UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Workshop 4

1. Let $Y_1, Y_2, Y_3, Y_4, Y_5 \sim \text{Exp}(\lambda)$ and $Y_6, Y_7 \sim \text{Exp}(\beta\lambda)$ be independent random variables with $\beta > 0$ and $\lambda > 0$ unknown parameters. Suppose we observe $(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7) = (y_1, 5, 2, 2.5, 4, 0.25, 0.75)$ where y_1 is missing. Use the EM algorithm to find maximum likelihood estimates for β and λ.

2. The following table gives the number of deaths to women 80 years and older reported by day during the years 1910-1912.

Number of deaths	0	1	2	3	4	5	6	7	8	9
Observed Frequency	162	267	271	185	111	61	27	8	3	1

   (a) If the deaths are unrelated (independent) and their frequency is constant over time, a Poisson distribution is appropriate to describe the data. Fit a Poisson distribution to the data. From your result, would you say that the data are Poisson distributed (you can try to plot the observed counts against the expected counts under the fitted model)?

   (b) There are likely to be different patterns of deaths in winter and summer, in which case a mixture of two Poisson distributions may provide a much better fit. Formulate an EM algorithm for this mixture model. For this, you need to specify the complete and the missing data and derive the E-step and the M-step.

   (c) Implement your solution in R and apply it to the data above. Compare the fit with the fit by a Poisson distribution in part (a). By simply visual inspection, which model describes the data better?

3. Suppose that there are four types of genes: AA, Aa, aA, aa, and that the probabilities of occurrence are, respectively, $\frac{1}{4}(1 + 2\theta^2)$, $\frac{1}{2}\theta(1 - \theta)$, $\frac{1}{2}\theta(1 - \theta)$, and $\frac{1}{4}(1 + 2(1 - \theta)^2)$, with $0 < \theta < 1$. Suppose further that the number of occurrences of each gene type is $y = (y_1, y_2, y_3, y_4) = (69, 29, 25, 104)$. Using the EM algorithm, determine the maximum likelihood estimate of θ.
   Hint: Analogously to the genetic linkage example in week 6, suppose that category AA can be divided into two subcategories, AA_1 and AA_2, with probabilities $\frac{1}{4}$ and $\frac{1}{2}\theta^2$, respectively. Similarly, suppose that category aa can be divided into two subcategories aa_1 and aa_2, with probabilities $\frac{1}{4}$ and $\frac{1}{2}(1 - \theta)^2$, respectively. Further, let z_1 and z_2 be the number of gene occurrences in subcategories AA_1 and aa_1, respectively.
   Note: We do not need the EM algorithm to find the mle in this particular problem, but it serves as an illustrative example.