MATH562 – Fall 2025

Problem Set: Week 14

1. Suppose that estimator T for parameter $\theta \in \mathbb{R}$ has expectation equal to $\theta(1+\gamma)$, so that the bias is $\theta\gamma$. The bias factor γ can be estimated by $C = \mathrm{E}^*(T^*)/T - 1$, where E^* denotes expectation over bootstrap sampling and T^* is based on the bootstrap samples.

   (a) Argue that the variance estimate $T = n^{-1}\sum(Y_j - \bar{Y})^2$ is such an estimator, and show that in this case C is exactly equal to γ. **Solution**: In this case θ is the population variance and the standard calculation (check it if unsure)

   $$\mathrm{E}\left\{\sum_{j=1}^{n}(Y_j-\bar{Y})^2\right\} = \mathrm{E}\left[\sum_{j=1}^{n}\{(Y_j-\mu)-(\bar{Y}-\mu)\}^2\right] = \mathrm{E}\left\{\sum_{j=1}^{n}(Y_j-\mu)^2\right\} - n\,\mathrm{E}\left\{(\bar{Y}-\mu)^2\right\} = n\theta - n\theta/n$$

   where $\mu = \mathrm{E}(Y_j)$ gives

   $$\mathrm{E}(T) = \mathrm{E}\left\{n^{-1}\sum(Y_j-\bar{Y})^2\right\} = \frac{(n-1)}{n}\theta = \theta - \theta/n,$$

   so $\gamma = -1/n$. Note that the only assumptions about the Y_j are they are independent with mean μ and variance θ.

   Now $T^* = n^{-1}\sum(Y_j^* - \bar{Y}^*)^2$, where the $Y_j^* i.i.d. \{y_1, \ldots, y_n\}$ with probabilities $1/n$, and this distribution has mean $\bar{y}$ and variance

   $$\frac{1}{n}\sum y_j^2 - \bar{y}^2 = \frac{1}{n}\sum(y_j-\bar{y})^2 = t.$$

   We can apply the computation above to see that T^* is also downwardly-biased as an estimate of its population variance t, with mean

   $$\mathrm{E}^*(T^*) = \frac{(n-1)}{n}t,$$

   so $C = \mathrm{E}^*(T^*)/t - 1 = -1/n = \gamma$, as stated.

   (b) If C were approximated from R resamples by C^*, what would be the simulation variance of C^*? **Solution**: The estimator of C would be $C^* = R^{-1}\sum_{r=1}^{R} T_r^*/t - 1$, which has variance $R^{-1}\mathrm{Var}^*(T^*)/t^2$, because each of the T_r^* has variance $\mathrm{Var}^*(T^*)$ and they are independent.

2. Let T be the median of a random sample of size $n = 2m+1$ with ordered values $y_{(1)} < \cdots < y_{(n)}$; the observed value of T is therefore $t = y_{(m+1)}$.

   (a) Show that $T^* > y_{(l)}$ if and only if fewer than $m+1$ of the Y_j^* are less than or equal to $y_{(l)}$, and deduce that

   $$\Pr{}^*(T^* > y_{(l)}) = \sum_{j=0}^{m}\binom{n}{j}\left(\frac{l}{n}\right)^j\left(1-\frac{l}{n}\right)^{n-j}.$$

   This specifies the exact resampling distribution of the sample median, and can be used to prove that the bootstrap estimate of $\mathrm{Var}(T)$ is consistent as $n \to \infty$. **Solution**: By definition, the median of $Y_1^*, \ldots, Y_n^*$ when $n = 2m+1$ is $Y_{(m+1)}^*$. Hence

   $$T^* > y_{(l)} \quad\Longleftrightarrow\quad Y_{(m+1)}^* > y_{(l)} \quad\Longleftrightarrow\quad Y_{(n)}^*, \ldots, Y_{(m+1)}^* > y_{(l)},$$

which is true if and only at most m of the Y^* are less than or equal to $y_{(l)}$. The probability that a single Y^* is less than or equal to $y_{(l)}$ is $p = l/n$, and as the Y^* are independent, this gives the stated binomial probability, because if we let I_j be the indicator of the event $Y_j^* \leq y_{(l)}$, then

$$\Pr{}^*(T^* > y_{(l)}) = \Pr{}^*\left(\sum_{j=1}^{n} I_j \leq m\right) = \sum_{j=0}^{m} \binom{n}{j} p^j (1-p)^{n-j},$$

as required.

(b) Use the resampling distribution in (a) to show that for $n = 11$,

$$\Pr{}^*(T^* \leq y_{(3)}) = \Pr{}^*(T^* \geq y_{(9)}) = 0.051,$$

and deduce that an approximate basic bootstrap 90% confidence interval for the population median is $(2y_{(6)} - y_{(8)}, 2y_{(6)} - y_{(4)})$. **Solution**: It is easy to check that $\Pr^*(T^* > y_{(8)}) \doteq 1 - \Pr^*(T^* > y_{(3)}) \doteq 0.05$, so $\Pr^*(y_{(4)} \leq T^* \leq y_{(8)}) \doteq 0.9$.

For the bootstrap confidence interval, note that

$$0.9 \doteq \Pr{}^*(y_{(4)} \leq T^* \leq y_{(8)}) = \Pr{}^*(y_{(4)} - t \leq T^* - t \leq y_{(8)} - t),$$

where t is the observed median $y_{(6)}$, so the basic bootstrap argument gives (approximate) 90% confidence interval $(2y_{(6)} - y_{(8)}, 2y_{(6)} - y_{(4)})$.