EXAMINATION JANUARY 2003: SOLUTIONS

1. (i) We have the count table

	A	A'	Total
B	54	139	193
B'	27	30	57
Total	81	169	250

so that

$$P(A) = \frac{81}{250} = 0.324 \qquad P(B|A) = 2P(B^{'}|A) \qquad P\left(A^{'} \cap B\right) = \frac{139}{250} = 0.556$$

6 MARKS

(ii) We have

$$P(B) = \frac{193}{250} = 0.772$$

2 MARKS

(iii) Using conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{54}{193} = 0.2798$$

to 4 d.p

. **3 MARKS**

(iv) Table is

	A	A'	
B	0.216	0.556	0.772
B'	0.108	0.120	0.228
	0.324	0.676	1

. **6 MARKS**

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

$$P(A|B) \neq P(A)$$

2 MARKS

(vi) Fitted values under the independence hypothesis are

	E	E'
F	62.532	130.468
F'	18.468	38.532

giving a test statistic

$$\chi^2 = 7.552971$$

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 exceeds the critical value, and the test of indepence is rejected at the $\alpha = 0.05$ level.

6 MARKS

2. (a) (i) The distribution is $Binomial(n, \theta)$

2 MARKS

The required probability is obtained from the Binomial mass function as

$$\begin{aligned} P[X \leq 2] &= P[X=0] + P[X=2] + P[X=2] \\ &= (1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^{9} + \binom{10}{2}\theta^2(1-\theta)^{8} \\ &= 0.9998862 \end{aligned}$$

2 MARKS

(ii) From tables

$$P[X > 12] \approx 1 - \Phi\left(\frac{12 - 1000 \times 0.01}{\sqrt{1000 \times 0.01 \times 0.99}}\right) = 1 - \Phi(0.6356) = 0.2625$$

3 MARKS

(iii) $Y \sim Geometric(\theta)$

2 MARKS

and for $y = 1, 2, 3, ...$

$$P[Y > y] = 1 - P[Y \leq y] = 1 - (1 - (1-\theta)^{y}) = (1-\theta)^{y}$$

2 MARKS

and thus if $\theta = 0.01$

$$P[Y > 10] = (1 - 0.01)^{10} = 0.9044$$

1 MARK

(iv) We have, by the conditional probability definition

$$P[Y > y_2 | Y > y_1] = \frac{P[Y > y_1 \cap Y > y_2]}{P[Y > y_1]} = \frac{P[Y > y_2]}{P[Y > y_1]} = \frac{(1-\theta)^{y_2}}{(1-\theta)^{y_1}} = (1-\theta)^{y_2 - y_1} = P[Y > y_2 - y_1]$$

5 MARKS

(v) The test statistic $Y_{\max}$ was observed to be $y_{\max} = 473$; under H_0, we know that

$F_{Y_{\max}}(y) = P[Y_{\max} \leq y] = \{1 - (1-\theta)^{y}\}^{n} \qquad \therefore \qquad P[Y_{\max} \geq y] = P[Y_{\max} > y - 1] = 1 - \left\{1 - (1-\theta)^{y-1}\right\}^{n}$

with $\theta = 0.01$ and $n = 20$. Thus, for a p-value,

$$\begin{aligned} P[Y_{\max} \geq y_{\max}] &= 1 - \left\{1 - (1-\theta)^{y_{\max}-1}\right\}^{n} \\ &= 1 - \left\{1 - (1-0.01)^{472}\right\}^{20} \\ &= 0.1604 \end{aligned}$$

This is not below the chosen significance level, and thus the null hypothesis **cannot be rejected**.

8 MARKS

3. (i) Test statistic is

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.53 - 18.24}{1\sqrt{\frac{1}{18} + \frac{1}{24}}} = 7.34$$

with the usual Z-test 95% critical values of ± 1.96; thus the test gives strong evidence to **reject** the null hypothesis.

5 MARKS

(ii)

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(18 - 1) \times 2.24 + (24 - 1) \times 1.96}{18 + 24 - 2} = 2.079$$

2 MARKS

(iii)Now need to use a two sample T test; test statistic is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim Student(n_1 + n_2 - 2)$$

if H_0 is true. Here

$$t = \frac{20.53 - 18.24}{(2.079)^{1/2}\sqrt{\frac{1}{18} + \frac{1}{24}}} = 5.09$$

with critical values from tables of ± 2.0211, so again the test is rejected.at $\alpha = 0.05$.

5 MARKS

(iv) Need a two sample F-test;

1 MARK

We have test statistic

$$q = \frac{s_1^2}{s_2^2} \sim Fisher(n_1 - 1, n_2 - 1)$$

if H_0 is true; here

$$q = \frac{2.24}{1.96} = 1.1428.$$

2 MARKS

Now, the 0.025 and 0.975 critical value from tables are available from the $Fisher(13, 17)$ table

$$C_{R_1} = 0.389 \qquad C_{R_2} = 2.466$$

Hence the critical region is defined as the complement of the interval $(0.389, 2.466)$. The test statistic lies in this interval, and not the critical region, and hence the null hypothesis of equal variances cannot be rejected at significance level $\alpha = 0.05$.

4 MARKS

(v) **Non-parametric methods** can be used (for example the Mann-Whitney-Wilcoxon test for equal medians, the Kolmogorov-Smirnov test for the equality of two distribution functions), as can **Monte Carlo exact** or **permutation testing** methods.

3 MARKS

(vi) Multiple testing corrections (Bonferroni, Step Procedures) are important to maintain the true significance level of a range of statistical tests; essentially, a significance test at level α allows for a rate of false positive results of α; if a large number of tests is done, we may by chance (actually, by design) reject the null hypothesis when it is TRUE on several occasions. The corrections adjust the significance level so that the rate of false positive results **across all tests** jointly is at the target level α; in practice, this means either reducing the level at which an individual test is deemed significant (Multiple testing corrections), or adjusting the p-value (Step procedures)

3 MARKS

4. (a)(i) The fitted values are obtained assuming $p = 0.25$ for all nucleotides; thus the fitted values are given by

$$\hat{n}_i = np = 0.25n \qquad i = 1, 2, 3, 4$$

giving the following table of fitted values:

Nucleotide	A	C	G	T	Total
Observed	135.00	202.00	148.00	116.00	601
Fitted	150.25	150.25	150.25	150.25	601

2 MARKS

(ii) Test statistic

$$LR = 2\sum_{j=1}^{4} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}} = 2\left[135 \log \frac{135}{150.25} + 202 \log \frac{202}{150.25} + 148 \log \frac{148}{150.25} + 116 \log \frac{116}{150.25}\right] = 26.18681$$

which we compare with a $\chi^2_{k-d-1} = \chi^2_3$ null distribution.

4 MARKS

(iii) The (upper) critical value in a one-sided test for $\alpha = 0.01$ is 12.838, thus the test is **rejected**.

2 MARKS

(iv) We can recast this by fitting the required model; the estimated parameters are

$$\widehat{\theta} = \frac{135 + 202}{601} = 0.5607 \qquad \widehat{\phi} = 1 - \widehat{\theta} = 0.4393$$

and the corresponding fitted values are

Nucleotide	A	C	G	T	Total
Observed	135.00	202.00	148.00	116.00	601
Fitted	168.50	168.50	132.00	132.00	601

and an LR test statistic

$$LR = 17.29799$$

The degrees of freedom here is $k - d - 1 = 4 - 1 - 1 = 2$, the critical value is 10.597 and the null hypothesis is rejected.

5 MARKS

(v) We would generate a large number of sequences of length 601 **according to the null model**, and count the proportion of times that a word of length at least 10 was observed. This proportion is a Monte Carlo estimate of the p-value in the test; if the proportion is less than significance level α then the test is rejected.

3 MARKS

(vi) Here, 20000 sequences were sampled, and thus the approximate percentiles in the null distribution are

Length	< 6	7	8	9	10	11	≥ 12	Total
Count	10538	5643	2123	1234	420	37	5	20000
Estimated cumulative prob.	0.5269	0.8091	0.9152	0.9769	0.9979	0.9975	1.000	

Thus probability of observing a repeated word of length at least 10 under the computed null distribution is

$$\frac{420 + 37 + 5}{20000} = 0.0231.$$

Thus we can reject the hypothesis at the $\alpha = 0.05$ significance level, but not at the $\alpha = 0.01$ significance level.

5 MARKS

(vii) A permutation test would generate all possible 601! permutations of the raw sequence, and the calculation of the **permutation exact** percentiles and p-value calculated in the way described above. However, 601! is extremely large, so instead of working out the maximum repeated word lengths for all of these permutations, a large sample of **randomly generated** permutations can be used to approximate the percentiles/p-values. This method does not rely on specifying or estimating the nucleotide probabilities, as these values are inferred by the permutation process.

4 MARKS