

# Solutions

Statistics 416

Exam 1

March 5, 2009

1. It is possible to use more than two dyes and, therefore, more than two samples on a single slide. Suppose that a four-dye system has been developed so that four distinct samples can be measured together on a single slide. Suppose researchers were interested in testing for differential expression among 4 treatments (A, B, C, and D). All comparisons among treatments are considered to be of equal interest. If a total of 4 four-color slides and a total of 16 experimental units were available, how would you assign treatments to experimental units and experimental units to slides? Complete the table below by inserting treatment labels into the blanks to show one possible assignment of treatments to experimental units and experimental units to slides that you would consider for this experiment.

*(8 points) Because both slides and dyes might affect our measures of gene expression, we would like to treat slides as blocks and dyes as blocks. A Latin square design will be appropriate here. One Latin square design is as follows.*

Slide	Dye			
	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

2. Use quantile normalization to normalize the “dataset” below.

gene	slide1	slide2	slide3	slide4
1	11	16	11	4
2	8	10	7	7
3	4	9	4	10
4	7	7	1	0

*(8 points)*

gene	slide1	slide2	slide3	slide4
1	12	12	12	6
2	8	8	8	8
3	3	6	6	12
4	6	3	3	3

3. a) Compute the MAD value for the original “data” from slide 4 in problem 2.

(5 points) 4, 7, 10, 0 ----> 0, 4, 7, 10

$median=(4+7)/2=5.5$

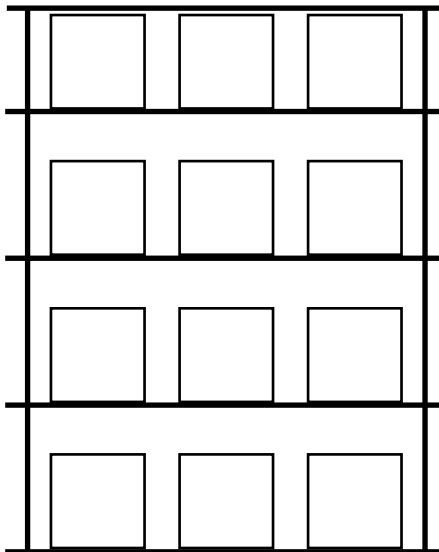
*absolute deviations* 5.5, 1.5, 1.5, 4.5 ----> 1.5, 1.5, 4.5, 5.5

$MAD=median\ of\ absolute\ deviations=(1.5+4.5)/2=3$

b) What are MAD values used for in the context of microarray data analysis?

(4 points) *MAD values are used for scale normalization of two-color array data. They are also computed and used as part of the Tukey’s bi-weight procedure for determining the MAS5.0 value of a probeset.*

4. Consider an experiment in which researchers are interested in studying the effects of four diets (A, B, C, and D) on gene expression in mice. All possible comparisons between pairs of diets are of equal interest to the researchers. Suppose 12 cages, each containing 4 mice, are available for use in the experiment. Furthermore, suppose that the mice in each cage eat from a common feeder so that all the mice in any particular cage must receive the same diet. Suppose also that the cages are fixed on a rack with 4 shelves and 3 cages per shelf as depicted below. A total for 48 Affymetrix GeneChips are available for this experiment so that one GeneChip will be used to measure expression in each mouse.



a) Past experiments conducted by the researcher have show that there can be differences among the responses of mice on different shelves. For example, the mice in cages on the top shelf may gain slightly more weight over the course of a study than mice on the bottom shelf. With this in mind, indicate one example assignment of diets to cages that you would consider for this experiment by writing the letter (A, B, C, or D) in each box in the figure above.

(8 points) We should treat each shelf like a block. Because we have room for only three treatments on each shelf, we will have to use an incomplete block design. In this case, we can use a balanced incomplete block design. For example,

shelf 1: A B C  
shelf 2: A B D  
shelf 3: A C D  
shelf 4: B C D

b) What are the experimental units in this experiment?

(5 points) cages

c) What are the observational units in this experiment?

(5 points) mice

d) Is blocking used in this experiment? If so, describe the blocks.

(5 points) Yes. Shelves are (incomplete) blocks in this case.

e) Would you describe this experiment as a completely randomized design? If yes, explain why. If not, name the design using the standard experimental design terminology.

(5 points) No. It is a balanced incomplete block design with multiple observations per experimental unit.

f) Write down a model for the 48 expression measures of a single gene in this experiment. You may use the abbreviated model notation that we discussed in class in which the model is specified with words rather than symbols.

(6 points)  $Y = \text{shelf cage diet}$

g) Would you consider any of the factors in your model to be random? If so, list the random factor or factors.

(5 points) The factor "cage" must be random because its levels correspond to the experimental units. The factor "shelf" could be considered either as fixed or random.

h) Now suppose that instead of storing the cages on the rack, the researchers are able to keep all 12 cages on one table so that differences among shelves are no longer an issue. As before, each cage will contain 4 mice, all mice in each cage must receive the same diet, and all possible comparisons between pairs of diets are of equal interest to the researchers. However, suppose that due to budget constraints, only 6 two-color microarray slides are available for use in the experiment rather than 48 Affymetrix GeneChips. Given these resources, clearly explain the microarray experimental design that you would recommend to the researchers. In particular, explain what should be hybridized to each slide, and provide information about dye assignments.

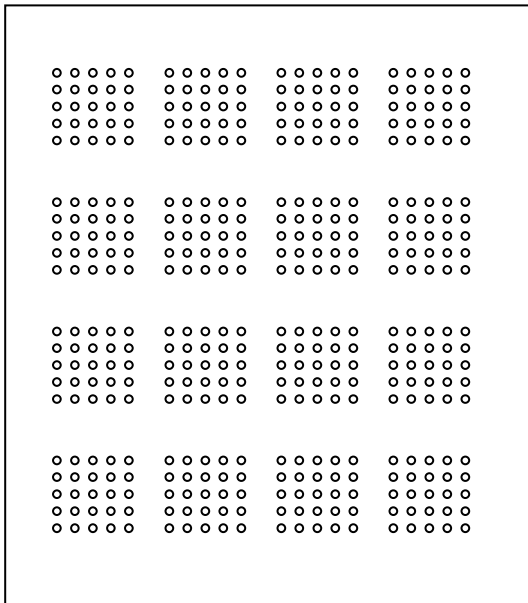
h) (Continued from page 3) A picture using our two-color microarray experimental design notation should help make your recommendations clear.

(8 points) We should form one pool of RNA from all four mice in any cage. We should pair treatments on slides using a design like the following.

A ----> B  
 A <---- C  
 A ----> D  
 B ----> C  
 B <---- D  
 C <---- D

Note that this is a balanced incomplete block design with slides as blocks. Dyes cannot be balanced within each treatment, but each treatment appears at least once with each dye.

5. Suppose the figure below represents a two-color microarray. Assume this slide was spotted in the manner discussed in class.



a) What is, most likely, the maximum number of pins that were used to print this slide?  
 (4 points) 16

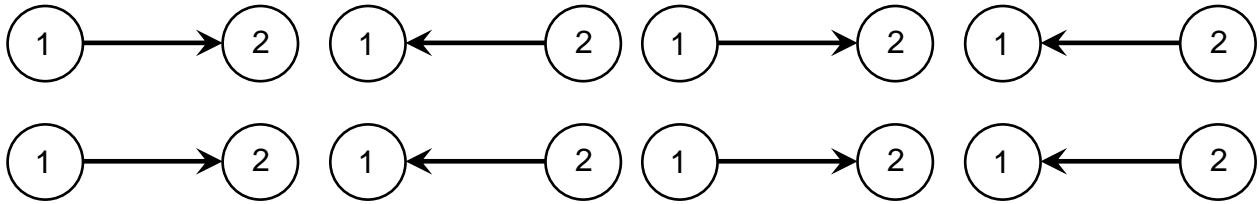
b) Color in any two spots that were printed at the same time.

(4 points) Color the spot in the upper left hand corner of each 5 x 5 grid, for example.

c) Circle any one set of spots that were printed with a single pin.

(4 points) Circle any of the 5 x 5 grids.

6. Suppose a two-treatment two-color microarray experiment was conducted using the design depicted below.



Suppose the normalized data for a single gene are available in R as follows.

	slide	dye	treatment	y
1	1	3	1	2.1
2	1	5	2	7.9
3	2	3	2	8.3
4	2	5	1	2.8
5	3	3	1	4.2
6	3	5	2	8.9
7	4	3	2	7.7
8	4	5	1	7.4
9	5	3	1	6.9
10	5	5	2	7.7
11	6	3	2	6.3
12	6	5	1	3.9
13	7	3	1	3.5
14	7	5	2	7.4
15	8	3	2	5.3
16	8	5	1	3.4

a) In each blank next to a section of R code below, write “dye,” “treatment,” or “neither” depending on whether the R code provides a valid p-value for testing the significance of the dye factor for this gene, the significance of the treatment factor for this gene, or neither, respectively.

(8 points)

dif=y[treatment==1]-y[treatment==2]  
t.test(dif[1:4],dif[5:8],var.equal=T) *neither*

dif=y[dye==5]-y[dye==3]  
t.test(dif,var.equal=T) *neither*

dif=y[dye==5]-y[dye==3]  
t.test(dif[1:4],dif[5:8],var.equal=T) *neither*

dif=y[dye==5]-y[dye==3]  
t.test(dif[c(1,3,5,7)],dif[c(2,4,6,8)],var.equal=T) *treatment*

```
dif=y[treatment==1]-y[treatment==2]
t.test(dif,var.equal=T) neither
```

```
m=matrix(y,byrow=T,ncol=2)
m=matrix(m[,2]-m[,1],byrow=T,ncol=2)
t.test(m[,1],m[,2],var.equal=T) treatment
```

```
dif=y[treatment==1]-y[treatment==2]
t.test(dif[c(1,3,5,7)],dif[c(2,4,6,8)],var.equal=T) dye
```

```
t.test(y[treatment==1],y[treatment==2],var.equal=T) neither
```

b) Fill in the blanks with actual numbers so that the R code below will provide a valid test of the significance of the treatment factor for this gene.

(8 points)

```
x=c(7.9-2.1, 8.9-4.2, 7.7-6.9, 7.4-3.5)
```

```
y=c(2.8-8.3, 7.4-7.7, 3.9-6.3, 3.4-5.3)
```

```
t.test(x,y,var.equal=T)
```

	slide	dye	treatment	y
1	1	3	1	2.1
2	1	5	2	7.9
3	2	3	2	8.3
4	2	5	1	2.8
5	3	3	1	4.2
6	3	5	2	8.9
7	4	3	2	7.7
8	4	5	1	7.4
9	5	3	1	6.9
10	5	5	2	7.7
11	6	3	2	6.3
12	6	5	1	3.9
13	7	3	1	3.5
14	7	5	2	7.4
15	8	3	2	5.3
16	8	5	1	3.4