STAT516 **Homework 4** **Due Thursday, March 24**

0. Write a complete solution to any Exam 1 problem for which you did not receive full credit. Please turn in your original Exam 1 answers with your answers to this homework assignment.

1. Suppose 3 dependent tests are conducted. Two of the three tests have a true null hypothesis, and one of the three has a false null hypothesis. Suppose the joint distribution of V and R is given by the following (partially complete) table, where V is the number of false positives and R is the number of rejected null hypotheses.

r	v	P(V=v, R=r)	$V/\max\{R,1\}$
0	0	0.20	0
1	0	0.40	0
1	1	0.02	1
2	0	0.00	0
2	1	0.20	1/2
2	2	0.10	1
3	0	0.00	0
3	1	0.00	1/3
3	2	0.08	2/3
3	3	0.00	1

$$FDR = (.02)(1) + (.2)(1/2) + (.1)(1) + (.08)(2/3) = \boxed{0.273}$$

a) Fill in the missing entries in the joint probability distribution of V and R.

b) Determine FDR for this testing scenario.

2. 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]	22	100	10
(0.1-0.2]	10	78	8 2/3
(0.2-0.3]	18	68	8.5
(0.3-0.4]	10	50	7 1/7
(0.4-0.5]	9	40	6 2/3
(0.5-0.6]	6	31	6.2
(0.6-0.7]	4	25	6.25
(0.7-0.8]	7	21	7
(0.8-0.9]	8	14	7
(0.9-1.0]	6	6	6

WORKING FROM SMALL P-VALUES TO LARGE, FIRST BIN WITH COUNT ≤ TAIL MEAN.

a) Estimate the number of true null hypotheses using the histogram-based estimator described in course notes. $10 \cdot 6.2 = \boxed{62}$

b) Estimate the number of true null hypothesis using the λ-threshold method discussed in course notes with λ=0.8. You may assume that no p-value equals exactly 0.8 when computing the estimator.

$$\frac{8+6}{1-0.8} = \boxed{70}$$

c) Using the estimate in part (b) for the number of true null hypotheses, provide as much information as possible about the q-value for a gene with a p-value of 0.2.

$$q\text{-VALUE} \leq \min\left\{\frac{.2(70)}{32}, \frac{.3(70)}{50}, \frac{.4(70)}{60}, \cdots, \frac{1.0(70)}{100}\right\}$$

$$= \frac{.3(70)}{50} = \frac{21}{50} = 0.42.$$

IF THE 18 p-VALUES IN THE BIN $(.2, .3]$ WERE ALL LESS THAN $.2+\varepsilon$ FOR SOME SMALL $\varepsilon > 0$, THE q-VALUE FOR $p = 0.2$ WOULD BE $>$ $\frac{(.2+\varepsilon)70}{50}$, WHICH CONVERGES TO 0.28 AS $\varepsilon \to 0$. THUS, WE KNOW

q-VALUE FOR $p = 0.2$ IS IN THE INTERVAL

$(0.28, 0.42]$

3. Consider 100 coins, each with two sides (heads and tails). 55 of the coins are fair coins that each land heads with probability 0.5 and tails with probability 0.5. The other 45 coins are unfair. Each of these coins lands heads with probability 0.9 and tails with probability 0.1. Suppose your job is to discover which of the coins are unfair. To do so, you may flip each coin n times and record the result of each flip.

a) Determine n and a rule for declaring a coin to be unfair based on the result of n flips so that the true positive rate (TPR) and expected discovery rate (EDR) are each greater than 0.85. To match the definitions of TPR and EDR we learned in class to this coin flipping scenario, note that unfair coins are like differentially expressed genes and fair coins are like equivalently expressed genes. The null hypothesis for the i^{th} coin is

$$H_{0i} : \text{P(coin i lands heads)=P(coin i lands tails)=0.5.}$$

b) Compute TPR and EDR for your choice of n and your rule presented in part (a). (Even if you can't find a rule that satisfies the requirement in (a), you can get full credit for this part if your computations are correct for any rule that you proposed.)

TAKE $n = 6$ AND DECLARE A COIN TO BE UNFAIR IF X = NUMBER OF HEADS IS 5 OR 6. THEN

$$EDR = P(X \geq 5 \mid \text{COIN UNFAIR}) = \binom{6}{1}(0.1)(0.9)^5 + (0.9)^6$$

$$= \boxed{0.885735}$$

AND

$$TPR = P(\text{COIN UNFAIR} \mid X \geq 5) = \frac{P(X \geq 5 \mid \text{COIN UNFAIR})\,P(\text{UNFAIR})}{P(X \geq 5)}$$

$$= \frac{(0.885735)(.45)}{(0.885735)(.45) + \left(6(.5)^6 + (.5)^6\right)(.55)}$$

$$= \boxed{0.868865}$$

4. Suppose researchers are interested in studying the effect of fertilizer and a viral infection on gene expression in plants. Three fertilizer amounts (low, medium, and high) are randomly assigned to a total of 12 individually potted plants, such that four plants are treated with each fertilizer amount. Suppose that two leaves of comparable developmental stage are identified for each plant. One of the leaves on each plant is randomly selected for infection with a plant virus. A gel containing the virus is rubbed on each selected leaf. The other leaf on each plant is rubbed with the gel without the virus to serve as an uninfected control. Suppose that after the fertilizer and viral infections have been applied for a relevant length of time, sufficient RNA can be extracted from each leaf to obtain a measure of expression for thousands of genes using microarray technology.

a) Name the treatment factor or factors in this experiment, and name the levels for each treatment factor. FERTILIZER (LOW, MEDIUM, HIGH) VIRUS (INFECTION, CONTROL)

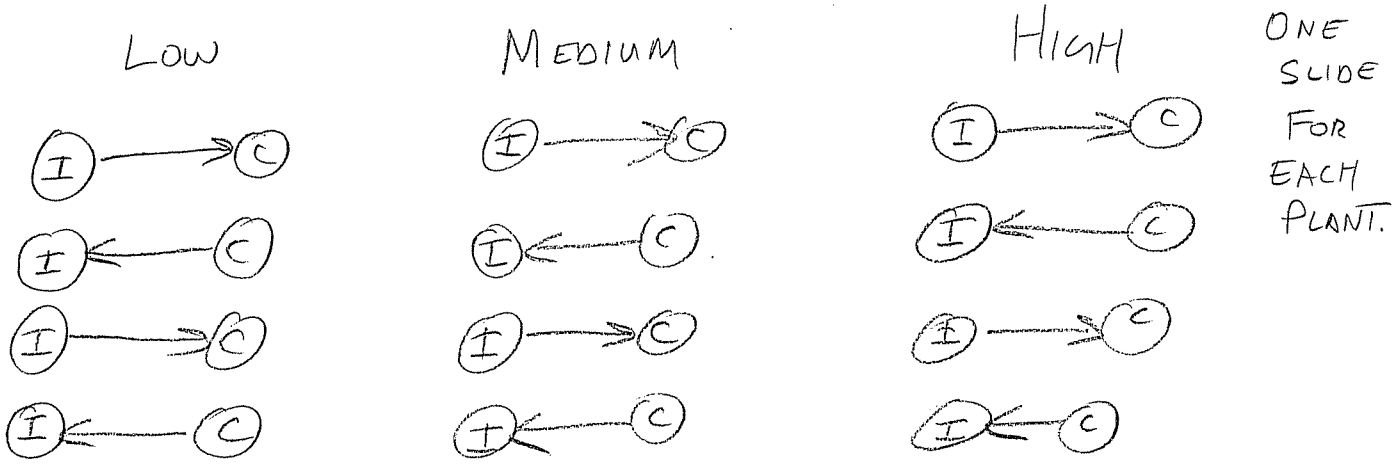
b) What are the experimental units in this experiment?
PLANTS (WHOLE-PLOT) AND LEAVES (SPLIT-PLOT)

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

d) Write a model that you would use for the analysis of expression data from a single gene assuming that one Affymetrix GeneChip is used to measure expression in each leaf. You may use the abbreviated model notation we discussed in class where a word is used for each factor. Please circle any terms in your model that you would choose to treat as random.

FERT VIRUS FERT:VIRUS (PLANT)

e) Draw a diagram using our circle and arrow notation and any other additional notation necessary to show how you would assign leaf RNA samples to microarray slides if 12 two-color microarray slides were available for this experiment. Suppose the researchers are primarily interested in testing for differences in expression caused by the virus and to determine if any differences caused by the virus vary with fertilizer amount.



ONE SLIDE FOR EACH PLANT.

f) Write the model that you would use for the analysis of the design as diagramed in part (e). You may use the abbreviated model notation we discussed in class where a word is used for each factor. Please circle any terms in your model that you would choose to treat as random.

DYE FERT VIRUS FERT:VIRUS (SLIDE)

INTENTIONALLY CONFOUNDED WITH PLANT

g) Suppose the following are normalized gene expression measures for the two-color microarray experiment you diagrammed in part (e). (Obviously these are not real normalized expression measures; the values have been selected to make computations easier.)

Fertilizer Level	Infected Leaf	Control Leaf	d
low	5	4	−1
low	2	2	0
low	4	3	−1
low	1	3	−2
med	8	3	−5
med	4	1	3
med	9	4	−5
med	12	5	7
high	9	5	−4
high	7	4	3
high	5	3	−2
high	8	6	2

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

Reduce this data to a vector of differences **d**. Provide a matrix X of full column rank and a corresponding vector of parameters **β** so that a standard multiple regression model of the form **d**=X**β**+**r** can be used to test the null hypothesis of no virus main effect and the null hypothesis of no interaction between virus and fertilizer.

$\delta_5 - \delta_3$ = DIFFERENCE IN DYE EFFECTS

θ_L = CONTROL-VIRUS FOR LOW FERT

θ_M = CONTROL-VIRUS FOR MED. FERT

θ_H = CONTROL-VIRUS FOR HIGH FERT

$$\underline{\beta} = \begin{bmatrix} \delta_5 - \delta_3 \\ \theta_L \\ \theta_M \\ \theta_H \end{bmatrix}$$

h) Write the null hypothesis of no virus main effect in the form A**β**=**0**, where A is a matrix with one or more rows and **0** is a vector of zeros with one or more elements.

NO VIRUS MAIN EFFECT $\Leftrightarrow \frac{\theta_L + \theta_M + \theta_H}{3} = 0$

$\Leftrightarrow \theta_L + \theta_M + \theta_H = 0$

$\therefore A = [0\ 1\ 1\ 1]$

i) Repeat part (h) for the null hypothesis of no interaction between virus and fertilizer. $\theta_L = \theta_M = \theta_H$

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

5. It is possible to use the EBarrays approach when there is only one experimental unit per treatment. (This is definitely not recommended.) Suppose a two-treatment experiment has been conducted with one experimental unit per treatment. Suppose model parameters have been estimated as follows:

$$p \approx 0.05 \quad \alpha \approx 10 \quad \alpha_0 \approx .5 \quad v \approx 50.$$

Which of the following genes would have the higher estimated posterior probability of differential expression? Back up your answer with appropriate calculations.

	Treatment 1 Expression (original scale)	Treatment 2 Expression (original scale)
Gene 1	100	200
Gene 2	1000	1500

$$\text{ESTIMATED PPDE} = \frac{p f_{DE}(\underline{x})}{p f_{DE}(\underline{x}) + (1-p) f_{EE}(\underline{x})}$$

$$= \frac{1}{1 + \frac{(1-p)}{p} \frac{f_{EE}(\underline{x})}{f_{DE}(\underline{x})}}$$

THUS, WHICHEVER GENE HAS HIGHER $\frac{f_{EE}(\underline{x})}{f_{DE}(\underline{x})}$ HAS LOWER EPPDE.

$$\text{COMPUTE} \quad \frac{f_{EE}(\underline{x}_1)}{f_{DE}(\underline{x}_1)} \Big/ \frac{f_{EE}(\underline{x}_2)}{f_{DE}(\underline{x}_2)} = \left(\frac{v+1000+1500}{v+100+200}\right)^{2\alpha+\alpha_0} \left[\frac{(v+100)(v+200)}{(v+1000)(v+1500)}\right]^{\alpha+\alpha_0}$$

$$> 1.$$

THUS, ESTIMATED PPDE IS HIGHER FOR GENE 2.