BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

MOCK EXAMINATION

Answer two questions. Questions carry 25 marks each.

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

1. A base-calling procedure is very accurate in determining the nucleotides in a DNA sequence, in that it correctly identifies each base with high probability and only rarely misclassifies bases. Let E_A, E_C, E_G and E_T be the events that the program **classifies** a base as A, C, G and T respectively, and let F_A, F_C, F_G and F_T be the events that base under study is **actually** an A, C, G and T respectively. Suppose that, prior to any analysis being carried out it is assumed that

$$\mathrm{P}(F_A) = p_A = 0.30 \qquad \mathrm{P}(F_C) = p_C = 0.20 \qquad \mathrm{P}(F_G) = p_G = 0.20 \qquad \mathrm{P}(F_T) = p_T = 0.30$$

and that the **conditional** probabilities of **(mis)classification** of a base, **given** its actual type are given by the following table

		Is Called As			
		A	C	G	T
	A	0.900	0.025	0.025	0.050
The Base	C	0.025	0.850	0.100	0.025
	G	0.025	0.100	0.850	0.025
	T	0.050	0.025	0.025	0.900

so that, for example, from the top row

$$\mathrm{P}(E_A|F_A) = 0.900 \qquad \mathrm{P}(E_C|F_A) = \mathrm{P}(E_G|F_A) = 0.025 \qquad \mathrm{P}(E_T|F_A) = 0.050$$

and so on.

(i) Using the Total Probability formula, compute the probability that an unknown base under analysis is classified as $i \in \{A, C, G, T\}$, that is, compute

$$\mathrm{P}(E_i) = \sum_{j \in \{A,C,G,T\}} \mathrm{P}(E_i|F_j)\,\mathrm{P}(F_j) \qquad \text{for each } i \in \{A, C, G, T\}$$

(ii) Compute, using Bayes Theorem or the conditional probability formula, the conditional probability that a base is **actually** an A, **given** that is classified as an A.

Compute also the three conditional probabilities that a base is **actually** an A, given that it is classified as C, G and T.

(iii) Express, in mathematical notation and in terms of the events above, the event that a base is misclassified (event M say), and compute the Total Probability that M occurs.

(iv) Using the answer from (iii), compute the Total Probability that a sequence of length k bases is completely correctly identified, assuming that successive bases appear in the sequence independently, and that the classification of successive bases is also an independent process.

(v) If the bases are classified sequentially from the beginning of the sequence, find an expression for the probability that the first misclassification occurs at base position x. To which probability distribution does this experimental context correspond ?

(vi) Compute the length L of the longest sequence that can be analyzed so that there is a probability of at least 0.95 of there being **at most one misclassification**.

2. (a) In the Poisson process model for events that occur at random in continuous time with constant rate λ, there are three related probability distribution results

- the numbers of events occurring in disjoint intervals of lengths $t_1, t_2, t_3, \ldots$ are independent random variables $X_1, X_2, X_3, \ldots$ with $X_i \sim \textit{Poisson}\,(\lambda t_i)$
- the times between the occurrences of events are independent continuous random variables $T_1, T_2, T_3, \ldots$ with $T_i \sim \textit{Exponential}\,(\lambda)$
- the time of the nth event is a continuous random variable Y_n with $Y_n \sim \textit{Gamma}\,(n, \lambda)$

Suppose that the locations of the occurrences of a small nucleotide pattern in a large genomic segment (approximately) follow a Poisson process with rate parameter $\lambda = 0.001$

(i) Let $t_1 = 5000$ (bases). Compute $\mathrm{P}[X_1 = 2]$, $\mathrm{P}[X_1 \geq 0]$ and $\mathrm{P}[X_1 < 4]$

(ii) Suppose that the first 20000 bases of the segment are split into 10 sections of equal size. Let $X_1, X_2, \ldots, X_{10}$ be the counts of the numbers of occurrences of the pattern in each of these sections. Let

$$S_{10} = \sum_{i=1}^{10} X_i$$

be the random variable recording the **total** number of occurrences of the pattern in the first 20000 bases.

Compute $\mathrm{P}[S_{20} \geq 3]$
(iii) Compute the probability that the first occurrence of the pattern occurs **beyond** base position 7500.

(iv) The $\textit{Gamma}\,(n, \lambda)$ distribution can be approximated by the $\textit{Normal}\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right)$ distribution, so that the cdf of Y_n can be approximated

$$F_{Y_n}(y) \approx \Phi\left(\frac{\lambda y - n}{\sqrt{n}}\right)$$

where Φ is the standard normal cdf. Using the statistical tables provided, compute an approximate probability that the 10th occurrence of the pattern occurs somewhere in the first 10000 bases.

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

$$\begin{aligned} H_0 &: \mu = c \\ H_1 &: \mu \neq c \end{aligned}$$

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

$$t = \frac{\bar{x} - c}{s/\sqrt{n}}$$

(where n is the sample size, $\bar{x}$ is the sample mean and s^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is the sample variance) has a $\textit{Student}(n-1)$ distribution **if H_0 is true**

(i) Carry out a test of $H_0 : \mu = 0$ vs $H_1 : \mu \neq c$ at an appropriately chosen significance level for the sample data

$$1.90, -1.10, 1.20, -0.11, -1.90, 2.20, -0.87, 0.76, -2.90, 2.70$$

(ii) Now carry out a test of the hypothesis

$$\begin{aligned} H_0 &: \sigma^2 = 1 \\ H_1 &: \sigma^2 > 1 \end{aligned}$$

using the test statistic

$$Q = \frac{(n-1)s^2}{\sigma^2}$$

that has a χ^2_{n-1} distribution if H_0 is true.

3. In a sequence alignment exercise for two DNA sequences of equal length, a binary numerical sequence can be constricted to record the positions of exact matches, that is for the two sequences

Sequence 1	A	A	C	C	T	C	A	A	G	A
Sequence 2	A	C	C	A	T	C	A	G	G	G
Match	1	0	1	0	1	1	1	0	1	0

(i) Let p_{MATCH} be the probability that a match is observed at a given location under the null hypothesis H_0 that the sequence are **unrelated** in evolutionary terms and that nucleotides are observed independently with probabilities 0.3, 0.2, 0.2 and 0.3 for A, C, G and T respectively.

Compute p_{MATCH}.

(ii) For two sequences of length $N = 100$, state the distribution (if H_0 is true) of the statistic X_{MATCH} that records the number of matches between the two sequences.

(iii) Explain how to use the observed value of X_{MATCH} , x_{MATCH} say, to test the hypothesis H_0, explaining how the different components of the hypothesis test are computed.

Refer to the fact that appropriate critical values for a two-sided test at the $\alpha = 0.05$ level are 18 and 35

(b) An alignment of 5 sequences of length $N = 100$ is required; a similar hypothesis test (where now H_0 states that the five sequences are evolutionarily unrelated and that nucleotides are sampled independently with the probabilities given above) to that above is to be carried out.

Explain in detail how this new hypothesis test should proceed, again explaining how the different components of the test are computed.

(c) In a binary sequence alignment, it can be shown (see lecture notes) that Y_n, the random variable recording the length of the **longest run** of matches in a collection of n such runs, is a **maximum order statistic** random variable with cumulative distribution function

$$F_{Y_n}(y) = \mathrm{P}\left[Y_n \leq y\right] = \left(1 - p_{MATCH}^{y-1}\right)^n \qquad y = 0, 1, 2, ...$$

(i) Explain how this result can be used as the basis of a hypothesis test relating to the alignment of the two sequences.

(ii) Using the approximation for the p-value when the observed maximum run length for two sequences of length N is y_n

$$p \approx 1 - (1 - p_{MATCH}^{y_n})^{(1-p_{MATCH})N} \approx 1 - \exp\left\{-(1 - p_{MATCH})N p_{MATCH}^{y_n}\right\}$$

find the approximate critical value for the test if $N = 1000$, using the hypothesized values for the nucleotide probabilities described above.

(iii) To what value does the critical value change if an alignment of five sequences is considered.

4. (a) Data from a large intron separating exons 17 and 18 in a genomic segment containing the BRCA2 gene reveals the following nucleotide distribution:.

Nucleotide				
A	C	G	T	Total
1821	1348	1304	2380	6853

(i) Compute estimates of the nucleotide probabilities $p_1 = p_A, p_2 = p_C, p_3 = p_G$ and $p_4 = p_T$ from these data

(ii) Test the hypothesis that the nucleotide probabilities are equal, that is

$$H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$$

using the test statistics

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

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

where $\hat{n}_i$ for $i = 1, 2, 3, 4$ are the fitted values assuming that H_0 is true; both of these statistics have a chi-squared distribution with 3 degrees of freedom if H_0 is true.

(b) Count data from two DNA sequences was collected

	Nucleotide				
	A	C	G	T	Total
Sequence 1	250	140	180	230	800
Sequence 2	320	270	310	300	1200
Total	570	410	490	530	2000

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

(i) Complete the table of **fitted values**

$$\hat{n}_{ij} = n_{i.}\hat{p}_j = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, ..., r,\ j = 1, ..., c$$

assuming H_0 is true, where $n_{i.}$ is the total of the ith row, $n_{.j}$ is the total of the jth column, and n is the total number of observations..

(ii) Compute the **Chi-squared statistic** χ^2 and the **Likelihood Ratio statistic** LR

(iii) Carry out a test of H_0 at the significance level of $\alpha = 0.01$

Recall that, here, the test statistics are defined as

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \qquad LR = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

both of which have a chi-squared distribution with 3 degrees of freedom if H_0 is true.

[Note: in lecture notes, the factor of 2 was omitted from the LR statistic definition - sorry]