Bootstrap Methods
Assignment 1 - Due February 9^{th} at 1:10pm

Instructions: Submit your code (all algorithms and functions defined) and only relevant output.

Policy: Assignments that are submitted within 24 hours after they are due have the grade reduced by twenty per cent and a further twenty percent for each day thereafter.

1. Complete question "6.10 a" from the text

2. Suppose $X_1, \ldots, X_n$ are IID random variables with distribution function F. Denote the empirical distribution function as F_n and recall that if $X^\star \sim F_n$ then

$$F_n(x) = P_{F_n}(X^\star < x) = \frac{\#X_i's < x}{n} = \frac{1}{n}\sum_1^n I\{X_i < x\}$$

.

- Derive the mean and variance of $F_n(x)$.
- Prove: $F_n(x) \longrightarrow F(x)$ as $n \to \infty$

3. Under resampling with replacement from a random sample $X_1, \ldots, X_n$ in which $T = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ has observed value t, show that

$$E_{F_n}[T^\star] = (n-1)t/n$$

4. There are $(2n-1)$ choose $(n-1)$ ways that $n-1$ red balls can be put in a line with n white balls. Explain the connection to the number of distinct bootstrap samples where the original sample consists of n distinct values $X_1, \ldots, X_n$.

5. The AML data give the remission time, in weeks, of patients with acute myelogeneous leukamia (AML) who are receiving maintenance chemotherapy. The data, X_i $i = 1, \ldots, n$, are {9, 13, **13**, 18, 23, **28**, 31, 34, **45**, 48,**161**} where the bold faced numbers indicate that the time was right censored, that is, the full time of remission was not observed because the patient was lost to follow up for some reason. We assume the remission times and censoring times are independent. For the i^{th} individual we denote the remission time by R_i and the censoring time by C_i where we assume $R_i \sim F$ and $C_i \sim G$ $\forall i$. Note that $X_i = min(R_i, C_i)$. Typically such data are recorded as pairs (X_i, d_i) where d_i is an indicator for the event $R_i < C_i$.

- For such data F is typically estimated by the product limit estimator $\hat{F}$ where

$$1 - \hat{F}(x) = \prod_{i:X_i \le x} \left(\frac{n-i}{n+1-i}\right)^{d_i}.$$

  For the above data tabulate and plot $\hat{F}$. Show that, in general, if $d_i = 1 \forall\ i$ the product limit estimator is simply the usual empirical distribution function.

- Similarly G may be estimated by $\hat{G}$ where

$$1 - \hat{G}(x) = \prod_{i:X_i \le x} \left(\frac{n-i}{n+1-i}\right)^{1-d_i}.$$

Consider the following bootstrap scheme where we sample n times from $\hat{F}$ to get $R_1^\star, \ldots, R_n^\star$ and n times from $\hat{G}$ to get $C_1^\star, \ldots, C_n^\star$. How would we compute $X_1^\star, \ldots, X_n^\star$? Show that

$$Pr(X_i^\star > x) = \prod_{i:X_i \leq x} \left(\frac{n-i}{n+1-i} \right).$$

- Implement the above resampling scheme and give the estimated bootstrap distribution for $\theta^\star$ when the parameter of interest is $\theta = Pr(R_i > 20 \text{ weeks})$.
- Finally, suppose we decide to simply resample the original data (X_i, d_i) with replacement. Show that this bootstrap resampling scheme is equivalent to the one suggested above.