Exercise 3

1. Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution F, having finite second moment.

a. Find the Bootstrap estimate (not the Jackknife estimate) of $\mathrm{Var}_F(X)$.

b. Describe an algorithm (computer program) to estimate the Bootstrap estimate of the variance of the α^{th} trimmed mean.

2*. For $n = 2m - 1$ odd and $Y_1, \ldots, Y_n$ any n real numbers, define the sample median as the middle order statistic

$$\hat{\theta} = Y_{(m)};$$

note that this definition applies even when the n values are not distinct. Now let $X_1, \ldots, X_n$ be a sample from a continuous distribution F, and denote its sample median by $\hat{\theta} = X_{(m)}$. Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample, and $\hat{\theta}^* = X_{(m)}^*$ the bootstrap median.

a. For $j = 1, \ldots, n$, let

$$M_j^* = \#\{i : X_i^* = X_{(j)}\},$$

the number of times the bootstrap sample contains the j^{th} order statistic. Show that M_j^* has a binomial distribution, and determine its parameters.

b. Argue that

$$\{X_{(m)}^* > X_{(k)}\} = \{\sum_{i=1}^{k} M_i^* \leq m - 1\}.$$

c. Determine, for all $k = 1, \ldots, n$,

$$P^*(\widehat{\theta}^* > X_{(k)}) \quad \text{and} \quad P^*(\widehat{\theta}^* = X_{(k)}).$$

d. Give an explicit formula for the variance of the bootstrap median, $\mathrm{Var}^*(\widehat{\theta}^*)$.

3.**: The Efron-Stein inequality for the Jackknife. Let $X_1, \ldots, X_n$ be iid and let θ be any function of $n-1$ variables having a second moment. Set

$$\theta_{(i)} = \theta(X_1, X_{i-1}, X_{i+1}, \ldots, X_n) \quad \text{and} \quad \theta_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \theta_{(i)}.$$

Our goal is to prove:

$$\mathrm{Var}(\theta(X_1, \ldots, X_{n-1})) \leq E\sum_{i=1}^{n}(\theta_{(i)} - \theta_{(\cdot)})^2. \tag{1}$$

Inequality (1) states that the Jackknife estimate of variance without the factor $(n-1)/n$, which is the term being averaged on the right hand side of (1), is on average an overestimate of the true variance. In different words, inequality (1) says that this variance estimate is 'conservative.'

For the following computation you may want to apply the identity

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i<j}(x_i - x_j)^2, \tag{2}$$

which holds for all real numbers $x_1, \ldots, x_n$. However, if you choose to use (2) then you must supply a proof.

The proof of (1) divided into a sequence of steps.

1. Define $c_k = E[\{E\theta(X_1, \ldots, X_{n-1}) \mid X_1, \ldots, X_k\}^2]$.

   Show that $E\theta_{(i)}^2 = c_{n-1}$, and does not depend on i, and that $E(\theta_{(i)}\theta_{(j)}) = E\{E[\theta(X_1, \ldots, X_{n-1})\theta(X_2, \ldots, X_n) \mid X_2, \ldots, X_{n-1})]\} = c_{n-2}$.

2. Show that the right-hand side of (1) contains only the terms c_{n-2} and c_{n-1}, and counting the numbers of terms carefully show that it equals $(n-1)(c_{n-1} - c_{n-2})$.

3. Show that the left-hand side of (1) equals $c_{n-1} - c_0$.

4. Using properties of conditional expectation, and the Cauchy-Schwarz inequality taking expectations with respect to the variable X_1, show that for any k, $c_{k+1} - c_k \geq c_k - c_{k-1}$ and explain the relation to convexity.

5. Using the convexity property of c_k shown in 4, show that $c_{n-1} - c_0 \leq (n-1)(c_{n-1} - c_{n-2})$, thus proving (1).