% Psych 216A: Statistics and data analysis in MATLAB
% Solutions to Homework 2 (Hypothesis testing and correlation)

% 1. For the data provided below, the mean of x is greater than the mean of y.
% Use the bootstrap method to calculate a two-tailed p-value for the null
% hypothesis that the two sets of data actually come from the same distribution.

x = [3 4 2 3 4 3 4];
y = [1 0 2 3 1];
pool = [x y];
dist = bootstrp(10000,@(sample) mean(sample(1:length(x))) - ...
    mean(sample(length(x)+(1:length(y)))),pool');
pval = sum(abs(dist) > abs(mean(x)-mean(y))) / length(dist);
fprintf('pval = %.10f
',pval);

% 2. For the data provided below, compute the correlation between x and y using
% basic MATLAB operations (arithmetic operations, mean, std, sum, length). Then
% use corr to compute the correlation (this should produce the same result).
% Finally, use bootstrapping to compute a 95% confidence interval on the
% correlation.

x = [2.5 2.8 3.4 2 1.2 2.1 1.8 2.8 1.1 2.2 3.2 0.61 3 0.53 2 0.45 2 ...
    1 -0.53 1.2 2.2 2.5 1.9 1.9 2.8 2.8 1.3 3.1 2.7 2.3 0.47 2.1 1.5 3.8 ...
    3 2.8 0.97 2.5 2.3 3.9 0.54 3.6 2.3 2.7 4.4 2.7 0.44 2.1 1.7 2.1].^2;
y = [5.9 6.4 3.7 3.1 5.1 3.6 6.2 5.1 3 4.7 6 1.2 6.1 -0.51 4.4 3 ...
    5.1 2.2 0.14 3.2 7.2 7.4 4 4.5 5.3 2 4.9 3 4.6 3.9 3.5 4 3.6 5.4 ...
    4.5 6.1 2.6 4.9 3.5 4.8 1 6.8 4.2 3.7 7.5 1.7 2.8 1.8 3.6 4.3].^2;
xz = (x - mean(x)) / std(x,1);
yz = (y - mean(y)) / std(y,1);
r = sum(xz .* yz) / length(xz);
fprintf('r = %.10f
',r);
r = corr(x',y');
fprintf('r = %.10f
',r);
dist = bootstrp(10000,@(x0,y0) corr(x0,y0),x',y');
ptiles = prctile(dist,[2.5 97.5]);
fprintf('CI (95%) = [%1.10f %1.10f]
',ptiles);

% 3. The data provided below represent the measurement of two different
% quantities, A and B, on the same group of subjects. Visualize the data
% and then use the bootstrap method to calculate a two-tailed p-value for the
% null hypothesis that there is no difference in the means of A and B.
% (Note that this question goes a little beyond what was covered in lecture.
% The comments in the code below will guide you through the solution.)
% specify the data
A = [2.67 3.87 3.37 1.72 3.4 3.07 2.72 2.46 2.78 4.73 2.42 4.17 3.51 3.31 ...
    4.77 2.96 3 2.54 3.16 3.36];
B = [4.66 5.12 4.06 3.73 3.29 5.59 1.74 4.53 3.63 4.95 4.43 4.41 3.85 4.19 ...
    2.73 4.31 3.31 3.45 3.31 3.61];

% visualize the data
figure; hold on;
scatter(A,B,'r.);
plot([0 7],[0 7],'b-');
axis square;
xlabel('A');
ylabel('B');

% we could treat A and B as if they came from independent subject pools,
% but such a move would reduce statistical power. Instead, let's compute
% the difference between A and B for each individual subject.
difference = B - A;

% let's now compute the mean difference. The result is positive, but
% is the result statistically different from zero?
actualvalue = mean(difference);

% let's pose the null hypothesis that differences between B and A are
% characterized by a probability distribution with mean zero. To test
% this hypothesis, let's create a distribution that is exactly like
% our observed distribution except that we've centered it on zero.
nulldistribution = difference - mean(difference);

% now draw bootstrap samples from this distribution and
% compute the mean of each sample. The result tells us what
% ranges of values are likely under the null hypothesis.
dist = bootstrp(10000,@(sample) mean(sample),nulldistribution);

% the proportion of values in the bootstrap distribution that are more
% extreme than our actually observed value is the p-value.
pval = sum(abs(dist) > abs(actualvalue)) / length(dist);

% visualize the results
figure; hold on;
scatter(A,B,'r.');</codetexthtml