Name: KEY

Statistics 516 **Exam 1** **March 3, 2010**

1. Suppose a test for differential expression is conducted for each of 100 genes. The following table provides information about the observed p-values.

Range	Number of p-values	TAIL SUM	TAIL MEAN
[0.0-0.1]	42	100	$10 < 42$
(0.1-0.2]	10	58	$58/9 = 6.\overline{4} < 10$
(0.2-0.3]	13	48	$6 < 13$
(0.3-0.4]	4	35	$5 \geq 4$
(0.4-0.5]	10	31	
(0.5-0.6]	3	21	
(0.6-0.8]	8	18	
(0.8-1.0]	10	10	

a) Estimate the number of true null hypotheses using the histogram-based estimator described in course notes. Note that the bins in the table above are not all of the same width.

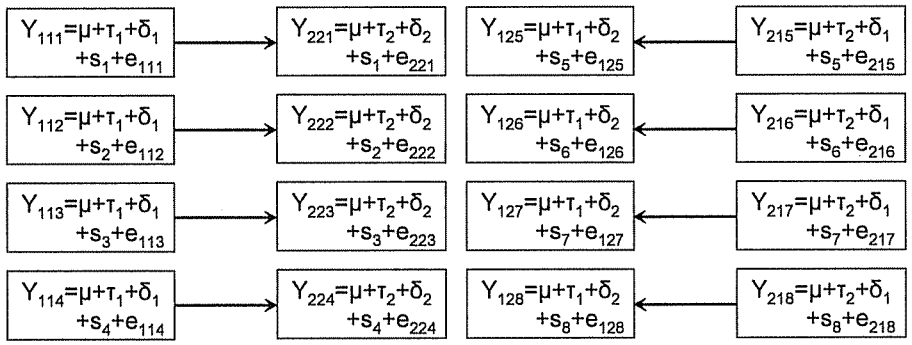
BY CALCULATIONS ABOVE,

$$\hat{m}_0 = 10 \times 5 = 50.$$

b) Estimate the number of true null hypothesis using the λ-threshold method discussed in course notes with $\lambda=0.6$. You may assume that no p-value equals exactly 0.6 when computing the estimate.

$$\hat{m}_0 = \frac{18}{1-0.6} = 45$$

2. Suppose a scientist would like to determine the effect of two treatments (1 and 2) on gene expression in mice. The scientist has a total of 16 mice and 8 two-color microarray slides. The scientist chooses to use the best available design for this situation and to model the data for a single gene using a linear mixed-effects model as discussed in class using the notation below.



$$Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$$

Provide unbiased estimators of the error and slide variance components in terms of the y_{ijk} data.

LET $U_k = Y_{11k} - Y_{22k}$ FOR $k=1,2,3,4$.

LET $V_{k-4} = Y_{12k} - Y_{21k}$ FOR $k=5,6,7,8$.

THEN $U_1, U_2, U_3, U_4 \overset{iid}{\sim} N(\tau_1-\tau_2+\delta_1-\delta_2, 2\sigma_e^2)$

AND $V_1, V_2, V_3, V_4 \overset{iid}{\sim} N(\tau_1-\tau_2+\delta_2-\delta_1, 2\sigma_e^2)$

THUS, $$\frac{\sum_{k=1}^{4}\left[(U_k-\bar{U})^2+(V_k-\bar{V})^2\right]}{2\left[(4-1)+(4-1)\right]} \equiv \hat{\sigma}_e^2$$

IS AN UNBIASED ESTIMATOR OF σ_e^2.

LET $a_k = \frac{Y_{11k}+Y_{22k}}{2}$ FOR $k=1,2,3,4$.

LET $a_k = \frac{Y_{12k}+Y_{21k}}{2}$ FOR $k=5,6,7,8$.

THEN $a_1,\ldots,a_8 \overset{iid}{\sim} N(\mu+\bar{\tau}+\bar{\delta}, \sigma_s^2+\sigma_e^2/2)$.

THUS, $\hat{\sigma}_s^2 = \frac{1}{7}\sum_{k=1}^{8}(a_k-\bar{a})^2 - \hat{\sigma}_e^2/2$ IS AN UNBIASED ESTIMATOR OF σ_s^2.

3. Suppose we test five null hypotheses and obtain p-values for the five tests as follows:

Null Hypothesis	H_{01}	H_{02}	H_{03}	H_{04}	H_{05}
p-value	0.200	0.033	0.002	0.010	0.036

a) Use Benjamini and Hochberg's method to determine which null hypotheses should be rejected if false discovery rate is to be controlled at the 0.05 level.

.002
.010
.033
.036 $\leq \frac{(m-1)\alpha}{m} = \frac{4 \times 0.05}{5} = .04$
.200 $> \frac{m\alpha}{m} = 0.05$

THUS, ALL NULLS EXCEPT H_{01} ($P = .2$) SHOULD BE REJECTED

b) Suppose that the total number of true null hypotheses is estimated to be 2. Using this estimate, convert the five p-values into q-values.

		q-VALUES
$.002 \times 2/1$	$= 0.004$	0.004
$.010 \times 2/2$	$= 0.010$	0.010
$.033 \times 2/3$	$= 0.022$	0.018
$.036 \times 2/4$	$= 0.018$	0.018
$.200 \times 2/5$	$= 0.080$	0.080

4. Suppose all steps in the RMA preprocessing algorithm up through quantile normalization have been completed for an experiment involving 5 GeneChips with 5 probe pairs per probe set. Suppose the quantile normalized values for one probe set are provided in the table below.

	Probe					ROW
GeneChip	1	2	3	4	5	M
1	2	4	4	4	8	4
2	1	3	5	3	6	3
3	4	6	2	6	12	6
4	0	2	6	2	4	2
5	3	5	3	5	10	5

COL M: −2 0 0 0 4

−2	0	0	0	4	
−2	0	2	0	3	
−2	0	−4	0	6	
−2	0	4	0	2	
−2	0	−2	0	5	

a) The researchers discover that one of the five probes in this probe set is actually measuring the expression of a gene different from the gene measured by the other four. Based on the data in the table above, identify the probe that you suspect measures a different gene, and briefly explain your reasoning.

PROBE 3 IS MOST LIKELY THE OUTLIER. IT IS NEGATIVELY CORRELATED WITH ALL OTHER PROBES.

b) Using all the data in the table above, compute the measure of expression the RMA procedure would assign to GeneChip 5.

USE MEDIAN POLISH (SEE ABOVE FOR 1ST STEPS)

$[\hat{\epsilon}]$ =

					M
0	0	0	0	0	0
0	0	2	0	−1	0
0	0	−4	0	2	0
0	0	4	0	−2	0
0	0	−2	0	1	0
M	0	0	0	0	0

$[\hat{y}] = [y] - [\hat{\epsilon}]$

$$\begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ 3 & 5 & 5 & 5 & 9 \end{bmatrix}$$

VALUE FOR GENECHIP 5 IS $\frac{3+5+5+5+9}{5} = \boxed{5.4}$

5. Suppose an investigator will use 24 plants (experimental units) and 12 two-color microarray slides to study gene expression in three plant genotypes. Genotypes A and B contain different genetic mutations that cause changes to leaf surfaces. Genotype C is the wild type genotype, meaning that it is the non-mutated form of Genotypes A and B and serves as a control genotype. The goal of the experiment is to identify genes that are differentially expressed between each mutant genotype and the control (i.e., differences between A and C and differences between B and C). Direct comparison of the two mutants (A vs. B) is not of interest. Consider the following two candidate designs that measure each of 24 experimental units exactly once.

Design 1: Use four slides to obtain a dye-balanced comparison of Genotypes A and B. Use four slides to obtain a dye-balanced comparison of Genotypes A and C. Use four slides to obtain a dye-balanced comparison of Genotypes B and C.

Design 2: Use six slides to obtain a dye-balanced comparison of Genotypes A and C. Use six slides to obtain a dye-balanced comparison of Genotypes B and C.

a) Ignoring the dye factor for this question, what experimental design terminology would you use to describe Design 1?

BALANCED INCOMPLETE BLOCK DESIGN

b) Suppose we are interested is separately testing the following two null hypotheses for each gene.

H_{01} : Mean Expression for Genotype A = Mean Expression for Genotype C
H_{02} : Mean Expression for Genotype B = Mean Expression for Genotype C

$$\underset{\sim}{\beta} = \begin{bmatrix} \delta_2 - \delta_1 \\ \tau_A - \tau_C \\ \tau_B - \tau_C \end{bmatrix}$$

Which design would you prefer? Support your answer with appropriate calculations.

$$X_1 = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

$$X_1'X_1 = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 8 & -4 \\ 0 & -4 & 8 \end{bmatrix} \qquad X_2'X_2 = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

$$(X_1'X_1)^{-1} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/6 & 1/12 \\ 0 & 1/12 & 1/6 \end{bmatrix} \qquad (X_2'X_2)^{-1} = \begin{bmatrix} 1/12 & 0 & 0 \\ 0 & 1/6 & 0 \\ 0 & 0 & 1/6 \end{bmatrix}$$

DESIGNS ARE EQUIVALENT BECAUSE $1/6 = 1/6$ AND $1/6 = 1/6$.

c) Suppose we are interested is estimating the two-element vector

Mean Expression for Genotype A – Mean Expression for Genotype C
Mean Expression for Genotype B – Mean Expression for Genotype C.

Which design would you prefer based on the D-optimality criterion we discussed in class? Support your answer with appropriate calculations.

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\text{DET}\left(M(X_1'X_1)^{-1}M' \right) = \begin{vmatrix} 1/6 & 1/12 \\ 1/12 & 1/6 \end{vmatrix}$$

$$= \frac{1}{36} - \frac{1}{144} = \frac{3}{144} = \frac{1}{48}$$

$$\text{DET}\left(M(X_2'X_2)^{-1}M' \right) = \begin{vmatrix} 1/6 & 0 \\ 0 & 1/6 \end{vmatrix} = \frac{1}{36}.$$

$\frac{1}{48} < \frac{1}{36}$ So DESIGN 1

IS PREFERRED. (SURPRISED?)