Exam 1

Steven Chiou

Due date:

1. **Textbook problem 2.1** Listed below are values of survival time in years for 6 males and 6 females from the *WHAS100* study. Right-censored times are denoted by a "+" as a superscript.

   Males: 1.2, 3.4, 5.0^+, 5.1, 6.1, 7.1

   Females: 0.4, 1.2, 4.3, 4.9, 5.0, 5.1^+

   Using these data, compute the following **without a software package**:

   a. (4 pts) The Kaplan-Meier estimate of the survival function for each gender.
   b. (4 pts) Pointwise 95 percent confidence intervals for the survival functions estimated in problem (1a).
   c. (4 pts) Pointwise 95 percent confidence interval estimates of the 50th percentile of the survival time distribution for each gender.
   d. (4 pts) The estimated mean survival time for each gender using all available times, upto 7.1.
   e. (4 pts) A graph of the estimated survival functions for each gender computed in problem (1a) along with the point wise and overall 95 percent limits computed in problem (1b).

2. Use the complete *WHAS100* dataset and gender as the group indicator, compute the log-rank statistic, Q, presented as Equation (9) on page 60 of note 2 with $\omega_i = 1$. Use Q to compute a p-value to test the null hypothesis of $H_o : S_0(t) = S_1(t)$. Do this **without a software package** (6 pts), and verify the calculation with `survdiff` (4 pts).

3. There are many ways to form a basis for survival curve comparison. Here are some:
   a. The numerator term in Q without the square:

$$D_1 = \sum_{i=1}^{D} \{d_{1i} - E(d_{1i})\}.$$

   b. The largest distance between the two curves:

$$D_2 = \max |S_1(t) - S_0(t)|.$$

   c. The difference between the median survival times:

$$D_3 = S_1^{-1}(0.5) - S_0^{-1}(0.5)$$

   d. The difference between the mean survival times:

$$D_4 = \int_0^{t_{(n)}} \{S_1(u) - S_0(u)\}\, du,$$

   where $t_{(n)}$ is the maximum observed survival time.

   Compute each of the above statistic for the $WHAS100$ dataset (5×4 pts).

4. The statistics computed in (3) do not provide meaningful interpretations when standing along. We will use a permutation approach to test for the null hypothesis of $H_o : S_0(t) = S_1(t)$ based on these statistics. The idea of a permutation test is simple. The general procedure can be summarized into the following steps:
   i. Compute the desired statistic based on the observed data; we will call this the observed statistic.
   ii. Permute the data under the null.
   iii. Compute the statistics for each possible permutation in Step ii.; we will call these permutation statistics.
   iv. Draw conclusion based on where the observed statistic stands among the permutation statistics.

   The statistics we computed in (3) are the observed statistics in Step i. If the null hypothesis of $H_o : S_0(t) = S_1(t)$ is true, then one can randomly shuffle the group indicator to generate different permutations (Step ii) and the statistics for these permutations should be similar (Step iii).

   a. (5×4 pts) Generate 5000 permutation and, for each of the permutation, compute the four statistics presented in (3). We will call the permutated statistics D^*_{1i}, D^*_{2i}, D^*_{3i} and D^*_{4i} for $i = 1, \ldots, 5000$.
   b. (5×4 pts) Compute the p-value based on these statistics by

$$p = 2 \cdot \frac{\min(N_1, N_2)}{5000},$$

   where $N_1 = \#\{D \geq D^*\}$, $N_2 = \#\{D \leq D^*\}$, and # means the "number of", e.g., N_1 is the number of these permutated statistics less than or equal to the observed statistic.

5. Another method to compare two survival curves is to consider a sign test. Suppose we have two groups of uncensored survival times:

   Males: $x_1, x_2, \ldots, x_{n_0}$.

   Females: $y_1, y_2, \ldots, y_{n_1}$.

   The sign test looks at the statistic

$$U = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \operatorname{sgn}(x_i - y_j),$$

   where $\operatorname{sgn}(\cdot)$ is the sign function. In the presense of right censoring, survival times can not be compared directly and a modified version of $U = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} U_{ij}$ is considered, where

$$U_{ij} = \begin{cases} 1 & \text{if } x_i > y_j, y_j \text{ is uncersored.} \\ -1 & \text{if } x_i < y_j, x_i \text{ is uncersored.} \\ 0 & \text{otherwise.} \end{cases}$$

   a. (5 pts) Compute U for the WHAS100 dataset.
   b. (5 pts) Obtain a permutation p-value based on 5000 permutations.