1. Comparing two populations

The file scores contains the results of an experiment to test whether directed reading activities in the classroom help elementary school students improve aspects of their reading ability. A treatment class of 21 third-grade students participated in these activities for eight weeks, and a control class of 23 third-graders followed the same curriculum without the activities. After the eight-week period, students in both classes took a Degree of Reading Power (DRP) test which measures the aspects of reading ability that the treatment is designed to improve. There are 44 cases and the variable name Treatment tells whether a student participated in activities (treated) or not (control). The Response is the score on the DRP test.

1. Plot these data; that is, produce boxplots and histograms for each group. Do the responses look different?

2. We are interested in knowing whether these activities help students read better. We shall make use of the permutation test to test whether or not the means of these populations are the same.

(a) Compute the observed mean difference \( \bar{Y}_2 - \bar{Y}_1 \).

(b) Make an histogram of the randomized distribution of \( \bar{Y}_2 - \bar{Y}_1 \) under the null hypothesis. Because there are too many groupings, you may want to use Monte Carlo simulations to compute an approximate sampling distribution of \( \bar{Y}_2 - \bar{Y}_1 \).

(c) Use your (approximate distribution) to test the hypothesis that these special activities help the children.

3. Test the same hypotheses as before but using a two-sample \( t \)-test now. State your assumptions carefully and check whether they are reasonable.

4. Finish your write-up by drawing some conclusions.

2. Random numbers

The file randu contains a matrix of 100 rows and 3 columns. The values stored in this matrix are obtained by the random number generator RANDU. Each number is presumably independent from the others and takes values between zero and one with uniform probability. We wish to investigate the properties of these random numbers.
1. Produce a histogram for each of the three columns. Does the data appear consistent with the uniform probability between 0 and 1?

2. One distribution that is often used to model random variables in the interval \([0, 1]\) is the Beta distribution whose density depends upon two parameters \(a > 0\) and \(b > 0\), and has the following form:

\[
f(x|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1}.
\]

The function \(\Gamma(a)\) is a generalization of the factorial that is defined on real numbers (if \(n\) is an integer, \(\Gamma(n) = (n - 1)!\)) This may look a bit mysterious, but you can get a sense of how this distribution looks like by plotting the values of the density for a series of different parameters. You can do this in R using the command `dbeta(x,a,b)` which returns the value of the beta density with parameters \(a, b\) at the point \(x\). (To plot the density, you may want to create a vector \(x\) of values between 0 and 1 and evaluate `dbeta` on this vector).

Try superimposing the histogram of each column and the beta density that best matches its shape. To do this, make sure that your histograms report relative rather than absolute frequencies (set `probability=T`), and use the function `lines` to superimpose a plot. In your write-up, report the values of the parameters \((a^*, b^*)\) you believe give the best match. [Note: you can do this matching graphically playing with parameter values, or follow any systematic method of your choice. In any case, please document carefully your line of reasoning to receive full credit.]

3. Produce a scatter plot for each pair of columns (you may want to use the function `pairs()`); do the data look independent?

4. Call the first column \(x\), the second \(y\), the third \(z\), and plot \(9x - 6y + z\) versus \(x\). What do you notice?