

Μη παραμετρική συμπερασματολογία με μεθόδους επαναδειγματοληψίας: το jackknife

Μη Παραμετρική Στατιστική

Παναγιώτης Παπασταμούλης
Επίκουρος Καθηγητής
Τμήμα Στατιστικής ΟΠΑ

papastamoulis@aueb.gr

Παρασκευή, 10/04/2020



Μέθοδοι επαναδειγματοληψίας

- Το jackknife (Quenouille 1949, Tukey 1958) και το bootstrap (Efron, 1979) είναι μη παραμετρικές μέθοδοι για
 - ▶ εκτίμηση τυπικών σφαλμάτων συναρτησιακών
 - ▶ υπολογισμό διαστημάτων εμπιστοσύνης συναρτησιακών
- Υπολογιστικά, το jackknife είναι λιγότερο απαιτητικό, αλλά το bootstrap έχει περισσότερα πλεονεκτήματα, πχ:
 - ▶ εκτίμηση της δειγματικής κατανομής εκτιμητών συναρτησιακών

To Jackknife



To Jackknife

Η μέθοδος Jackknife προσφέρει

- Εκτιμητές με μικρότερη μεροληψία σε απόλυτη τιμή
- Εύκολο υπολογισμό του τυπικού σφάλματος του εκτιμητή

Συμβολισμοί-ορισμοί

- Έστω $T_n = T(X_1, \dots, X_n)$ εκτιμητής μιας ποσότητας θ : $T_n = \hat{\theta}_n$.
- Η *μεροληψία* του εκτιμητή T_n ορίζεται ως

$$\text{bias}(T_n) = ET_n - \theta, \quad \theta \in \Theta$$

- Ορίζουμε ως

$$T(-i) = T(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n), \quad i = 1, \dots, n$$

τον εκτιμητή που προκύπτει αν **αφαιρέσουμε** την i παρατήρηση στο δείγμα μας.

Το Jackknife

Η βασική ιδέα του Jackknife:

- Αφαιρούμε παρατηρήσεις από το αρχικό δείγμα και εκτιμούμε ξανά την παράμετρο που μας ενδιαφέρει.
- Αυτό παρέχει πληροφορία για τη μεταβλητότητα του εκτιμητή.
- Επομένως αν αφαιρούμε κάθε φορά από μία παρατήρηση και εξετάζοντας πόσο αλλάζουν οι τιμές του εκτιμητή παίρνουμε μια εικόνα για τη διασπορά του εκτιμητή.

Ορισμός (Εκτιμητής Jackknife)

Ο **εκτιμητής Jackknife** ορίζεται ως

$$T_{\text{jack}} := T_n - b_{\text{jack}} \quad (1)$$

όπου με b_{jack} ορίζουμε την **jackknife εκτίμηση της μεροληψίας**

$$b_{\text{jack}} := (n - 1) (\bar{T}_n - T_n) \quad (2)$$

και

$$\bar{T}_n = \frac{1}{n} \sum_{i=1}^n T(-i). \quad (3)$$

Εναλλακτικές εκφράσεις:

- $T_{\text{jack}} = nT_n - (n - 1)\bar{T}_n$.
- $T_{\text{jack}} = \frac{1}{n} \sum_{i=1}^n \tilde{T}_i$, όπου οι **ψευδοτιμές** \tilde{T}_i ορίζονται ως

$$\tilde{T}_i := nT_n - (n - 1)T(-i), \quad i = 1, \dots, n.$$

Παράδειγμα 1

- Έστω τυχαίο δείγμα: 4, 3, 7, 6, 5, 9 από πληθυσμό με κατανομή F .
Να υπολογιστούν οι εκτιμητές jackknife για τα συναρτησιακά $T_1(F) = \mu$ και $T_2(F) = \sigma^2$.
- Ξέρουμε ότι οι εκτιμητές αντικατάστασης είναι $\hat{\mu} = T_1(\hat{F}_n) = \bar{X}_n$
 $\hat{\sigma}^2 = T_2(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, αντίστοιχα.
- Προκύπτουν οι εκτιμήσεις $\hat{\mu} = 5.67$ και $\hat{\sigma}^2 = 3.89$
- Jackknife εκτίμηση των μ και σ^2

| Παρατηρήσεις | | Μέσος | | Διασπορά | |
|--------------|-------|---------|---------------|----------|---------------|
| i | x_i | $T(-i)$ | \tilde{T}_i | $T(-i)$ | \tilde{T}_i |
| 1 | 4.00 | 6.00 | 4.00 | 4.00 | 3.33 |
| 2 | 3.00 | 6.20 | 3.00 | 2.96 | 8.53 |
| 3 | 7.00 | 5.40 | 7.00 | 4.24 | 2.13 |
| 4 | 6.00 | 5.60 | 6.00 | 4.64 | 0.13 |
| 5 | 5.00 | 5.80 | 5.00 | 4.56 | 0.53 |
| 6 | 9.00 | 5.00 | 9.00 | 2.00 | 13.33 |

Παράδειγμα 1

- Για τη μέση τιμή

- ▶ Αρχικός εκτιμητής $T_n = \bar{X}_n = 5.67$
- ▶ Εκτιμητής jackknife

$$T_{\text{jack}} = \frac{1}{n} \sum_{i=1}^n \tilde{T}_i = 5.67$$

- ▶ Παρατηρούμε ότι οι δύο εκτιμήσεις συμπίπτουν.

- Για τη διασπορά

- ▶ Αρχικός εκτιμητής $T_n = \hat{\sigma}^2 = 3.89$
- ▶ Εκτιμητής jackknife

$$T_{\text{jack}} = \frac{1}{n} \sum_{i=1}^n \tilde{T}_i = 4.67$$

- ▶ Παρατηρούμε ότι οι δύο εκτιμήσεις διαφέρουν.
- ▶ Η δειγματική διασπορά είναι $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = 4.67$
- ▶ Δηλαδή: η jackknife εκτίμηση αφαιρέσει τη μεροληψία του $\hat{\sigma}^2$ και κατέληξε στον αμερόληπτο εκτιμητή S_n^2 .

Εκτίμηση τυπικών σφαλμάτων

- Αμερόληπτος εκτιμητής για τη διασπορά των ψευδοτιμών

$$\tilde{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\tilde{T}_i - \frac{1}{n} \sum_{j=1}^n \tilde{T}_j \right)^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\tilde{T}_i - T_{\text{jack}} \right)^2.$$

- Η jackknife εκτίμηση της διασποράς του T_n είναι

$$v_{\text{jack}} = \frac{\tilde{s}_n^2}{n} \quad (4)$$

- Κάτω από συνθήκες ομαλότητας στον T , μπορεί να δειχθεί ότι ο v_{jack} είναι συνεπής εκτιμητής της $\text{Var}(T_n)$.
- Η v_{jack} δεν είναι συνεπής για μη «ομαλά» στατιστικά, όπως τα δειγματικά ποσοστιαία σημεία $T_n = \tilde{F}_n^{-1}(p)$.

Παράδειγμα: εκτίμηση τυπικού σφάλματος δειγματικού μέσου

- Γνωρίζουμε ότι $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$
- που εκτιμάται με $\widehat{\text{Var}}(\bar{X}_n) = \frac{S_n^2}{n}$
- Η εκτίμηση του τυπικού σφάλματος του δειγματικού μέσου:
$$\widehat{\text{se}} = \sqrt{\frac{S_n^2}{n}}$$
- Στο προηγούμενο παράδειγμα έχουμε $\widehat{\text{se}} = \sqrt{0.78}$
- Μέσω του jackknife εκτιμούμε τη διασπορά του δειγματικού μέσου από την εξίσωση (4).

- ▶ $\tilde{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\tilde{T}_i - T_{\text{jack}} \right)^2 = 4.67$

- ▶ Άρα

$$v_{\text{jack}} = \frac{\tilde{s}_n^2}{n} = 0.78$$

- ▶ οπότε $\widehat{\text{se}}_{\text{jack}} = \sqrt{v_{\text{jack}}} = \sqrt{0.78}$

Μέση τιμή

- Είδαμε στο παράδειγμα ότι η jackknife εκτίμηση για τη μέση τιμή συμπίπτει με τον δειγματικό μέσο.
- Συμβαίνει πάντα αυτό;
- Οι ψευδοτιμές είναι

$$\begin{aligned}\tilde{T}_i &= nT_n - (n-1)T(-i) \\ &= n\frac{1}{n}\sum_{i=1}^n X_i - (n-1)\frac{1}{n-1}\sum_{j \neq i}^n X_j \\ &= \sum_{i=1}^n X_i - \left(\sum_{j=1}^n X_j - X_i \right) \\ &= X_i\end{aligned}$$

- $T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^n \tilde{T}_i = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$
- Άρα δεν υπάρχει λόγος να χρησιμοποιήσουμε jackknife για να εκτιμήσουμε μία μέση τιμή.

Διασπορά

- Η jackknife εκτίμηση της διασποράς είναι

$$T_{\text{jack}} = S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

που είναι αμερόληπτη.

- Hint για την απόδειξη:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n^2} \left\{ (n-1) \sum_{i=1}^n X_i^2 - \sum_{i=1}^n \sum_{j \neq i} X_i X_j \right\} \end{aligned}$$

Παράδειγμα

- Προσομοιώστε $n = 30$ παρατηρήσεις από την $\mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right)$ και υπολογίστε την *jackknife* εκτίμηση του συντελεστή συσχέτισης καθώς και το τυπικό σφάλμα της εκτίμησης.
- Δες `jackknife.R`

Πρόταση (Ιδιότητες εκτιμητή jackknife)

- ❶ Αν T_n είναι αμερόληπτος, το ίδιο ισχύει και για τον T_{jack} :

$$ET_n = \theta, \quad \forall \theta \in \Theta \quad \Rightarrow \quad ET_{\text{jack}} = \theta, \quad \forall \theta \in \Theta.$$

- ❷ Έστω ότι η μεροληψία του T_n γράφεται ως

$$\text{bias}(T_n) = \frac{\alpha_1(\theta)}{n} + \frac{\alpha_2(\theta)}{n^2} + \dots \quad (5)$$

όπου $\alpha_1, \alpha_2, \dots$ συναρτήσεις που δεν εξαρτώνται από το n . Τότε

- 2.1 Η jackknife εκτίμηση της μεροληψίας εκτιμά την μεροληψία του T_n με ακρίβεια της τάξης $1/n^2$.
- 2.2 Η μεροληψία του T_{jack} είναι μικρότερης τάξης της μεροληψίας του T_n κατά παράγοντα $1/n$.

Απόδειξη ιδιότητας 1

- Έστω ότι ο T_n είναι αμερόληπτος: $ET_n = \theta, \forall \theta \in \Theta$.
- Το ίδιο θα ισχύει και για τους $T(-i), i = 1, \dots, n$:

$$ET(-i) = \theta, \quad \forall \theta \in \Theta.$$

- Συνεπώς

$$\begin{aligned} ET_{\text{jack}} &= E \{ nT_n - (n-1)\bar{T}_n \} \\ &= nET_n - (n-1)E\bar{T}_n \\ &= n\theta - (n-1)\theta \\ &= \theta, \quad \theta \in \Theta. \end{aligned}$$

- Δείξαμε ότι και ο T_{jack} θα είναι αμερόληπτος.

Απόδειξη ιδιοτήτων 2

- Από την (5) έχουμε

$$\text{bias}(T_n) = \frac{\alpha_1(\theta)}{n} + \frac{\alpha_2(\theta)}{n^2} + O\left(\frac{1}{n^3}\right)$$

όπου $O\left(\frac{1}{n^3}\right)$ είναι μία ποσότητα η οποία είναι $\leq M \frac{1}{n^3}$ για κάποια σταθερά M .

- Η μεροληψία των $T(-i)$ γράφεται ως

$$\text{bias}(T(-i)) = \frac{\alpha_1(\theta)}{n-1} + \frac{\alpha_2(\theta)}{(n-1)^2} + O\left(\frac{1}{n^3}\right)$$

διότι $O(1/(n-1)^3) = O(1/n^3)$.

- Άρα

$$\text{bias}(\bar{T}_n) = \text{bias}\left(\frac{1}{n} \sum_{i=1}^n T(-i)\right) = \frac{\alpha_1(\theta)}{n-1} + \frac{\alpha_2(\theta)}{(n-1)^2} + O\left(\frac{1}{n^3}\right)$$

Απόδειξη ιδιοτήτων 2

- Συνεπώς η μέση τιμή της jackknife εκτίμησης της μεροληψίας είναι

$$\begin{aligned}Eb_{\text{jack}} &= (n-1) \{E\bar{T}_n - ET_n\} \\ &= (n-1) \{(E\bar{T}_n - \theta) - (ET_n - \theta)\} \\ &= (n-1) \{\text{bias}(\bar{T}_n) - \text{bias}(T_n)\} \\ &= (n-1) \left\{ \left(\frac{1}{n-1} - \frac{1}{n} \right) \alpha_1(\theta) + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2} \right) \alpha_2(\theta) + O\left(\frac{1}{n^3}\right) \right\} \\ &= \frac{\alpha_1(\theta)}{n} + \frac{(2n-1)\alpha_2(\theta)}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\ &= \frac{\alpha_1(\theta)}{n} + O\left(\frac{1}{n^2}\right) \\ &= \text{bias}(T_n) + O\left(\frac{1}{n^2}\right)\end{aligned}$$

Απόδειξη ιδιοτήτων 2

- Άρα το b_{jack} εκτιμά την μεροληψία του T_n με ακρίβεια της τάξης του $1/n^2$.
- Παρόμοια μπορούμε να δείξουμε ότι

$$\text{bias}(T_{\text{jack}}) = -\frac{\alpha_2(\theta)}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right)$$

- που σημαίνει ότι η μεροληψία του T_{jack} είναι μία τάξης μεγέθους μικρότερη της μεροληψίας του αρχικού εκτιμητή T_n .