Stat 416 Homework 3

Due Date: Tuesday, February 24

1. Suppose researchers were interested in studying the effects of two diets (A and B) on gene expression in muscle tissue of hogs. Eight pens containing four hogs each were available for an experiment. For logistical reasons, all hogs in any one pen had to be fed the same diet. The researchers randomly assigned 4 of the pens to diet A and the other 4 to diet B. At the time of slaughter, one RNA sample was taken from muscle tissue of each of the 32 hogs. These 32 samples were measured using 32 Affymetrix GeneChips (one for each sample).

   (a) Does this experiment involve blocking? If so, name the blocks.

   (b) What are the experimental units in this experiment?

   (c) What are the observational units in this experiment?

   (d) Write down a statistical model for this experiment based on the experimental design. You may use abbreviated notation like that discussed on slide 45 of "Introduction to Mixed Linear Models in Microarray Experiments."

   (e) Is this experiment best described as a completely randomized design, randomized complete block design, split-plot design, incomplete block design, or Latin square design? Explain.

2. Suppose researchers were interested in studying the effects of two diets (A and B) and four doses of a drug (0, 10, 20, and 30 mg/kg of body weight) on gene expression in muscle tissue of hogs. Eight pens containing four hogs each were available for an experiment. For logistical reasons, all hogs in any one pen had to be fed the same diet. The researchers randomly assigned 4 of the pens to diet A and the other 4 to diet B. Within each pen, the four hogs were randomly assigned to the four doses of the drug in a completely randomized manner with one hog for each dose. Each hog was injected with its assigned dose once each week prior to slaughter. At the time of slaughter, one RNA sample was taken from muscle tissue of each of the 32 hogs. These 32 samples were measured using 32 Affymetrix GeneChips (one for each sample).

   (a) Name the treatment factors considered in this experiment.

   (b) Name the levels of each treatment factor.

   (c) Name the experimental units in this experiment.

   (d) Is this experiment best described as a completely randomized design, randomized complete block design, split-plot design, incomplete block design, or Latin square design? Explain.

   (e) Write down a model for the data based on this experimental design. You may use the abbreviated model notation discussed in class.

   (f) Suppose that instead of using Affymetrix GeneChips, the researchers decided to measure expression using a total of 16 two-color microarray slides. Furthermore, suppose the researchers

were primarily interested in understanding differences in gene expression between the two diets for each dose of the drug. Draw a picture (using the design notation that we have used in class) to illustrate how you would recommend pairing samples on slides and assigning dyes.

(g) Write down a model for the data based on this two-color microarray experimental design. You may use the abbreviated model notation discussed in class.

3. Suppose a researcher is interested in comparing the effect of three treatments (A, B, and C) on gene expression in maize seedlings. A total of 18 seedlings are available for a two-color microarray experiment. A total of 9 slides can be used for the experiment. The researcher uses a balanced and completely randomized design to randomly assign the three treatments to the 18 seedlings. Draw a picture (using the design notation that we have used in class) to illustrate how you would pair seedlings on slides and assign dyes if a balanced incomplete block design with dye balance is desired.

4. This question is intended to help your gain a better understanding of variance components and calculations used to determine standard errors.

(a) The expected value of a random variable Y is denoted by $E(Y)$. It is the mean or average value taken by Y. If $E(Y) = \mu$, then $E(a + bY) = a + b\mu$ for any real constant values a and b. This says that if Y is μ on average, the $a + bY$ is $a + b\mu$ on average. If $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$, then $E(Y_1 + Y_2) = \mu_1 + \mu_2$. You can put these rules together if necessary to complete the following problems. Suppose Y_i is a random variable with mean μ_i for all i.

i. Determine $E(3Y_1 - 9)$.

ii. Determine $E(Y_1 + Y_2 + Y_3)$

iii. Determine $E(2Y_1 - Y_2 + .5Y_3)$.

(b) Consider the random variable $\{Y - E(Y)\}^2$, i.e., the squared deviation of the random variable Y from its mean value. The mean of this random variable is defined as the variance of Y; i.e., the variance of a random variable Y is $\mathrm{Var}(Y) = E[\{Y - E(Y)\}^2]$. Thus the variance of Y is the expected squared deviation of a random variable from its mean. Random variables that tend to yield values close to their means will have small variances. Random variables that tend to yield values far from their means will have large variances.

A constant real number denoted c can be viewed as a random variable that always takes the value c. Determine $E(c)$ and $\mathrm{Var}(c)$.

(c) If the random variables Y_1 and Y_2 are independent and c_1 and c_2 are any real constants, then

$$\mathrm{Var}(c_1Y_1 + c_2Y_2) = c_1^2\mathrm{Var}(Y_1) + c_2^2\mathrm{Var}(Y_2).$$

If $\mathrm{Var}(Y_i) = \sigma^2$ for $i = 1, 2$; determine

i. $\mathrm{Var}(Y_1 + Y_2)$.

ii. $\mathrm{Var}(Y_1 - Y_2)$

iii. $\text{Var}\{(Y_1 + Y_2)/2\}$

iv. $\text{Var}(2Y_1)$

v. $\text{Var}(Y_1 + 6)$

vi. $\text{Var}(2Y_1 + 6)$

(d) If the random variables $Y_1, Y_2, \ldots, Y_n$ are independent and $c_1, c_2, \ldots, c_n$ are any real constants, then

$$\text{Var}(\sum_{i=1}^{n} c_i Y_i) = \text{Var}(c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n) = c_1^2 \text{Var}(Y_1) + c_2^2 \text{Var}(Y_2) + \cdots + c_n^2 \text{Var}(Y_n).$$

If $\text{Var}(Y_i) = \sigma^2$ for $i = 1, 2, \ldots, n$; determine

i. $\text{Var}(\sum_{i=1}^{n} Y_i)$

ii. $\text{Var}(\frac{1}{n} \sum_{i=1}^{n} Y_i) = \text{Var}(\bar{Y})$

(e) The covariance between random variables Y_1 and Y_2 is the mean of the random variable

$$\{Y_1 - E(Y_1)\} \cdot \{Y_2 - E(Y_2)\}.$$

We write the covariance between Y_1 and Y_2 as $\text{Cov}(Y_1, Y_2)$. Note that if Y_1 tends to be below its mean when Y_2 is above its mean and vice versus, then the covariance between Y_1 and Y_2 will be negative. If both variables tend to be above their averages together or below their averages together, then $\text{Cov}(Y_1, Y_2)$ will be positive. The correlation between Y_1 and Y_2 is defined as

$$\frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}}.$$

Thus the correlation between two random variables will have the same sign as the covariance. If random variables are independent, their covariance (an thus their correlation) is 0.

i. Determine $\text{Cov}(Y, c)$ where Y is a random variable and c is any constant value.

ii. How is $\text{Cov}(Y, Y)$ related to $\text{Var}(Y)$?

iii. How is $\text{Cov}(Y_1, Y_2)$ related to $\text{Cov}(Y_2, Y_1)$?

(f) Suppose Y_i is a random variables, and c_i is a constant value. Then

$$\text{Cov}(c_1 Y_1, Y_2) = c_1 \text{Cov}(Y_1, Y_2), \quad \text{Cov}(Y_1, c_2 Y_2) = c_2 \text{Cov}(Y_1, Y_2),$$

$$\text{Cov}(c_1 Y_1, c_2 Y_2) = c_1 c_2 \text{Cov}(Y_1, Y_2), \quad \text{Cov}(Y_1 + Y_2, Y_3) = \text{Cov}(Y_1, Y_3) + \text{Cov}(Y_2, Y_3),$$

$$\text{and } \text{Cov}(Y_1, Y_2 + Y_3) = \text{Cov}(Y_1, Y_2) + \text{Cov}(Y_1, Y_3).$$

Let $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$ for all $i \neq j$, and let $\text{Var}(Y_i) = \sigma_i^2$ for all i. Find expressions for the following:

i. $\text{Cov}(Y_1, Y_1 - Y_2)$

ii. $\text{Cov}(Y_1 + Y_2, Y_1 - Y_2)$

iii. $\mathrm{Cov}(3Y_1 + 4Y_2, Y_3 - 2Y_2)$

(g) Consider the mixed model for a single gene for the experiment depicted on page 49 of the notes "Introduction to Mixed Linear Models in Microarray Experiments." Let σ_m^2 and σ_e^2 denote the variance components for the mouse random effects and the residual random effects, respectively. Determine the following in terms of these variance components.

i. $\mathrm{Var}(Y_{111})$.

ii. The covariance between any two observations from a single mouse.

iii. The correlation between any two observations from a single mouse.

iv. The covariance between any two observations from different mice in the same treatment group.

v. The covariance between any two observations from mice in different treatment groups.

vi. The variance of the average of the two observations from a single mouse.

vii. The variance of the average of all the observations from a single treatment group.

viii. $\mathrm{Var}(\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot})$.

(h) Now suppose the data in the file *hw3data.txt* has been observed for a single gene, where y denotes the log-scale normalized signal intensity. You can read this file in R using the following commands:

```
d=read.delim("http://www.public.iastate.edu/~dnett/microarray/hw3data.txt")
d=as.matrix(d)
```

i. For each mouse, find the sample variance of the two observations for that mouse. This should give you 8 sample variances. Find the average of those sample variances. This provides an estimate of σ_e^2.

ii. Now find the average of the two observations for each mouse. Compute a two-sample t-test using these averages (4 for each treatment) to test for differential expression between treatments. Report a p-value and the degrees of freedom for your test.

iii. Within each treatment group, find the sample variance of the 4 averages used in the previous question. Average these two sample variances to obtain an estimate of your answer to problem 1 (g) (vi). Use this together with your answer to problem 1 (h)(i) to find an estimate of σ_m^2.