# MSc BIOINFORMATICS

# PROBABILITY AND STATISTICS

# EXAMINATION 2004-2005: SOLUTIONS

1. (a) **p-value:** a p-value is a quantity that facilitates the completion of a statistical hypothesis test. Let $T$ be a (scalar) test statistic for testing hypothesis $H_0$, and let $t$ be the observed value of the test statistic. Let the distribution of $T$ be dependent on parameter $\theta$, and consider tests about $\theta$. The p-value, $p$, is defined as

$$p = Pr[T \text{ at least as extreme (numerically) as } t|H_0].$$

In the most straightforward case, when $H_0$ is a simple null hypothesis and alternative $H_1$ is a two-sided general alternative, that is

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta \neq \theta_0 \end{aligned}$$

then the p-value is a tail-probability in the null distribution of $T$ under $H_0$, that is, $p = Pr[|T| \geq t|\theta_0]$. This calculation is adjusted if the nature of the alternative hypothesis changes. The p-value is compared with the significance level, $\alpha$, in order to complete the test; if $\alpha \geq p$, then $H_0$ is rejected. The p-value is calculated in every form of statistical testing; for example, in tests for differential expression in microarray analysis.

***5 MARKS***

(b)Family-wise error rate: the FWER is the probability of falsely rejecting at least one of a sequence or family of null hypotheses, under the assumption that all of the null hypothesis are in fact true. That is, the FWER is the composite Type I error probability across the family of tests. It is considered whenever a number of statistical hypothesis tests are carried out, as the FWER is generally higher than the single test significance level $\alpha$. Several standard procedures are available for the control of the FWER; most adjust the single-test $\alpha$ so that the overall FWER is maintained at the original target level. The simplest procedure is the Bonferroni correction, a slightly less conservative correction is the Bonferroni-Sidak. The control of FWER is important in gene expression analysis, where large numbers of genes are being inspected simultaneously.

***5 MARKS***

(c)Randomization or permutation test: a randomization or permutation test is a statistical hypothesis test in which the key quantity of interest, the distribution of a specified test statistic under the null hypothesis, is computed exactly by considering all possible configurations of the data that are equally likely under the null hypothesis. For example, in a two-group comparison, under the null hypothesis that data from the two groups have no distributional difference, all possible allocations of data of to groups are considered, with a suitable test statistic computed for each allocation. Such tests are useful whenever distributional assumptions about the data need to be relaxed, or whenever the distribution of the test statistic cannot be computed analytically.

***5 MARKS***

(d)Hidden Markov model: a HMM is a stochastic model for a latent process, $\{X_t\}$ say, that underlies a discrete-time/space sequence of observations, $\{Y_t\}$. The latent process is Markov, in that

$$Pr[X_t|X_s, s < t] \equiv Pr[X_t|X_{t-1}].$$

Typically, in discrete time, the stochastic nature of $\{X_t\}$ is determined by an initial state, $X_0$, and a transition matrix, $P$, that determines the probability of moving from one state to another state. HMMs are used in biological sequence analysis to make inference about hidden structure; for example, for gene-finding, CpG island detection in DNA sequence analysis, or the discovery of motifs or secondary structure in protein sequence analysis.

***5 MARKS***

(e) Principal components analysis: PCA is a technique used in high dimensional data analysis, for example in the analysis of microarray data. It is used to reduce the dimensionality of a data set or data matrix. Iit describes the data set in terms of its components of variance. Each principal component describes a percentage of the total variance of a data set, and computes loadings or weights that each variate contributes to this variance. For example, the first principal component of a data set describes the dimension which accounts for the greatest amount of variance of the data set. The coefficients of the principal components quantify the loading or weight of each variate to that amount of variance. The mathematical assumptions behind PCA include multivariate normality of the underlying observations. PCA can be used as the basis for classification procedures.

***5 MARKS***

2. (a) Let $A$ indicate population A, $E$ indicate genotype 0 and $F$ indicate unaffected status. The total number of individuals in the study is

$$(72+12+30+15)+(108+67+22+78) = 129+275 = 404$$

(a) (i) the probability that the person is Affected

$$P(F') = \frac{30+15+22+78}{404} = \frac{145}{404} = 0.3589$$

***2 MARKS***

(ii) the probability that the person is from Population B

$$P(A') = \frac{108+67+22+78}{404} = \frac{275}{404} = 0.6807$$

***2 MARKS***

(iii) the conditional probability that the person is Affected, given that they are from Population A

$$P(F'|A) = \frac{30+15}{72+12+30+15} = \frac{45}{129} = 0.3488$$

***3 MARKS***

(iv) the conditional probability that the person is Affected, given that they are from Population A and have genotype code 1.

$$P(F'|A,E') = 1512+15 = 1527 = 0.5556$$

***3 MARKS***

(b) Pooled and fitted table given below

Pooled over Popn.

| | Genotype | |
|---|---|---|
| Disease Status | 0 | 1 |
| Unaffected | 180 | 79 |
| Affected | 52 | 93 |

Fitted Under Independence

| | Genotype | |
|---|---|---|
| Disease Status | 0 | 1 |
| Unaffected | 148.73 | 110.26 |
| Affected | 83.27 | 61.73 |

which yields a test statistic $\chi^2 = 43.02$ and thus a highly significant result (with $p < 0.001$ from tables of the $Chi-squared(1)$ distribution).

***8 MARKS***

(c) (i) Using the facts given, we have that, approximately

$$\log\widehat{\psi}_A - \log\psi_A \sim N(0, s_A^2) \qquad \log\widehat{\psi}_B - \log\psi_B \sim N(0, s_B^2)$$

so that

$$(\log\widehat{\psi}_A - \log\psi_A) - (\log\widehat{\psi}_B - \log\psi_B) \sim N(0, s_A^2 + s_B^2)$$

suggesting a test statistic

$$z = \frac{(\log\widehat{\psi}_A - \log\widehat{\psi}_B) - (\log\psi_A - \log\psi_B)}{\sqrt{s_A^2 + s_B^2}}$$

which, under this null hypothesis, should resemble an observation from an N(0,1) distribution. Under $H_0$, we compute

$$z = \frac{(\log\widehat{\psi}_A - \log\widehat{\psi}_B)}{\sqrt{s_A^2 + s_B^2}} = \frac{\log 3 - \log 5.72}{\sqrt{0.1972 + 0.0824}} = -1.218$$

which is not extreme under the null, and hence the null hypothesis cannot be rejected.

***4 MARKS***

(ii) Given this result, it seems that the chi-squared analysis in (b) might be valid, and hence we can reject independence between the genotype and the trait, that is, there seems to be a degree of association. Note that assessing the compatibility for pooling in this way is not necessarily completely satisfactory; it may be more appropriate to inspect various conditional probabilities (as in the Simpson's paradox/confounding examples). However if the study design is not known (case-control or cohort), the comparison of odds-ratios might be all that is possible.

***3 MARKS***

3. (a) We will use two-sample t-tests, under the assumption of equal variances in the two groups. The degrees of freedom in the null distribution is (16.1)+(16.1) = 30. The equal-variance assumption can be tested using a two-sample F-test; the sample variance ratio $s_1^2/s_2^2$ is $Fisher(15, 15)$ distributed under the null hypothesis of equal variances. NOTE: We will use the significance level of $\alpha = 0.01$; using Bonferroni, this controls the FWER to be at most 0.05 in each family of tests (one family for mean comparison, one family for variance comparison). Thus the critical values for the F-test are the 0.005 and the 0.995 quantiles of the $Fisher(15, 15)$ distribution. Now, the tables given only have the 0.025 and 0.975 quantiles, which are 0.349 and 2.862 respectively; but this is sufficient, as all test statistics lie between these two critical values.

|  | $s_P^2$ | $t$ | p-val | $s_1^2/s_2^2$ | Reject Equal Variances |
|---|---|---|---|---|---|
| GENE 1 | 1.516 | 6.216 | **0.000** | 0.438 | NO |
| GENE 2 | 0.582 | -0.163 | 0.872 | 0.459 | NO |
| GENE 3 | 0.843 | -5.357 | **0.000** | 0.371 | NO |
| GENE 4 | 0.921 | -0.557 | 0.582 | 0.665 | NO |
| GENE 5 | 0.541 | 12.582 | **0.000** | 0.590 | NO |

Thus genes 1, 3 and 5 are differentially expressed between the two tumour types in terms of mean level. However, it may be that the distribution of observations in one sample is different from that in the other - we cannot check this with only summary statistics.

***15 MARKS***

(b)The biggest likely difficulty is the systematic difference that is often observed between arrays; due to different spotters, scanners operators etc., the gene expression measures in the two channels are often dramatically different between supposedly replicate arrays. Therefore efforts must be made to carry out appropriate normalization; using boxplots the need for normalization can be assessed, and then, for example quantile normalization methods can be used to correct for differences between arrays.

***5 MARKS***

(c) Can either use non-parametric methods (Mann-Whitney-Wilcoxon or Kolmogorov-Smirnov two sample tests), or randomization or permutation procedures, or both in combination. The former relies on asymptotic properties of rank-sums, and so may be inappropriate for such small samples. The latter computes the exact (or Monte Carlo exact) null distribution of a specified test statistic. Note that in both cases, the original data and not merely the sample summary statistics must be used.

***5 MARKS***

4. (a) **Hierarchical Algorithms:** A hierarchical algorithm yields an entire hierarchy of clusterings for the given data set. *Agglomerative methods* start with the situation where each object in the data set forms its own cluster, and then successively merges clusters until only one large cluster (the entire data set) remains. *Divisive methods* start by considering the whole data set as one cluster, and then splits up clusters until each object is separated. Data sets for clustering of $N$ observations can have either of the following structures:

- an $N \times p$ **data** matrix, where rows contain the different observations, and columns contain the different variables.
- an $N \times N$ dissimilarity matrix, whose $(i, j)$th element is $d_{ij}$, the **distance** or **dissimilarity** between observations $i$ and $j$ that has the properties
  - $d_{ii} = 0$
  - $d_{ij} \geq 0$
  - $d_{ji} = d_{ij}$
- Typical data distance measures between two data points $i$ and $j$ with measurement vectors $x_i$ and $x_j$ include
  - the *Euclidean distance* for continuous measurements

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_{\boldsymbol{j}})^{\mathsf{T}} (\boldsymbol{x}_i - \boldsymbol{x}_{\boldsymbol{j}})}$$

  - the *Manhattan distance* for continuous or discrete measurements

$$d_{ij} = \sum_{k=1}^{p} |(x_{ik} - x_{jk})^2| = \sum_{k=1}^{p} \sqrt{(x_{ik} - x_{jk})^2}$$

In conventional hierarchical clustering, the method of agglomeration or combining clusters is determined by the distance between the clusters themselves, and there are several available choices. For merging two clusters $C_i$ and $C_j$ , with $N_1$ and $N_2$ elements respectively, the following criteria can be used

- In *average* (or *average linkage*) clustering, the two clusters that have the smallest *average distance between the points in one cluster and the points in the other*

$$d(C_i, C_j) = \frac{1}{N_1 N_2} \sum_{k \in C_i, l \in C_j} d_{kl}$$

  are merged .

- In *connected* (*single linkage, nearest-neighbour*) clustering, the two clusters that have the *smallest distance between a point in the first cluster and a point in the second cluster*

$$d(C_i, C_j) = \min_{k \in C_i, l \in C_j} d_{kl}$$

  are merged

- In *compact* (*complete linkage*, *furthest-neighbour*) clustering, the two clusters that have the l*argest distance between a point in the first cluster and a point in the second cluster*

$$d(C_i, C_j) = \max_{k \in C_i, l \in C_j} d_{kl}$$

  are merged.

***10 MARKS***

(b) One advantage of a model-based approach to clustering is that it allows the use of statistical model assessment procedures to assist in the choice of the number of clusters. A common method is to use approximate **Bayes factors** to compare models of different orders (i.e. models with different numbers of clusters). This method gives a systematic means of selecting the parameterization of the model, the clustering method, and also the number of clusters. The Bayes factor is the posterior odds for one model against the other assuming neither is favored a priori. Two methods based on the Bayes factor have been used.

- The **Approximate Weight of Evidence** (**AWE**) This is a heuristically derived approximation to twice the log Bayes factor
- The **Bayesian Information Criterion** (**BIC**) A more reliable approximation to twice the log Bayes factor called the *Bayesian Information Criterion*, which, for model $M$ is given by

  $$BIC_M = 2 \log L_M + const \approx 2 \log L_M(\widehat{\theta}) - d_M \log N$$

  where LM is the model-based marginal likelihood from, $L_M(\widehat{\theta})$ is the maximized log likelihood of the data for the model $M$, and $d_M$ is the number of parameters estimated in the model. The number of clusters is not considered a parameter for the purposes of computing the BIC. The **larger** the value of the BIC, the stronger the evidence for the model.

In a non model-based setting, an approximate likelihood measure, and either of these two measures can be used. Choice of the approximate likelihood may need some careful consideration.

***6 MARKS***

(c) (i) If a model-based procedure is used, $p(y|k)$ can be obtained using marginal likelihood type-arguments. For example, in a Gaussian model, the marginal likelihood is analytically available, and can be computed in a routine fashion. Approximate marginal likelihood calculations can also be attempted. For non model-based procedures, the classification procedure can proceed by using a transformed heuristic measure of distance between $y$ and the cluster $k$, perhaps using the linkage measures described above, for example, could set

$$p(y|k) \propto \exp\{-\lambda(y, C_k)\}$$

for some parameter $\lambda$ and distance measure $d(y, C_k)$ between $y$ and cluster $k$, with summary measure $C_k$.

***5 MARKS***

(ii) The classification procedure could be assessed.using within sample prediction error or cross-validation; both methods rely on training the method on data in the original clustering with a subset of genes held back for testing. The proportion of correct classifications of the test data gives an indication of the out-of-sample prediction accuracy, and this indication is more accurate using cross-validation.

***4 MARKS***