PhD Prelim Exam
METHODS

Summer 2005
(Given on 7/19/05)

Fertilizers increase crop yield, but the response to fertilization is not linear. Adding 400 lb/ac does not produce twice the yield as adding 200 lb/ac. Eventually, yield reaches a maximum (see figure 1a, below); adding additional fertilizer provides little to no additional yield. Since fertilizer is expensive and environmentally damaging, farmers want to avoid excess fertilization.

The Michaelis-Menton equation is often used to relate the mean yield to the amount of fertilizer.

$$\mu_{Y|X} = \frac{\alpha_0}{1 + \alpha_1/X}, \tag{1}$$

where $\mu_{Y|X}$ is the expected yield from a plot with X amount of fertilizer. The parameter α_0 is the maximum yield; the parameter α_1 controls the steepness of the curve. One of the many ways to estimate (α_0, α_1) is to linearize the curve:

$$\frac{1}{\mu_{Y|X}} = \frac{1}{\alpha_0} + \left(\frac{\alpha_1}{\alpha_0}\right)\frac{1}{X} \tag{2}$$

and consider the regression model:

$$Y_i^* = \beta_0 + \beta_1 X_i^* + \epsilon_i, \tag{3}$$

where:

$$\begin{aligned} Y_i^* &= 1/Y_i \\ X_i^* &= 1/X_i \\ \beta_0 &= 1/\alpha_0 \\ \beta_1 &= \alpha_1/\alpha_0 \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

The data for this problem are from a study of corn response to nitrogen fertilizer in Washington state. Plots were randomly assigned to one of 11 fertilizer amounts, with three replicates of each fertilizer amount. There are a total of 33 observations. Data (X,Y) are plotted in figure 1a; transformed values (X^*,Y^*) are plotted in figure 1b.

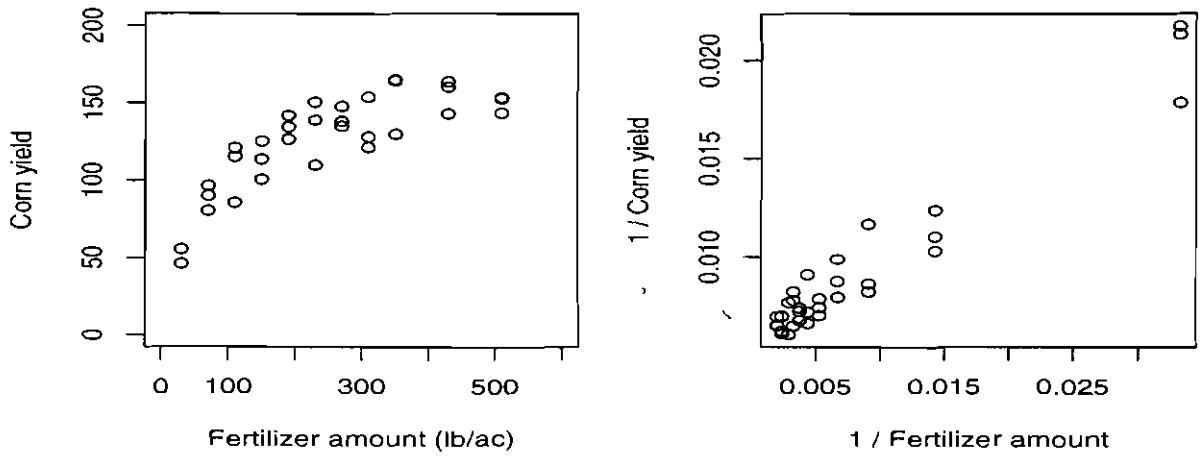

Figure 1: a) Corn yield as a function of amount of applied fertilizer. b) Plot of transformed values (1/Fertilizer and 1/Yield)

a. The researchers estimated β_0 and β_1 in equation 3 by ordinary least-squares regression. They want to know whether this is a reasonable approach. Figure 2 gives various diagnostic plots:
a) a normal quantile-quantile plot of the observations,
b) a plot of residuals vs predicted values,
c) a plot of Cook's distance against fertilizer amount, and
d) a plot of h_{ii}, the diagonal elements of the "hat" matrix, $X(X'X)^{-1}X'$, vs fertilizer amount.
For each of these plots, what does that plot tell you, if anything, about the appropriateness of model (3) for these data? One or two sentences for each plot should be sufficient.

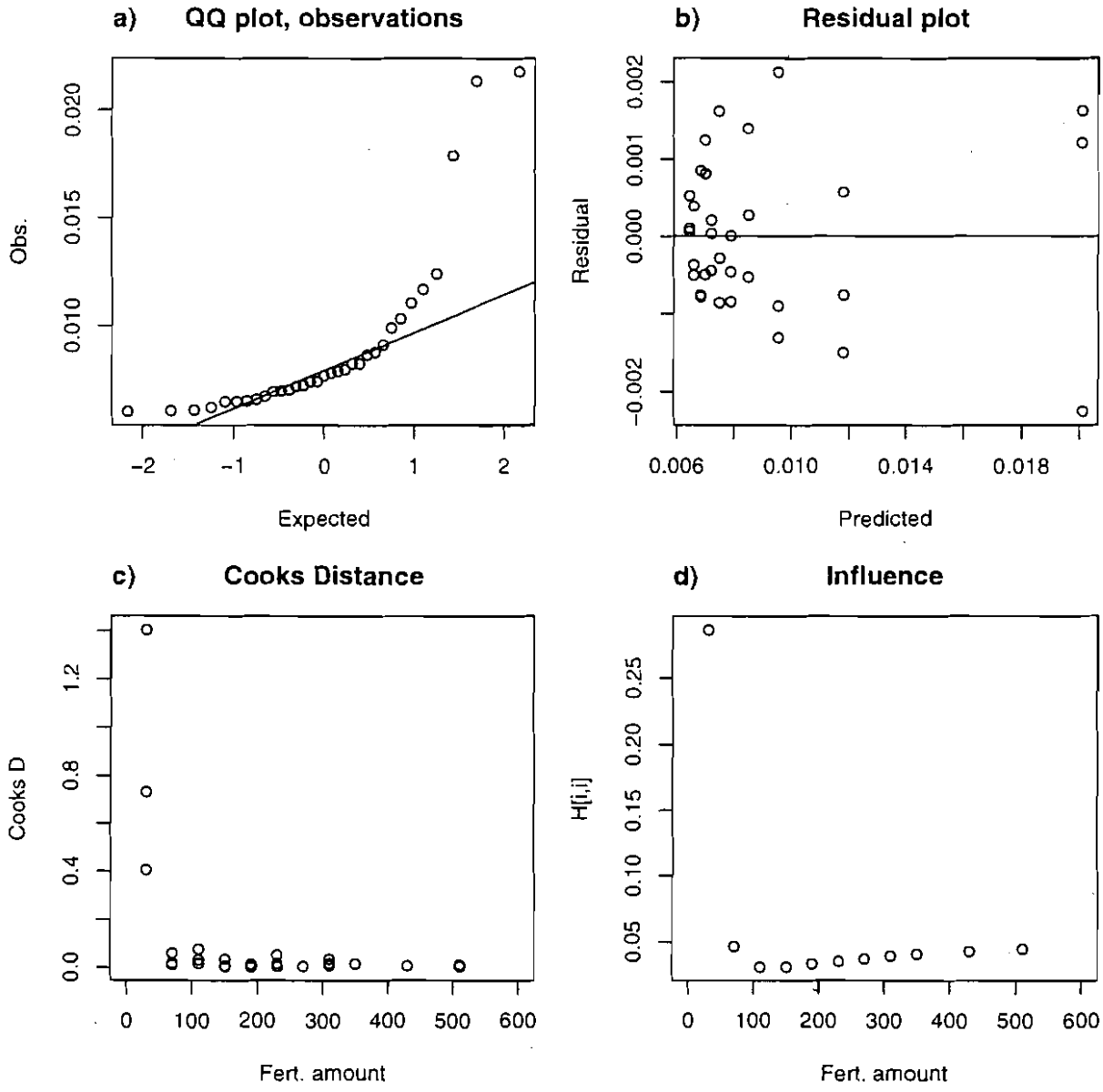

Figure 2: Diagnostic plots.

For subsequent questions, assume that OLS regression is reasonable.

b. Estimates of (β_0, β_1) and their estimated variance-covariance matrix are:

Parameter	estimate	Variance-covariance matrix	
β_0	0.00560	5.608E-8	-3.190E-6
β_1	0.435	-3.190E-6	0.0004031

Note: the notation 5.608E-8 is "scientific notation", representing the number 0.00000005608.

Construct a 95% confidence interval for β_1.

c. The researchers are especially concerned about linearity of the regression (equation 3) and ask you to construct a formal test of lack of fit of equation 3. I have fit various models that might be useful in constructing such a test.

model	equation	Error d.f.	Error SS
A	$Y_i^* = \beta_0 + \epsilon_i$	32	0.0005102
B	$Y_i^* = \beta_0 + \beta_1 X_i^* + \epsilon_i$	31	0.00003154
C	$Y_i^* = \beta_0 + \beta_1 X_i^* + \beta_2 (X_i^*)^2 + \epsilon_i$	30	0.00003091
D	$Y_i^* = \beta_1 I(X_i = 40) + \beta_2 I(X_i = 80) +$ $\cdots + \beta_{11} I(X_i = 520) + \epsilon_i$	22	0.00002818

Note that I() is the indicator function, so model D is a one-way ANOVA model that fits a separate mean to each amount of fertilizer.

Use the relevant information to construct the most appropriate test of lack of fit. Report your test statistic, the p-value (at least approximately), and a short conclusion.

d. You eventually discover that the study used a randomized complete block design. The experimental field was divided into three blocks according to elevation and soil type. Each of the 3 blocks had 11 plots with similar elevation and the same soil type. The 11 fertilizer amounts were randomly assigned to the 11 plots in a block. The data are plotted, with letters (A,B,C) to indicate the block, in figure 3.

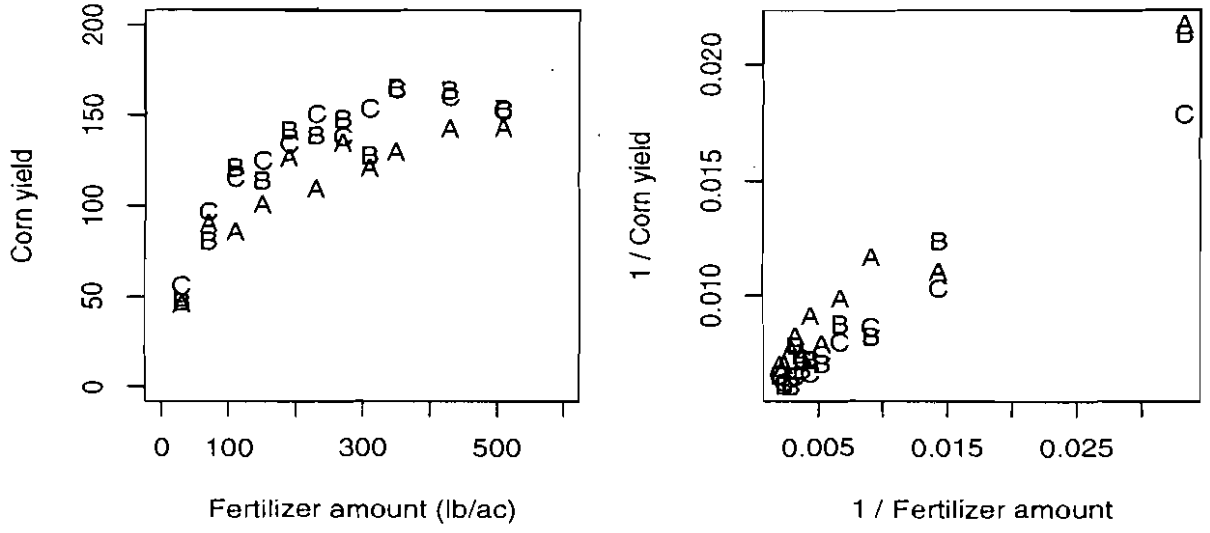

Figure 3: a) Corn yield as a function of amount of applied fertilizer. b) Plot of transformed values (1/Fertilizer and 1/Yield), with letters indicating the block.

The maximum corn yield (α_0) in the Michaelis-Menton model (equation 1) varies with elevation and soil type, so it is likely to differ between the three blocks. You do not know whether α_1 varies between blocks.

Write out a linearized regression model (i.e. an extension of equation 3), that accounts for block-block variation in maximum yield. Define all subscripts and variables.

e. For this study, should block effects be treated as fixed effects or as random effects? Explain.

The remaining questions are based on the OLS model (equation 3) fit only to the 11 data points from block C. The OLS estimates and the variance-covariance matrix are:

Parameter	estimate	Variance-covariance matrix	
β_0	0.00539	1.598E-8	-9.094E-7
β_1	0.369	-9.094E-7	0.0001149

f. The parameters of the Michaelis-Menton model (α_0 and α_1) are transformations of the parameters in the regression model (β_0, and β_1).

Estimate α_0 and α_1 for block C.

Are the regression estimates ($\hat{\beta}_0$, $\hat{\beta}_1$) maximum-likelihood estimates? Are the transformed estimates ($\hat{\alpha}_0$, $\hat{\alpha}_1$) maximum-likelihood estimates? Briefly explain why or why not.

g. Estimate an approximate variance for $\hat{\alpha}_1$.

h. Farmers are often interested in estimating the amount of fertilizer that gives 80% of the maximum yield. Denote this amount $X_{0.80}$.

Derive an estimator of $X_{0.80}$,
estimate $X_{0.80}$, and
calculate an asymptotic 95% confidence interval for $X_{0.80}$.

Note: If you could not answer the two previous questions, use $\hat{\alpha}_1 = 70$ and Var $\hat{\alpha}_1 = 9$ to answer this question.

i. The data set is small (11 observations), so you are not sure whether the asymptotic confidence interval calculated in part h is reasonable. Describe how you could construct a "small sample" confidence interval. No calculations needed here.

j. Carry out a small sample, exact test of the hypothesis $X_{0.80} = 290$. Report your test statistic and an approximate p-value.

Hint: Consider the distribution of $4\hat{\beta}_1 - 290\hat{\beta}_0$.

Mark and Steve,

These are not the final answers to be distributed. However, they should be sufficient to indicate what I'm looking for in each question.

a. The model seems generally appropriate.
There may be very slight inequality of variance (residual plot, b), but there are no signs of outliers.
I am a bit concerned about the response at the lowest fertilizer levels. Those three points have unusually high influence (influence plot, d) and large residuals (Cooks D plot, c).

Note: the QQ plot of the observations is irrelevant.

b. s.e. $\hat{\beta}_1 = \sqrt{0004031} = 0.020$. error d.f. = 33-2=31, $t_{31,0.975} = 2.039$, so 95% ci is $0.435 \pm (2.039)(0.020) = (0.394, 0.476)$.

c. Since there are replicate observations at the same fertilizer amount, and the researchers have not indicated a specific interest in one type of deparature from linearity, the most appropriate test is to compare the regression to a one-way ANOVA model. The test statistic is an F statistic calculated from the change in error SS.

$$F = \frac{(0.00003154 - 0.00002818)/(31 - 22)}{0.00002818/22} = 0.291$$

This has 9,22 d.f. The F-value is smaller than any tabulated value, so p> 0.50 (or 0.10, depending on what's in the tables).

There is no evidence of lack of fit.

d. Both the slope and intercept of the regression model (β_1 and β_0) depend on the maximum yield α_0, so both must vary with the block. The most reasonable model is:

$$Y^*_{ij} = \beta_{0i} + \beta_{1i} X^*_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

where i indicates the block, j the plot within the block, β_{0i} is the intercept for block i, and β_{1i} is the slope for block i.

Note: an answer like

$$Y^*_{ij} = \beta_0 + \alpha_i + \beta_1 X^*_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

will get partial credit.

e. The blocks here are specific elevations and soil types. They should be considered fixed. Also, the number of blocks is too small to make reliable inferences about a random effect.

f. Using the relationships between parameters in model (1) and model (3), the inverse transformations are: $\alpha_0 = 1/\beta_0$ and $\alpha_1 = \beta_1/\beta_0$. The "plug-in" estimators are:

$$\hat{\alpha}_0 = 1/\hat{\beta}_0 = 1/0.00539 = 185, \text{ and}$$

$$\hat{\alpha}_1 = \hat{\beta}_1/\hat{\beta}_0 = 0.369/0.00539 = 68.5$$

The OLS estimates of β_0, β_1 for model (3) are mle's because the error distribution is iid normal. $(\hat{\alpha}_0, \hat{\alpha}_1)$ are also mles, because they are an invertible function of mle's.

g. Get this by delta method (more details in final version of the answer). I get Var $\hat{\alpha}_1$ = 10.77

h. The maximum yield is α_0, so we need to find $X_{0.80}$ that solves

$$0.80\alpha_0 = \frac{\alpha_0}{1 + \alpha_1/X_{0.80}}$$

This is $X_{0.80} = \frac{0.8}{0.2}\alpha_1$, so the mle of $X_{0.80}$ is $4\hat{\alpha}_1 = (4)(68.5) = 274$.

Var $\hat{X}_{0.80} = (4^2)$Var $\hat{\alpha}_1 = (16)(10.77) = 172.3$.

A 95% ci is $274 \pm t_{31,0.975}\sqrt{172.3} = 274 \pm (2.039)(13.1) = (247, 301)$.

One could also use a normal quantile, since the inference is asymptotic.

i. Use a regression bootstrap. That is: fit the regression, estimate the residuals and predicted values for each X value. The bootstrap data sets are constructed by randomly sampling with replacement from the set of residuals and adding those the the predicted values for each of the 11 fertilizer amounts. Estimate $X_{0.80}$ for each bootstrap data set and derive the bootstrap confidence interval from the estimated distribution of $\hat{X}_{0.98}$. The simple bootstrap c.i. is the 0.025 and 0.975 quantiles of the bootstrap distribution. More complicated bootstrap estimators are often better.

j. The following hypotheses are all equivalent:

$$\begin{aligned} X_{0.80} &= 290 \\ 4\alpha_1 &= 290 \\ 4\beta_1/\beta_0 &= 290 \\ 4\beta_1 - 290\beta_0 &= 0 \end{aligned}$$

Define $U = 4\hat{\beta}_1 - 290\hat{\beta}_0$. The estimate is $U = 1.476 - 1.563 = -0.087$. Under H0: $X_{0.80}$ = 290, E $U = 0$ and
Var $U = 4^2$Var $\hat{\beta}_1 + 290^2$Var $\hat{\beta}_0 + 2(4)(-290)$Cov $\hat{\beta}_0, \hat{\beta}_1 = 0.00529$. U is normally distributed since $\hat{\beta}_0$ and $\hat{\beta}_1$ are bivariate normal. Hence, the small sample exact test uses the test statistic:

$$T = \frac{U - 0}{\sqrt{\text{Var } U}},$$

which has a t distribution with 31 d.f. I get

$$T = \frac{-0.087 - 0}{\sqrt{0.00529}} = -1.19$$

which has a p-value > 0.20.

Note: It is possible to calculate a small-sample exact confidence interval for $X_{0.80}$ by inverting a set of hypothesis tests of the form H0: $4\beta_1 - k\beta_0 = 0$.

A process engineer wishes to determine whether a change made to a chemical process has an important impact on the mean yield associated with a run of the process. A complicating issue in this regard is that each batch of raw material is sufficient to make only a few process runs, and different batches can be expected to have different characteristic yields.

Throughout this question, we will use a model for

$$y_{ijk} = \text{yield for run } k \text{ made using process } i \text{ and raw material batch } j$$

of the form

$$y_{ijk} = \mu_i + \beta_j + \varepsilon_{ijk} \qquad (*)$$

where the μ_i $(i = 1, 2)$ are unknown constants, the β_j are iid $\mathrm{N}\left(0, \sigma_\beta^2\right)$ independent of the ε_{ijk} that are themselves iid $\mathrm{N}\left(0, \sigma^2\right)$, and the variance components σ_β^2 and σ^2 are unknown constants.

Suppose initially that each raw material batch is sufficient to make 2 runs, and that 4 batches of raw material are available to the engineer for the study. Two possible plans for data collection are:

Plan I	**Plan II**
1 run from each raw material batch is made with each process (a total of 4 runs are made with each process)	2 raw material batches are dedicated to each process (a total of 4 runs are made with each process)

1. For

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

write out matrices $\mathbf{X}$ and $\mathbf{Z}$ so that the model (*) for 8 observations can be represented in usual matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Do this first for **Plan I** and then for **Plan II**. In the first case, write the observations in the order

$$\mathbf{Y} = \begin{pmatrix} y_{111} \\ y_{211} \\ y_{121} \\ y_{221} \\ y_{131} \\ y_{231} \\ y_{141} \\ y_{241} \end{pmatrix}$$

In the second case, write the observations in the order

$$\mathbf{Y} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{231} \\ y_{232} \\ y_{241} \\ y_{242} \end{pmatrix}$$

2. What is the covariance matrix for **Y** (in the order indicated above) under **Plan 1**? Under **Plan 2**?

3. If all 4 unknown parameters were of some interest one might consider comparing **Plan 1** and **Plan 2** using appropriate 4×4 Fisher information matrices. Use the notation

$\mathbf{D}\left(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2\right)$ = the 4×4 Fisher Information matrix for $\mathbf{U} \sim \text{MVN}_2\left(\begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma^2 + \sigma_\beta^2 \end{pmatrix}\right)$

$\mathbf{E}\left(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2\right)$ = the 4×4 Fisher Information matrix for $\mathbf{V} \sim \text{MVN}_2\left(\begin{pmatrix} \mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma^2 + \sigma_\beta^2 \end{pmatrix}\right)$

$\mathbf{F}\left(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2\right)$ = the 4×4 Fisher Information matrix for $\mathbf{W} \sim \text{MVN}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma^2 + \sigma_\beta^2 \end{pmatrix}\right)$

and write the Fisher Information matrices associated with **Plan I** and with **Plan II** in terms of these matrices.

4. Let $\bar{y}_{1..}$ be the arithmetic average of the process 1 observations and $\bar{y}_{2..}$ be the arithmetic average of the process 2 observations. What are the mean and standard deviation of

$$\bar{y}_{1..} - \bar{y}_{2..}$$

under **Plan 1**? Under **Plan 2**? (These can be found without appeal to matrix representations.)

5. For purposes of comparing μ_1 and μ_2, what does your answer to part 4 indicate about which of the two plans will typically be most effective?

Below are two hypothetical data sets, one corresponding to **Plan 1** and one corresponding to **Plan 2**.

Plan I		
Process	Batch	Yield
1	1	82.0
2	1	78.6
1	2	71.8
2	2	75.6
1	3	80.0
2	3	78.8
1	4	77.6
2	4	77.8

Plan II		
Process	Batch	Yield
1	1	82.0
1	1	79.2
1	2	71.9
1	2	76.3
2	3	79.3
2	3	78.9
2	4	77.0
2	4	77.8

6. For **both plans**, show the simple "by hand" calculations necessary to make valid/exact 95% t confidence intervals for $\mu_1 - \mu_2$.

7. Simple valid/exact 95% χ^2 confidence limits for σ^2 can be made from either set of hypothetical data above. Choose one of the plans and show the "by hand" calculations needed.

In the real application motivating this problem, practical constraints dictated that all runs from a given raw material batch had to be made consecutively, batches were of different sizes, and all runs from process 1 had to be made before runs from process 2. In fact, 4 small batches were dedicated to process 1, 1 larger batch was split between the two processes, and 1 batch of moderate size was dedicated to process 2. Attached to this question is an R printout useful in the analysis of the engineer's data. Use it in answering the following questions.

8. Is there a statistically significant difference between the processes? Explain, referring carefully to appropriate items on the printout.

9. How does run-to-run variability in yield appear to compare with batch-to-batch variability? Explain, again referring to appropriate items on the printout.

10. The engineer in charge of this study says to you "We need to redo this study. We'll need to run process 1 before process 2. I can get raw material batches big enough to make as many as $r = 10$ runs per batch. We'll run the same number of batches, l, with each process (splitting no batch between processes). I want to estimate $\mu_1 - \mu_2$ to within .5. I'd like to minimize the total number of runs made

$$\text{total runs made} = 2lr$$

in meeting this goal." How many batches should we use for this study, and how many runs per batch should we make?

Find this person appropriate values of r and l on the basis of the estimates on the printout.

R Printout

```
> data
   process batch     y
1        1     1 82.72
2        1     1 78.31
3        1     1 82.20
4        1     1 81.18
5        1     2 80.06
6        1     2 81.09
7        1     3 78.71
8        1     3 77.48
9        1     3 76.06
10       1     4 87.77
11       1     4 84.42
12       1     4 84.82
13       1     5 78.61
14       1     5 77.47
15       1     5 77.80
16       1     5 81.58
17       1     5 77.50
18       2     5 78.73
19       2     5 78.23
20       2     5 76.40
21       2     6 81.64
22       2     6 83.04
23       2     6 82.40
24       2     6 81.93
25       2     6 82.96

> Process<-as.factor(process)

> Batch<-as.factor(batch)

> output.1<-lme(y~1+Process,random=~1|Batch)

> summary(output.1)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  108.6438 113.1858 -50.32191

Random effects:
 Formula: ~1 | Batch
        (Intercept) Residual
StdDev:    2.927192 1.467032

Fixed effects: y ~ 1 + Process
               Value Std.Error DF  t-value p-value
(Intercept) 81.05442  1.260345 18 64.31128  0.0000
Process2    -0.67123  1.019483 18 -0.65841  0.5186
```

```
 Correlation:
         (Intr)
Process2 -0.19


Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-1.901566862 -0.557726122 -0.005590905  0.505906835  2.018904889

Number of Observations: 25
Number of Groups: 6

> intervals(output.1)
Approximate 95% confidence intervals

 Fixed effects:
                lower       est.     upper
(Intercept) 78.406534 81.0544212 83.702308
Process2    -2.813087 -0.6712329  1.470621
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Batch
                    lower     est.    upper
sd((Intercept)) 1.501453 2.927192 5.706776

 Within-group standard error:
   lower     est.    upper
1.059645 1.467032 2.031042


> predict(output.1,level=0:1)
   Batch predict.fixed predict.Batch
1      1       81.05442      81.09966
2      1       81.05442      81.09966
3      1       81.05442      81.09966
4      1       81.05442      81.09966
5      2       81.05442      80.62849
6      2       81.05442      80.62849
7      3       81.05442      77.69771
8      3       81.05442      77.69771
9      3       81.05442      77.69771
10     4       81.05442      85.31342
11     4       81.05442      85.31342
12     4       81.05442      85.31342
13     5       81.05442      78.61820
14     5       81.05442      78.61820
15     5       81.05442      78.61820
16     5       81.05442      78.61820
17     5       81.05442      78.61820
18     5       80.38319      77.94697
```

19	5	80.38319	77.94697
20	5	80.38319	77.94697
21	6	80.38319	82.29782
22	6	80.38319	82.29782
23	6	80.38319	82.29782
24	6	80.38319	82.29782
25	6	80.38319	82.29782

1. Plan I

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad Z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Plan II

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \qquad Z = \text{as above}$$

2. In both cases

$$\text{Cov } Y = Z \underset{4\times 4}{\sigma_\beta^2 I} Z' + \sigma^2 \underset{8\times 8}{I} = \sigma_\beta^2 \underset{4\times 4}{I} \otimes \underset{2\times 2}{J} + \sigma^2 I$$

(variances are all $\sigma_\beta^2 + \sigma^2$, covariances for observations from the same lot are σ_β^2, and all other covariances are 0)

3. Fisher Information for independent observations adds so for Plan I the FI is

$$4F(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2)$$

While for Plan II the FI is

$$2D(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2) + 2E(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2)$$

4. For Plan I

$$\bar{y}_{1..} - \bar{y}_{2..} = \left(\mu_1 + \bar{\beta}_. + \tfrac{1}{4}(\epsilon_{111} + \epsilon_{121} + \epsilon_{131} + \epsilon_{141})\right) - \left(\mu_2 + \bar{\beta}_. + \tfrac{1}{4}(\epsilon_{211} + \epsilon_{221} + \epsilon_{231} + \epsilon_{241})\right)$$

$$= \mu_1 - \mu_2 + \epsilon_{1..} - \epsilon_{2..}$$

So $E(\bar{y}_{1..} - \bar{y}_{2..}) = \mu_1 - \mu_2$ and

$$\text{Var}(\bar{y}_{1..} - \bar{y}_{2..}) = 2\left(\frac{\sigma^2}{4}\right) = \frac{1}{2}\sigma^2$$

So $\sqrt{\text{above}} = \frac{1}{\sqrt{2}}$

For Plan II

$$\bar{y}_{1..} - \bar{y}_{2..} = \left(\mu_1 + \tfrac{1}{2}(\beta_1 + \beta_2) + \epsilon_{1..}\right) - \left(\mu_2 + \tfrac{1}{2}(\beta_3 + \beta_4) + \epsilon_{2..}\right)$$

$$= \mu_1 - \mu_2 + \tfrac{1}{2}(\beta_1 + \beta_2 - \beta_3 - \beta_4) + \epsilon_{1..} - \epsilon_{2..}$$

So $E(\bar{y}_{1..} - \bar{y}_{2..}) = \mu_1 - \mu_2$

$$\text{Var}(\bar{y}_{1..} - \bar{y}_{2..}) = \left(\tfrac{1}{2}\right)^2 4(\sigma_\beta^2) + \frac{1}{2}\sigma^2$$

$$= \sigma_\beta^2 + \frac{1}{2}\sigma^2$$

So $\sqrt{\text{above}} = \sqrt{\sigma_\beta^2 + \frac{1}{2}\sigma^2}$

5. Plan I is better, producing the smaller variance for the estimated difference in means

6. For Plan I, we can base an interval on 4 paired differences. These are

lot 1: $82.0 - 78.6 = 3.4$
lot 2: $71.8 - 75.6 = -3.8$

lot 3 : $80.0 - 78.8 = 1.2$

lot 4 : $77.6 - 77.8 = -.2$

So $\bar{d} = .15$, $s_d = 3.0216$ and the (3 d.f.) t interval is

$$.15 \pm 3.182 \frac{3.0216}{\sqrt{4}} \quad \text{i.e.} \quad .15 \pm 4.807$$

For Plan II, we may make 4 batch mean responses. The first 2 with mean μ_1 and the 2nd 2 with mean μ_2. These can be used to make a difference in 2 sample means and a pooled (2 d.f.) estimate of variance (of a single lot mean) and thus a 2-sample interval

batch means are:

lot 1 : 80.6	lot 3 : 79.1
lot 2 : 74.1	lot 4 : 77.4
mean 77.35	mean 78.25
variance 21.125	variance 1.445

So a (2 d.f.) interval is

$$(77.35 - 78.25) \pm 4.303 \sqrt{\frac{21.125 + 1.445}{2}} \sqrt{\frac{1}{2} + \frac{1}{2}}$$

$$-.9 \pm 14.46$$

7. The 4 within batch differences are independent and have variance $2\sigma^2$. For Plan I these have 0 mean, while for Plan II they each have mean $\mu_1 - \mu_2$. So for Plan I, a 4 d.f. interval for $2\sigma^2$ can be made based on the

sum of squares of these differences. For Plan II a 3df interval for $2\sigma^2$ can be made based on their sample variance. In either case, then dividing limits by 2 gives limits for σ^2.

8. There is no statistically significant difference. The interval

$$(-2.813,\ 1.471)$$

is for $\mu_2 - \mu_1$ (The R convention sets level 1 as a baseline, so the "effect" of "process 2" is the incremental effect.) The interval covers 0.

9. $\hat{\sigma}_\beta = 2.33$ and $\hat{\sigma} = 1.47$ So appearances are that batch-to-batch variability is larger than the run-to-run variability

10. r runs from a given batch produce a sample mean with variance

$$\sigma_\beta^2 + \frac{\sigma^2}{r}$$

An average of l (independent) such sample means has variance

$$\frac{1}{l}\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right)$$

So, $\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$ will have mean $\mu_1 - \mu_2$ and variance

$$\frac{2}{l}\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right)$$

Then, in rough terms, the engineer wants

$$2\sqrt{\frac{2}{l}\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right)} \approx .5$$

That is,

$$\frac{2}{l}\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right) \approx \frac{1}{16}$$

$$l = 32\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right)$$

And the engineer wants to minimize (over choice of $r = 1, 2, \ldots, 10$)

$$rl = 32\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right) r$$

$$= 32\left(r\sigma_\beta^2 + \sigma^2\right)$$

Obviously, $r = 1$ is best. So then, what remains is the choice of

$$l = 32\left(\sigma_\beta^2 + \sigma^2\right)$$

and plugging in the estimates from R, I'd suggest

$$l = 32\left((2.93)^2 + (1.47)^2\right) = 344$$

OUCH! This is "too big" something will have to give ... either the roughly 95% "confidence level" or the "to within .5" requirement.

PhD Preliminary Examination – 2005

Methods Question 3

1 Problem Background

This question concerns a series of studies conducted as part of a preliminary investigation into potential health benefits of a genetically modified strain of alfalfa. While the history of how these studies came to be conducted is interesting, we will condense the presentation here for purposes of brevity. The work concerns a particular type of genetically modified alfalfa, know to contain a compound called Resveratrol that might help protect against colon cancer. Regular alfalfa does not contain this compound. For various reasons, over a period of about 2 years, 5 studies were conducted to offer preliminary evidence that the Resveratrol contained in this "transgenic" strain of alfalfa might help prevent colon cancer. Note that these were all preliminary *investigative* studies, not meant to result in confirmatory inference relative to the effects of either Resveratrol or its presence in the genetically modified alfalfa.

The response of interest is called "aberrant crypt foci" (ACF), essentially a cluster of abnormal cells (or a tumor) in the colon. The studies involved mice that were all fed a standard diet for one week and were then injected with a single dose of azoxymethane at 5mg/kg body weight; azoxymethane is known to induce tumors of the type of concern. Mice were then put on a variety of (randomly) assigned diets for 5 weeks, sacrificed, and the number of ACF counted in three regions of the colon; "Proximal" (near the stomach), "Medial" (mid-colon) and "Rectal" (near the end of the colon). The total (i.e., sum) of these three counts was then also easily recorded. Thus, data were recorded for counts of total ACF, proximal ACF, medial ACF, and rectal ACF.

It was also known (prior to the conduct of studies 3, 4, and 5, but not studies 1 and 2) that the Resveratrol contained in the genetically modified alfalfa is "encapsulated" in (i.e., coated with) a galactose compound. Galactose is a sugar, like glucose, but is not very soluble in water. That is, galactose does not dissolve easily.

Thus, the encapsulation of Resveratrol in galactose may prohibit this compound from becoming biologically available to the mice fed the genetically modified alfalfa. Alpha-galactosidase is an enzyme that breaks down galactose. The thought is that a diet containing the genetically modified alfalfa plus galactosidase may result in Resveratrol being available to the mice fed that diet. In total, the treatments that were employed in the five studies under consideration are described as follows.

1. Basal Diet (BD).

   This was your standard mouse pellet, of the type you might find at pet stores for gerbils and hamsters (i.e., "Purina Mouse Chow").

2. Control Alfalfa (CA).

   A diet that contained regular (not genetically modified) alfalfa, which is known to not contain Resveratrol.

3. Transgenic Alfalfa (TA).

   A diet that contained the genetically modified alfalfa, which is known to contain some Resveratrol.

4. Control Alfalfa plus Resveratrol (CAR).

   A diet that contained control alfalfa, but to which was added directly the compound Resveratrol.

5. Basal Diet plus Resveratrol (BDR).

   A diet that contained the basal diet (standard mouse food), but to which was added directly the compound Resveratrol.

6. Transgenic Alfalfa plus Alpha-Galactosidase (TA-α).

   A diet that contained the genetically modified alfalfa, but to which was added the enzyme alpha-galactosidase.

7. Basal Diet plus Alpha-Galactosidase (BD-α).

   A diet consisting of the basal diet, but to which was added the enzyme alpha-galactosidase.

8. Control Alfalfa plus Alpha-Galactosidase (CA-α).

   A diet consisting of control alfalfa, but to which was added the enzyme alpha-galactosidase.

A schematic summary of which treatments were included in which studies is presented in Table 1 (tables and figures are given starting on page 13). Brief summaries of the results from these studies are presented in Table 2 (study 1) through Table 6 (study 5). In addition, plots of the observed values of number of ACF (total, proximal, medial, and rectal) for Study 1 and Study 2 are presented in Figures 1 and 2, respectively, so that you can get a feel for the raw data. In examination of these two figures, note that it is known that these types of tumors should be most prevalent in the rectal portion of the colon, and least prevalent in the proximal region.

2 The Statistical Problem

Given all of the above results and, in particular, that of Study 5, the investigators plan to apply for a major grant from the National Institutes of Health (NIH) to investigate the problem in a scientifically rigorous fashion, rather than the piecemeal collection of studies that had been conducted. To have a good chance of success in such a grant application however, they need to present some evidence for their working hypothesis. That working hypothesis is that (1) Resveratrol does reduce the occurrence of tumors of the type considered, (2) in a diet containing the genetically modified alfalfa, Resveratrol is unavailable because it is encapsulated in galactose, (3) the Resveratrol in the genetically modified alfalfa can be "'released"' through the addition of the enzyme alpha-galactosidase, and (4) alpha-galactosidase by itself (i.e., added to a diet with no Resveratrol) has no effect.

The investigators seek out the help of a panel of statisticians, of which you are a member. Information they can provide these statisticians includes what they have seen in the data as summarized by Table 2 through Table 6 and Figures 1 and 2, and several other points based on what is believed about the problem. Specifically, they can offer the following points.

1. The total ACF count is of primary interest, although if a method was suggested

that could be used across observations on all three of the regions in which counts were made, that would be welcome.

2. All treatment/study combinations contained small counts of total ACF, and many of these combinations had at least one zero value.

3. The investigators expected some variation in the same treatment across studies. This is expected because the studies were conducted over a period of about 2 years, each study involves its own technicians, and the measurement operation involves a complex procedure.

4. To illustrate the difficulties of treatment comparison, of the possible 28 treatment pairs, 3 appear in no studies, 9 appear in 1 study, 10 appear in two studies, 5 appear in 3 studies, and 1 appears in 4 studies. In addition, note that none of the treatments appears in all of the studies.

READ AND ANSWER QUESTIONS 1 and 2 AT ABOUT THIS POINT – Questions are given in Section 5.

3 Some Exploratory Work

Given that total ACF is of primary interest, but indications of whether a similar statistical structure could be used in individual regions (i.e., proximal, medial, rectal) as well was of interest, a series of exploratory plots were constructed by an unnamed statistician and presented to our statistical panel, of which you are a member. These plots are described here.

Because only the treatment pair of treatments 1 (BD) and 2 (CA) appeared in 4 studies (see Table 1), and because effect of study was identified as likely to be present by the investigators, a plot of mean total ACF for these two treatments across studies was constructed and is presented in Figure 3. Because for any nearly any distribution other than the normal there is a relation between expected value and variance, it is natural for statisticians to wonder about the mean-variance relation in a set of data. Our unseen statistician thus also produces plots of "group" variances versus "group means", where group is defined as a given treatment within a given study. Table 1

indicates there are a total of 23 such groups, but only those containing 5 or more mice are included in the plots. Figure 4 presents group variance versus group mean for each measure of ACF (total, rectal, medial, and proximal) with study identified by different symbols. Figure 5 also presents group variance versus group mean, but in this plot study is not identified, and symbols are used to denote the ACF measure (total, proximal, medial, rectal). Note that the fact that values with the highest group means are all total ACF is not particularly indicative of anything, since total ACF is a sum of the other measures for each mouse. Finally, Figure 6 presents a plot of log group standard deviation (the logarithm of the square root of the group variance) versus log group mean, with structure otherwise similar to that of Figure 5. To assist with interpretation, our unnamed statistician has produced ordinary least squares (ols) fits for some of the figures. Let v_g be the variance for group g and m_g the mean for group g. Table 7 presents some simple models of the relation between v_g and m_g, along with the ols parameter estimates.

READ AND ANSWER QUESTIONS 3 and 4 AT ABOUT THIS POINT

4 A Little Help from the Unnamed Statistician

The unnamed statistician in our story issues a memo to the panel of statisticians suggesting the following:

1. One reasonable approach, rather than leaping straight to overall model structures, might be to consider the type of response distributions appropriate for what you gave in your answer to Question 1(a).

2. Consider defining the basic responses of interest to be the total ACF counts, and developing a model for these responses. Then, consider whether the same model structure might also be applied to the other ACF counts (rectal, medial, proximal) and whether there are any "connections" that render these models a consistent description of the problem.

3. One reasonable supposition might be that counts exhibiting greater variance than mean could be modeled with a Gamma-Poisson mixture for appropriately

chosen groups of observations.

4. One possibility might be to model the gamma mixing distributions for various treatments so that they are allowed to differ in mean, but with constant mean-variance relation.

To further assist you, the unnamed statistician delivers an additional plot, presented in Figure 7. This plot is essentially a reproduction of Figure 5 but with curves depicting ols fits from the second and fifth rows of Table 7 overlaid.

READ AND ANSWER QUESTIONS 5, 6, 7, 8, 9, and 10 AT ABOUT THIS POINT

5 Questions

1. What is a reasonable first step in the process of developing a statistical model for analysis of this problem?

   (a) Do it.

   (b) Keeping in mind the points offered by the investigators and given in Section 2, a list of statistical issues that must be faced would include the following, along with the associated statistical modeling implications.

      i. Sample sizes within individual studies are small. The statistical implications of this are that it would prove difficult to analyze each study separately and obtain much power for detecting treatment differences.

      ii. The observed data consist of counts, many of which are small. The statistical implications of this are that an appropriate discrete distribution will need to be specified for responses, or a clever transformation will need to be found.

      iii. Treatments appear in studies in quite an unbalanced fashion. The statistical implications of this are that adding study as a factor in a typical ANOVA approach to combining information across studies

would lead to difficulty in both estimation of treatment effects, and interpretation of those effects.

Add three additional issues to this list, along with the statistical implications of each. Don't try to provide solutions yet, just identify issues that are likely to arise.

2. Statistician 1 on the panel immediately suggests that the situation presented by Table 1 is hopeless for construction of an overall analysis. His recommendation is to take square roots of the responses of total ACF (since they are counts), conduct a series of individual t-tests for treatment pairs, forget about difficulties with simultaneous inference (i.e., multiple tests and the associated problems of controlling overall Type I error rates), and combine p-values across studies with the same comparisons using, for example, what is called "Fisher's Method for Combining p-values" or a similar method. Comment on this possibility. Which, if any, of the issues listed in question 1(b) (those given plus those of your answer) does this suggestion deal with in a potentially adequate (not necessarily "best") manner? Which, if any, of those issues might adversely effect the proposed procedure? Which, if any, of those issues are simply not dealt with by this proposal?

3. Consider the exploratory analysis presented to you by the unnamed statistician in Section 3.

   (a) What, if anything, is suggested about the effect of study?

   (b) What, if anything, might you conclude about the effect of different studies on the relation between mean and variance in ACF counts?

   (c) What, if anything, might you suggest about the relation between mean and variance in responses relative to the ACF measure used (i.e., total, proximal, medial, and rectal)?

   (d) What, if anything, would you conclude about modeling the large variances seen in the summaries of Table 2 through Table 6?
   Hint: the unnamed statistician suggests considering what a linear relation

between log standard deviation and log mean implies for the relation of variance to mean.

4. Statistician 2 suggests the following linear mixed model structure for total ACF (and also suggests that it might be applied to the region-specific ACF counts as well).

$$Y_{i,j,k} = \mu + \alpha_j + \gamma_i + \epsilon_{i,j,k}, \tag{1}$$

where $Y_{i,j,k}$ is the response for mouse k in treatment j and study i, μ is an overall mean response, α_j is the effect of treatment j, and

$$\gamma_i \sim iidN(0, \tau^2); \quad \epsilon_{i,j,k} \sim iidN(0, \sigma^2)$$

such that γ_i are random study effects and $\epsilon_{i,j,k}$ are random error terms. She also suggests that the response variables $Y_{i,j,k}$ be taken as either the square root or logarithm of the total ACF counts.

Comment on this proposal, with reference to the issues identified in Question 1(b) (those given plus those of your answer), similar to the way you used these issues in your answer to Question 2 concerning the proposal of Statistician 1.

5. Consider the suggestions offered by the unnamed statistician.

   (a) Which of the issues identified in Question 1(b) are directly addressed by these modeling suggestions? Which are ignored (or at least put off until later)?

   (b) Why might the unnamed statistician have suggested that gamma mixing distributions be restricted in the manner suggested in the fourth comment offered?

   (c) What phenomena, if any, have you seen in the exploratory work that might suggest the type of "connection" eluded to in the second comment of the unnamed statistician? (If you have already included this in your answer to Question 5(a) simply indicate that).

6. Statistician 3 on the panel suggests the following model. Let mice be indexed by k, and combinations of treatments and studies (i.e., combinations of i and

j from the notation of Statistician 2 in Question 4) be indexed by g (for group). Take the distribution of responses (consider first total ACF) $Y_{g,k}$ to be conditionally independent (given $\lambda_{g,k}$) having Poisson distributions with parameters $\lambda_{g,k}$ as,

$$f(y_{g,k}|\lambda_{g,k}) = \frac{1}{y_{g,k}!}\lambda_{g,k}^{y_{g,k}} \exp(-\lambda_{g,k}). \tag{2}$$

Further, for a given group g, model the $\lambda_{g,k}$ as independent and identically distributed following a gamma distribution with parameters μ_g and ϕ. He develops this model by beginning with a gamma distribution represented as

$$g(\lambda_{g,k}|\alpha, \beta_g) = \frac{\beta_g^{\alpha}}{\Gamma(\alpha)}\lambda_{g,k}^{\alpha-1} \exp\{-\beta_g\lambda_{g,k}\}, \tag{3}$$

and then re-parameterizes using $\mu_g = \alpha/\beta_g$ and $\phi = \alpha$.

Statistician 4 follows the model development of Statistician 3 up through equation (2) above but then replaces equation (3) with

$$g(\lambda_{g,k}|\alpha_g, \beta) = \frac{\beta^{\alpha_g}}{\Gamma(\alpha_g)}\lambda_{g,k}^{\alpha_g-1} \exp\{-\beta\lambda_{g,k}\}, \tag{4}$$

and re-parameterizes using $\mu_g = \alpha_g/\beta$ and $\phi = \beta$.

(a) What are the implications of the model of Statistician 3 given in expressions (2) and (3) for the relation between mean and variance across groups (treatment/study combinations)?

Hint: the goal here, as in the next question, is to represent variance as a function of the mean, which may vary over groups, but with any additional terms or coefficients remaining constant over groups.

(b) What are the implications of the model of Statistician 4 given in expressions (2) and (4) for the relation between mean and variance across groups (treatment/study combinations)?

(c) Which would you prefer, given the information available so far in this problem?

7. The suggestion of the unnamed statistician was actually to model sets of responses for all mice in a given treatment as coming from single gamma-Poisson mixture, not treatment/study combinations.

(a) Re-formulate the model of either Statistician 3 or Statistician 4 from Question 6 for this suggestion (the choice, for this question, is unimportant).

(b) How does, if at all, this new model account for the potential study effect which was made explicit in the model of Statistician 2 (e.g., equation (1))?

8. Suppose that the gamma-Poisson mixture structure suggested by the unnamed statistician is to be followed, regardless of whether the parameterization is to be that suggested by Statistician 3 or Statistician 4 in Question 6. A number of specific models may then be suggested, including the following:

Model 1. Model 1 takes every treatment to have a potentially different value of μ in the gamma mixing distribution, with the other parameter (either α or β) held constant across treatments and studies.

Model 2. Model 2 takes treatments 1 (BD), 2 (CA), 3 (TA), 7 (CA-α), and 8 (BD-α) to all have a common value of μ (say μ_0) in the gamma mixing distribution, and treatments 4 (CAR), 5 (BDR), and 6 (TA-α) to all have a common, but different value of μ (say μ_1). The other parameter (either α or β) is taken as constant across all treatments and studies.

Model 3. Model 3 takes treatments 1 (BD), 2 (CA), 3 (TA), 7 (CA-α), and 8 (BD-α) to all have a common value of μ (say μ_0) in the gamma mixing distribution, and treatments 4 (CAR), and 5 (BDR) to have a common but different value of μ (say μ_1), and treatment 6 (TA-α) to have yet a different value of μ (say μ_2). The other parameter (either α or β) is taken as constant across all treatments and studies.

Model 4. Model 4 takes all treatments to have a common value of μ and a common value of the other parameter (either α or β) across all treatments and studies.

Suppose, further, that either a likelihood analysis or a Bayesian analysis is to be pursued with these models.

(a) Why might one consider Model 2 within the context of this problem – that is, what hypothesis does Model 2 represent?

(b) What difference is represented between Model 2 and Model 3 in terms of the problem of interest?

9. Statistician 2 has not abandoned the idea of explicitly modeling random study effects. She now suggests an alternative form for a model in which responses for mouse k in study i and treatment j (denoted $Y_{i,j,k}$) are taken as Poisson $(\lambda_{i,j,k})$, but now with,

$$\log(\lambda_{i,j,k}) = \mu + \alpha_j + \gamma_i + \epsilon_{i,j,k} \tag{5}$$

where $\gamma_i \sim iidN(0, \tau^2)$ and $\epsilon_{i,j,k} \sim iidN(0, \sigma^2)$

You are given that, if $log(X) \sim N(0, \sigma^2)$, then X has a lognormal distribution with $E(X) = \exp(\mu + 0.5\sigma^2)$ and $var(X) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$.

(a) What does this model imply about the relation between means and variances for treatment/study groups of mice? Is this more similar to the model of Statistician 3 given by equation (3), or the model of Statistician 4 given by equation (4)?

(b) Is this mean-variance relation a direct result of the random term for studies (γ_i) in the model? That is, would it still be true even if we simply dropped the random study terms, or if we changed to model to $\log(\lambda_{i,j,k}) = \mu + \alpha_j$ where $\alpha_j \sim iidN(0, \kappa^2)$?

10. Our focus to this point has been modeling total ACF counts, with the idea that we might want to "keep in mind" the possibility of either applying a common model to ACF counts in regions of the colon (proximal, medial, rectal) or constructing a model to deal with these responses simultaneously. The unnamed statistician now suggests a model for the latter situation (modeling counts in the three regions simultaneously). This model is presented for a single mouse as follows:

Let Y_p, Y_m, and Y_r be random variables associated with ACF counts in the proximal, medial, and rectal colon regions, respectively. Assume that, given a

parameter λ, these variables are independent with probability mass functions $Y_p \sim Po(\gamma_p\lambda)$, $Y_m \sim Po(\gamma_m\lambda)$ and $Y_r \sim Po(\gamma_r\lambda)$ for fixed and known constants γ_p, γ_m, and γ_r, such that $\gamma_p + \gamma_m + \gamma_r = 1$. Also assume that λ is a random variable having a gamma distribution with parameters α and β.

(a) Relate this model to the models discussed previously for total ACF counts. Is it a completely different way to conceptualize the problem in a statistical sense, or is it similar to one or more of the previous models?

(b) If we were to apply this model over groups of mice (either treatment/study combinations or just treatment groups in total) would the implications for modeling the mean-variance relation change from those discussed in Question 6?

(c) Does this model incorporate dependence among the three variables Y_p, Y_m, and Y_r within mice? If so, how?

Table 1: Description of which treatments were included in which studies. For treatments, 1=BD, 2=CA, 3=TA, 4=CAR, 5=BDR, 6=TA-α, 7=CA-α, and 8=BD-α.

	Treatments							
Study	1	2	3	4	5	6	7	8
1	X	X	X	X	X			
2	X	X	X	X	X			
3			X			X		
4	X	X					X	X
5	X	X	X		X	X	X	X

Table 2: Means, variances and sample sizes for number of ACF in Study 1.

Trt	Mean	Variance	Number of Mice
1 (BD)	8.3	42.68	10
2 (CA)	4.8	31.82	10
3 (TA)	6.9	38.00	10
4 (CAR)	2.5	12.06	10
5 (BDR)	4.4	22.04	10

Table 3: Means, variances and sample sizes for number of ACF in Study 2.

Trt	Mean	Variance	Number of Mice
1 (BD)	9.5	19.67	4
2 (CA)	8.8	32.78	8
3 (TA)	6.1	30.70	8
4 (CAR)	5.8	14.62	10
5 (BDR)	5.4	19.78	9

Table 4: Means, variances and sample sizes for number of ACF in Study 3.

Trt	Mean	Variance	Number of Mice
3 (TA)	3.0	12.00	10
6 (TA-α))	0.8	1.73	10

Table 5: Means, variances and sample sizes for number of ACF in Study 4.

Trt	Mean	Variance	Number of Mice
1 (BD)	6.9	32.99	10
2 (CA)	3.3	7.79	10
7 (CA-α)	5.1	8.32	10
8 (BD-α)	9.9	44.99	10

Table 6: Means, variances and sample sizes for number of ACF in Study 5.

Trt	Mean	Variance	Number of Mice
1 (BD)	8.6	17.82	10
2 (CA)	10.2	12.92	4
3 (TA)	10.3	57.33	3
5 (BDR)	3.4	5.60	10
6 (TA-α)	1.3	1.33	3
7 (CA-α)	8.0	2.67	4
8 (BD-α)	11.7	29.00	9

Table 7: Ordinary least squares fits for some simple models relating group variance to group mean.

Model	Parameter Estimates	Corresponding Figure
$v_g = \alpha_0 + \alpha_1 m_g$	$\alpha_0 = -0.79$ $\alpha_1 = 3.59$	Figure 5
$v_g = \alpha_1 m_g$	$\alpha_1 = 3.44$	Figures 5, 7
$v_g = \alpha_0 + \alpha_1 m_g + \alpha_2 m_g^2$	$\alpha_0 = -1.16$ $\alpha_1 = 3.88$ $\alpha_2 = -0.03$	Figure 5
$v_g = \alpha_1 m_g + \alpha_2 m_g^2$	$\alpha_1 = 3.36$ $\alpha_2 = 0.01$	Figure 5
$v_g = m_g + \alpha_2 m_g^2$	$\alpha_2 = 0.304$	Figures 5, 7
$log(v_g) = \alpha_0 + \alpha_1 log(m_g)$	$\alpha_0 = 0.35$ $\alpha_1 = 0.54$	Figure 6

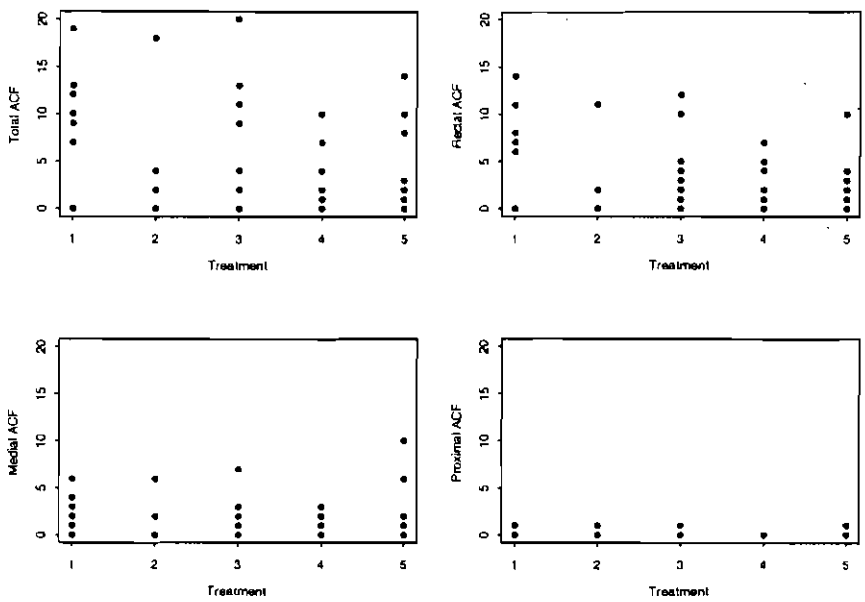

Figure 1: Individual data values of total number of ACF, number ACF in rectal region, number ACF in medial region and number of ACF in proximal region for mice in Study 1. Treatments are 1=BD, 2=CA, 3=TA, 4=CAR, 5=BDR.

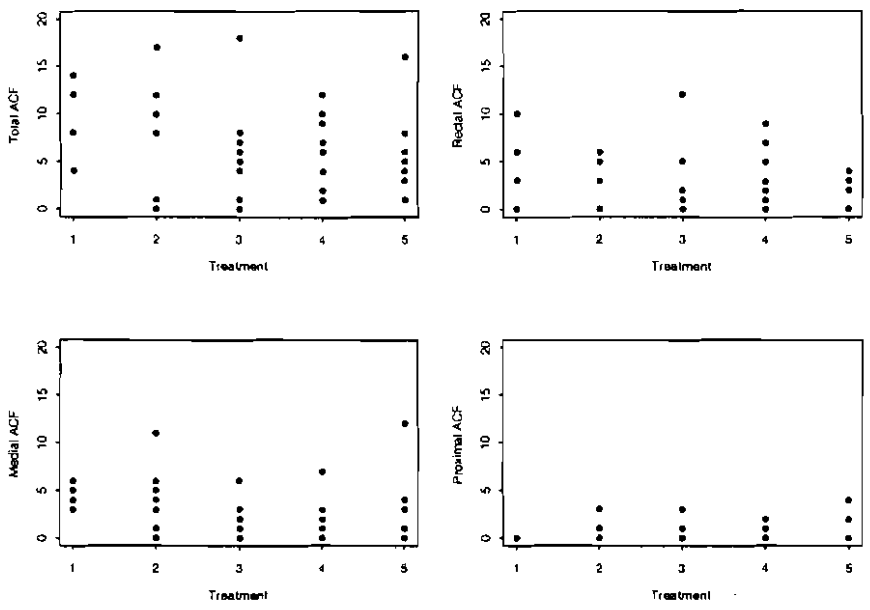

Figure 2: Individual data values of total number of ACF, number ACF in rectal region, number ACF in medial region and number of ACF in proximal region for mice in Study 2. Treatments are 1=BD, 2=CA, 3=TA, 4=CAR, 5=BDR.

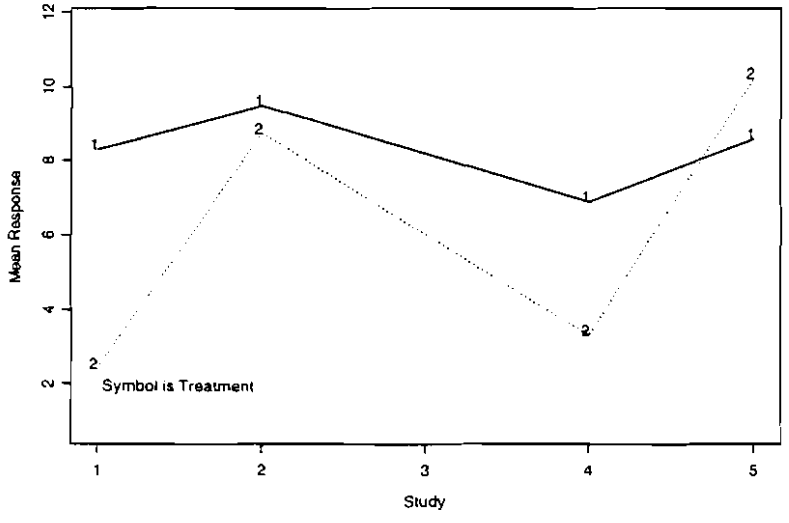

Figure 3: Mean number of total ACF in mice receiving treatments 1 (BD) and 2 (CA) in studies for which these treatments were included. Plotted symbol is treatment number.

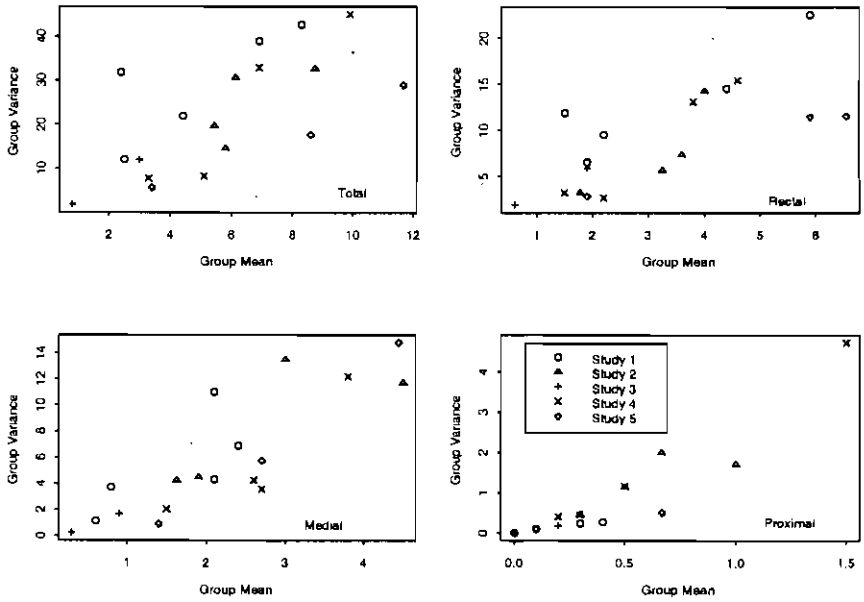

Figure 4: Group variances versus groups means for total ACF and ACF counts in regions of the colon (as labeled). Plotting symbol is study.

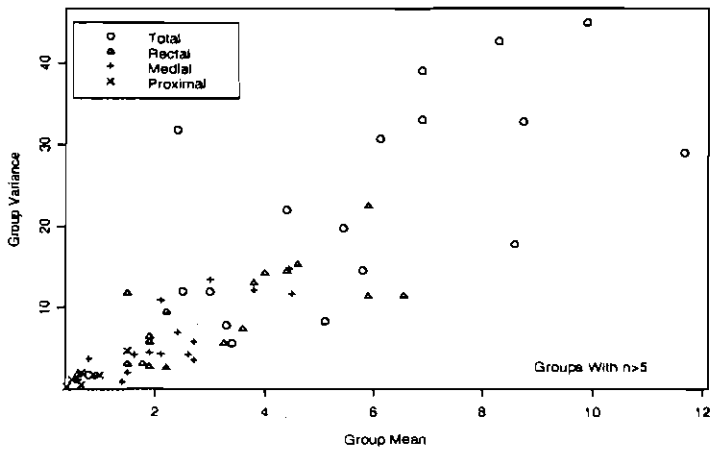

Figure 5: Group variances versus groups means for over all studies, with plotting symbols used to denote ACF measure (total, proximal, medial, or rectal).

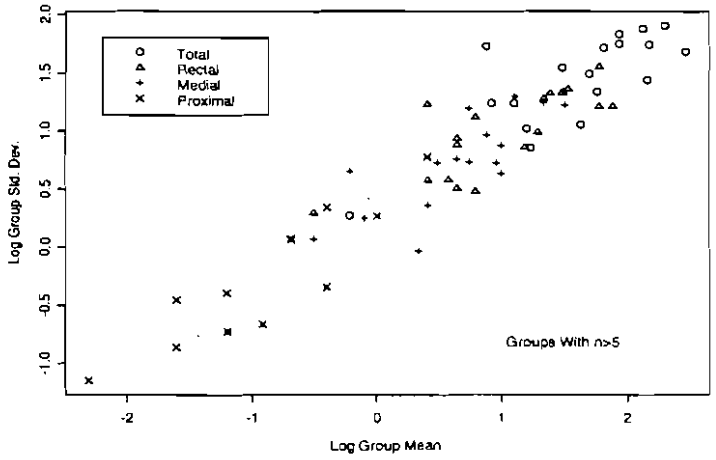

Figure 6: Log group standard deviations versus log groups means for over all studies, with plotting symbols used to denote ACF measure (total, proximal, medial, or rectal).

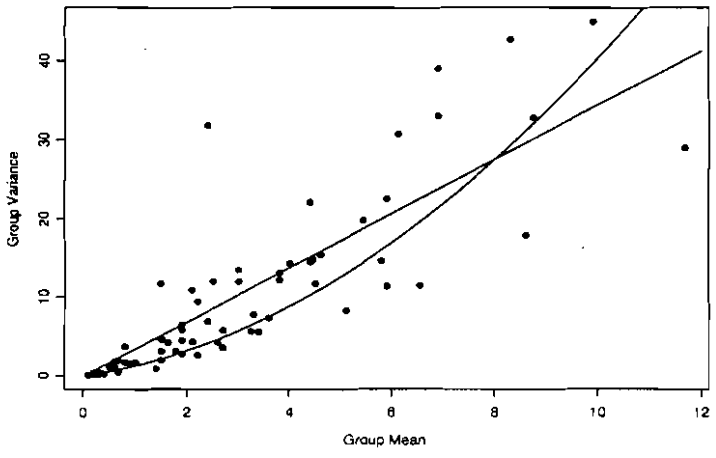

Figure 7: Group variances versus group means over all studies and ACF measures with overlaid curves from rows 2 and 5 of the models in Table 7.

PhD Preliminary Examination – 2005

Answers – Methods Question 3

1. A reasonable first step in formulating a model is to define appropriate random variables and non-random quantities (i.e., determine what is to be modeled through random quantities and what is not).

   (a) Here, the responses of interest are ACF counts in each of three regions of the colon for individual mice in fixed treatment groups and studies. Random variables associated with these counts can be appropriately defined in a number of ways, as long as there is consistency exhibited. For example, we might first consider an individual mouse in a given treatment and study group. For mouse k then, define the random variable Y_k to be associated with a given measure of ACF (total, rectal, medial, or proximal). If we wish to cover all of these possibilities, let $Y_k(t)$ be associated with total ACF count, $Y_k(r)$ be associated with rectal ACF count, $Y_k(m)$ be associated with medial ACF count, and $Y_k(p)$ be associated with proximal ACF count. Note that, in this situation, we would have the restriction that

   $$Y_k(t) = Y_k(r) + Y_k(m) + Y_k(p); \quad \text{for all } k$$

   We could provide additional indices for study and treatment if we wished. Consider only one of the ACF measures (e.g., total) and define $Y_{i,j,k}$ to be associated with the total ACF count for mouse k in treatment j of study i, for example.

   Alternatively, we could attempt a "grand" structure to account for all possible non-random variables using a random field representation with

$Y(s_k)$ being associated with the ACF count in mouse k, where

$$s_k = (i, j, v); \quad \text{for study } i, \text{ treatment } j, \text{ colon region } v$$

For any of the above definitions of random variables, we can assume independence across mice (index k in all of the above) and note that the set of possible values for any individual random variable will be $\Omega \equiv \{0, 1, \ldots\}$.

(b) Among the possible issues that could be identified in addition to those already given, the following seem the most prominent:

i. Given that the investigators believe there is inherent variability among studies, it would be unreasonable to dismiss study as a factor that should be taken into account in any analysis. Finding an appropriate manner to take "study effect" into consideration then becomes an important issue.

ii. The summary information in Table 2 through Table 6 indicate the presence of substantial variability among mice within a given study and treatment combination. Determining an appropriate way to account for this variability will be an issue in any model structure developed.

iii. The structure of the observations provides values in 3 regions and the total. While indications from the investigators are that the focus can be total ACF count, an underlying issue is whether a structure (i.e., model) can be formulated that could be applied to counts in any of the regions, and/or whether there are any consistencies that can be found in the behavior of the data among regions.

2. This suggestion is not entirely unreasonable. It attempts to deal with the issues of unbalanced treatment structure across studies and study-to-study vari-

ability by combining p-values, rather than incorporating study as an explicit "factor" in a model structure. It does not really take into account the issues of small sample sizes within studies or high variability among mice within a treatment/study combination in any way, and would be expected to have low discriminatory power. The form of the data as small counts would remain a concern here; on the other hand, this would probably not lead the procedure to give misleading results affecting, again, primary the power of the procedure to distinguish among treatments. Perhaps the most damaging criticism of this suggestion is that it seems to represent a case of "doing statistics for the sake of statistics". That is, it is doubtful that this procedure would lead to any greater understanding of the problem than simply visual examination of the summary values in Table 2 through Table 6.

3. (a) There is not a great deal of information on which to rely for even a preliminary assessment of study effect. Figure 3, which is constructed for the two treatments expected *a priori* to be the most similar, suggests that there may well be an effect of study, but that there may also be an "interaction" between study and treatment which, given the available sample sizes and unbalanced treatment design over studies, would be difficult to model explicitly. A glance back at Figures 1 and 2 suggests that, while study may be a source of variability, it could well be "swamped" by variability among mice within a given treatment and study combination.

   (b) The relation between group mean and variance appears remarkable stable across different studies and different treatments. Figure 4 might hint that variance increases more rapidly as a function of mean for study 1 than for the other studies, but this would be a fairly "fine grained" interpretation of that figure, which might not be warranted, and would be difficult to support based on that figure alone.

(c) Both Figure 5 and Figure 6 seem to suggest that the ACF measure used (total, rectal, medial, proximal) is not a major factor in the mean-variance relation exhibited by the data. Although treatment groups are not identified in these plots, all of the groups represented appear to follow basically the same relation for mean and variance.

(d) Following the hint, let v_g be the variance for a group of observations, and m_g the mean. If there is a linear relation between log standard deviation and log mean then,

$$\begin{aligned} \log[(v_g)^{1/2}] &= \alpha_0 + \alpha_1 \log(m_g) \Rightarrow \\ (1/2)\log(v_g) &= \alpha_0 + \alpha_1 \log(m_g) \Rightarrow \\ v_g &= exp(2\alpha_0) m_g^{2\alpha_1}, \end{aligned}$$

Group variances are then proportional to group means to the power $2\alpha_1$. Because Figure 6 seems to suggest a fairly linear relation between log standard deviation and log mean, the coefficient for α_1 in the last row of Table 7 (i.e., 0.54) implies that we have a situation in which we might model variance as proportional to mean. Indications from 3(b) and 3(c) above are that this model for variance as a function of mean is likely adequate across all treatment/study group combinations.

4. This proposal is targeted almost entirely at the issue of potential variability among studies, here modeled explicitly through random treatment effects. Otherwise, it seems highly misguided. In particular,

(a) Transformation of responses using square roots will not dispense entirely with the phenomenon of small counts and could be adversely effected by zero counts. While for that procedure of Statistician 1 (question 2) this difficulty might not be of immense concern (since it was a means

comparison procedure), here this is likely to have an adverse effect since the focus is estimation of variances (e.g., τ^2 and σ^2). The suggestion of a logarithmic transformation is also clearly untenable, as there are more than one or two zero values that would need to be replaced with small numbers even to allow computation.

(b) The model takes treatment effects as simple additive values. The (admittedly scant) evidence of Figure 3 suggests that this may not be adequate, but data availability may not allow estimation of a more elaborate model with interaction terms (which would need to be interpreted in any case).

(c) Perhaps most damaging criticism of this model proposal, it completely fails to take into account the apparent relation between means and variances for treatment/study combinations. Under either model (1) or model (2), variance among mice within a group is constant ($\tau^2 + \sigma^2$) regardless of the mean value.

(d) The structure of model (2), while pleasing from the viewpoint of what we "might wish could be done" is nearly totally unsupported by the data, and ignores potential complications due to the unbalanced structure of treatments across studies completely.

5. (a) Issues that are directly addressed by these suggestions include, first, the issue of data consisting of small counts (including zeros) through the use of a conditional Poisson structure, which matches the needed support and deals with the presence of zero values in a natural manner. Also, mixing these Poisson distributions over gammas attempts, at least, to deal with the issue of high variability among responses. Combining across studies within the mixture structure also may assist in combining results over studies, as indicated as necessary due to the small sample sizes within studies, and avoiding complications (at least potentially) caused

by the unbalance of treatment structure across studies. The possibility that a similar model structure might be used with ACF counts in different regions, as well as total ACF, is facilitated by the additive property of the Poisson distribution, and the suggestion of exploratory analysis that the mean-variance relation may be stable across all regions, studies, and treatments provides the possible "connection" eluded to in comment number 2 of the unnamed statistician. The one issue seemingly ignored by these suggestions is modeling a study effect, at least in an explicit manner. Variability among studies is folded in with variability among mice within treatments.

(b) Restricting the gamma distributions in the manner suggested serves primarily to reduce the number of parameters that must be estimated in the model. The small sample sizes within studies indicate that this is a necessary feature in any model developed. The indications of the exploratory analysis of Section 3 are that this may be possible here.

(c) See answer to Question 5(a).

6. To provide answers to both Question 6(a) and 6(b), note first that a standard conditioning argument provides,

$$\begin{aligned} E(Y_{g,k}) &= E[E(Y_{g,k}|\lambda_{g,k})] = E(\lambda_{g,k}) \\ var(Y_{g,k}) &= E[var(Y_{g,k}|\lambda_{g,k})] + var[E(Y_{g,k}|\lambda_{g,k})] \\ &= E(\lambda_{g,k}) + var(\lambda_{g,k}) \end{aligned}$$

(a) The implication of the model of Statistician 3 in equations (2) and (3) is that,

$$\begin{aligned} var(Y_{g,k}) &= \frac{\alpha}{\beta_g} + \frac{\alpha}{\beta_g^2} \\ &= \mu_g + \frac{1}{\alpha}\mu_g^2 \end{aligned}$$

Thus, the variance of the $Y_{g,k}$ should be a linear combination of the group mean and square of the group mean.

(b) The implication of the model of Statistician 4 in equations (2) and (4) is that,

$$\begin{aligned} var(Y_{g,k}) &= \frac{\alpha_g}{\beta} + \frac{\alpha_g}{\beta^2} \\ &= \mu_g + \frac{1}{\beta}\mu_g \\ &= \mu_g\left(1 + \frac{1}{\beta}\right). \end{aligned}$$

Thus, the variance of the $Y_{g,k}$ should be proportional to the group mean.

(c) Given the evidence of Figures 4, 5, 6, and (especially) 7, it seems clear that the model of Statistician 4 with constant β across groups would be preferred, as it implies group variance proportional to group mean rather than a function of the square of the mean.

7. (a) Using the model of Statistician 4 given by expressions (2) and (4), reformulation is accomplished by simply defining groups to be treatments rather than treatment/study combinations. Thus, let $Y_{j,k}$ be associated with the ACF count (e.g., total ACF count) for mouse k in treatment j. Take, conditional on $\lambda_{j,k}$, $Y_{j,k} \sim indepPo(\lambda_{j,k})$ and take $\lambda_{j,k} \sim iid$ gamma with parameters α_j and β, or, using $\mu_j = \alpha_j/\beta$ and $\phi = \beta$,

$$g(\lambda_{j,k}|\mu_j, \phi) = \frac{\phi^{\mu_j\phi}}{\Gamma(\mu_j\phi)}\lambda_{j,k}^{\mu_j\phi-1}\ \exp(-\phi\lambda_{j,k})$$

(b) This model does not explicitly represent variability due to studies. Rather, it folds variability among studies in with variability among mice within a treatment. That this may be a reasonable approach was suggested as far back as Question 3(a), where it was noted that variability among studies might well be "overwhelmed" by variability among individual mice.

8. (a) Model 2 represents the situation in which the treatments with no biologically available Resveratrol (BD, CA, TA, CA$-\alpha$ and BD-α) have a common mean, while the treatments with biologically available Resveratrol (CAR, BDR, and TA-α) have a different mean. This represents the substantive hypothesis that Resveratrol is an active compound that can reduce the rate of tumor formation, but also takes the amount of available Resveratrol to be similar among treatments CAR, BDR, and TA-α.

   (b) The difference between Model 2 and Model 3 is that Model 3 allows the amount of biologically available Resveratrol to differ between treatments to which it has been artificially added (CAR and BDR) and that in which it was simply "released" by the enzyme alphagalactosidase (TA-α). Since there is no *a priori* reason to believe that the amount of Resveratrol released by galactosidase from the genetically modified alfalfa is the same as that artificially added to the basal and control alfalfa diets, this is also a potentially viable model.

9. (a) In this model, $\log(\lambda_{i,j,k})$ is distributed as a normal random variable with mean $\mu + \alpha_j$ and variance $\tau^2 + \sigma^2$. Thus, $\lambda_{i,j,k}$ is lognormal with

$$\begin{aligned} E(\lambda_{i,j,k}) &= \exp\{\mu + \alpha_j + (1/2)(\tau^2 + \sigma^2)\} \\ var(\lambda_{i,j,k}) &= \exp\{2\mu + 2\alpha_j + \tau^2 + \sigma^2\}\left(\exp\{\tau^2 + \sigma^2\} - 1\right) \end{aligned}$$

   or,

$$\begin{aligned} E(\lambda_{i,j,k}) &= \mu_{i,j,k} \\ var(\lambda_{i,j,k}) &= \mu_{i,j,k}^2\left(\exp\{\tau^2 + \sigma^2\} - 1\right) \end{aligned}$$

   Using the conditioning argument, but with $\lambda_{i,j,k}$ now lognormal,

$$\begin{aligned} var(Y_{i,j,k}) &= E\{\lambda_{i,j,k}\} + var\{\lambda_{i,j,k}\} \\ &= \mu_{i,j,k} + \mu_{i,j,k}^2\left(\exp\{\tau^2 + \sigma^2\} - 1\right) \end{aligned}$$

which shows that the mean-variance relation implied by this model is more similar to that of equation (3) than it is that of equation (4).

(b) No, this mean-variance relation is induced by the assignment of a lognormal distribution to the Poisson parameters, not the presence of the random study effects *per se*. It would occur with any model that included this lognormal distributional assignment, regardless of the mean structure used.

10. (a) This model is quite similar to the gamma-Poisson mixture models discussed in Questions 6 and 7. In fact, those models result from the current suggestion by taking the response variable to be modeled as the sum $Y_p + Y_m + Y_r$ here. That is, with Y_p, Y_m and Y_r conditionally independent, and $\gamma_p + \gamma_m + \gamma_r = 1$, $Y = Y_p + Y_m + Y_r$ has a Poisson distribution with parameter λ. If λ is then modeled as a gamma random variable, the previous gamma-Poisson models result.

(b) No, the mean-variance relation would still be that of variances proportional to means. Consider, for example, a model in which gamma mixing distributions for treatment groups (as in Question 7) are parameterized with α_t and common β. The expected values and variances implied by the model of this question are (apply standard conditioning argument):

$$\begin{aligned}
E\{Y_p\} &= \frac{\gamma_p \alpha_t}{\beta} \\
E\{Y_m\} &= \frac{\gamma_m \alpha_t}{\beta} \\
E\{Y_r\} &= \frac{\gamma_r \alpha_t}{\beta} \\
var\{Y_p\} &= \frac{\gamma_p \alpha_t}{\beta}(1 + 1/\beta) \\
var\{Y_m\} &= \frac{\gamma_m \alpha_t}{\beta}(1 + 1/\beta) \\
var\{Y_r\} &= \frac{\gamma_r \alpha_t}{\beta}(1 + 1/\beta)
\end{aligned}$$

(c) Yes, this model results in marginal dependence as given by

$$\begin{aligned} cov(Y_p, Y_m) &= \gamma_p \gamma_m \frac{\alpha_t}{\beta^2} \\ cov(Y_p, Y_r) &= \gamma_p \gamma_r \frac{\alpha_t}{\beta^2} \\ cov(Y_m, Y_r) &= \gamma_m \gamma_r \frac{\alpha_t}{\beta^2} \end{aligned}$$