

Due: Tuesday, Apr 1, by 5pm. I'll probably be in 1436 Wilson, but you may give it to Norma Elwick (reading room, 3rd floor Wilson), to put in my mailbox.

Remember, you are to do 3 of the following 4 problems. You can choose. My intent was to write 1 easier theory problem, 1 easier data analysis problem, 1 harder theory problem, and 1 harder data analysis problem. Your perception may differ; even so, choose 3 problems. If you do all 4, I'll grade them all and drop the lowest. Further information and suggestions for organizing answers to open ended data analysis problems (like problem 4) are on the 'homework information' part of the class web site.

I believe you can do everything in R or everything in SAS, but a combination of SAS and R will probably get answers more easily.

1. Two “just answer the questions” data analyses

- (a) The data in `deguelin.txt` are from a study of the effect of the toxicity of the “organic” insecticide deguelin on the Chysanthemum aphid, *Macrosiphoniella sanborni*. A photo, two drawings, and some background natural history can be found at <http://www.entomology.umn.edu/cues/inter/inmine/Aphidsc.html>; Chysanthemum aphid. The study included 6 doses from 5.12 to 50.12 (units unknown). Between 48 and 50 aphids were exposed to each dose; the response is the number of dead aphids.
 - i. Consider a linear logistic dose response model for these data. The model is:
 $\text{logit } \pi_i = \beta_0 + \beta_1 D_i$. Estimate the intercept, slope and LD_{50} . Calculate the standard errors for each.
 - ii. Consider a total of eight alternative models by combining all combinations of (logit or probit link function) \times (dose or log dose) \times (no control mortality or non-zero control mortality). The model in part a is one of these 8. Which model is the most reasonable choice for these data? Support your choice.
 - iii. Using the most appropriate model from part b, construct a reasonable 95% confidence interval for the LD_{50} .
 - iv. Using the most appropriate model from part b, estimate the Benchmark Dose for a Benchmark Risk = 5% and the lower 95% confidence bound. Because of the substantial apparent control mortality, the Benchmark Risk is defined in terms of the excess risk, that is 5% of the risk that is attributable to deguelin.
 - v. Find the NOEL (No observed effect level) and LOEL (Lowest observed adverse effect level). For the purpose of this part (and this part only), assume that the mortality for a dose of 0 is 14 aphids out of 48 tested.
- (b) The data in `ethyglycol.txt` are from a developmental toxicity test of ethylene glycol (the major component of antifreeze for car radiators) in rats. In the experiment, between 27 and 29 pregnant rabbits per dose were exposed to 0, 1250, 2500, or 5000 mg/kg/day of ethylene glycol. The response is the total number of dead or malformed babies in each litter. Each row in the data file represents one litter; the columns are the dose, the total number of deaths and malformations, and the litter size. If you prefer the totals per dose, they are:

	total dose	total deaths and litter size	malignancies
	0	412	25
	1250	382	46
	2500	369	110
	5000	363	270

- i. Is there evidence of overdispersion in these data? Explain why or why not.
 - ii. Fit a 3 parameter linear logistic model using log dose and a non-zero control mortality. Estimate the slope and LD_{50} and their standard errors. Adjust for overdispersion if appropriate.
 - iii. Test whether the slope = 0 after adjusting for overdispersion if appropriate.
2. In class, I said that the $1 - \alpha$ coverage confidence interval for the LCp was obtained by considering the pivotal quantity $\theta = \log\left(\frac{p}{1-p}\right) - \beta_0 - \beta_1 D$ and inverting tests of $\theta = 0$. Another way of saying the same thing is to compute

$$z = \frac{\log\left(\frac{p}{1-p}\right) - \hat{\beta}_0 - \hat{\beta}_1 D}{\sqrt{\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 D)}}$$

and find the values of D that reject $H_0: \log\left(\frac{p}{1-p}\right) - \beta_0 - \beta_1 D = 0$ at exactly $p = 1 - \alpha/2$.

I talked about the concepts and provided the answer in the Encycl. of Stats. article, but didn't derive the formula.

- (a) Assume that the sample sizes are sufficiently large that the estimated coefficients are approximately normal, i.e.

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} V_0 & C \\ C & V_1 \end{bmatrix}\right)$$

Please derive the $1 - \alpha$ confidence interval for LC_p .

Note: This derivation does not depend on whether the model is logit or probit or whether dose enters linearly or as log dose.

- (b) In class, our discussion has focused on the linear logistic dose-response model, $\text{logit } \pi_i = \beta_0 + \beta_1 D_i$. In some combinations of species and chemicals, there appears to be a threshold below which lower amounts of the chemical have similar effects,

$$\text{logit } \pi_i = \begin{cases} \beta_0 + \beta_1 \theta & D_i \leq \theta \\ \beta_0 + \beta_1 D_i & D_i > \theta \end{cases}$$

All three parameters (β_0 , β_1 , and θ) and their asymptotic variance-covariance matrix

$\begin{bmatrix} V_0 & C & C_{0t} \\ C & V_1 & C_{1t} \\ C_{0t} & C_{1t} & V_t \end{bmatrix}$ can be estimated by maximum likelihood and the inverse observed in-

formation. LC_{50} is still the concentration at which $P[\text{event}] = 0.5$. This is one difference between the threshold model and the 'background mortality' model discussed in class, in which the LC_{50} is usually defined in terms of excess risk, ignoring the background mortality. For the threshold model, what is the estimator of the LC_{50} given estimates of (β_0 , β_1 , and θ).

- (c) You are asked to compute a confidence interval for the LC_{50} in the threshold model. If the LC_{50} is sufficiently larger than the threshold, θ , can the confidence interval from part (a) be used without modification? If not, explain what sorts of modifications need to be made. Note: I'm not asking you to derive the confidence interval for the threshold model. Just say what changes (if any) need to be made to the approach used in part (a).

3. In class, I said that overdispersion models can be interpreted as a consequence of correlation between individuals in a litter. This is explored in this problem. Consider a beta-binomial model for responses Y_{ij} of individuals j in litter i . For this problem, there are no treatment effects (i.e. consider data from a single dose). The model is:

$$\begin{aligned} Y_{ij} | p_i &\sim \text{independent Bernoulli}(p_i) \\ p_i &\sim \beta(\alpha, \beta) \end{aligned}$$

A Bernoulli distribution is a Binomial distribution with 1 event, i.e. $\text{Bernoulli}(p_i)$ is the same as $\text{Binomial}(1, p_i)$.

- (a) Derive the mean and variance of Y_{ij} .

Note: I am interested in the marginal distribution, not the conditional distribution. It may help to know that $E p_i = \frac{\alpha}{\alpha+\beta}$ and $\text{Var } p_i = \frac{\alpha}{\alpha+\beta} \frac{\beta}{\alpha+\beta} \frac{1}{\alpha+\beta+1}$.

- (b) Derive the covariance of Y_{ij} and Y_{ik} , that is the covariance of responses from two individuals in the same litter.

Hint: The conditional variance formula used a couple of times in class has a conditional covariance extension:

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(E Y_{ij} | p_i, E Y_{ik} | p_i) + E \text{Cov}(Y_{ij}, Y_{ik} | p_i)$$

- (c) Derive the correlation between responses from two individuals in the same litter

- (d) Derive the correlation between responses from two individuals in different litters.

4. The data in chlorpyrifos.txt are from a sediment toxicity test measuring the effect of the pesticide chlorpyrifos on the harpacticoid copepod *Amphiascus tenuiremis*. Sediment tends to accumulate chemicals present in the overlying water because many chemicals bind to clay particles. *Amphiascus* is a benthic invertebrate commonly used in sediment toxicity tests because they live in the top 1-2 cm of the sediment (where chemicals from the water tend to accumulate) and they are an important food source for fish and other aquatic life. A picture of a closely related critter is at <http://entomology.tfrec.wsu.edu/pearent/images/misc/Crustacea/harpact.jpg>

These copepods have life stages, rather like frogs. Frogs have 2 stages, tadpoles and adults that are very different. Copepods have more, but I'm focusing on two: nauplii and adults. The life stages of copepods differ in their sensitivity to chemicals. I am providing data from nauplii (stage = N), which are very young individuals, and adults (stage = A). These tests are done by putting 25 adults or 20 nauplii in a small beaker, adding sediment with different concentrations of the contaminant, and recording mortality after 96 hours. There are replicate beakers at each dose. The doses used for adults are not the same as those used for the nauplii. I don't know why, but my guess is because it is hard to dissolve most pesticides in water, so it is difficult to prepare a solution at a specific concentration. Once a solution is prepared, the concentration of chlorpyrifos can be measured accurately. The data file has one row for each beaker, with the stage, the dose, the number of live animals (count) and the number of individuals put in that beaker (n). Note that the % mortality is $100(1-\text{count}/n)$.

The investigators want to know:

For the adults: The LC_{50} and a 95% confidence interval

For the adults: The Benchmark Dose for a 5% excess risk (i.e. 5% mortality above background), using a 95% confidence bound.

Whether the LC_{50} 's for the adults and nauplii are the same. They really need a formal test (e.g. Wald or likelihood ratio test). Comparing confidence intervals is not sufficient.