.

(c) Once an expression data set has been partitioned into a collection of K clusters (labelled $1, 2..., K$ say), classification of new expression measurements/profiles is possible.

Suppose that n_k observations in the original sample have been clustered into cluster k. Suppose that y represents a new expression measurement (or profile). Then y may be classified to one of the clusters using the Bayes rule: let

$$q_k = \frac{p\left(y|k\right) p_k}{\sum\limits_{l=1}^{K} p\left(y|l\right) p_l}$$

where

- $p\left(y|l\right)$ is the probability density of observing measurement y from cluster l, for $l = 1, 2, ..., K$.
- p_l is the prior probability that the new measurement arises from cluster l, for $l = 1, 2, ..., K$

The set of values $\{q_1, ..., q_K\}$ specify a (posterior) probability distribution on cluster membership, in light of the new data. The classification is completed by choosing the largest q_k.

(i) Describe a method of constructing a suitable $p\left(y|k\right)$ from the results of the original clustering.

(ii) Discuss how the adequacy of the classification procedure could be assessed.