VCU Bioinformatics and Bioengineering Summer Symposium
Microarray Statistics Problem Set

1. The BBSI admission committee, hoping to reduce it's work by half, decided to focus primarily on applications from those whose last names are in the first half of the alphabet.

X. Actually, I made that up: we paid no attention to first letters, neither this year nor any other year.

Y. Actually, I made ***that*** up: We ***did*** select for people this year with first names beginning with "A" through "M".

What do you think? Given the resources you have available (which do not include transcripts of the admission committee), what insight can you provide regarding scenarios **X** and **Y**?

The chief resource you have available is the list of names of participants for the current year and for previous years. To save you the bother of scraping the BBSI web page, I've created a file called "bbsi-names.txt", in the SHARED-FILES directory in CyanoBIKE. The first 13 names are yours and the rest are your predecessors.

We'll say that there is evidence of selection if the average of the letters starting your last names is significantly different from the average of the letters starting the last names of everyone else. How to tell?

A. Use the READ function and the SHARED option to DEFINE a variable containing all the names of BBSI participants, past and present. Verify that the list is what it ought to be by listing the FIRST 13 items in the list (use the *number* token of the FIRST function to specify how many you want to list.

B. DEFINE a variable containing just the last names of everyone. Do this by SPLITting the list of names every " " (the space that separates the first from the last name. Then take the SECOND element IN-EACH of the resulting first-name/last-name pairs. (Note that SECOND has a an *in* token that enables you to switch between the second element of a list and the second element in each given list)

C. Read in a function LETTER-TO-NUMBER, that will be useful to calculate an average of letters. Do this using the RUN-FILE file function (which executes functions and possibly adds them to your list of functions). The file name is "sam-functions.bike" in the SHARED directory. If you execute RUN-FILE, you should see three new functions appear under your FUNCTIONS tab, including LETTER-TO-NUMBER.

D. Play with LETTER-TO-NUMBER. What does it return when you give it "a"? "A"? "Z"? "Zodiac"? ("Alpha" "bravo" "Charlie") ?

E. DEFINE set1 as your last names, i.e. the first 13 last-names in the variable you defined in **B**. DEFINE set2 as the remaining last names, using the SUBTRACT-SET function and either the FROM or BY option, as you wish. DEFINE *bbsi-2009* as set1 and DEFINE *bbsi-rest* as set2, because you're going to fool around with set1 and set2 and you don't want to lose track of reality.

F. What is the MEAN first letter of your last names? What is the MEAN first letter of everyone else's last names? SUBTRACT the first mean from the second, and DEFINE *true-difference* to be this value. What is the difference between these two means? Is this difference significant?

G. To determine significance, do the same operations as you did in **E** and **F**, but with a randomized list of names. DEFINE *shuffled-last-names* as the list of last-names SHUFFLEd. Change the definition of set1 and set2 to use *shuffled-last-names* instead of the original list. Recalculate the means and the difference of the means as in **F**. Is the new difference greater or smaller than the true difference?

H. Repeat these operations 100 times. Do this by means of a FOR-EACH loop, going through the operations for each iteration from 1 To 100. Drag into the ***Body*** section of the loop, the definitions of the shuffled last names, set1, set2, the two means, and their difference. For the ***Result*** section, WHEN the ORDER difference > *true-difference* **COUNT** the iteration. DEFINE the result of the loop as the variable called *expected-by-chance*. How many times in 100 shufflings would you expect to see a distribution of names at least as divergent as the names in your class?

I. Do you think that observed distribution of last names might have arisen by chance or is it evidence of a selection process?

That took a fair bit of work. Surely there's a simpler way to find out whether the mean last name in your class is significantly different from the mean last name in other classes. Well, there is!

J. Enter into the T-TEST function *bbsi-2009* and *bbsi-rest* and execute the function. Never mind what the number means for the moment. Go to the t-test probability calculator (*http://www.danielsoper.com/statcalc/calc08.aspx*, also available on the simulation web site) and plug in the t-value you just calculated. For degrees of freedom, give a crude estimate as (m + n – 2), where m is the number of people in your class and n is the total number in the other classes. Click **Calculate**. What is the significance of the numbers you get?

K. What is a t-test?

2. Can a t-test be used to determine whether the mean expression value given for a gene from a microarray experiment for a certain condition is significantly different from the mean expression value given for the same gene for a control condition? What do you need in order to use it?

3. Apply a t-test to microarray values and compare it with your own sense of what's right.

3a. Use INSIDE-MICROARRAY to examine the experimental (or target) values and control values for two genes measured in the experiment of Hihara et al (2001). Bring down Hihara_2001 from the Data/Microarray menu. For *gene* use SGL0001. For *condition* use 2 (indicating the second condition, i.e. 1 hr of high intensity light). Select from options +CONTROL-VALUES, +TARGET-VALUES, +CONTROL-MEAN, and +TARGET-

MEAN . Execute the function. Re-execute it replacing sgl0001 with SLL1185. What do you think of the significance of the difference between the two means?

3b. Now see what a t-test has to say. Use T-TEST to get a t-value comparing the CONTROL-VALUES and TARGET-VALUES for each gene. What is the probability that differences of these magnitudes could have arisen by chance due to selection from a common pool of values?