# STA 410/2102, Spring 2002 — Assignment #1

Due at **start** of class on February 15. Worth 16% of the final mark.

*Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own.*

Consider an experiment with paired subjects, such as the following:

> We wish to test whether the type of lighting — incandescent or fluorescent — affects student performance on tests such as the SAT. Suppose we think test scores are likely to vary a lot from school to school (perhaps due to demographics, or to quality of school administration). To reduce the effect of this variability, we conduct the experiment on pairs of students. We choose one pair from each of 20 schools, selecting the pair for a school randomly from among the volunteers at that school. We set up two rooms that are as identical as possible except for the lighting. One student from each school writes the test in the room with incandescent lighting, the other in the room with fluorescent lighting. We assign rooms to the students in a pair randomly. For each pair, we compute the difference in the test scores of the two students, and use the average difference over pairs to test whether lighting has an effect.

The data from such an experiment will consist of $n$ pairs ($n = 20$ above), which we can write as $(x_1, y_1), \ldots, (x_n, y_n)$. From these pairs, we obtain differences, $d_i = x_i - y_i$. The average difference, $(1/n) \sum d_i$, will be used to judge whether we have reason to believe that the mean of the distribution from which the $x_i$ are drawn differs from the mean of the distribution from which the $y_i$ are drawn. This will be done by a hypothesis test.

The null hypothesis for this test might be different for different situations. For the experiment described above, we might consider the null hypothesis to be that the lighting has *no effect*. If so, the distribution of the $x_i$ should be identical to the distribution of the $y_i$. In another experiment — for instance, comparing two different styles of classroom instruction — we might be sure that the $x_i$ and $y_i$ will have distributions that differ in some way, at least slightly, and will therefore use a null hypothesis that states only that these distributions have the *same mean* (but might differ in skewness, for example).

The standard approach with either null hypothesis is to use a paired $t$ test, which is the same as a one-sample $t$ test of the null hypothesis that the mean of the differences is zero. This test will be exactly correct if the distribution of the differences is normal, and the difference for one pair is independent of the difference for another pair. (We'll assume here that this independence assumption holds, even though it might not in practice.)

What if the distribution of differences isn't normal? As long as the distribution has finite variance, the $t$-test will work correctly when $n$ is large enough, but how about when $n$ is small? Are there other tests that work better when $n$ is small, and the data are not normal?

In this assignment, we will consider two alternative tests. Both tests are intended for use when the null hypothesis is that the distributions for $x_i$ and $y_i$ are identical, from which it follows that the distribution of $d_i = x_i - y_i$ is symmetrical around zero. (If our null hypothesis is only that the means are the same, the distribution of the differences might not be symmetrical.)

The *trimmed t test* is based on the idea that it might be a good idea to get rid of the most extreme data points, on the grounds that they introduce lots of undesirable variability. We drop the smallest 5% and the largest 5% of the $n$ differences, $d_i$ (rounding $0.05n$ to the nearest integer if necessary). We then perform a $t$ test on the remaining differences (testing the null hypothesis that their mean is zero). Discarding the most extreme observations would clearly bias the results if the distribution were not symmetric, but when it is symmetric, as we're assuming, we might think that it is a good idea.

The *permutation test* is based on the idea that if the distribution of differences is symmetric around zero, changing the signs of the differences at random has no effect on the distribution. We can therefore test the null hypothesis by comparing the actual mean difference, $\bar{d}^*$, with the mean differences for $K$ data sets obtained by randomly setting the signs of the $d_i$ to be positive or negative. Suppose that the means of these randomly changed data sets are $\bar{d}^{(1)}, \bar{d}^{(2)}, \ldots, \bar{d}^{(K)}$. If $|\bar{d}^*|$ is less than or equal to $J$ of the $|\bar{d}^{(k)}|$, we declare that the $p$-value for the test to be $(J+1)\,/\,(K+1)$. If the null hypothesis is true, the distribution of the $\bar{d}^{(k)}$ is the same as the distribution of $\bar{d}^*$, and $J$ will be equally likely to be any integer from 0 to $K$. Apart from its discreteness, the $p$-value will therefore have a distribution that is uniform between 0 and 1, as a valid $p$-value should.

In this assignment, you will test how well these procedures, and the standard $t$ test, work in two simulated situations, which are defined by the following procedures for generating pairs, $(x_i, y_i)$:

1) Generate $c_i \sim N(1,1)$, $a_i \sim N(0,1)$, $b_i \sim N(0,1)$, independently. Then let $x_i = c_i + a_i$ and $y_i = c_i + b_i + \mu$.

2) Generate $c_i \sim \text{Exp}(1)$, $a_i \sim \text{Exp}(1)$, $b_i \sim \text{Exp}(1)$, independently. Then let $x_i = c_i\,a_i$ and $y_i = c_i\,b_i + \mu$.

The parameter $\mu$ will be zero when simulating data from the null hypothesis. It will be some non-zero value when simulating data from an alternative in which the distributions of $x_i$ and $y_i$ are different.

You should first evaluate how well the standard $t$ test, trimmed $t$ test, and permutation test work for data generated under the null hypothesis ($\mu = 0$) by methods (1) and (2) above. Test this using sample sizes of $n = 20$ and $n = 100$. For each test, you should generate at least 1000 data sets, and find the $p$-values for each data set using each of the three methods. For the permutation test, use at least $K = 99$ permutations (preferably more, if the computations don't take too long). Plot a histogram of $p$-values for each test and each situation, and also find the probability of a Type I error when testing at a significance level of 0.05.

You should then repeat these simulations with $\mu = 0.2$, in order to evaluate how powerful the tests are at detecting a difference of this size. In particular, find the Type II error rate for each test when testing at a significance level of 0.05.

You should hand in a properly formatted and documented program, the plots and other outputs from your tests, and a discussion of the meaning of the results.

Some R functions you may find useful: `rnorm, rexp, sample, sort, t.test, hist.`