STA 621: Nonparametric Statistics
Breheny

Assignment 3
Due: Thursday, October 18

Mathematical concepts and derivations

1. Define the "accuracy" of a Monte Carlo approximation to be the standard error of $\widehat{ASL}$ over the true 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 are needed to achieve 10% accuracy when $ASL = .1$?
   (b) How many permutations are needed to achieve 10% accuracy when $ASL = .01$?
2. Let $x_i \sim f(x - \Delta g_i)$, where g_i is an indicator denoting group membership. Show that
   $$\mathbb{P}_0(\mathbf{r}) + \Delta \frac{\partial}{\partial \Delta} \mathbb{P}_\Delta(\mathbf{r})|_{\Delta=0} = \frac{1}{n!}\{1 + \Delta T(\mathbf{r})\},$$
   where
   $$T(\mathbf{r}) = \sum_i g_{(i)} \mathbb{E}\left\{\frac{-\partial \log f(X_{(i)})}{\partial X_{(i)}}\right\}.$$
   To accomplish this, you will need to interchange differentiation and integration. This cannot always be done – in general, certain regularity conditions regarding f need to hold. Assume that these conditions hold and that interchanging the two is possible.

   Hint: You may wish to consult Section 5.4 of Casella & Berger to refresh your memory concerning joint densities of order statistics.
3. Show that the sign test is the locally most powerful rank test when X follows a double exponential distribution.
4. Show that for the Mann-Whitney test statistic,
   $$\sigma_N(\theta_0) = \sqrt{\frac{N+1}{12mn}},$$
   where $N = m + n$ is the combined sample size of the two groups, and $\sigma_N(\theta_0)$ satisfies
   $$\frac{T - \mu_n(\theta_0)}{\sigma_n(\theta_0)} \xrightarrow{\text{d}} N(0,1).$$
5. At a conference once, I heard someone say that Wilcoxon rank-sum tests are not valid level α tests for the two-group comparison problem when the variances of the two groups are different. Is this correct? Why or why not?

6. In lecture, 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 distribution has density

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

mean μ and variance $2\beta^2$.

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 and c_2. Show that the p-value of the test based on T is the same as that based on U; specifically, show that

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

where the random variables $\hat{T}^*$ and $\hat{U}^*$ follow the null distribution.

Simulation

8. Conduct a simulation comparing the relative power of the Wilcoxon and t-tests for $n = 6, \ldots, 100$ (you can choose the intervals) with an equal number of observations in each group (*i.e.*, 3 in each group, ..., 50 in each group). Use the following progression of Δ values: $\Delta = \sqrt{2/n}$. 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. Comment on how well the asymptotic results seem to agree with your finite-n results.

Application

9. The website has data 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.

   (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 a 0-1 indicator of group membership.

   (c) Test the same null hypothesis using a permutation test with test statistic similar to the above, only which compares the absolute difference of medians of the two groups.

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

10. Refer to the previous problem, which describes a study of the relationship between driving and illegal drug use. 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 the Monte Carlo approximate for the ASL. Justify your choice of Monte Carlo replications based on Problem 1.

    (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 crossover experiment involving cystic fibrosis that was mentioned in lecture, carry out a van der Waerden test "by hand". By this I mean that you may of course use a computer to carry out the calculations and Monte Carlo integration, but you may not use an `R` package that implements the test for you. You may, however, check your answers against the results given by a package to make sure your answer is correct.

    (a) Calculate the normal-approximation ASL.
    (b) Calculate a Monte Carlo ASL.