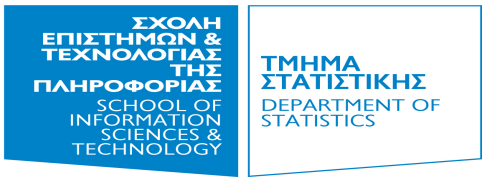
Μη παραμετρική συμπερασματολογία με μεθόδους επαναδειγματοληψίας: το jackknife

Μη Παραμετρική Στατιστική

Παναγιώτης Παπασταμούλης
Αναπληρωτής Καθηγητής
Τμήμα Στατιστικής ΟΠΑ

papastamoulis@aueb.gr

2026



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

Το Jackknife

Το Jackknife

Η μέθοδος Jackknife προσφέρει

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

Συμβολισμοί-ορισμοί

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

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

- Ορίζουμε ως

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

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*

$$T_{\text{jack}} := T_n - b_{\text{jack}} \tag{1}$$

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

$$b_{\text{jack}} := (n-1)\left(\overline{T}_n - T_n\right) \tag{2}$$

και

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

Εναλλακτικές εκφράσεις:

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

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

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 $\widehat{\mu} = T_1(\widehat{F}_n) = \overline{X}_n$ $\widehat{\sigma}^2 = T_2(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, αντίστοιχα.
- Προκύπτουν οι εκτιμήσεις $\widehat{\mu} = 5.67$ και $\widehat{\sigma}^2 = 3.89$
- Jackknife εκτίμηση των μ και σ^2

Παρατηρήσεις		Μέσος		Διασπορά	
i	x_i	$T(-i)$	$\widetilde{T}_i$	$T(-i)$	$\widetilde{T}_i$
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 $T_n = \overline{X}_n = 5.67$
  - Εκτιμητής jackknife

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

  - Παρατηρούμε ότι οι δύο εκτιμήσεις συμπίπτουν.
- Για τη διασπορά
  - Αρχικός εκτιμητής $T_n = \widehat{\sigma}^2 = 3.89$
  - Εκτιμητής jackknife

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

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

Εκτίμηση τυπικών σφαλμάτων

- Αμερόληπτος εκτιμητής για τη διασπορά των ψευδοτιμών

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

- Η jackknife εκτίμηση της διασποράς του T_n είναι

$$v_{\text{jack}} = \frac{\widetilde{s}_n^2}{n} \tag{4}$$

- 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 $T_n = \widetilde{F}_n^{-1}(p)$.

Παράδειγμα: εκτίμηση τυπικού σφάλματος δειγματικού μέσου

- Γνωρίζουμε ότι $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$
- που εκτιμάται με $\widehat{\mathrm{Var}}(\overline{X}_n) = \frac{S_n^2}{n}$
- Η εκτίμηση του τυπίκου σφάλματος του δειγματικού μέσου:
  $\widehat{\mathrm{se}} = \sqrt{\frac{S_n^2}{n}}$
- Στο προηγούμενο παράδειγμα έχουμε $\widehat{\mathrm{se}} = \sqrt{0.78}$
- Μέσω του jackknife εκτιμούμε τη διασπορά του δειγματικού μέσου από την εξίσωση (4).
  - $\widetilde{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(\widetilde{T}_i - T_{\mathrm{jack}}\right)^2 = 4.67$
  - Άρα
    $$v_{\mathrm{jack}} = \frac{\widetilde{s}_n^2}{n} = 0.78$$
  - οπότε $\widehat{\mathrm{se}}_{\mathrm{jack}} = \sqrt{v_{\mathrm{jack}}} = \sqrt{0.78}$

Μέση τιμή

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

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

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

Διασπορά

- Η jackknife εκτίμηση της διασποράς είναι

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

που είναι αμερόληπτη.

- Hint για την απόδειξη:

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

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)

1. *Αν T_n είναι αμερόληπτος, το ίδιο ισχύει και για τον T_{jack}:*

$$\mathrm{E}T_n = \theta, \quad \forall \theta \in \Theta \quad \Rightarrow \quad \mathrm{E}T_{\text{jack}} = \theta, \quad \forall \theta \in \Theta.$$

2. *Έστω ότι η μεροληψία του T_n γράφεται ως*

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

όπου $\alpha_1, \alpha_2, \ldots$ 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: $\mathrm{E}T_n = \theta$, $\forall \theta \in \Theta$.
- Το ίδιο θα ισχύει και για τους $T(-i)$, $i = 1, \ldots, n$:

$$\mathrm{E}T(-i) = \theta, \quad \forall \theta \in \Theta.$$

- Συνεπώς

$$\begin{aligned} \mathrm{E}T_{\text{jack}} &= \mathrm{E}\left\{nT_n - (n-1)\overline{T}_n\right\} \\ &= n\mathrm{E}T_n - (n-1)\mathrm{E}\overline{T}_n \\ &= n\theta - (n-1)\theta \\ &= \theta, \quad \theta \in \Theta. \end{aligned}$$

- 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 $\leqslant M\frac{1}{n^3}$ για κάποια σταθερά M.

- Η μεροληψία των $T(-i)$ γράφεται ως

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

διότι $O(1/(n-1)^3) = O(1/n^3)$.

- Άρα

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

Απόδειξη ιδιοτήτων 2

- Συνεπώς η μέση τιμή της jackknife εκτίμησης της μεροληψίας είναι

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

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.

Λέμε ότι ο T_n είναι *τετραγωνική στατιστική συνάρτηση*, αν μπορεί να εκφραστεί ως:

$$T_n = E(T_n) + \frac{1}{n}\sum_{i=1}^{n} \alpha^{(n)}(x_i) + \frac{1}{n^2}\sum_{1 \leqslant i \leqslant j \leqslant n} \beta^{(n)}(x_i, x_j), \tag{6}$$

όπου $\alpha^{(n)}(x)$ και $\beta^{(n)}(x, y)$ είναι συναρτήσεις των x και x, y, αντίστοιχα.

Πρόταση

Αν ο αρχικός εκτιμητής T_n είναι τετραγωνική στατιστική συνάρτηση της μορφής (6)*, ο b_{jack}, εκτιμά αμερόληπτα τη μεροληψία του T_n.*

Εκτίμηση Διασποράς Jackknife

Ο Tukey (1958) , θεωρώντας ότι οι ψευδοτιμές είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, πρότεινε την εκτίμηση της διασποράς του εκτιμητή jackknife:

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

όπου

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

- Το v_{jack} εκτιμά τόσο τη $\text{Var}(T_n)$ όσο και τη $\text{Var}(T_{\text{jack}})$.
- Εκτίμηση τυπικού σφάλματος:

$$\text{se}_{\text{jack}} = \sqrt{v_{\text{jack}}}$$

Ιδιότητες της Εκτίμησης

- Η εκτίμηση v_{jack} είναι γενικά **μεροληπτική**.
- Τείνει να **υπερεκτιμά** την πραγματική διασπορά $\text{Var}(T_n)$.
- Παρόλα αυτά, σε ορισμένες περιπτώσεις είναι **ασυμπτωτικά συνεπής**.

Ασυμπτωτική Συνέπεια[1]

Θεώρημα

Έστω ότι

$$E(X_1) = \mu, \qquad \mathrm{Var}(X_1) = \sigma^2 < \infty$$

και $T_n = g(\bar{X})$ *όπου* g *είναι συνεχώς παραγωγίσιμη με* $g'(\mu) \neq 0$. *Τότε*

$$\frac{T_n - g(\mu)}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1)$$

και

$$\frac{v_{\text{jack}}}{\sigma_n^2} \to^{\sigma.\text{β.}} 1$$

όπου

$$\sigma_n^2 = n^{-1}[g'(\mu)]^2\sigma^2$$

[1] δες: Miller (1964). A trustworthy jackknife. *The Annals of Statistics*.

Περιορισμοί της Μεθόδου

Το προηγούμενο θεώρημα καλύπτει κυρίως στατιστικές συναρτήσεις της μορφής:

$$T_n = g(\bar{X})$$

- Παρόμοια αποτελέσματα ισχύουν και για:

$$T_n = g(S_n^2)$$

- Ωστόσο, η μέθοδος **δεν είναι συνεπής για μη ομαλά στατιστικά**.

Μη Συνέπεια για Ποσοστιαία Σημεία

Θεώρημα

Ο jackknife εκτιμητής της διασποράς για το p ποσοστιαίο σημείο

$$T(F) = F^{-1}(p)$$

δεν είναι συνεπής.

δες Efron (1982). The jackknife, the bootstrap and other resampling plans. *SIAM review*

Παράδειγμα

Στόχος: Εκτίμηση του τυπικού σφάλματος του δειγματικού μέσου με jackknife.

$$\widetilde{s}_n^2 = 4.67$$

Άρα

$$v_{\text{jack}} = \frac{4.67}{n} = 0.78$$

και

$$\text{se}_{\text{jack}} = \sqrt{0.78}$$

Παρατήρηση

Για τον δειγματικό μέσο γνωρίζουμε ότι $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ και $\widehat{\text{se}} = \sqrt{\frac{S_n^2}{n}} = \sqrt{0.78}$.

Άρα το αποτέλεσμα συμπίπτει με τη θεωρητική εκτίμηση.

Διάστημα Εμπιστοσύνης Jackknife

Ο Tukey (1958) πρότεινε το διάστημα

$$T_{\text{jack}} \pm t_{n-1;\alpha/2}\, \text{se}_{\text{jack}}$$

ως προσεγγιστικό $100(1-\alpha)\%$ διάστημα εμπιστοσύνης.

Βασική ιδέα:

- Οι ψευδοτιμές $\tilde{T}_i$ σε ορισμένες περιπτώσεις μπορεί να θεωρηθούν i.i.d.
- Τότε

$$\frac{\sqrt{n}(T_{\text{jack}} - \theta)}{\tilde{s}_n}$$

ακολουθεί κατά προσέγγιση κατανομή t_{n-1}.

Σύνδεση jackknife και συνάρτησης επιρροής

Μπορεί να δειχθεί ότι

$$IF(X_i) \approx (n-1)(T_n - T_{(-i)})$$

- Η **συνάρτηση επιρροής**: μετρά την επίδραση που έχει η **προσθήκη** (μικρής) μάζας πιθανότητας στο X_i
- Το **jackknife**: μετρά την πεπερασμένη επίδραση της **αφαίρεσης** του X_i από το δείγμα.
- Το jackknife είναι μια εμπειρική εκτίμηση της συνάρτησης επιρροής κάθε μίας παρατήρησης