Stat 516 Homework 3 Solutions

1. (a) *(5 points)* Expression data were simulated for an experiment involving $g = 10,000$ genes and two treatment groups with five independent experimental units in each treatment group. Gene-specific variances $\sigma_1^2, \ldots, \sigma_g^2$ were simulated as independent and identically distributed draws from an inverse gamma distribution with parameters $\alpha = 6$ and $\beta = 2$ (gamma mean $= \alpha/\beta = 3$). The difference between expression means (treatment 2 minus treatment 1) for gene j, denoted by δ_j, was set to 0 for 8,000 randomly selected genes. For a randomly selected 1,000 of the other 2,000 genes, $\delta_j|\sigma_j^2$ was drawn from the normal distribution with mean $-\sigma_j$ and variance σ_j^2. For the remaining 1,000 genes, $\delta_j|\sigma_j^2$ was drawn from the normal distribution with mean σ_j and variance σ_j^2. The treatment differences $\delta_1, \ldots, \delta_g$ were mutually independent. Let y_{ijk} denote the log-scale expression measure for treatment i, gene j, and experimental unit k within treatment i ($i = 1, 2; j = 1, \ldots, g; k = 1, \ldots, 5$). Then, conditional on $\sigma_1^2, \ldots, \sigma_g^2$ and $\delta_1, \ldots, \delta_g$,

$$y_{ijk} = \mu + \phi_{ik} + \gamma_j + (-1)^i \delta_j/2 + \varepsilon_{ijk},$$

where $\mu = 10$; $\phi_{11}, \ldots, \phi_{25} \overset{iid}{\sim} N(0, 2.5^2)$; $\gamma_1, \ldots, \gamma_g \overset{iid}{\sim} N(0, 2^2)$; and $\varepsilon_{ijk} \sim N(0, \sigma_j^2)$ for all $i = 1, 2; j = 1, \ldots, g$; and $k = 1, \ldots, 5$ with conditional mutual independence among all ϕ, γ, δ, and ϵ terms.

(b) *(2 points)*

```
> get.pvalue=function(d)
+ {
+   t.test(d[1:5],d[6:10],var.equal=T)$p.value
+ }
>
> p=apply(y,1,get.pvalue)
>
> hist(p)
>
> sum(p<0.001)
[1] 1
> sum(p[de.genes]<0.001)
[1] 1
```

(c) *(2 points)*

```
> y=sweep(y,2,apply(y,2,median),"-")
> boxplot(y)
> y[1,]
 [1] -3.061299 -3.409843 -2.622891 -3.135267 -4.510781 -3.776967 -4.491191
 [8] -4.326745 -4.809676 -5.725496
```

(d) *(2 points)*

```
> p=apply(y,1,get.pvalue)
>
> hist(p)
> hist(p[de.genes])
> hist(p[!(1:10000%in%de.genes)])
>
```

```
> sum(p<0.001)
[1] 157
> sum(p[de.genes]<0.001)
[1] 152
```

(e) *(2 points)* The normalized data provides much more information about differential expression.

(f) *(2 points)*

```
> mean(p[de.genes]<0.001)
[1] 0.076
```

2. (a) *(2 points)* $\mathrm{Var}(Y_{111}) = \mathrm{Var}(\mu + \tau_1 + m_{11} + e_{111}) = \mathrm{Var}(m_{11} + e_{111}) = \sigma_m^2 + \sigma_e^2$

(b) *(3 points)*

$$\begin{aligned}
\mathrm{Cov}(Y_{ij1}, Y_{ij2}) &= \mathrm{Cov}(\mu + \tau_i + m_{ij} + e_{ij1}, \mu + \tau_i + m_{ij} + e_{ij2}) \\
&= \mathrm{Cov}(m_{ij} + e_{ij1}, m_{ij} + e_{ij2}) \\
&= \mathrm{Cov}(m_{ij}, m_{ij}) + \mathrm{Cov}(m_{ij}, e_{ij2}) + \mathrm{Cov}(e_{ij1}, m_{ij}) + \mathrm{Cov}(e_{ij1}, e_{ij2}) \\
&= \mathrm{Cov}(m_{ij}, m_{ij}) \\
&= \mathrm{Var}(m_{ij}) = \sigma_m^2
\end{aligned}$$

(c) *(2 points)*

$$\frac{\sigma_m^2}{\sigma_m^2 + \sigma_e^2}$$

(d) *(2 points)* 0

(e) *(2 points)* 0

(f) *(2 points)* $\mathrm{Var}(\bar{Y}_{ij\cdot}) = \mathrm{Var}(m_{ij} + \bar{e}_{ij\cdot}) = \sigma_m^2 + \sigma_e^2/2$

(g) *(2 points)* $\mathrm{Var}(\bar{Y}_{i\cdot\cdot}) = \mathrm{Var}(\bar{m}_{i\cdot} + \bar{e}_{i\cdot\cdot}) = \sigma_m^2/4 + \sigma_e^2/8$

(h) *(2 points)* $\sigma_m^2/2 + \sigma_e^2/4$

(i) *(3 points)*

```
d=read.table("hw3data.txt",header=T)
mouse.vars=tapply(d$y,factor(d$mouse),var)
mouse.vars
      1       2       3       4       5       6       7       8
0.00845 0.00245 0.00080 0.01445 0.00125 0.50000 0.04805 0.08820
mean(mouse.vars)
0.08295625
```

(j) *(4 points)*

```
mouse.means=tapply(d$y,factor(d$mouse),mean)
mouse.means
mouse.means
    1     2     3     4     5     6     7     8
7.575 7.335 7.420 6.115 4.185 6.710 5.115 4.470

t.test(mouse.means[1:4],mouse.means[5:8],var.equal=T)
t = 3.0314, df = 6, p-value = 0.02306
```

(k) *(3 points)*

```
(var(mouse.means[1:4])+var(mouse.means[5:8]))/2
0.8629698
```

An estimate of $\sigma_m^2 + \sigma_e^2/2$ is 0.8629698. Thus and estimate of σ_m^2 is

$$0.8629698 - 0.08295625/2 = 0.8214917.$$

(l) *(2 points)* We get the same results for our t-test. Squaring the standard deviations from the lme output, we get almost the exact same variance estimates.

3. (a) *(2 points)* Blocking is not used in this experiment. Blocking was defined in our notes as "grouping similar experimental units together and assigning different treatments within such groups of experimental units." Thus the pens of pigs are not blocks because all pigs in a pen received the same treatment.

(b) *(2 points)* Each pen of pigs is an experimental unit because the treatments were randomly assigned and independently applied to the pens.

(c) *(2 points)* An observational unit would be a muscle sample from a hog. There is a one-to-one correspondence between hogs and observational units.

(d) *(3 points)*

$$Y_{ijk} = \mu + \delta_i + p_{ij} + e_{ijk}$$

where Y_{ijk} is the response for the k^{th} hog in the j^{th} pen that received treatment i; μ is the overall mean; δ_i is the effect of the i^{th} diet; p_{ij} is the random effect of the j^{th} pen for the i^{th} diet; and e_{ijk} is the residual random effect for the k^{th} hog in the j^{th} pen that received treatment i. The p_{ij} are assumed to be independent and normally distributed with mean 0 and variance σ_p^2. The e_{ijk} are independent and normally distributed with mean 0 and variance σ_e^2. All the random effects are independent of one another. In abbreviated form, we have

$$Y = \text{ diet pen.}$$

(e) *(2 points)* This is a completely randomized design (with multiple observations per experimental unit).

4. (a) *(2 points)* diet and dose of drug

(b) *(2 points)* diet: A, B
dose of drug: 0, 10, 20, 30 mg/kg body weight

(c) *(3 points)* There are two types of experimental units in this example. Pens are the experimental units with respect to the treatment factor diet because diets were randomly assigned and independently applied to pens. Hogs are the experimental units with respect to the treatment factor dose of drug because the drug dose was randomly assigned and independently applied to individual hogs.

(d) *(3 points)* Based on the answer to (c), this experiment is best described as a split-plot design. The whole-plot factor is diet. The whole-plot experimental units are pens. The split-plot factor is dose of drug. The split-plot experimental units are hogs.

(e) *(3 points)* Y=diet dose diet:dose pen

(f) *(4 points)* Four pairs of pens should be formed, where each pair has one pen that received diet A and one that received diet B. Within each pair of pens, an arrow should connect the hogs from different pens that received the same dose of the drug (0 hog to 0 hog, 10 hog to 10 hog, etc.). There are several different ways for choosing the arrow direction that are equally valid. For each dose, half the arrows should point from a diet A pig to a diet B pig. Half should point from a diet B pig to a diet A pig.

(g) *(4 points)*
Y=diet dose diet:dose dye pen slide

5. (a) *(3 points)*

$$X = \begin{bmatrix} 1 & -1 & 1 & 0 \\ 1 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}$$

(b) *(9 points)* Consider the design

A $\rightarrow$ D B $\rightarrow$ D C $\rightarrow$ D
A $\leftarrow$ D B $\leftarrow$ D C $\leftarrow$ D

The corresponding X matrix is

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \end{bmatrix}.$$

$$X'X = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad (X'X)^{-1} = \begin{bmatrix} 1/6 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}.$$

Comparing the last three diagonal elements of this inverse matrix to those of the previous inverse matrix shows that this proposed design dominates the previous design because none of the last three diagonals immediately above are larger than the corresponding values for the previous design, while some are strictly less than the corresponding values from the previous design.

Recall that each experimental unit is represented by a circle and each slide by an arrow. Thus every drawing should have had 12 circles and 6 arrows. (I left the circles out of my drawing but I do have 12 characters – 2 As, 2 Bs, 2 Cs, and 6 Ds.) There was no restriction on the number of experimental units for each treatment, given that 12 experimental units total were used. Some of you seemed to incorrectly assume that it was necessary to have 3 for each treatment.

6. (a) *(2 points)* $\sigma_s^2/2 + \sigma_u^2 + \sigma^2/2$

(b) *(2 points)* $\sigma_s^2/4$

(c) *(2 points)* $\sigma_s^2/4$

(d) *(2 points)* 0

(e) *(4 points)* According to the model in part (e), the covariance between observations from the same block is σ_b^2. Thus we have $\sigma_b^2 = \sigma_s^2/4$ from part (b). According to the model in part (e), the variance of Y_{ij} is $\sigma_b^2 + \sigma_e^2$. Thus, from part (a) and the expression for σ_b^2, we have

$$\sigma_b^2 + \sigma_e^2 = \sigma_u^2 + \sigma_s^2/2 + \sigma^2/2 \Rightarrow \sigma_e^2 = \sigma_u^2 + \sigma_s^2/4 + \sigma^2/2.$$