STA 3000, Fall 2014 — Assignment #3

Due December 8. 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. In particular, you should not leave any discussion of this assignment with any written notes or other recordings, nor receive any written or other material from anyone else by other means such as email.

Question 1: [50 marks] Consider the null hypothesis that n observed points, $(x_1, y_1), \ldots, (x_n, y_n)$, are independently drawn from a distribution that is uniform over some convex region of the (x, y) plane. Devise a pure significance test of this hypothesis — ie, a way of computing a p-value from the data for which the p-value will be uniformly distributed over $(0, 1)$ if the null hypothesis is true. Design your test so that it will be sensitive to departures from the null hypothesis in which the data is uniformly distributed over some non-convex region, or non-uniformly distributed over some convex region. You needn't worry about departures from the assumption of independence.

If you can, think of more than one pure significance test, and compare their properties, such as whether or not they are invariant to translating, rotating, and re-scaling the data (including scaling x and y differently). Also, discuss what one would have to do to address this problem using Bayesian inference, and how easy or hard that would be.

Don't worry about computational issues for this problem. It is OK if your significance test doesn't reduce to some simple formula, and it is OK if the only way you can think of computing the p-value by computer is inefficient. (However, there must be *some* way of computing the p-value for a data set, with arbitrarily small error, even if it is inefficient.) I don't expect you to be able to derive quantitative properties of the tests analytically.

Hint: Find a sufficient statistic for the model assumed by the null hypothesis, and then base a test statistic on the conditional distribution of the full data given the sufficient statistic, which does not depend on the unknown parameter (in this case, the convex region), as discussed in lecture, and in Section 3.3 of Cox and Hinkley's *Theoretical Statistics*.

Question 2: [50 marks] Consider an indefinite sequence of binary (0/1) observations, $y_1, y_2, \ldots$. We decide to model these as a realization of a sequence of random variables $Y_1, Y_2, \ldots$ whose joint distribution is defined by a formula for the probability that Y_{n+1} is 1 given values for $Y_1, \ldots, Y_n$ of the following form:

$$P(Y_{n+1} = 1 \,|\, y_1, \ldots, y_n) \;=\; \frac{a_1 b_1 e^{c_1 d(y_1,\ldots,y_n)} \;+\; a_2 b_2 e^{c_2 d(y_1,\ldots,y_n)}}{b_1 e^{c_1 d(y_1,\ldots,y_n)} \;+\; b_2 e^{c_2 d(y_1,\ldots,y_n)}}$$

Here, $d(y_1, \ldots, y_n) = \#\{i : y_i = 1\} - \#\{i : y_i = 0\}$. Note: $P(Y_1 = 1) = (a_1 b_1 + a_2 b_2) \,/\, (b_1 + b_2)$.

A formula of this form gives valid probabilities if the constants a_1 and a_2 are in $[0, 1]$, the constants b_1 and b_2 are positive, and the constants c_1 and c_2 are any real numbers. Find as large a subset of such values for $a_1, a_2, b_1, b_2, c_1, c_2$ as you can that result in the distribution for $Y_1, Y_2, \ldots$ being infinitely exchangeable.

By de Finetti's exchangeability theorem, each $a_1, a_2, b_1, b_2, c_1, c_2$ that results in an exchangeable distribution for $Y_1, Y_2, \ldots$ must correspond to a Bayesian model in which $Y_1, Y_2, \ldots$ are i.i.d. given the value of some parameter, which has some prior distribution. Show what these models are like explicitly for the exchangeable distributions for $Y_1, Y_2, \ldots$ that you found above.