Statistics 347 **April 11: Due April 24–26** **Homework 2**

Question A0, A1 to be handed in for credit April 24; Questions A2, A3 due on April 26; In addition, questions A4 for statistics PhD students should be handed in on April 26.

You may ask me or the course assistants, Han Han and Joe Guinness, for help with any of these questions. Discussion of homework problems among students is encouraged. However, all material handed in for credit must be your own work. It is a violation of the honour code to share homework-related documents, files, computer code, plots or text with another student. Collaboration on homeworks is not permitted.

A0. In this exercise, you are asked to write a brief report on the data analysis in the paper by Sharon *et al.* (2010) titled *Commensal bacteria play a role in mating preference of Drosophila Melanogaster,* Proceedings of the National Academy of Sciences, vol 107, No. 46, 20052–20056. The file `commensal.txt` has five columns of data, generation number, followed by mating counts for four couple types, CxC, CxS, SxC, SxS. Here SxC denotes matings of male flies raised on diet S (starch) with females raised on diet C (corn-molasses-yeast). All of these flies were raised without antibiotic additives.

My understanding is that two breeding populations were raised generation after generation on the same diet C or S, but the flies destined for experimental purposes were removed from the breeding population and raised for one intermediate generation on the CMY diet before testing. Thus the testing for generation 6 was done on the offspring, so generation 6 is really 6+1. See Fig. 1.

Each experiment consists of a varying number of wells, from 20 to 70, with four flies in each well. The number of wells is not reported for the first three generations, 2, 6, 7, but the values for subsequent generations are 24, 39, 20, 24, 36, 23, 70, 46, 24, 45, 23, 48, 48, 48, 48. The last three rows of data may be taken fom a parallel experiment run under a similar protocol, so the generation numbers are not relevant.

Your report should address the various formulae and the associated numbers and graphs reported in the paper. Is the design sound? Are there any anomalies in the data or in the formulae? Does the SII index increase with generation number, or can it reasonably be taken as constant? Is the number of matings related to the number of wells? Is the pattern of variation different for the last three rows? Are the data consistent with the assumption of independence of mating events in successive generations?

A1. The data for example B in Applied Statistics p. 58 are given in `.../edu/~pmcc/glm/ExB.txt`. The response is the mean interval in months between successive births in the same family. Thus a family of size k gives rise to $k-1$ intervals, 1–2, 2–3,..., each of which corresponds to a particular sex sequence MM, MF, FM or FF. Any systematic pattern is of potential interest.

Calculate the residual sum of squares for the additive models

$$\textit{fam size} + \textit{birth order} + \textit{sex seq} \quad \text{and} \quad \textit{fam size} + \textit{reverse birth order} + \textit{sex seq}$$
$$\textit{fam size.birth order} + \textit{sex seq} \quad \text{and} \quad \textit{fam size.reverse birth order} + \textit{sex seq}$$

Comment on the difference between the first two models, and how this difference affects the interpretation. Comment also on any unexpected or unusual effects. (As part of this analysis, you should first check to see whether transformation might be helpful.)

Give a one-paragraph summary *in your own words* of the main patterns of variation in these data.

A2. Let Y be a square array of observations with rows and columns indexed by the same set of levels, as in Q. 5 from HW1. The linear model formulae `y~row`, `y~col`, and `y~row+col` are expressed algebraically as

$$E(Y_{ij}) = \alpha_i, \quad E(Y_{ij}) = \beta_j, \quad \text{and} \quad E(Y_{ij}) = \alpha_i + \beta_j.$$

When the array is square, other non-factorial models suggest themselves in a natural way, depending on the application. Consider the following:

```
XA <- model.matrix(~row-1)
XB <- model.matrix(~col-1)
XAB <- model.matrix(~row:col-1)
```

`XBA <- model.matrix(~col:row-1)`

Explain what subspace is spanned by the columns of the following matrices. Describe each subspace algebraically, give the dimension as a function of the number of rows, and explain how it might be used. (These are matrices, not model formulae, so you need to type `V <- XA+XB` and use `y~V` in the model formula.)

(i) `XA + XB`

(ii) `XA - XB`

(iii) `XAB + XBA`

(iv) `XAB - XBA`

A3. Estimating the LD_{50}. In toxicology, the LD_{50} is the dose that causes a 50% mortality rate (lethal dose 50%). Experiments are often carried out at a sequence of dose levels, $x_0, x_1, \ldots$ each dose being twice the preceding dose. The model most commonly used in toxicology is linear in log dose. Suppose that the following results have been obtained in an experiment at various multiples of the baseline dose.

Dose $\log_2(x)$	0	1	2	3	4	5
Mortality y/m	0/7	2/9	3/8	5/7	7/9	10/11

Here y/m is the number of deaths occurring in a sample of m individuals.

Plot the data, i.e. the mortality fraction against log dose.

Fit the linear logistic model in which the logit of the mortality rate is linear in log dose. Superimpose the fitted probabilities on the plot.

Obtain the estimate of the $\log_2 \mathrm{LD}_{50}$, and use Fieller's method to generate a confidence interval.

Consider the null hypothesis that $\log_2 \mathrm{LD}_{50} = 4$ as a sub-model or restriction of the linear logistic model. Fit the sub-model and compute the log likelihood ratio statistic $\mathrm{LR}(4)$, i.e. twice the difference between the unrestricted and the restricted log likelihood. If the null hypothesis is correct, this difference should be distributed approximately as χ^2_1. Compute the p-value.

By plotting the restricted log likelihood against the hypothesized value of the LD_{50}, construct a likelihood-based 95% confidence set for $\log_2 \mathrm{LD}_{50}$.

A4. An equivalence relation on a finite non-empty set $[n] = \{1, \ldots, n\}$ is a Boolean function or matrix B that is reflexive, symmetric and transitive. Reflexive means $B(i,i) = 1$ for each i; symmetry means $B(i,j) = B(j,i)$ for each pair; transitive means $B(i,j) = 1$ and $B(j,k) = 1$ implies $B(i,k) = 1$. A partition of a finite non-empty set is a set of non-empty non-overlapping subsets whose union is the entire set. A partition, an equivalence relation and a block factor are synonyms, different words meaning the same thing, but perhaps suggesting different images or representations. The subsets are called the blocks of the partition, and the number of blocks is denoted by $\#B$, which is also the matrix rank. A partition of $\{1, \ldots, 6\}$ is usually written in the form 13|26|45 or 135|26|4 as a list of blocks, but the order of the blocks and the order within blocks are ignored. Thus 62|54|13 is the same as 13|26|45 or $\{\{1,3\},\{2,6\},\{4,5\}\}$, but different from 14|25|36. Let $\mathcal{B}_n$ be the set of partitions of $\{1, \ldots, n\}$, and let $\#\mathcal{B}_n$ be the number of elements in $\mathcal{B}_n$.

Show that the first few values of $\#\mathcal{B}_n$ are $1, 2, 5, 15, 52, \ldots$, and that these are the moments of the Poisson distribution with mean 1.

The Ewens distribution on $\mathcal{B}_n$ with parameter $\lambda > 0$ is given by

$$p_n(B;\lambda) = \frac{\Gamma(\lambda)\,\lambda^{\#B}}{\Gamma(n+\lambda)} \prod_{\text{blocks}} (\#b - 1)!,$$

where $\#b$ denotes number of elements in the set b. Show that this is in fact a probability distribution for each $n = 1, \ldots, 5$.

Find the probability that two distinct elements belong to the same block.

Show that the Ewens family is an exponential family. Hence or otherwise, find the mean and variance of the number of blocks as a function of n and λ

Deduce from the cumulant function that the random variable $\#B$ is a sum of independent Bernoulli variables. Hence or otherwise, show that the number of blocks is approximately Poisson. Find the mean, and the deficit of the variance over the mean.

Explain what it means for p_n to be the marginal distribution of p_{n+1}. Show that this property holds, and explain the implications.

A cocktail party has 27 guests who arrange themselves in conversational groups, one group of size 7, two of size 4, three of size 3 and three of size one. Fit the Ewens model and compute the conditional probability that the next arriving guest joins each of the existing groups.

A5. Random permutations. A permutation of $[n]$ is a 1–1 function $\pi\colon [n] \to [n]$, and a random permutation is a probability distribution on the set of $n!$ permutations of $[n]$. A permutation π may be written either as a set of ordered pairs $\{(i, \pi(i))\}$ or in cycle format. The cycle that includes the element 1 is $\{1, \pi(1), \pi^2(1), \pi^3(1), \ldots\}$; the cycle that includes 2 is $\{2, \pi(2), \pi^2(2), \ldots\}$, which is either the same as the cycle that includes 1 or disjoint from it. Here is a permutation written in both formats:

$$\pi = \begin{pmatrix} 1 & \cdots & n \\ \pi(1) & \cdots & \pi(n) \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 6 & 7 & 8 & 5 & 2 & 3 & 1 \end{pmatrix} \quad = \quad (1,4,8)(2,6)(3,7)(5)$$

This permutation has four cycles, one fixed point or uni-cycle, two bi-cycles and one tri-cycle. The number of cycles is denoted by $\#\pi$.

Stirling's number S_n^r (of the first kind) is the nmber of permutations of $[n]$ having exactly r cycles. By considering the options available for inserting a new element into an existing permutation, show that these numbers satisfy the recurrence relation $S_{n+1}^r = S_n^{r-1} + nS_n^r$. Hence deduce that $\sum_r \lambda^r S_n^r = \lambda^{\uparrow n}$ is the ascending factorial function $\lambda(\lambda+1)\cdots(\lambda+n-1)$.

The uniform distribution puts mass $1/n!$ on each of the permutations, and the associated exponential-family distribution with canonical statistic $\#\pi$ is

$$p_n(\pi) = \frac{e^{\theta \#\pi}}{n!\, M(\theta)},$$

where $M(\theta)$ is the moment generating function of the number of cycles in a uniformly distributed random permutation. Find the normalizing constant $M(\theta)$.

The cycles of a permutation determine a partition of $[n]$, and the mapping from permutations to partitions of $[n]$ is onto. For example, the permutation shown above corresponds to the partition 148|26|37|5 by ignoring within-cycle order. Given a partition $B = \{b_1, b_2, \ldots\}$ with block sizes $\#b_r$, what is the associated number of permutations? Hence deduce the marginal distribution on partitions induced by the exponential-family distribution on permutations.

For the Ewens distribution with parameter λ, what is the marginal distribution of the number of blocks? What is the conditional distribution of the block sizes given that $\#B = k$?

A6. A partition of the integer n is an additive integer decomposition such as $11 = 1+1+1+2+2+4$, usually written in the form $1^3\,2^2\,3^0\,4^1\,\cdots$, with trailing zeros ignored. The superscripts are called the multiplicities. A generic integer partition y of n is $1^{y_1}2^{y_2}3^{y_3}\cdots n^{y_n}$, and the components of y are the multiplicities satisfying $\sum jy_j = n$. Consider the following probability distribution on the partitions of a fixed integer $n \geq 1$. The components of Y are independent Poisson variables with means $E(Y_j) = 1/j$ subject to the condition that $\sum jy_j = n$. Show that the joint distribution of Y is

$$p_n(Y = y) = \frac{1}{\prod_j j^{y_j}\, y_j!}.$$

Derive the associated exponential family whose canonical statistic is the sum $y_\bullet$ of the components. Find the joint distribution including the normalizing constant.

Describe briefly two methods of estimating the parameter from a partition $(y_1, y_2, \ldots)$ of a large integer such as $n = 1000$. One of these methosd may be maximum likelihood. The other should be such that the computations can be done in R.

T1. GLM 2.3.

T2. GLM 2.7.

T3. GLM 2.8.

T4. Find the m.g.f. of the density $\exp(y - e^y)$ and obtain the associated exponential family.

T5. GLM 2.14.

T6. Let Y be a random variable having the logistic distribution. Deduce that there exist independent and identically distributed random variables X_1, X_2 such that $Y = X_1 - X_2$. What is the distribution of X_1?

T7. Let $Y_1, \ldots, Y_n$ be Gaussian with mean zero and covariance matrix $\mathrm{cov}(Y_i, Y_j) = \sigma^2 \exp(-\beta|i-j|)$ with $|\beta| < 1$. Show that $E(Y_{i+1} \,|\, Y_1, \ldots, Y_i) = \beta Y_i$. Hence or otherwise deduce that $\hat{\beta}$ is approximately equal to $\sum Y_i Y_{i-1} / \sum Y_{i-1}^2$.

Suppose that you wish to obtain a more accurate estimate of β, but the length of the series is only 10, and there is no opportunity to extend this. However, there are ten series in parallel, all with the same autocorrelation function, so pooling the information is a natural strategy. All series have zero mean, and the covariances for series r, s are $\mathrm{cov}(Y_{ri}, Y_{sj}) = \sigma_{rs} \exp(-\beta|i-j|)$. For fixed β, find the maximum likelihood estimate $\hat{\Sigma}_\beta$ of the covariance matrix $\Sigma = \{\sigma_{rs}\}$, and hence obtain the profile log likelihood $l(\hat{\Sigma}_\beta, \beta; y)$ for β. Plot this function and discuss briefly the wisdom of pooling the data in this way.

Your answers to questions T2–T6 should occupy no more than one page each.