BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

EXAMINATION JANUARY 2002: SOLUTIONS

1. (i) We have

$$\mathrm{P}(E|F) = 0.85 \qquad \mathrm{P}(E|F') = 0.10 \qquad \mathrm{P}(F) = 0.001$$

3 MARKS

(ii) Using the Total Probability formula

$$\mathrm{P}(E) = \mathrm{P}(E|F)\,\mathrm{P}(F) + \mathrm{P}(E|F')\,\mathrm{P}(F') = 0.85 \times 0.001 + 0.10 \times 0.999 = 0.10075$$

4 MARKS

(iii) Using Bayes Theorem to compute $\mathrm{P}(F|E)$

$$P(F|E) = \frac{\mathrm{P}(E|F)\,P(F)}{\mathrm{P}(E)} = \frac{0.85 \times 0.001}{0.85 \times 0.001 + 0.10 \times 0.999} = 0.00844$$

to 5 d.p

.

4 MARKS

Comment: Although the program is reasonably accurate (with low rates of false negatives and false positives) it is of little use overall as the prognostic probability is very small

2 MARKS

(v) The events are **not independent** as, for example

$$P(F|E) \neq P(F)$$

2 MARKS

(vi) As

$$P(E \cap F) = \mathrm{P}(E|F)\,P(F) = 0.85 \times 0.001 = 0.000085$$

and so on, we have

	E	E'	Sum
F	0.00085	0.00015	0.00100
F'	0.09990	0.89910	0.99900
Sum	0.10075	0.89925	

6 MARKS

(v) We have

$$X \sim Binomial(n, p)$$

where

$$p = \mathrm{P}(\text{“correct classification”}) = P(E \cap F) + P(E' \cap F') = 0.00085 + 0.89910 = 0.89995$$

4 MARKS

2. (a) (i) Due to the **symmetry** of the standard normal pdf around zero, we must have that

$$\phi(z) = \phi(-z)$$

and consequently

$$\Phi(z) = 1 - \Phi(-z)$$

Hence for $z \leq 0$, we can evaluate using this formula.

2 MARKS

(ii) From tables

$$\begin{aligned} \mathrm{P}\left[Z \leq 1.2\right] &= \Phi(1.2) = 0.8849 \\ \mathrm{P}\left[Z > 2.0\right] &= 1 - \Phi(1.2) = 1 - 0.9772 = 0.0228 \\ \mathrm{P}\left[-0.5 \leq Z < 1.0\right] &= \Phi(1.0) - \Phi(-0.5) = \Phi(1.0) - [1 - \Phi(0.5)] = 0.8413 - [1 - 0.6915] \\ &= 0.5328 \end{aligned}$$

4 MARKS

(iii) If $X = \mu + \sigma Z$ then

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

and hence the cdf of random variable X at point x is available by evaluating the standard normal cdf at $(x-\mu)/\sigma$. Hence ALL normal cdf calculations can be obtained from the standard normal cdf.

3 MARKS

(b) The $\alpha = 0.01$ critical value in a ONE_SIDED test lies at the 0.01 quantile of the standard normal, that is, at -2.3263.

2 MARKS

(i) First sample

$$z = \frac{\bar{x} - c}{\sigma/\sqrt{n}} = \frac{18.20 - 20}{2.5/\sqrt{12}} = -2.494$$

Test statistic **more extreme** than critical value $\Longrightarrow$ REJECT H_0

5 MARKS

(ii) Second sample

$$z = \frac{\bar{x} - c}{\sigma/\sqrt{n}} = \frac{19.60 - 20}{2.5/\sqrt{28}} = -0.847$$

Test statistic **not more extreme** than critical value $\Longrightarrow$ CANNOT REJECT H_0

5 MARKS

If σ is **not known**, need to use an estimate s of σ(derived from the sample data) and instead use a **one-sample t-test** with test statistic

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

which has a **Student(n-1)** distribution if the null hypothesis is true. The two critical values for the tests above would be

$$n = 12 \quad C_R = -2.718 \qquad\qquad n = 28 \quad C_R = -2.472$$

4 MARKS

3. Given

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

(i) This formula assumes that the nucleotides in each sequence are sampled **independently** from the same (multinomial) distribution within each sequence and between sequences, and uses the Total Probability result

$$p_{MATCH} = \Pr(\text{Match at given position}) = \sum_{i=1}^{d} \mathrm{P}(\text{Match at given position} \cap \text{Character is } i) = \sum_{i=1}^{d} (p_i \times p_i) = \sum_{i=1}^{d} p_i^2$$

3 MARKS

For the probabilities given,

$$p_{MATCH} = 0.275^2 + 0.225^2 + 0.20^2 + 0.30^2 = 0.25625$$

1 MARK

(ii) For such a run, we need x consecutive matches followed by a non-match, which corresponds to a $Geometric(\theta)$ probability calculation with parameter $\theta = 1 - p_{MATCH}$. Hence the required formula is

$$p_{RUN}(x) = (1-\theta)^x \theta = p_{MATCH}^x (1 - p_{MATCH}) \qquad x = 1, 2, 3$$

4 MARKS

So for $x = 0$, the formula should be

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

which is sensible as a run of length 0 corresponds to an immediate non-match.

2 MARKS

For $x = 5$, we have

$$p_{RUN}(x) = p_{MATCH}^5 (1 - p_{MATCH}) = (0.25625)^5 (1 - 0.25625) = 0.00082 \qquad \text{to 5 dp}$$

2 MARKS

(b) (i) If X counts the number of occurrences, then $X \sim Poisson(\mu)$ where $\mu = \lambda t = 0.00005 \times 100000 = 5$, and hence using the Poisson mass function formula

$$\mathrm{P}[X > 2] = 1 - \mathrm{P}[X \leq 2] = 1 - [\mathrm{P}[X = 0] + \mathrm{P}[X = 1] + \mathrm{P}[X = 2]] = 1 - \left[e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!}\right] = 0.87535$$

3 MARKS

The expected number of occurrences is $\mu = \lambda t$ (from notes).

2 MARKS

(ii) From notes, for $y > 0$

$$\begin{aligned} F_{Y_{\min}}(y) = \quad & \mathrm{P}[Y_{\min} \leq y] = 1 - \mathrm{P}[Y_{\min} > y] = 1 - \mathrm{P}[X_1 > y, ...X_k > y] \\ & = 1 - \prod_{i=1}^{k} \mathrm{P}[X_i > y] = 1 - \prod_{i=1}^{k} \{1 - \mathrm{P}[X_i \leq y]\} \\ & = 1 - \prod_{i=1}^{k} \{1 - F_{X_i}(y)\} = 1 - \{1 - F_X(y)\}^k \end{aligned}$$

and hence, as from notes, $F_X(y) = 1 - e^{-\lambda y}$ for the exponential,

$$F_{Y_{\min}}(y) = 1 - \left\{1 - (1 - e^{-\lambda y})\right\}^k = 1 - e^{-k\lambda y} \qquad y > 0$$

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

8 MARKS

4. (a)(i) The fitted values are

$$\hat{n}_{ij} = n_{i.}\hat{p}_j = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, 2,\ j = 1, 2, 3, 4$$

giving the following table of fitted values:

	Nucleotide				
	A	C	G	T	Total
Exon 1	427.542	180.095	197.994	310.369	1116
Exon 2	1889.458	795.905	875.006	1371.631	4932
Total	2317	976	1073	1682	6048

8 MARKS

(ii) Test statistics

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 2.558 \qquad LR = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij}\log\frac{n_{ij}}{\hat{n}_{ij}} = 2.566$$

4 MARKS

(iii) For both tests, compare with the 0.95 quantile of the χ^2_3 distribution, that is 7.81. Clearly, **both** tests indicate that there is **no evidence to reject** the hypothesis that the nucleotide probabilities are identical for the two exons..

3 MARKS

(b) For the new comparison, test against the **pooled** table where the fitted values are;

	Nucleotide				
	A	C	G	T	Total
Exons	2116.529	1034.883	1099.267	1797.321	6048
Intron	1006.471	492.117	522.733	854.679	2876
Total	3123	1527	1622	2652	8924

giving

$$\chi^2 = 94.221 \qquad LR = 95.891$$

which clearly now IS a (very) significant result, so that we can conclude that there is evidence to **reject** the hypothesis that the exon and intron distributions are identical.

10 MARKS