Advanced Statistical Methods Hw8

Do Hyup Shin

2021-11-23

Problem 10.4

Verify formula (10.38) for the number of distinct bootstrap samples.

Solution

We'll show that the number of distinct bootstrap samples = $\binom{2n-1}{n}$. This problem is a duplicate combination problem. Let $(x_1, x_2, \ldots, x_n)$ be the sample and the size of sample is n. Let the number of times each observation is chosen is $a_i \ \forall i = 1, 2, \ldots, n$. Then, $\sum_{i=1}^n a_i = n$ with $\forall 0 \le a_i \le n$ and $\forall a_i$ are nonnegative integer. We should find the number of combination a_i satisfying above condition. This problem is the same as following problem. Suppose that there exist n-1 bars($= |$) and n dots($= \cdot$). Let's arrange the two types of symbols in a row. Then, we can express the arranged line in this way ___ | ___ | ___ $\ldots$ ___ | ___ | ___ and ___ means where $\cdot$ can enter. There exists n seperation which is ___.

Thus, we can correspond $\forall a_i$ to the number of $\cdot$ in ith ___. We know that the number of permutations n-1 bars($= |$) and n dots($= \cdot$) is $\dfrac{(2n-1)!}{(n-1)!n!} = \binom{2n-1}{n}$.

Therefore, the number of distinct bootstrap samples is $\binom{2n-1}{n}$.

Problem 10.5

A normal theory least squares model (7.28)-(7.30) yields $\hat{\beta}$ (7.32). Describe the parametric bootstrap estimates for the standard errors of the components of $\hat{\beta}$.

Solution

The distribution of $\hat{\beta}$ is $\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$. If we know the σ^2, the standard errors of components of $\hat{\beta}$ are $se(\hat{\beta}_i) = \sigma(e_i^t (X^T X)^{-1} e_i)^{1/2} \quad \forall i = 1, 2, \ldots, p$ where e_i is the standard basis vector with ith element zero. But, if we don't know the σ^2, then we replace $s^2 = MSE = \dfrac{1}{n-p-1} y^t (I - H) y = \dfrac{1}{n-p-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ instead of σ^2. Let the design matrix X be fixed.

Problem 10.7

Verify formula (10.70).

Solution

We'll show that the variance of sample mean of bootstrap sample $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$ is $\sum_{i=1}^n (x_i - \bar{x})^2/n^2$. Let $X = (x_1, x_2, \ldots, x_n)$ be random sample from popuplation F and define $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Let $\hat{F}$ be the empirical probability distribution that puts probability 1/n on each point x_i. The bootstrap sample with replace from $\{x_1, x_2, \ldots, x_n\}$ is $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$, i.e $x_i^* \overset{iid}{\sim} \hat{F}$. Then, $P(x_i^* = x_j) = \frac{1}{n} \quad \forall 1 \le i, j \le n$. So the expectation of x_i^* is $E_{\hat{F}}(x_i^*) = \sum_{j=1}^n x_j P(x_i^* = x_j) = \sum_{j=1}^n x_j \frac{1}{n} = \frac{1}{n}\sum_{j=1}^n x_j = \bar{x}$.

Define $\bar{x}^* = \frac{1}{n}\sum_{j=1}^n x_j^*$ which is sample mean of bootstrap sample. Then, the variance of $\bar{x}^*$ is

$$\begin{aligned}
var_{\hat{F}}(\bar{x}^*) &= E_{\hat{F}}((\bar{x}^* - E(\bar{x}^*))^2) = E_{\hat{F}}((\bar{x}^* - \bar{x})^2) \\
&= E_{\hat{F}}(\sum_{j=1}^n \frac{1}{n}(x_j^* - \bar{x}))^2 = \frac{1}{n^2}E_{\hat{F}}(\sum_{j=1}^n (x_j^* - \bar{x}))^2 \\
&= \frac{1}{n^2}E_{\hat{F}}(\sum_{j=1}^n (x_j^* - \bar{x})^2 + \sum_{i\neq j}(x_i^* - \bar{x})(x_j^* - \bar{x})) \\
&= \frac{1}{n^2}E_{\hat{F}}(\sum_{j=1}^n (x_j^* - \bar{x})^2) \quad (\because E_{\hat{F}}(x_i^* - \bar{x})(x_j^* - \bar{x}) = 0\ \forall i \neq j) \\
&= \frac{1}{n^2}nE_{\hat{F}}(x_1^* - \bar{x})^2 \quad (\because \forall(x_j^* - \bar{x}) \text{ are following independently identical distribution}) \\
&= \frac{1}{n}\sum_{j=1}^n (x_j - \bar{x})^2 P(x_1^* = x_j) = \frac{1}{n}\sum_{j=1}^n \frac{1}{n}(x_j - \bar{x})^2 \\
&= \frac{1}{n^2}\sum_{j=1}^n (x_j - \bar{x})^2
\end{aligned}$$

Therefore, $var_{\hat{F}}(\bar{x}^*) = \frac{1}{n^2}\sum_{j=1}^n (x_j - \bar{x})^2$. Suppose that there exist B bootstrap samples. Then, We can calculate $\bar{x}^{*j}$ for each jth bootstrap sample. So the estimate of $var_{\hat{F}}(\bar{x}^*)$ is $v\hat{a}r_{boot}(\bar{x}^*) = \frac{1}{B-1}\sum_{j=1}^B (\bar{x}^{*j} - \bar{x}_{(\cdot)})^2$ where $\bar{x}_{(\cdot)} = \frac{1}{B}\sum_{j=1}^B \bar{x}^{*j}$.

In conclusion, $v\hat{a}r_{boot}(\bar{x}^*) = \frac{1}{B-1}\sum_{j=1}^B (\bar{x}^{*j} - \bar{x}_{(\cdot)})^2 \to var_{\hat{F}}(\bar{x}^*) = \frac{1}{n^2}\sum_{j=1}^n (x_j - \bar{x})^2$ as $B \to \infty$.

Problem 10.9

A survey in a small town showed incomes $x_1, x_2, \ldots, x_m$ for men and $y_1, y_2, \ldots, y_n$ for women. As an estimate of the differences,

$$\hat{\theta} = median\{x_1, x_2, \ldots, x_m\} - median\{y_1, y_2, \ldots, y_n\}$$

was computed.

(a) How would you use nonparametric bootstrapping to assess the accuracy of $\hat{\theta}$?
(b) Do you think your method makes full use of the bootstrap replications?

Solution

(a)

Let $X = (x_1, x_2, \ldots, x_m)$ and $Y = (y_1, y_2, \ldots, y_n)$ be the samples of men and women, respectively. Some large number B of bootstrap samples are independently drawn. Let $X^{*j} = (x_1^{*j}, x_2^{*j}, \ldots, x_m^{*j})$ and $Y^{*j} = (y_1^{*j}, y_2^{*j}, \ldots, y_n^{*j})$ be the Bth bootstrap sample of X and Y, respectively. The corresponding bootstrap replications are calculated, say $\hat{\theta}^{*j} = median\{x_1^{*j}, x_2^{*j}, \ldots, x_m^{*j}\} - median\{y_1^{*j}, y_2^{*j}, \ldots, y_n^{*j}\}$. Then, the bootstrap estimate of standard error for $\hat{\theta}$ is $\hat{se}_{boot}(\hat{\theta}) = (\frac{1}{B-1}\sum_{i=1}^B (\hat{\theta}^{*i} - \hat{\theta}_{(\cdot)})^2)^{1/2}$. So, we can assess the accuracy of $\hat{\theta}$ by above process.

(b)

Problem 11.1

We observe $y \sim \lambda G_{10}$ to be $y = 20$. Here λ is an unknown parameter while G_{10} represents a gamma random variable with 10 degrees of freedom ($y \sim G(10, \lambda)$ in the notation of Table 5.1). Apply the Neyman constructions as in Figure 11.1 to find the confidence limit endpoints $\hat{\lambda}(0.025)$ and $\hat{\lambda}(0.975)$.

Solution

The pdf of y is $f_\lambda(y) = \frac{1}{\Gamma(10)\lambda^{10}} y^9 e^{-\frac{y}{\lambda}}$. The loglikelihood function is $l(\lambda) = log(f_\lambda(y)) = -log9! - 10log\lambda + 9logy - \frac{y}{\lambda}$. We can find the mle of λ satisfying $\frac{\partial l}{\partial \lambda} = -\frac{10}{\lambda} + \frac{y}{\lambda^2} = 0$. Then, the mle of λ is $\hat{\lambda} = \frac{y}{10} = 2$. Since $y \sim G(10, \lambda)$, $\hat{\lambda} = \frac{y}{10} \sim G(10, \frac{\lambda}{10})$.

We'll show that $\lambda_1 \leq \lambda_2 \Rightarrow P_{\lambda_2}(\hat{\lambda} \leq r) \leq P_{\lambda_1}(\hat{\lambda} \leq r)$. The cdf of $\hat{\lambda}$ is

$$
\begin{aligned}
F_\lambda(r) &= \int_0^r \frac{1}{9!(\lambda/10)^{10}} x^9 e^{-\frac{10x}{\lambda}} dx \\
&= \int_0^{10r/\lambda} \frac{1}{9!} t^9 e^{-t} dx \quad (10x/\lambda = t, dx = \lambda/10dt) \\
&= \frac{1}{9!} \int_0^{10r/\lambda} t^9 e^{-t} dx
\end{aligned}
$$

Then, $F_\lambda(r)$ is decreasing function of λ because the integral interval $(0, 10r/\lambda)$ is reduced when λ is incresing. Thus, $\lambda_1 \leq \lambda_2 \Rightarrow F_{\lambda_2}(r) \leq F_{\lambda_1}(r)$. Define the function of α-quantile of $\hat{\lambda}$ for λ denoted $g_\alpha(f_\lambda)$ satisfying $P_\lambda(\hat{\lambda} \leq g_{\frac{\alpha}{2}}(f_\lambda)) = \frac{\alpha}{2}$. Then, $g_\alpha(f_\lambda)$ is increaing function for λ.

So, we'll find $\hat{\lambda}_{(up)}$ and $\hat{\lambda}_{(lo)}$ such that $g_{0.025}(f_{\hat{\lambda}_{(up)}}) = \hat{\lambda}$ and $g_{0.975}(f_{\hat{\lambda}_{(lo)}}) = \hat{\lambda}$. This means $P_{\hat{\lambda}_{lo}}(X \geq \hat{\lambda}) = \int_{\hat{\lambda}}^\infty f_{\hat{\lambda}_{lo}}(x)dx = 0.025$ and $P_{\hat{\lambda}_{up}}(X \leq \hat{\lambda}) = \int_0^{\hat{\lambda}} f_{\hat{\lambda}_{up}}(x)dx = 0.025$ where $f_\lambda(x)$ is the pdf of $\hat{\lambda}$.

We know that $\hat{\lambda} \sim G(10, \frac{\lambda}{10}) \Leftrightarrow \frac{20}{\lambda}\hat{\lambda} \sim G(10, 2) = \chi^2(20)$. Using this, $\frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda} \sim \chi^2(20)$ and $\frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda} \sim \chi^2(20)$ respectively. Then,

$$
\begin{aligned}
P_{\hat{\lambda}_{lo}}(X \geq \hat{\lambda}) &= P_{\hat{\lambda}_{lo}}(\frac{20}{\hat{\lambda}_{(lo)}} X \geq \frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda}) \\
&= P(Y \geq \frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda}) \quad (Y = \frac{20}{\hat{\lambda}_{(lo)}} X \sim \chi^2(20)) \\
&= 0.025
\end{aligned}
$$

So, $\frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda} = \chi^2_{0.025}(20) \rightarrow \hat{\lambda}_{(lo)} = \frac{20\hat{\lambda}}{\chi^2_{0.025}(20)} = \frac{40}{\chi^2_{0.025}(20)}$

```r
#mle of lambda
hat_lambda = 2

#the value of lambda_lo
hat_lam_lo = 20*hat_lambda/qchisq(0.025, 20, lower.tail = F)
hat_lam_lo
```

```
## [1] 1.170631
```

$\hat{\lambda}_{(lo)} = \frac{40}{\chi^2_{0.025}(20)} = 1.170631$.

Similarly,

$$
\begin{aligned}
P_{\hat{\lambda}_{up}}(X \leq \hat{\lambda}) &= P_{\hat{\lambda}_{up}}(\frac{20}{\hat{\lambda}_{(up)}} X \leq \frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda}) \\
&= P(Y \leq \frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda}) \quad (Y = \frac{20}{\hat{\lambda}_{(up)}} X \sim \chi^2(20)) \\
&= 0.025
\end{aligned}
$$

Thus, $\frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda} = \chi^2_{0.975}(20) \rightarrow \hat{\lambda}_{(up)} = \frac{20\hat{\lambda}}{\chi^2_{0.975}(20)} = \frac{40}{\chi^2_{0.975}(20)}$.

```
#the value of lambda_up
hat_lam_up = 20*hat_lambda/qchisq(0.975, 20, lower.tail = F)
hat_lam_up
```

```
## [1] 4.170673
```

$\hat{\lambda}_{(up)} = \frac{40}{\chi^2_{0.975}(20)} = 4.170673$.

Therefore, $\hat{\lambda}(0.025) = 1.170631$ and $\hat{\lambda}(0.975) = 4.170673$.

The 95% confidence interval of Neyman constructions is (1.170631, 4.170673).

Problem 11.3

Suppose $\hat{G}$ in (11.33) was perfectly normal, say $\hat{G} \sim N(\hat{\mu}, \hat{\sigma}^2)$. What does $\hat{\theta}_{BC}(\alpha)$ reduce to in this case, and why does this make intuitive sense?

Solution

Suppose that $\hat{G}$ is cdf of $N(\hat{\mu}, \hat{\sigma}^2)$. Then,

$$\hat{\theta}_{BC}[\alpha] = \hat{G}^{-1}[\Phi(2z_0 + z^{(\alpha)})] = 2z_0 + z^{(\alpha)}$$

where $z_0 = \Phi^{-1}(p_0)$ and $z^{(\alpha)} = \Phi^{-1}(\alpha)$.

Problem 11.5

Suppose $\hat{\theta} \sim Poisson(\theta)$ is obeserved to equal 16. Without employing simulation, compute the 95% central BCa interval for θ. (You can use the good approximation $z_0 = a = 1/(6\hat{\theta}^{1/2})$.)

Solution

Problem 11.6

Use the R program bcajack (available with its help file from efron.web.stanford.edu under "Talks") to find BCa confidence limits for the student score eigenratio statistic as in Figure 10.2.

Solution