BASIC ASSIGNMENT 6
CORE METHODS IN EDUCATIONAL DATA MINING
PROFESSOR BAKER
CORRELATION MINING
DUE NOON, WEDNESDAY NOVEMBER 12

In this assignment, you need to apply a set of metrics to two data sets, B6-data-v2-set1.csv and B6-data-v2-set1.csv

These are not real data sets, but are simulated data, generated for the purpose of this assignment.

You can use any tool you want to complete this assignment. This includes Microsoft Excel, and statistical or data mining package you choose, and other tools available on the internet.

The goal of this assignment is to conduct post-hoc methods to determine which of a large set of correlations can be trusted – and in the process to understand how to validly conduct analyses involving large numbers of correlations in big data sets.

Question 1:

Data set 1 represents a set of distinct studies conducted on small populations of students – you might see these types of results if you administered a survey to the students of just one classroom teacher. Within data set 1, how many correlations are statistically significant (according to the customary $p<0.05$ definition) if you do not apply any sort of post-hoc control?

Question 2:

If you apply a post-hoc Bonferroni control to these results, how many correlations remain statistically significant?

Question 3:

If you apply Benjamini & Hochberg's FDR Correction to these results, how many correlations remain statistically significant?

Question 4:

What is the correlation with the lowest p-value that comes up significant for B&H but not for Bonferroni?

Question 5:

Why is the answer for Question 4 correct?

A) That study has a higher N, so the p value is lower than a study with higher correlation
B) That study has a lower N, so the p value is lower than a study with higher correlation
C) That study has a higher N, so the p value is higher than a study with higher correlation
D) That study has a lower F value, so the p value is higher than a study with higher correlation

Question 6: Data set 2 represents a larger set of correlations within data from a larger population of students – for example, the entire population of students using a medium-sized online learning environment. Within data set 2, how many of the 1,112 correlations are NOT statistically significant (according to the customary $p<0.05$ definition) if you do not apply any sort of post-hoc control?

Question 7:

If you apply a post-hoc Bonferroni control to these results, how many correlations are now NOT statistically significant?

Question 8:

If you apply Benjamini & Hochberg's FDR Correction to these results, how many correlations are now NOT statistically significant?

Question 9:

What is the lowest correlation that is still statistically significant, according to Bonferroni's test?

Question 10:
Now do you see why Professor Baker says in the video that statistical significance doesn't matter much for really big data sets? (and this is NOT a big data set by reckoning in other fields)

A) Yes – Bonferroni is ridiculously conservative with 1,112 tests, and yet correlations that are absurdly small still come up statistically significant.
B) No. I think the answer to Question 9 is a fine correlation, perfectly likely to represent a large effect.
C) No. Big data is a FAD. No one should ever expect to work with a data set over 150 data points, after the societal collapse predicted by James Howard Kunstler occurs.