**UNIVERSITY OF LONDON**
**IMPERIAL COLLEGE LONDON**

**Examination for the Master of Science**

**in**

**Bioinformatics**

# PROBABILITY AND STATISTICS

**Wednesday 9th March 2004**
**10.00am to 12.00pm**

**Answer any two questions**

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

1. Linkage disequilibrium (LD) testing can be used to assess whether there is any evidence of association between the alleles at different marker loci, which may indicate meaningful haplotype structure.

In a study of a genetically isolated (and therefore genetically homogeneous) human population, a two-marker, biallelic marker system, with genotypes $\{A, a\}$ at marker locus 1, and $\{B, b\}$ at marker locus 2, was to be studied. Tissue samples from 24 individuals were obtained, genotyped at these two loci, and by thorough molecular investigation, the haplotype information was also discovered. A summary of haplotype information for the sample is contained in the table below:

| | Marker 2 | | |
|---|---|---|---|
| Marker 1 | $B$ | $b$ | Total |
| $A$ | 20 | 16 | 36 |
| $a$ | 4 | 8 | 12 |
| Total | 24 | 24 | 48 |

that is, for example, there were 20 $AB$ haplotypes, 16 $Ab$ haplotypes etc, and therefore 36 $A$ alleles observed in the sample.

(a) A single haplotype unit is selected at random from the 48 in the study. Evaluate

(i) the probability that it is $ab$,

(ii) the probability that it contains a $b$ allele

(iii) the conditional probability that it contains a $b$ allele, if it is observed to contain an $a$ allele.

Express your solutions in formal probability notation, giving full details and justfication of the computations used.

(b) To test for LD in this sample, a Chi-squared test for association can be used on the haplotype data. The test, based on the test 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 the alleles assort independently, 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 $n_{..} = 48$ is the total number of haplotypes. If $H_0$ is true, $\chi^2$ 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) Carry out a test of the same hypothesis based on the Likelihood Ratio (LR) statistic

$$LR = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

which has the same asymptotic Chi-squared(1) null distribution, where log corresponds to natural log, or ln.

(d) The null distribution, crucial in the significance tests in (b) and (c), is only appropriate when the sample size is large. Explain how tests based on the same test statistics could be used in an **exact** test of the same hypothesis where the row and column totals are held **fixed** in the sample.

2. (a) The Poisson process is deemed to be an appropriate model for the locations of occurrences of a particular four nucleotide motif. The motif is believed to occur, on average, once every 5000 nucleotides in the genomic region under study.

(i) State the distribution of the discrete random variable, $N$, that counts the number of occurrences of the motif in given genomic segment of length 10000 nucleotides, and show that

$$P[N < 2] = e^{-\lambda}(1 + \lambda)$$

for some parameter $\lambda > 0$ to be identified.

(ii) Evaluate $P[N < 2]$.

(iii) The distances between occurrences of the motif are continuous random variables $X_1, X_2, ...$with cdf

$$F_1(x) = 1 - e^{-\lambda x} \qquad x > 0$$

that is, the distances are independent $Exponential(\lambda)$ random variables. Let

$$M_n = \max\{X_1, ..., X_n\}$$

be the maximum observed between the first $n$ occurrences of the motif. Using the result for the cdf of the maximum of $n$ independent and identically distributed random variables,

$$F_{M_n}(x) = \{F_1(x)\}^n$$

find the formula for $F_{M_n}$.

(iv) Find $P[M_n > 30000]$ if $n = 20$ motifs are considered.

(b) An alternative probability model for the inter-occurrence distances is based on the *Power Law* distribution, defined by the cdf

$$F_2(x) = 1 - \left(\frac{\beta}{\beta + x}\right)^{\nu} \qquad x > 0$$

for parameters $\nu, \beta > 0$. In addition, if

$$L_n = \min\{X_1, ..., X_n\}$$

is the minimum derived from a sample of distances of size $n$, then it can be shown that

$$F_{L_n}(x) = 1 - \{1 - F_2(x)\}^n.$$

(i) Show that, for $x > 0$

$$P[L_n > x] = \left(\frac{\beta}{\beta + x}\right)^{n\nu}$$

and hence deduce that $L_n$ also has a power-law distribution.

(ii) Explain how $L_n$ can be used as a test statistic in a test of the hypothesis that the Power-Law model is adequate to explain the pattern of inter-occurrence distances. If $L_n$ is observed to be 15000 for $n = 20$, and $\nu = 1$ and $\beta = 95000$ characterize the hypothesized Power-Law model, carry out a test at the $\alpha = 0.05$ significance level for the observed data.

3. The differential expression of a gene in two tissue types is measured using cDNA microarray experiments. The expression, relative to some common baseline is measured and recorded on the log scale, for $n_1 = 9$ Type 1 tissue samples, and $n_2 = 12$ Type 2 tissue samples.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TYPE 1 : | 2.30 | 2.85 | 1.67 | 2.44 | -0.07 | 3.03 | 2.40 | 0.70 | -0.94 | | | |
| TYPE 2 : | 0.24 | -0.03 | -1.60 | -1.43 | -1.94 | -0.62 | 0.28 | -1.68 | -0.18 | 1.79 | -2.69 | 0.46 |

(a) Carry out tests, at the $\alpha = 0.05$ significance level, of the hypotheses

(i) the gene is **up-regulated** in the type I tissue relative to the baseline

(ii) the gene is **differentially expressed** in the type II tissue, relative to baseline

(iii) the gene is **differentially expressed** in the type II tissue, relative to the type I tissue.

Assume that the two samples are independent normal samples. Explain in full the tests used, and the assumptions made.

(b) The normality assumption for the data in (a) is critical if parametric tests are to be used.

(i) Give two features of a data sample that might indicate that the data are not well-modelled by a normal distributon

(ii) Briefly outline a method for hypothesis testing for data that are not normally distributed

(c) A typical cDNA microarray experiment has information on many thousands of genes. Discuss the issues involved in statistical hypothesis testing for differential expression of the genes in such an experiment, giving details of one specific technique for adjusting the testing procedure in this context.

4. A simple classification rule to discriminate between two classes of proteins uses a single predictor $X$ (derived from a number of physical and chemical features, and measured on a continuous scale) and states that

$$X \leq C_R \Longrightarrow \text{Class 1}$$

$$X > C_R \Longrightarrow \text{Class 2}$$

for some threshold $C_R$. The predictor is believed to have a normal distribution within each class

$$\text{Class 1} \quad : \quad X \sim N\left(\mu_1, \sigma_1^2\right)$$

$$\text{Class 2} \quad : \quad X \sim N\left(\mu_2, \sigma_2^2\right)$$

(a) Using Bayes Theorem, and presuming the parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ are known prior to the classification, show that a protein for which $X = x$ is classified to class 1 if

$$\frac{f_1(x)\, p_1}{f_1(x)\, p_1 + f_2(x)\, p_2} > c$$

for some $c$, $p_1$ and $p_2$ with $0 \leq c, p_1, p_2 \leq 1$, and functions $f_1$ and $f_2$. Explain carefully each term in this expression.

(b) Using this expression, give an expression for constant $C_R$

(c) Explain how such a classification procedure would be implemented when the parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ are not known, but training data (where the class labels are known for given predictor values) are available.

(d) Explain how the accuracy of the classification procedure should be assessed. In particular, explain the role of **cross-validation** in the accuracy assessment.