UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY & MEDICINE

Examination for the Master of Science
in
Computational Genetics and Bioinformatics
2002

PROBABILITY AND STATISTICS

Students in Master of Science in Computational Genetics and Bioinformatics

Friday 18th January 2002 (2pm to 3.30pm)

Answer any two questions.
A Formula Sheet and Statistical Tables are provided. Calculators may be used

1. An exon prediction program is reasonably accurate in predicting the presence or absence of an exon in a given DNA sequence. In fact, the program **correctly** reports that an exon **is** contained in the sequence (event E), **given** that an exon actually is present (event F), with probability 0.85, but **incorrectly** reports that an exon is contained in the sequence when in fact no exon is present with probability 0.10. It can be assumed that the probability that an exon is actually contained in a novel sequence of a given length is 0.001.

(i) Express the three pieces of information contained above using the mathematical notation of probability and conditional probability.

(ii) Compute, using the Total Probability result, the probability $\mathrm{P}(E)$ that the program reports that an exon is present when it is used to analyze a novel sequence.

(iii) Compute, using Bayes Theorem or otherwise, the conditional probability that a novel sequence actually contains an exon, **given** that the program reports that an exon is present. Comment on your answer.

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

	E	E'	Sum
F	$\mathrm{P}(E \cap F)$	$\mathrm{P}(E' \cap F)$	$\mathrm{P}(F)$
F'	$\mathrm{P}(E \cap F')$	$\mathrm{P}(E' \cap F')$	$\mathrm{P}(F')$
Sum	$\mathrm{P}(E)$	$\mathrm{P}(E')$	

(v) Are events E and F **independent** ? Justify your answer.

(vi) The program is to be used to analyze a large number n of novel sequences. Let X be the discrete random variable that counts the total number of **correctly classified** sequences (out of n), that is, the number of sequences for which the program correctly reports the presence **or** absence of an exon.

Identify the probability distribution of X. Give the name and parameters of the distribution.

PLEASE TURN OVER

2. (a) The **Normal** distribution plays a central role in statistical analysis. The **standard normal** density function (pdf) for random variable Z is denoted $\phi(z)$ and is given by

$$\phi(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

and the standard normal distribution function (cdf) is denoted Φ and defined by

$$\Phi(z) = F_Z(z) = \mathrm{P}\left[Z \leq z\right] = \int_{-\infty}^{z} \phi(t)\ dt$$

(i) Statistical tables of Φ only contain values of $\Phi(z)$ for $z > 0$. Explain, using a sketch if necessary, why this is sufficient for all probability calculations involving the standard normal distribution, even for $z \leq 0$.

(ii) Using the probability tables provided, evaluate $\mathrm{P}[Z \leq 1.2]$, $\mathrm{P}[Z > 2.0]$ and $\mathrm{P}[-0.5 \leq Z < 1.0]$.

(iii) Now suppose that a general Normal random variable X is defined using a transformation of Z by $X = \mu + \sigma Z$. Show, using probability arguments that

$$F_X(x) = \mathrm{P}\left[X \leq x\right] = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

so that the tables of the standard Normal cdf are sufficient for **all** probability calculations involving the Normal distribution.

(b) In a **one sample** hypothesis test

$$H_0 : \mu = c$$
$$H_1 : \mu < c$$

(that is, with a **one-sided** alternative) for a Normal sample with σ **known,** the test statistic

$$z = \frac{\bar{x} - c}{\sigma/\sqrt{n}}$$

(where n is the sample size and $\bar{x}$ is the sample mean) has a standard normal distribution **if H_0 is true**

Carry out the hypothesis test at the $\alpha = 0.01$ significance level if

(i) $n = 12, \bar{x} = 18.20, \sigma = 2.5$ and $c = 20.0$

(ii) $n = 28, \bar{x} = 19.6, \sigma = 2.5$ and $c = 20.0$

Explain how to proceed with a test of the same hypothesis if σ is **unknown**

Identify the critical values (for $\alpha = 0.01$, in the same one-sided test) if σ is unknown for the data in (b) (i) and (ii)

3. The probability of a "character" (nucleotide/amino acid) match at any given position in a sequence alignment calculation, assuming that the two sequences have unrelated evolutionary origin is given by

$$p_{MATCH} = \sum_{i=1}^{d} p_i^2$$

where d is the number of characters (so that $d = 4$ for DNA sequences and $d = 20$ for protein sequences) and $p_i, i = 1, ..., d$ are the probabilities of observing each character type at any individual position. The alignment of the two sequences can be coded as a binary sequence using "1" (success) to indicate a match and "0" (failure) to indicate a non-match.

(i) Explain briefly the probabilistic assumptions underlying the formula for p_{MATCH}, and compute p_{MATCH} for a DNA sequence where the probabilities for the four nucleotides A, C, G and T, are 0.275, 0.225, 0.20 and 0.30 respectively.

(ii) Find a formula in terms of p_{MATCH} for the probability that a **run** of matched nucleotides of length x begins at position 1 in the sequence and ends at position x (with a non-match at position $x + 1$), denoted $p_{RUN}(x)$, and explain why, for $x = 0$

$$p_{RUN}(0) = 1 - p_{MATCH}$$

is a sensible definition.

(iii) Find the probability of such a run when $x = 5$. Show all working.

(b) (i) For a long DNA sequence of length 100000 bases, the positions at which a particular nucleotide pattern are detected within the sequence form (approximately) a Poisson process with parameter $\lambda = 0.00005$.

Compute the probability that **more than two** patterns are detected in the sequence What is the **expected** number of occurrences in a sequence of this length ?

(ii) For the sequence in (b)(i), the distances (measured in numbers of positions) between successive occurrences of the pattern are independent and identically distributed continuous random variables each having an *Exponential* (λ) distribution.

Show, by calculating the probability distribution of the **minimum order statistic**, that the **shortest** distance between k successive occurrences of the pattern is a random variable $Y_{\min}$ say with

$$Y_{\min} \sim \textit{Exponential}(k\lambda)$$

[*Recall that the minimum order statistic*

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

has distribution function

$$F_{Y_{\min}}(y) = 1 - \{1 - F_X(y)\}^n$$

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

PLEASE TURN OVER

4. (a) Two of the longest exons of the BRCA2 gene are exons 9 and 10, for which a summary of total numbers of nucleotides is included in the following table:

	Nucleotide				
	A	C	G	T	Total
Exon 9	433	192	200	291	1116
Exon 10	1884	784	873	1391	4932
Total	2317	976	1073	1682	6048

A test of the null hypothesis H_0, that the marginal probabilities of the four nucleotides are identical for both exons, is required.

(i) Complete the table of **fitted values** assuming H_0 is true, using the usual estimates of the nucleotide probabilities.

(ii) Compute either the **Chi-squared statistic** χ^2 or the **Likelihood Ratio statistic** (LR) for these data.

(iii) Carry out a (one-sided) test of H_0 at the significance level of $\alpha = 0.05$

(b) Sequence data for the intron separating exons 9 and 10 is also available. If the two exons are pooled, then we have the following table:

	Nucleotide				
	A	C	G	T	Total
Exons 9 and 10	2317	976	1073	1682	6048
Intron	806	551	549	970	2876
Total	3123	1527	1622	2652	8924

Is there any evidence that the nucleotide probabilities are **different** for this intron when compared with the probabilities for exons 9 and 10 ? Justify your answer by using a hypothesis test.

[Show all steps in your working]