**Math 180 Fall, 2014 Assignment 9**

This assignment is due on Tuesday, November 11.

The first two of these exercises are taken from *Statistics,* by Freedman, Pisani, and Purves.

1. Three sets of data are collected and the correlation coefficient is computed in each case. The variables are

1. grade point average in freshman year and in sophomore year
2. grade point average in freshman year and in senior year
3. length and weight of two-by-four boards.

Possible values for the correlation coefficients are

-0.50 0.00 0.30 0.60 0.95

Match the correlations with the data sets; two will be left over. *Explain your answers.*

2. Compute the correlation coefficient for each of the following three data sets.

x: 1 1 1 1 2 2 2 3 3 4

y: 5 3 5 7 3 3 1 1 1 1

x: 1 1 1 1 2 2 2 3 3 4

y: 1 2 1 3 1 4 1 2 2 3

x: 1 1 1 1 2 2 2 3 3 4

y: 2 2 2 2 4 4 4 6 6 8

*Extra credit*: As with anything involving sampling, another sample of 10 x—y pairs from the same data may produce a different correlation coefficient. We will see that a correlation coefficient may be legally significant, particularly in discrimination cases. The plaintiff usually alleges that a certain correlation coefficient is significantly different from 0. Starting with the null hypothesis that the true correlation coefficient is 0, there is a t-test to determine the probability that a correlation as large or larger (in absolute value) than the one at hand will be observed. It depends on the number N of pairs, which is 10 in each of our three data sets. Look up the test and compute the probability for each of the three data sets.

3. A test sometimes used to determine if the entries in a set of data have been fabricated is to take the least significant digit (usually the right-most) in each item of data and see if these are randomly, and in particular uniformly, distributed. (We have already seen in class that humans generally are quite unable to create a string of truly random digits; it is not easy even for machines.) In the famous case of Dr. Thereza Imanishi-Kari, she was accused of fabricating data that appeared in an article in the April 25, 1986 issue of the journal *Cell.* The investigators extracted the least significant digit from 47 items of data in that paper. Generally the least significant digit would be any from 0 to 9, inclusive, but in this case, because of the way in which the data were rounded (a very significant factor in the case) they only ranged from 1 to 9. In a very large set of random digits one would expect the digits 1 to 9 each to appear very close to 1/9 of the time, but course, that is impossible here since a digit can not appear a fractional number of times. To get some idea of how variable the results might be, they did the same with 47 items of data from a control. Here is a table of the digit frequencies of 1 to 9 in Dr. Imanishi-Kari’s paper (I-K) and in the control (Ctl).

digit: 1 2 3 4 5 6 7 8 9

I-K: 0 14 2 3 4 4 6 11 3

Ctl: 7 6 3 10 3 2 7 7 2

Using a chi-squared test, under the null hypothesis that in both cases the digits should have been randomly distributed, determine the probability, both in Dr. Imanishi-Kari’s case and in the control case, that one would see as large a chi-square value as that observed. In your opinion, does the result justify bringing charges against Dr. Imanishi-Kari?

Initially, Dr. Imanishi-Kari was found guilty of scientific misconduct, but this was reversed by an appeals panel that found the statistical aspect of the investigation seriously flawed. In the meantime, however, Dr. David Baltimore, the Nobel-prize winning head of the Rockefeller Institute, was forced to resign because he had co-authored the article and supported Dr. Imanishi-Kari.

Note: This might be an interesting case for some team to report.