STA 410S/2102S: Homework #1
Due February 4, 2005

Note: The answers to each question should be written or typed as a report to a non-statistician. The numerical results should be summarized in tables, and the text should provide the interpretation. All computations and code should be included as appendices. Code should be commented. Highlighting selected pieces of R text is sometimes helpful, but is **not** a suitable way to answer the question. You are encouraged to discuss the assignments with each other, but please write up your work on your own.

1. *Model matrices for designed experiments*

   (a) A randomized block design has observations on t treatments in b blocks. The blocks are chosen so that units within a block are as similar as possible, and each treatment occurs the same number of times in each block (often once). By comparing treatments within blocks, variation due to differences between experimental units is eliminated, or at least greatly reduced. A classical example from Cochran and Cox's 1958 book on experimental design has the following layout:

		Pounds of potash per acre					
		36	54	72	108	144	Mean
Block	I	7.62	8.14	7.76	7.17	7.46	7.63
	II	8.00	8.15	7.73	7.57	7.68	7.83
	III	7.93	7.87	7.74	7.80	7.21	7.71
Mean		7.85	8.05	7.74	7.51	7.45	7.72

   This experiment tested the effects of five levels of application of potash on the strength of cotton fibres; the response is the strength index of a sample of fibres taken from the indicated block The treatment is amount of potash, and the blocks were determined as homogenous plots of soil to which the treatment was applied. The model for the RB design is usually written as

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}; \quad i = 1, \ldots, t; j = 1, \ldots b$$

   where the τ_i are treatment effects and the β_j are block effects; usually the ϵ_{ij} assumed to be independent with mean 0 and variance σ^2.

   i. Fit this model using the `lm` command in R, and again using the `aov` command, and summarize the results by describing the effect of various amounts of potash fertilizer on the strength index of cotton.

   ii. Extract the model matrix from the `lm` object, and deduce from this what constraints are used in estimating the τ_i's and β_j's.

iii. Refit both models after specifying the sum constraint by:

`options{contrasts=c("contr.sum","contr.poly");`

note that this option stays in effect until it is re-assigned or until R is restarted. Does this change in options have any effect on your conclusions of part (i)?

(b) *2102 (bonus for 410):* Two-way control of blocking can sometimes be achieved with a Latin square design. This requires that the number of levels of blocking factor 1, and the number of levels of blocking factor 2, are the same as the number of treatments to be tested. An example of a 4×4 Latin square after randomization is

$$\begin{array}{cccc} T_4 & T_2 & T_3 & T_1 \\ T_2 & T_4 & T_1 & T_3 \\ T_1 & T_3 & T_4 & T_2 \\ T_3 & T_1 & T_2 & T_4. \end{array}$$

In an application in experimental psychology the rows might correspond to subjects and the columns to periods within an experimental session. The arrangement ensures that constant differences between subjects or between periods affect all treatments equally and thus do not induce error in estimated treatment contrasts. Describe how to fit a Latin square using `lm` and explain the structure of the model matrix.

2. *Permutation t-tests* Suppose we have two independent normal samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$. A test for equality of the means, assuming the variances are equal, is typically based on the t-statistic

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

where $s^2 = (n+m-2)^{-1}\{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2\}$ is the pooled estimate of the common variance. In R this is carried out using `t.test(x,y,var.equal=TRUE)`.

If the observations are samples from a normal distribution then the t-statistic follows a T_{n-2} distribution in repeated sampling from this model. A two-sample test of equality of means that does not depend on the normality assumption, but only assumes that the x's and y's come from the same distribution, is the permutation test. The idea behind this test is that if the means are equal, then the observations $x_1, \ldots x_m$ and $y_1, \ldots, y_n$ are independent samples from the same distribution, and the assignment of m observations to be "x"s, and n to be "y"s is random. The randomization distribution of the t statistic is obtained by recomputing t for all possible assignments of the $m+n$ observations to two groups of size m and n respectively.

(a) Using `rnorm`, generate a sample of size 10 from $N(0,1)$ and a sample of size 12 from the same distribution. This is your data set. Write an R function that takes as input your two vectors x and y, and takes 1000 samples from the $\binom{n+m}{m}$ possible assignments of the observations to two groups of size m and n, respectively.

(b) For each of the possible assignments generated in (i), compute the t-statistic. This is the *permutation* distribution of the t-statistic. Compute the permutation p-value.

(c) Construct a density estimate of the permutation distribution, and compare it to the exact t-distribution.

(d) *2102 (bonus for 410)*: Simulate the performance of the permutation t-test and the usual t-test when x and y come from a non-normal distribution. A suggested candidate non-normal distribution is a mixture of a $N(0, 1)$ and a $N(0, 9)$ with mixing probabilities 0.9 and 0.1, respectively. (This should give occasional outliers.)