STAT 5372 Assignment 3

(Lectures 9–13)

Mathematical concepts and derivations

1. Recall that for $\theta = T(F) = \mathbb{E}(X)$, the log of the profile nonparametric likelihood ratio statistic (see Lecture 9 on Empirical Likelihood) has the form:

$$\log \mathcal{R}(\theta) = \sup_F \{\log \mathcal{R}(F) | T(F) = \theta\} = \sup_{\{w_i\}} \left\{ \sum_i \log(n w_i) \Big| \theta = \sum_i w_i x_i \right\},$$

with the optimal solution $w_i = n^{-1}[1 + \lambda(x_i - \theta)]^{-1}$, where λ satisfies the equation (*) given in part(a). Show that for this setup we have the following limit distribution:

$$-2 \log \mathcal{R}(\theta) \xrightarrow{d} \chi^2_1.$$

Do this in three parts as follows.

   (a) Take a Taylor series expansion of

$$\frac{1}{n} \sum_i \frac{x_i - \theta}{1 + \lambda(x_i - \theta)} = 0 \qquad (*)$$

   about $\lambda = 0$ to show that $\lambda \approx (\bar{x} - \theta)/S$, where $S = \sum_i (\bar{x} - \theta)^2/n$.

   (b) Use the approximation in (a) to show that $-2 \log \mathcal{R}(\theta) \approx n(\bar{x} - \theta)^2/S$, while making use of the fact that for $x \approx 0$, $\log(1 + x) \approx x - x^2/2$.

   (c) Deduce the final result.

2. Define the "accuracy" of a Monte Carlo approximation to be the standard error of the Monte Carlo (or empirical) ASL divided by the true ASL, when B permutations are used (see Lecture 10 for definitions). That is:

$$\text{accuracy} = SE(\widehat{ASL})/ASL.$$

(Note: the Monte Carlo standard error is defined with respect to the Monte Carlo random permutations, holding the data fixed.)

   (a) How many permutations B are needed to achieve 10% accuracy when ASL = 0.1?

   (b) How many permutations B are needed to achieve 10% accuracy when ASL = 0.01?

3. Consider the vector of ranks $\boldsymbol{R}$ for an i.i.d. sample $X_1, \ldots, X_n$ from a continuous density. If R_i denotes the rank of X_i, show that:

   (a) $\mathbb{E}(R_i) = (n+1)/2$.
   (b) $\mathbb{V}(R_i) = (n-1)(n+1)/12$.
   (a) $\text{Cov}(R_i, R_j) = -(n+1)/12$.

4. Show that the Median Test is the LMPR test of H_0 when X follows a double exponential distribution (the pdf is displayed in Question 6, set $\mu = 0$ and $\beta = 1$). That is, show that the scores are:

$$a(i) \approx \text{sign}\left(i - \frac{n+1}{2}\right),$$

   where the approximation comes from using the Delta Method to approximate the expectation of a function: $\mathbb{E}[g(X)] \approx g(\mathbb{E}X)$.

5. Consider the Mann-Whitney (a.k.a. Wilcoxon rank sum) test statistic T ("Relative Efficiency", slides 8–10, Lecture 12).

   (a) Show that:

$$\frac{T - \mu_N(\theta_0)}{\sigma_N(\theta_0)} \xrightarrow{d} N(0,1), \qquad \text{where} \quad \mu_N(\theta_0) = \frac{N+1}{2m}, \quad \text{and} \quad \sigma_N^2(\theta_0) = \frac{N+1}{12mn}.$$

   (Note: you will need to multiply the asymptotic variance by the "finite population correction factor", $(N-n)/(N-1)$, in order to get this result.)

   (b) Is the test a valid level α test for the two-group comparison problem when the variances of the two groups are different? Why or why not?

6. In "Relative Efficiency" slide 13 (Lecture 12) we stated that the Wilcoxon rank sum test is (asymptotically) 1.5 times as efficient as a two-sample t-test when the data follows a double exponential distribution. Derive this result. Recall that the double exponential has density

$$f(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right).$$

7. Consider a linear rank statistic $T = \sum_i z_i a(r_i)$. Suppose $U = \sum_i z_i b(r_i)$, where $b(r_i) = c_1 + c_2 a(r_i)$ for some constants $c_1 \in \mathbb{R}$ and $c_2 > 0$. Show that the p-value of the test based on T is the same as that based on U; specifically, show that

$$ASL = \mathbb{P}_0(T^* \geq \hat{T}) = \mathbb{P}_0(U^* \geq \hat{U}),$$

   where $T^* = \sum_i z_i a(r_i^*)$ and $U^* = \sum_i z_i b(r_i^*)$ follow their respective null distributions.

Simulation

8. Conduct a simulation comparing the relative power of the Wilcoxon rank sum and t-tests for $N = 6, \ldots, 100$ (you can choose the intervals) with an equal number of observations in each group (i.e., $n = 3$ in each group when $N = 6$, etc.). For any given N, let the true difference between the group means be $\Delta = \sqrt{2/N}$. Conduct two simulations, one in which the true distribution of the data is normal, the other in which it is double exponential. Plot the relative power of the Wilcoxon test with respect to the t-test versus sample size N. Comment on how well the asymptotic results seem to agree with your finite sample results.

Application

9. The dataset `driving.txt` is from a study which examined the driving habits of illegal drug users as compared to non-illegal drug users. The outcome we will look at is following distance (specifically, the average following distance over the duration of the drive). It may be hypothesized that drug users like to engage in risky behavior and follow at closer speeds than other drivers. In the permutation tests of parts (b) and (c), you must use your own R code (which can be adapted from Lecture10.R); do not use a package! Be sure to submit your code.

   (a) Test the null hypothesis that the mean following distance of drug users is the same as that of non-illegal drug users using a t-test.

   (b) Test the null hypothesis that the distribution of following distance is the same in both groups using a permutation test with test statistic:

   $$T = \left| \frac{\sum_i g_i x_{(r_i^*)}}{\sum_i g_i} - \frac{\sum_i (1-g_i) x_{(r_i^*)}}{\sum_i (1-g_i)} \right|,$$

   where g_i is an indicator of group membership (0 or 1).

   (c) Test the same null hypothesis with a permutation test that only compares the absolute difference in medians between the two groups.

   (d) Briefly comment on the results of the three tests.

10. Carry out the tests below for the `driving.txt` data (see Problem 9). For (a)-(c), report the estimated ASL based on the normal approximation as well as the exact ASL. If the exact ASL is computationally infeasible to calculate, report a Monte Carlo approximation for the ASL. Justify your choice of Monte Carlo replications based on Problem 2.

    (a) Test the null hypothesis that the distribution of following distance is the same in both groups using the Wilcoxon rank-sum test.

    (b) Test the null hypothesis that the distribution of following distance is the same in both groups using the van der Waerden test.

    (c) Test the null hypothesis that the distribution of following distance is the same in both groups using the Median test.

    (d) Briefly comment on the results of the three tests in (a)-(c).

    (e) Which of the three tests in Problem 9 are most similar to the rank tests in (a)-(c)? Why?

11. For the dataset `cysticfibrosis.txt` carry out a test of independence (H_2) between the Drug and Placebo groups, using the capabilities of the R package `boot`. Specifically do the following.

    (a) Calculate the ASL from a permutation test with test statistic the usual (Pearson) correlation coefficient.

    (b) Calculate the ASL from a bootstrap test with test statistic the usual (Pearson) correlation coefficient.

    (c) Calculate the ASL from a (either permutation or bootstrap) test with test statistic as given by the van der Waerden test.