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.

Questions 0 and 3 to be handed in on May 3; question1 1 and 2 on May 8

0. Anomalous dispersion: Consider the context of Q0 from HW2, and suppose that the experimental set-up is as follows. In one connubial well the observed number of matings is zero, one or two, each with probability 1/3. If the number of matings is one, the mating type is homogamic with probability π_1, otherwise heterogamic with probability $1 - \pi_1$. If the number of matings is two, either both are homogamic with probability π_2, or both are heterogamic with probability $1 - \pi_2$. Events in distinct wells are independent and identically distributed. For one well, let N be the number of matings, and $Y \leq N$ the number of homogamic matings. You may assume that $\pi_2 \simeq \pi_1$ if necessary for simplification.

(i) What is the joint distribution of (Y, N)?

(ii) Calculate the conditional mean and variance of the ratio Y/N given $N = 1$ and $N = 2$. If $\pi_1 = \pi_2$, are the variables N and Y/N correlated?

(iii) In one generation, a single mating was observed in 26 wells, a double mating in 15 wells, the remaining 9 wells having no matings. Compute the mean and variance of $Y_\cdot$, the total number of homogamic matings in that generation, given the number of wells of each type.

(iv) Let F be a distribution on the integers $0, \ldots, m$ with mean $\mu = m\pi$ and variance $\sigma^2 m\pi(1 - \pi)$. The variance ratio σ^2 is called the dispersion factor of F relative to the binomial distribution, or the *dispersion factor* for brevity. Calculate the dispersion factor for the conditional distribution in the preceding question, given the number of wells of each type.

(v) Suppose that, on average, half of the matings occur in single-mating wells and half in double-mating wells. The total number of matings of each type (homogamic and heterogamic) is recorded, but the number of wells of each type is not. Compute the relevant dispersion factor for $Y_\cdot$.

(vi) Is there a possibility that the mechanism described here could explain the anomalous value of the Pearson chi-squared statistic for these data? Explain.

1. Toxicity of insecticides: Flour beetles *Tribolium castaneum* were sprayed with one of three insecticides in solution at different doses. The number of insects killed after a six-day period is recorded below:

	Deposit of insecticide (mg/10 cm^2)					
Insecticide	2.00	2.64	3.48	4.59	6.06	8.00
DDT	3/50	5/49	19/47	19/38	24/49	35/50
γ-BHC	2/50	14/49	20/50	27/50	41/50	40/50
DDT + γ-BHC	28/50	37/50	46/50	48/50	48/50	50/50

(a) Investigate graphically the relationship between the dose, either in original units or in log units, and the kill rate.

(b) On the graph for part (a), plot the linear logistic fitted curve for each of the insecticides plus the combination.

(c) Consider the two models, one in which the relationship is described by three parallel straight lines in the log dose and and one in which the three lines are straight but not parallel. Assess the evidence against the hypothesis of parallelism.

(d) Let *chem* be a 3-level factor, and let *ldose* be the log dose. Explain the relationship between the regression coefficients in the model formulae *chem + ldose* and *chem + ldose - 1*. Explain the relationship between the two covariance matrices.

(e) On the assumption that three parallel straight lines suffice, estimate the potency of the combination relative to each of the components. Use Fieller's method to obtain a 90% confidence interval for each of these relative potencies.

(f) Check to see if one of the alternative link functions probit, c-log log or log log, gives an appreciably better fit. Give the answer to part (e) for the c-log log model.

(g) Under the linear logistic model, estimate the combination dose required to give a 99% kill rate, and obtain a 90% confidence interval for this dose.

(h) Give a brief summary of your conclusions regarding the effectiveness of these three insecticides.

2. Fijian fertility: Table 1 gives the mean number of children born per woman, the women being classified by place, education, and years since first marriage. Any systematic variation in the number of children is of interest.

Table 1: Mean number of children born to women in Fiji of Indian race, by marital duration, type of place and education. Observed mean values and sample sizes.

Years since first marriage	Type of place: Urban, Education (1)	(2)	(3)	(4)	Type of place: Rural, Education (1)	(2)	(3)	(4)
<5	1.17	0.85	1.05	0.69	0.97	0.96	0.97	0.74
	12	27	39	51	62	102	107	47
5–9	2.54	2.65	2.68	2.29	2.44	2.71	2.47	2.24
	13	37	44	21	70	117	81	21
10–14	4.17	3.33	3.62	3.33	4.14	4.14	3.94	3.33
	18	43	29	15	88	132	50	9
15–19	4.70	5.36	4.60	3.80	5.06	5.59	4.50	2.00
	23	42	20	15	114	86	30	1
20–24	5.36	5.88	5.00	5.33	6.46	6.34	5.74	2.50
	22	25	13	3	117	68	23	2
25+	6.52	7.51	7.54	—	7.48	7.81	5.80	—
	46	45	13	0	195	59	10	0

Education categories are: (1) none, (2) lower elementary, (3) upper elementary, (4) secondary or higher. Lower figures give the number of women in the sample.

Fit an appropriate model describing how the number of children varies with marital age, mother's abode and education. Give a brief synopsis of the arguments justifying your formulation and choice of model, including checks for model adequacy.

Explain the meaning of all parameters in your model. Comment on the major factors affecting fertility.

Construct a 95% confidence interval for the mean number of children born to an urban woman with upper elementary education after ten years of marriage.

Estimate the lifetime average number of children born to rural women with secondary education. Give 90% confidence limits.

3. Incidence of byssinosis: The file `byss.txt` contains information, obtained from a survey conducted by a large textile company, on the prevalence of byssinosis, a lung disease to which cotton workers are subject. The file lists the observed prevalence of byssinosis `(affected, not affected)`, by `race` (white = 1; non white = 2), `sex` (male = 1; female = 2), `smok`ing habits (two levels), length of `empl`oyment (three levels), and `dust`iness of the work environment (three levels). In the last three cases, higher-numbered categories denote larger values (more smoking, longer employment and increased dustiness). Parts (a) and (b) are based on the assumption that the main-effects linear logistic model is substantially correct.

(a) Fit the main-effects linear logistic model. Explain how the residual degrees of freedom is calculated for the deviance.

(b) Interpret the coefficient of `sex(2)`. Construct an approximate 90% confidence interval for the odds ratio (males vs females) of contracting byssinosis.

(c) Drop the least significant factor from the model, proceeding until all the remaining factors are significant at the 5% level. Interpret the reduced model thus obtained.

(d) Beginning with the complete main-effects model, look for significant interactions by fitting each of the ten models *main effects + one interaction*. In judging the significance of interactions, you should bear in mind, at least informally, the effects of selection. After detecting the significant interactions, remove insignificant main effects as described in (c), except for those that are included in interactions. Interpret the model thus obtained.

(e) You are required to write a short report giving details of the excess risk associated with cotton dust. How fast does the risk increase with dust level? If necessary, give separate figures for males and females or for smokers and non-smokers.

(f) Does this analysis suggest that the aetiology of byssinosis is related to sex or race? Explain.

4. Let $Y_1, \ldots$ be independent and identically distributed according to the log series distribution

$$p_\theta(y) = \frac{\theta^y}{(y+1)\,M(\theta)}$$

on the non-negative integers $y = 0, 1, 2, \ldots$.

(i) Find the normalizing constant $M(\theta)$ as a function of θ. Show that these distributions determine a natural exponential-family of the GLM type. What is the range of θ-values for which this is a probability distribution?

(ii) For $\nu > 0$ and $\theta \in \Theta$, define the coefficients M_r^ν by

$$(M(\theta))^\nu = \sum_{r=0}^{\infty} M_r^\nu \theta^r$$

and assume that $M_r^\nu \geq 0$ for $\nu > 0$. Subject to this positivity condition, show that

$$M_y^\nu\, \theta^y \;/\; (M(\theta))^\nu \tag{1}$$

is a probability distribution of the exponential dispersion type on the non-negative integers. Find the cumulant function, and hence obtain the mean and variance of this distribution. Find the limit distribution as $\nu \to \infty$ and $\theta \to 0$ with $\lambda = \theta\nu$ fixed.

If needed, you can compute the first n coefficients numerically in Mathematica by typing

```
CoefficientList[Normal[Series[M[theta]^nu, {theta, 0, n}]], theta]
```

For fixed ν and large n, the coefficients behave as $M_n^\nu \sim \nu(\gamma + \log(n+\nu))^{\nu-1}/(n+\nu)$, where $\gamma = 0.57721..$ is Euler's constant. For fixed n and large ν, the coefficients behave as $M_n^\nu \sim (\nu/2)^{\uparrow n}/1^{\uparrow n}$ in the sense that the ratio tends to one as $\nu \to \infty$ for fixed n.

(iii) Show that

$$\sum_{y:y_\bullet=n} \frac{M_{y_1}^{\nu_1} \cdots M_{y_k}^{\nu_k}}{M_n^{\nu_\bullet}} = 1 \tag{2}$$

where the sum extends over non-negative integer vectors $(y_1, \ldots, y_k)$ whose components sum to n. Show that for all positive $\nu_1, \ldots, \nu_k$ the approximation $M_n^\nu \simeq (\nu/2)^{\uparrow n}/1^{\uparrow n}$ in (2) yields a probability distribution on non-negative integer vectors. Obtain the joint limit distribution as $\nu_1, \ldots, \nu_k$ tend to infinity in fixed proportion.

(iv) Stirling's number of the first kind S_m^c, the number of permutations of the set $[m] = \{1, \ldots, m\}$ having exactly c cycles, satisfies $S_m^m = 1$ for $m \geq 1$, since the identity permutation is the only one having m cycles. Show that $S_m^1 = (m-1)!$. By considering the possibilities for inserting a new element into an

existing permutation, show that $S^c_{m+1} = mS^c_m + S^{c-1}_m$, and hence deduce the first of the two generating functions below.

$$x^{\uparrow m} = x(x+1)\cdots(x+m-1) = \sum_{c=0}^{m} S^c_m x^c,$$

$$(-\log(1-x))^c = c! \sum_{m=c}^{\infty} S^c_m x^m/m!$$

Use the second generating function to find the relation between M^c_r and S^c_m for positive integer c.

(v) Consider the model in which $Y_1, \ldots, Y_k$ are independent and identically distributed with distribution (1), the parameter pair (θ, ν) being unknown. Show that for each fixed ν, $Y_\cdot$ is sufficient for θ. From the conditional distribution given $Y_\cdot$, obtain the maximum-likelihood estimate of ν, and the conditional likelihood-ratio statistic for testing the hypothesis that $Y_1, \ldots, Y_n$ are iid Poisson variables.

(vi) Consider a two-sample version of this problem in which $Y_{1,1}, \ldots, Y_{1,n_1}$ are iid with parameter θ_1, ν and $Y_{2,1}, \ldots, Y_{2,n_2}$ are iid with parameter θ_2, ν, all components being independent, and $\nu = 1$ known. Obtain the conditional distribution of $Y_{1\cdot}$ given $Y_{\cdot\cdot}$ under the assumption that $\theta_1 = \theta_2$, and more generally. Explain how you might use the conditional distribution to test the hypothesis that $\theta_1 = \theta_2$.

5. Let $q_1, \ldots, q_n$ be given weight functions $\mathcal{R} \to (0, \infty)$. Consider the statistical model with independent observations $Y_1, \ldots, Y_n$ such that Y_i has distribution

$$\mathrm{pr}(Y_i \in A) = \int_A q_i(x)\, d\mu(x) \Big/ \int_{\mathcal{R}} q_i(x)\, d\mu(x).$$

The parameter space consists of all probability distributions μ defined on Borel subsets on the real line.

(a) For each permutation π, show that the observations $(y_1, \ldots, y_n)$ and $y_{\pi(1)}, \ldots, y_{\pi(n)}$ determine the same likelihood function. Show that the empirical distribution function $\hat{P}_n$ is sufficient for the parameter. Deduce that the way in which the weight functions are associated with the observations is irrelevant.

(b) Obtain the maximum-likelihood estimator of μ.

(c) Suppose that you want to approximate the value of the integral over $(-1, 1)$ of the function

$$\frac{(1-y^2)^{3/2}}{(1-y+y^2)^2}$$

You have available ten observations $Y_i = U_i - V_i$ where U_i, V_i are independent and uniformly distributed on $(0, 1)$. The values are

$$-0.184,\ 0.264,\ -0.675,\ 0.243,\ -0.310,\ 0.222,\ -0.437,\ 0.624,\ -0.647,\ -0.529$$

Express this problem as a statistical model with μ as the parameter. Find the maximum-likelihood estimate of μ, and hence estimate the integral.

6. Let F_0, F_1 be distributions on the real line such that the log density ratio is a linear function $f_1(y)/f_0(y) = \exp(\alpha + \beta y)$. Let $\{(Y_i, x_i)\}$, $i = 1, \ldots, n$ be independent observations in which x_i is the indicator function for the population, and $Y_i \sim F_{x_i}$. Take as unknown parameters, the triple (F_0, α, β). Find the suffucient statistic S for F_0 when (α, β) is known, and obtain an equation for the maximum-likelihood estimate of F_0. Given S, derive the maximum-likelihood estimate of β. If necessary, you may assume that the xs are i.i.d. such that $x_i = 1$ with probability π. Comment on the nature of $\hat{F}_0$ and on the conditional likelihood for (α, β) given S.