BIONFORMATICS MSc.
EXAMINATION 2004
SOLUTIONS

1. (a) We have the count table

	B	b	Total
A	20	16	36
a	4	8	12
Total	24	24	48

so that

(i) $P(ab) = \frac{8}{48} = \frac{1}{6}$

2 MARKS

(ii) $P(*b) = P(Ab) + P(ab) = \frac{24}{48} = \frac{1}{2}$

3 MARKS

(iii) $P(ab|a*) = \frac{P(ab)}{P(a*)} = \frac{P(ab)}{P(aB)+P(ab)} = \frac{8/48}{4/48+8/48} = \frac{2}{3}$

3 MARKS

(b) Fitted values under the independence hypothesis are

	B	b
A	$\frac{36 \times 24}{48}$	$\frac{36 \times 24}{48}$
a	$\frac{12 \times 24}{48}$	$\frac{12 \times 24}{48}$

$\Longrightarrow$

	B	b
A	18	18
a	6	6

giving a test statistic

$$\chi^2 = \frac{(20-18)^2}{18} + \frac{(16-18)^2}{18} + \frac{(4-6)^2}{6} + \frac{(8-6)^2}{6} = 1.778$$

The critical value in this (one-sided) test is the 0.95 quantile of the χ^2_1 distribution, that is, 3.841 from tables. Thus the test statistic does not exceed the critical value, and the test of independence is not rejected at the $\alpha = 0.05$ level.

6 MARKS

(c) The LR test statistic is

$$\begin{aligned} LR &= 2\left[20\log\frac{20}{18} + 16\log\frac{16}{18} + 4\log\frac{4}{6} + 8\log\frac{8}{6}\right] \\ &= 2\left[2.107 - 1.885 - 1.622 + 2.301\right] = 1.804 \end{aligned}$$

The critical value in this (one-sided) test is the 0.95 quantile of the χ^2_1 distribution, that is, 3.841 from tables. Thus the test statistic does not exceed the critical value, and the test of independence is not rejected at the $\alpha = 0.05$ level.

6 MARKS

(d) An **exact** testing procedure can be used; the exact distribution of the test statistic (either chi-squared or LR) can be computed by evaluating the test statistic for **all** tables in which the row and column tables are held at their observed values. This is akin to a permutation test, but with row and column constraints. The test statistic for the actual data table, t, is compared with the (discrete) distribution of evaluated test statistics for all possible, legitimate tables, and the exact p-value is computed as

$$\frac{\text{Number of tables for which the test statistic is not less than } t}{\text{Total number of legitimate tables.}}$$

The number of such tables is finite, but may be large; if this is the case, a simulation-based Monte Carlo estimate of the null distribution and p-value may be used.

5 MARKS

2. (a) (i) The distribution is $Poisson(10000\mu)$ with $\mu = 1/5000$. Hence $N \sim Poisson\,(2)$

2 MARKS

The required probability is obtained from the Poisson mass function as

$$P\,[N < 2] = P\,[N = 0] + P\,[N = 1] = e^{-\lambda} + \lambda e^{-\lambda} = e^{-\lambda}\,(1 + \lambda)$$

with $\lambda = 2$.

3 MARKS

(ii) $P\,[N < 2] = 0.406$.

1 MARK

(iii) By the given result

$$F_{M_n}\,(x) = \{F_1\,(x)\}^n = \left(1 - e^{-\mu x}\right)^n$$

2 MARKS

(iv) We have, with $n = 20$,

$$P\,[M_n > 30000] = 1 - F_{M_n}\,(30000) = 1 - \left(1 - e^{-6}\right)^{20} = 0.048$$

4 MARKS

(b) By the given result, if $X_i \sim PowerLaw\,(\nu, \beta)$

$$F_{L_n}\,(x) = 1 - \{1 - F_2\,(x)\}^n = 1 - \left\{\left(\frac{\beta}{\beta + x}\right)^{\nu}\right\}^n = 1 - \left(\frac{\beta}{\beta + x}\right)^{n\nu} \tag{1}$$

and hence

$$P\,[L_n > x] = 1 - F_{L_n}\,(x) = \left(\frac{\beta}{\beta + x}\right)^{n\nu}$$

Now by (1), it is evident that $L_n \sim PowerLaw\,(n\nu, \beta)$

5 MARKS

(v) The test statistic L_n was observed to be $l_n = 15000$; under H_0, we know from above that

$$\begin{aligned} P\,[L_n > l_n] &= \left(\frac{\beta}{\beta + l_n}\right)^{n\nu} = \left(\frac{5000}{5000 + l_n}\right)^{20\times 1} \\ &= \left(\frac{95000}{95000 + 15000}\right)^{20\times 1} = \left(\frac{95}{110}\right)^{20} = 0.053 \end{aligned}$$

This is a p-value in the test of the stated hypothesis; we have evaluated, under the null hypothesis, the probability of observing a test statistic more extreme that the onw we did observe. The chosen significance level is $\alpha = 0.05$; thus the hypothesized model cannot be rejected.

8 MARKS

3. (a) Use standard hypothesis testing framework

(i) One sample test of the hypotheses

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu > 0 \end{aligned}$$

-that is, a **one-sided test -** use a one sample t-test as variance is unknown. RESULT : reject H_0 at $\alpha = 0.05$ ($p = 0.009$)

4 MARKS

(ii) One sample test of the hypotheses

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0 \end{aligned}$$

- a **two-sided** test. Use a one sample t-test as variance is unknown. RESULT : cannot reject H_0 at $\alpha = 0.05$ ($p = 0.069$)

4 MARKS

(iii) A two sample test of the hypothesis

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

- a **two-sided** test. Use a two sample t-test as variance is unknown. Might also wish to test equal variances (see (b)). RESULT : reject H_0 at $\alpha = 0.05$ ($p = 0.002$)

4 MARKS

(b) (i) Presence of outliers, presence of skewness.

2 MARKS

(ii) Can use **non-parametric** test (i.e. Mann-Whitney/Wilcoxon, or Kolmogorov-Smirnov). We evaluate a test statistic based on ranks/empirical cdf, and use asymptotic theory to compute the null distribution, or exact methods to achieve the same end.

3 MARKS

(c) The key issue is multiple testing; we wish to maintain a **family-wise error rate (FWER)** across the family of tests below the target significance level. The Type I error rate needs to be controlled, and this is typically done using step methods, a special case of which is the Bonferroni correction. If a FWER of $\alpha = 0.05$ is required, then the significance level on each individual test, in a family of k, should be set to

$$\frac{\alpha}{k}$$

This is a conservative correction, but ensures that the target FWER is maintained.

8 MARKS

4. (a) Using the theorem of total probability, the predictor X has density

$$f(x) = f(x|\text{Class 1})P[\text{Class 1}] + f(x|\text{Class 2})P[\text{Class 2}]$$

and by Bayes Theorem, we have the posterior probability

$$\begin{aligned} P[\text{Class 1} \mid X = x] &= \frac{f(x|\text{Class 1})P[\text{Class 1}]}{f(x)} = \frac{f(x|\text{Class 1})P[\text{Class 1}]}{f(x|\text{Class 1})P[\text{Class 1}] + f(x|\text{Class 2})P[\text{Class 2}]} \\ &= \frac{f_1(x)\,p_1}{f_1(x)\,p_1 + f_2(x)\,p_2} \end{aligned}$$

where

$f_1(x)$ is the **class conditional density** for class 1
$f_2(x)$ is the **class conditional density** for class 2
p_1 is the prior probability for class 1
p_2 is the prior probability for class 1

Thus we classify to class 1 if the posterior probability is greater than some threshold c, which must take a value between zero and 1, and which is dependent on the relative losses due to misclassification.

8 MARKS

(b) Obtain C_R by manipulation of the normal densities f_1 and f_2, and solution of the equation

$$\frac{f_1(C_R)\,p_1}{f_1(C_R)\,p_1 + f_2(C_R)\,p_2} = c$$

6 MARKS

(c) These parameters would be estimated from training data, either by maximum likelihood (some details of the estimation procedure must be given), method of moments, or by using prior information in addition to the likelihood information in a Bayesian analysis.

5 MARKS

(d) The accuracy can be assessed using, for example, leave-k-out, or m-fold cross validation, where a proportion of the training data are held back, and the remainder are used to build the model (estimate the parameters, and find C_R), and then the prediction accuracy of the learned rule can be assessed by predicting the class of the training data that have been held back. This can be done for all subsets of size k from n, or a sample of such subsets, with an average accuracy computed.

The rule can also be assessed using **ROC curves**.

6 MARKS