

Answer questions 1 and 4. PhD students should also answer question 5.

1. The data commonly referred to as Fisher’s iris data `.../iris.txt` were collected by Edgar Anderson and analyzed by Fisher using a taxonomic method known as discriminant analysis, or Fisher discriminant analysis. (This is a good example of the ‘Matthew effect’ in scientific attribution: Matthew 25:29.) The data consist of four variables measured on 150 iris specimens taken from Anderson’s study of the geographic variation of iris flowers in the Gaspé peninsula on the south shore of the St. Lawrence river in Eastern Quebec. The four variables, also called features, are sepal length and width plus petal length and width, so Y is a matrix of order 150×4 . The 150 specimens consist of three species, 50 *setosa*, followed by 50 *versicolor* and 50 *virginica*, so $species$ is a partition B of the set of specimens into three blocks or classes.

According to the Gauss-Ewens process, B is a random partition, and the conditional distribution given B is Gaussian with the same mean vector μ for each specimen regardless of the variety. The conditional covariance matrix given B is $I_n \otimes \Sigma + B \otimes \Sigma'$, in which Σ is the within-blocks covariance matrix of order 4, and Σ' is the between-blocks covariance matrix. For two specimens u, u' the vector difference $Y(u) - Y(u')$ has zero mean and covariance either 2Σ or $2(\Sigma + \Sigma')$ depending on whether the two specimens are of the same species or different species. In the discussion that follows, it is assumed for simplicity that $\Sigma' = \theta\Sigma$ for some scalar $\theta > 0$.

Let u' be a new iris specimen whose feature measurements are $y(u')$. According to the Gauss-Ewens cluster process, the conditional probability given the data (Y, B) plus the features measured on the new specimen, that the new specimen belongs to block b is

$$\text{pr}(u' \mapsto b \mid \dots) \propto \begin{cases} n_b \phi_4(y(u') - \tilde{\mu}_b; \tilde{\Sigma}_b) & b \in B, \\ \lambda \phi_4(y(u') - \mu; \Sigma(1 + \theta)) & b = \emptyset, \end{cases}$$

where n_b is the block size, \bar{y}_b is the sample mean vector, $\tilde{\mu}_b = (\mu + n_b\theta\bar{y}_b)/(1 + n_b\theta)$ is a weighted average of the parameter and the sample mean for block b , $\tilde{\Sigma}_b = \Sigma(1 + \theta/(1 + n_b\theta))$, and $\phi_k(y, \Sigma)$ is the density at $y \in \mathcal{R}^k$ of the zero-mean Gaussian distribution with covariance matrix Σ . The parameters that need to be estimated are the mean vector μ , the within-blocks covariance matrix Σ , and the variance ratio $\theta > 0$. In the calculations that follow, take $\theta = 10$ and $\lambda = 1$. Use the sample mean vector and pooled within-blocks sample covariance matrix as estimates for μ, Σ .

Show your estimates of μ and Σ . For $y(u') = (6.3, 2.9, 4.9, 1.7)$ find the conditional probability distribution given (Y, B) of the variety of u' , i.e. the probability that the variety is *setosa*, *versicolor*, *virginica* or other. Ditto for $y(u'') = (5.5, 3.1, 2.9, 0.8)$.

2. The following data were collected in a study of the relation between parental socioeconomic status (SES) and the mental health of children. Any systematic variation is of interest.

Parent’s SES	Mental health status			
	well	Mild	Moderate	Impaired
A (high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F (low)	21	71	54	71

Analyze the data and write a brief report (1 page at most).

3. The following data were collected in a food-tasting experiment, with four packaged food mixes presented as pairs in various orders. Analyze the data and give a brief summary of your conclusions. Comment on the magnitude of the order effect relative to difference between the four food types.

Pair (s, t)	Order	Frequency of response			Total
		Prefer s	No preference	Prefer t	
1, 2	1, 2	23	8	11	42
	2, 1	6	8	29	43
1, 3	1, 3	27	5	11	43
	3, 1	14	6	22	42
1, 4	1, 4	35	1	6	42
	4, 1	11	4	27	42
2, 3	2, 3	34	1	6	41
	3, 2	16	3	23	42
2, 4	2, 4	29	2	9	40
	4, 2	15	5	22	42
3, 4	3, 4	26	5	11	42
	4, 3	14	5	24	43

Explain why the assumption of independence of observations is not tenable for these data. Explain also why the assumption of independence is unlikely to lead to misleading conclusions regarding the estimated effects.

You may assume for simplicity that all observations are independent. To construct a model, assume that the quality of each food type is characterized by a real-valued parameter α , such that the probability of a being preferred to b is governed by the difference in qualities $\alpha_a - \alpha_b$.

4. *Baseball scores*: In the Bradley-Terry model for ranking k competitors, parameters $\theta_1, \dots, \theta_k$ representing ‘abilities’ are introduced in such a way that the probability π_{ij} that competitor i beats j is a function of the difference in their abilities. In the logit model, we have

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \theta_i - \theta_j.$$

Suppose that seven teams compete in a round-robin tournament, with each pair competing at least once.

(a) Write out the 21×7 model matrix \mathbf{X} for the Bradley-Terry model. The data below give the home-team win/loss record of seven American-League baseball teams (Milwaukee, Detroit, Toronto, New York, Boston, Cleveland and Baltimore) in the 1987 season. (New York beat Toronto in two of seven games at NY, and in four of six games at Toronto.)

Home Team	Away Team						
	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	4-3	4-2	4-3	6-1	4-2	6-0
Detroit	3-3	—	4-2	4-3	6-0	6-1	4-3
Toronto	2-5	4-3	—	2-4	4-3	4-2	6-0
New York	3-3	5-1	2-5	—	4-3	4-2	6-1
Boston	5-1	2-5	3-3	4-2	—	5-2	6-0
Cleveland	2-5	3-3	3-4	4-3	4-2	—	2-4
Baltimore	2-5	1-5	1-6	2-4	1-6	3-4	—

- (b) Fit the Bradley Terry model to these data to obtain a ranking of the teams. Extend this model by including a home-team advantage effect (equal for all teams). Obtain the likelihood-ratio statistic. Comment briefly on the magnitude of home-field advantage.
- (c) Estimate the probability that Detroit beats Boston (i) at Boston, (ii) at Detroit, and (iii) on neutral territory.

(d) Does the extended model fit the data? Comment briefly on any patterns in the residuals.

Statistics PhD students: (Q6 due Tuesday May 27)

5. The hypergeometric distribution is a probability distribution on contingency tables or non-negative integer-valued matrices of order $r \times c$ having the given marginal totals $n = (n_1, \dots, n_r)$ and $m = (m_1, \dots, m_c)$ with $m_\cdot = n_\cdot$. The distribution is given by

$$\text{pr}(T = t; m, n) = \frac{\prod_i n_i! \prod_j m_j!}{n_\cdot! \prod_{ij} t_{ij}!}$$

for non-negative integer-valued matrices t_{ij} having the given row and column totals.

Show that this is a probability distribution.

Let A be a factor with r levels and n_\cdot components in which level i occurs n_i times, and let B be another factor with c levels in which level j occurs m_j times. The indicator matrix generated by the expression `model.matrix(~A-1)` is denoted by \tilde{A} , and likewise for \tilde{B} . Let P be a random permutation [matrix] uniformly distributed on the permutations of $\{1, \dots, n_\cdot\}$. Show that the matrix $T = \tilde{A}'P\tilde{B}$ has the hypergeometric distribution.

The file `...~pmcc/datasets/birth-death.txt` contains data taken from Phillips and Feldman (1973) recording the month of birth and month of death of famous Americans. Is there any association between month of death and month of birth? The classical rule of thumb for the validity of Pearson's chi-square test, that the expected value in each cell should be at least 3–5, is badly violated here. Compute Pearson's chi-squared test statistic and the deviance statistic for independence of birth month and death month. The exact p -value is the hypergeometric exceedance probability. Estimate this p -value by simulation of 10000 independent hypergeometric tables, and counting the fraction of tables in which the observed value is exceeded. (`sort.list(runif(n))` generates a uniform random permutation of $\{1, \dots, n\}$.) Show the histogram of the simulated values with the nominal χ^2 density superimposed, and comment on the difference between the actual distribution of each statistic and the limiting chi-squared approximation.

For a more focused test statistic, aggregate the data by counting the number N_x of deaths occurring in birth month $+ x$ for $-5 \leq x \leq 6$, contrasting these values with the expected counts under independence. Discuss briefly the statistical significance of any patterns observed.

Explain how to extend the hypergeometric simulation scheme to three-way tables in which each one-dimensional table of marginal totals is fixed. Hence show how to test the hypothesis of complete independence in sparse contingency tables.

Explain how to extend the simulation scheme to three-way tables in which two two-dimensional tables of marginal totals are fixed. Hence show how to test the hypothesis of conditional independence in sparse contingency tables.

Explain how to extend the simulation scheme to three-way tables in which each two-dimensional table of marginal totals is fixed. Hence show how to test the hypothesis of no three-factor interaction in sparse contingency tables.

6. Let \mathcal{G}_k be the group generated by multiplication of k letters A_1, \dots, A_k according to the rules $A_r A_s = A_s A_r$ and $A_r^2 = 1$ for all letters A_r, A_s . If $k = 3$ and the letters are A, B, C , the group elements are

$$1, A, B, C, AB, AC, BC, ABC.$$

Formally, \mathcal{G}_k is the k -fold direct product of the group ± 1 . A real vector in the context of this question is a function $v: \mathcal{G}_k \rightarrow \mathcal{R}$, a point in the real vector space $\mathcal{R}^{\mathcal{G}_k}$ of dimension 2^k . A probability distribution π is a vector having non-negative components whose sum is one. Such a vector can be written as a list of numbers, $(\pi_g)_{g \in \mathcal{G}}$, but it is more convenient for groups to write π as a formal linear combination, such as

$$\pi = \pi_1 1 + \pi_A A + \pi_B B + \pi_C C + \pi_{AB} AB + \pi_{AC} AC + \pi_{BC} BC + \pi_{ABC} ABC$$

in the particular case $k = 3$. For example $0.3 + 0.7AB$ is a probability distribution on \mathcal{G} . Two formal linear combinations can be added, so the set of formal linear combinations is a vector space. They can also be multiplied like polynomials, so they have additional algebraic structure. Denote by $\mathcal{P}(\mathcal{G})$ the simplex of probability distributions on \mathcal{G} .

(i) Let X, X' be independent \mathcal{G}_k -valued random variables with distributions π, π' . Show that the group product $Y = XX'$ has distribution $\pi\pi'$, the formal product of formal linear combinations.

Let θ be a parameter vector such that $\theta_1 = 0$ and $0 \leq \theta_g \leq 1$ for each $g \neq 1$. The parameter space Θ_k is thus $[0, 1]^{2^k - 1}$. Associate with each component θ_g , the probability distribution $(1 - \theta_g) + \theta_g g$, and with the vector $\theta \in \Theta$ the probability distribution defined by the formal product

$$\pi_\theta = \prod_{g \in \mathcal{G}} ((1 - \theta_g) + g\theta_g).$$

This expression should be multiplied out and simplified using the group composition rules. A distribution π that can be factored into a product of linear factors is called factorizable.

(ii) What is the dimension of $\mathcal{P}(\mathcal{G}_k)$? What is the dimension of Θ ? Can every distribution in $\mathcal{P}(\mathcal{G}_k)$ be associated with a point $\theta \in \Theta$ and expressed as a formal product? If yes, give a proof. If not, give a counter-example.

(iii) A distribution π is infinitely divisible if for each integer $n \geq 1$ there exists a distribution ϕ depending on n such that $\pi = \phi^n$ is the n -fold formal product of the formal linear combination ϕ . Under what conditions is $1 - \theta + \theta g$ infinitely divisible?

Let $S \subset \mathcal{G}_k$ be a subset of \mathcal{G}_k , for example the identity and the k letters A_1, \dots, A_k . Consider the exponential family generated from the uniform distribution by exponential tilting using the operator

$$\exp\left(\sum_{g \in S} \lambda_g (g - 1)\right) \times 2^{-k}$$

with canonical parameter $\lambda \in \mathcal{R}^S$.

(iv) Show that every exponential model is factorizable, and give the relation between the two parameterizations, λ and θ .

(v) An iid sequence of 108 values resulted in the following counts for the group elements in \mathcal{G}_3 :

$$1 : 40, \quad A : 16, \quad B : 14, \quad AB : 8, \quad C : 15, \quad AC : 6, \quad BC : 9, \quad ABC : 0$$

The main-effects sub-model is the exponential-family model such that $\lambda_g = 0$ except for $g = 1, A, B, C$. Fit the main-effects exponential model and find the parameter estimates $\lambda_A, \lambda_B, \lambda_C$.

(vi) repeat part (v) with $S = \{1, A, B, C, AB\}$.