UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY & MEDICINE

Examination for the Master of Science in Bioinformatics 2003

PROBABILITY AND STATISTICS

Thursday 6th March 2003
10.00am to 12.00pm

Answer any two questions

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

1. Proteins are to be classified according to aspects of their secondary structure, namely

   (a) whether or not the total number of α-helices is less than five
   (b) whether or not the primary sequence length is at least sixty residues.

   In a collection of 250 proteins extracted from a protein database, it was discovered that

   - 81 had fewer than five α-helices,
   - of those that had fewer than five α-helices, **twice as many** proteins were observed to be greater than or equal to sixty residues in length than the number which were shorter than sixty residues in length, and
   - 139 had more than α-alpha helices **and** were greater than or equal to 60 residues in length.

The various probabilities for a randomly selected protein from this collection are to be computed.

(i) Express the information contained above using the mathematical notation of probability, conditional probability and joint probability. Denote by A the event that a randomly selected protein from the 250 selected has fewer than five α-helices, and by B the event that the protein is greater than or equal to sixty residues in length.

(ii) Compute the probability $P(B)$.

(iii) Compute the conditional probability a protein has fewer than five α-helices, **given** that it is greater than or equal to sixty residues in length.

(iv) Complete the following table with numerical values for the eight probabilities

	A	A'	
B	$\mathrm{P}(A \cap B)$	$\mathrm{P}(A' \cap B)$	$\mathrm{P}(B)$
B'	$\mathrm{P}(A \cap B')$	$\mathrm{P}(A' \cap B')$	$\mathrm{P}(B')$
	$\mathrm{P}(A)$	$\mathrm{P}(A')$	

(v) Are events A and B **independent** ? Justify your answer.

(vi) Suppose now that the collection of 250 proteins is to be used to make inference about the relationship between α-helix composition and primary sequence length in general. After forming the appropriate 2×2 table of counts, use the χ^2 test statistic

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

to test the null hypothesis H_0 that row and column classification are independent, 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 $n_{..} = 250$ is the total number of proteins in the collection. Recall that, if H_0 is true, then χ^2 has an approximate *Chisquared* distribution with 1 degree of freedom.

2. This question concerns the discrete probability distributions that arise in the context of a search of a genomic segment for a specific motif. The search is carried out on adjacent blocks, and the motif is either **present** in a block, or **not present;** the probability that it is present in a block is θ, and the blocks can be regarded as probabilistically independent.

(i) If n blocks are inspected, state the probability distribution of random variable X that records the total number of blocks in which a motif is found. If $n = 10$, compute

$$P\left[X \leq 2\right]$$

if $\theta = 0.01$. Report the result exactly with no rounding.

(ii) If, instead, $n = 1000$, probability calculations relating to X can be approximated using the Central Limit Theorem, that is, for any x in the range $0 \leq x \leq n$

$$P\left[X \leq x\right] \approx \Phi\left(\frac{x - n\theta}{\sqrt{n\theta\left(1-\theta\right)}}\right)$$

where Φ is the standard normal cumulative distribution function given in tables. Using this approximation for $n = 1000$ and $\theta = 0.01$, find the approximate probability that $X > 12$. Here, round all calculations to 4 decimal places.

(iii) If the number of blocks inspected (starting from block 1 and proceeding along the sequence) before a motif is found is random variable Y, state the distribution of Y, and find the functional form of the probability

$$P\left[Y > y\right]$$

for general $y = 1, 2, 3, ...$ Evaluate this probability for $y = 10$ if $\theta = 0.01$.

(iv) Using the conditional probability definition, show that for whole numbers $y_2 > y_1$

$$P\left[Y > y_2 | Y > y_1\right] = P\left[Y > y_2 - y_1\right]$$

This is the *lack of memory* property.

(v) Suppose that the first $n = 20$ motifs discovered are separated by the following numbers of blocks:

57	74	5	69	93	155	77	50	34	82
34	183	70	255	1	473	5	158	76	102

Let $y_{\max}$ denote the maximum observed between-motif distance in this sample. Using the observed value of $y_{\max}$ as a test statistic, and the result from (iii), carry out a hypothesis test of

$$H_0 : \theta = 0.01$$
$$H_1 : \theta > 0.01$$

at significance level $\alpha = 0.05$

[*Recall that the maximum order statistic*

$$Y_{\max} = \min\left\{X_1, ..., X_n\right\}$$

has distribution function

$$F_{Y_{\max}}(y) = \left\{F_X(y)\right\}^n$$

where $X_1, ..., X_n$ are independent and identically distributed random variables with distribution function F_X.]

3. In a **two sample** hypothesis test

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

for normally distributed data samples with common standard deviation σ **known** of sizes n_1 and n_2 respectively **,** the test statistic is

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

and has a standard normal distribution **if** H_0 **is true**

(i) Carry out the hypothesis test at the $\alpha = 0.05$ significance level if the summary statistics for the two samples are

	Sample Size n	Sample mean $\overline{x}$	Adjusted sample variance s^2
SAMPLE 1	18	20.53	2.24
SAMPLE 2	24	18.24	1.96

and $\sigma = 1$.

(ii) Suppose now that σ is not known. Compute the pooled estimate of σ^2

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

(iii) Again, test the hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

at the $\alpha = 0.05$ significance level, assuming that the common variance σ^2 is unknown.

(iv) Is there any evidence that σ is not equal in the two samples ? Justify your answer by completing a two sample test of the hypotheses

$$H_0 : \sigma_1 = \sigma_2$$
$$H_1 : \sigma_1 \neq \sigma_2$$

at the $\alpha = 0.05$ significance level.

(v) Describe methods for hypothesis testing for statistical differences between the two samples if all the 42 individual observations are available, but the data **cannot** be assumed to be normally distributed.

(vi) Briefly describe the importance of **multiple testing corrections** in statistical testing.

4. The 601 nucleotide sequence for chicken $\beta-$globin is available from NCBI (accession number J00860), and has the following summary statistics:

Nucleotide	A	C	G	T	Total
Count	135	202	148	116	601

(i) Complete the table of **fitted values** under a model where it is assumed

$$p_A = p_C = p_G = p_T$$

for the four nucleotide proportions.

(ii) Compute the **Likelihood Ratio statistic** (LR)

$$LR = 2\sum_{i=1}^{4} n_i \log \frac{n_i}{\widehat{n}_i}$$

to test the hypothesis

$$\begin{aligned} H_0 &: p_A = p_C = p_G = p_T \\ H_1 &: \text{not } H_0 \end{aligned}$$

(iii) Carry out a (one-sided) test of H_0 at the significance level of $\alpha = 0.01$. The degrees of freedom for the required null distribution is given by the usual $k - d - 1$ formula, where k is the number of categories, and d is the number of estimated parameters.

(iv) A test is required of the hypotheses

$$\begin{aligned} H_0 &: p_A = p_C = \theta \qquad p_G = p_T = \phi \\ H_1 &: \text{not } H_0 \end{aligned}$$

with probabilities (θ, ϕ), with $2(\theta + \phi) = 1$, not specified explicitly. Test these hypotheses using any appropriate testing procedure.

(v) The longest repeated nucleotide "word" (that is, a sequence which appears twice in its entirety in the complete sequence) in the chicken $\beta-$globin sequence is

$$GCCAGGCTGC$$

Explain how **simulation-based** (Monte Carlo) test procedures can used to assess whether the appearance of a repeated word of this length is surprising in such a segment, assuming a null model that the chicken $\beta-$globin sequence is generated by random and independent inclusion of nucleotides with equal probabilities.

(vi) In a Monte Carlo test of the hypothesis in (v), 20000 simulated sequences were generated, and the lengths of the longest repeated nucleotide sequence were recorded for each simulation, and tabulated as follows:

Length	< 6	7	8	9	10	11	≥ 12	Total
Count	10538	5643	2123	1234	420	37	5	20000

Is there sufficient evidence in the results of the Monte Carlo test to reject the hypothesis at the $\alpha = 0.05$ significance level ? Justify your answer.

(vii) Explain how a **permutation test** can be used to test whether the chicken $\beta-$globin sequence is randomly and independently generated with the nucleotide frequencies **unspecified**.