CSC 121 — Lab Exercise 11

This is a non-credit exercise, which you do not hand in.
You may work on your own or together with another student, as you please.

In this lab, you will use random permutations to assess informally whether or not some apparent relationship between two variables might plausibly be due just to chance, without there being any real relationship. This is a form of "significance test", but without any formulas. (This idea is due to Andreas Buja.)

You might also want to try using the `knitr` package for this lab exercise, producing an HTML file in the end that shows your analysis of the data.

A permutation of some items is just a rearrangement of the items in some order. R can produce a random permutation of the elements of a vector, `v`, with `sample(v)`.

For this lab, you should read the data at this URL:

```
http://www.cs.utoronto.ca/~radford/csc121/lab11-data
```

using `read.table` with the `header=TRUE` option. This data file has 50 observations of three numeric variables (A, B, and C) and a binary variable D. These 50 observations are assumed to be a random sample from some larger set of possible observations.

At first, we'll ignore the D variable. We'd like to see if any of the variables A, B, and C are related, not just in the 50 observations we have, but in the larger set of possible observations.

You should start by calling `plot` with the data frame you read, but with only variables A, B, and C. You should be able to extract just these three variables with a suitable form of indexing, using their names (which is better than using the column numbers). The call of `plot` will show you pairwise scatterplots of each of these three variables versus the others. You can then decide what relationships these variables *seem* to have to each other.

To check how sure you should be that a relationship you see is real, you can look at the plot of the actual data along with several plots of the same data, but with the values of one of the variables randomly permuted. Because the permutation is random, it's not possible for one of these permuted variables to have any real relationship with the other variable. But for any of these permutations, the variables might seem to be related just by chance. By comparing the plot of the real data with these plots of permuted data, you can see whether any relationship you see in the real data is stronger than what happens by chance in the plots of permuted data.

To do this, you should write a function called `perm_test`, which takes two arguments that are numeric vectors of the same length. This function should produce nine plots, arranged in a 3×3 grid (you can get this by calling `par(mfrow=c(3,3))` before producing the nine plots, which get filled in by row in this 3×3 array). The centre plot should be a scatterplot of the data in the two vectors. Each of the other eight plots should a scatterplot of the first

vector versus a randomly permuted form of the second vector. Using the `pch=20` option to `plot` may produce plots that are more readable when they are small.

Once all nine plots are visible, you should judge intuitively whether or not the relationship you see in the centre plot (of the real data) is stronger than the relationships you see in the eight randomly permuted plots. If it's not stronger, you might doubt whether the relationship is real (and hence would also be seen in future data from the same source).

Now, you should look at the variable D, which has two possible value, "X" and "Y". We'd like to know whether there may be relationships between any of the variables A, B, and C in the subset of data for which D has the "X" value. Similarly, we'd like to know whether there are relationships among A, B, and C in the subset for which D has the value "Y". Use your `perm_test` function and suitable indexing operations to investigate these questions.