Υπολογιστική Στατιστική

Κατερίνα Ορφανογιαννάκη

Τμήμα Μαθηματικών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
korfanog@math.uoa.gr

2020-2021

Μέθοδος Jackknife Εισαγωγή

Η μέθοδος Jackknife αναπτύχθηκε για τον περιορισμό της μεροληψίας ορισμένων εκτιμητριών. Επίσης η μέθοδος αυτή βρίσκει εφαρμογή σε περιπτώσεις σύνθετων εκτιμητριών για τις οποίες ο θεωρητικός υπολογισμός του τυπικού σφάλματος είναι αρκετά πολύπλοκος. Συγκεκριμένα η μέθοδος Jackknife προσφέρει:

1. Εκτιμήτριες συναρτήσεις με μικρότερη μεροληψία σε απόλυτη τιμή
2. Εύκολο υπολογισμό του τυπικού σφάλματος εκτιμήτριας.

Συμβολισμός

Πριν προχωρήσουμε στην ανάπτυξη της μεθόδου χρειάζεται να κάνουμε μία εισαγωγή στον συμβολισμό τον οποίο θα χρησιμοποιήσουμε και ο οποίος έχει μια ιδιαιτερότητα:
Έστω ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ μεγέθους n και μια εκτιμήτρια συνάρτηση $\hat{\theta}$ της πραγματικής παραμέτρου θ του πληθυσμού η οποία είναι μια συνάρτηση του δείγματος ($\hat{\theta} = T(X_1, X_2, \ldots, X_n)$).
Ορίζουμε ως $\hat{\theta}_{(i)}$ την τιμή της εκτιμήτριας που προκείπτει αν αφαιρέσουμε την i παρατήρηση από το δείγμα μας. Δηλαδή $\hat{\theta}_{(i)} = T(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$

Η Jackknife εκτιμήτρια

Η Jackknife εκτιμήτρια $\hat{\theta}_J$ της απλής εκτιμήτριας $\hat{\theta}$ υπολογίζεται ως

$$\hat{\theta}_J = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\bullet)} \quad ,$$

όπου $\bar{\hat{\theta}}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$

Διαδικασία:

- Για να υπολογίσουμε λοιπον την Jackknife εκτιμήτρια αρχικά αφαιρούμε την πρώτη παρατήρηση X_1 και υπολογίζουμε την εκτιμήτρια $\hat{\theta}_{(1)}$. Μετά επιστρέφουμε στο δείγμα την πρώτη παρατήρηση και αφαιρούμε τη δεύτερη X_2 και ούτω καθεξής μέχρι να αφαιρέσουμε και την τελευταία παρατήρηση. Σε κάθε επανάληψη της διαδικασίας το πλήθος των παρατηρήσεων είναι $n-1$.
- Αφαιρώντας παρατηρήσεις από το αρχικό δείγμα και εκτιμώντας ξανά την παράμετρο που μας ενδιαφέρει, παίρνουμε πληροφορία σχετικά με τη σταθερότητα, και άρα τη μεταβλητότητα, της εκτιμήτριας. Επομένως αν αφαιρούμε κάθε φορά μια παρατήρηση εξετάζοντας πόσο αλλάζουν οι τιμές της εκτιμήτριας παίρνουμε μια εικόνα σχετικά με τη διακύμανση της εκτιμήτριας.
- Η Jackknife εκτιμήτρια κάνει διόρθωση της εκτιμήτριας από το συνολικό δείγμα με τη χρήση των εκτιμητριών από τα δείγματα όπου έχουμε αφαιρέσει μια τιμή.

Ψευδοτιμές

Ένας εναλλακτικός τρόπος με τον οποίο μπορεί να υπολογιστεί η Jackknife εκτιμήτρια είναι με τη χρήση ψευδοτιμών. Συγκεκριμένα ορίζονται οι ψευδοτιμές p_i ως:

$$p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$$

Και η εκτιμήτρια Jackknife δίνεται από τη σχέση:

$$\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} p_i$$

.

Παράδειγμα Ι

Έστω οι εξής 6 παρατηρήσεις που αφορούν το χρόνο αναμονής σε λεπτά σε μια στάση λεωφορείου: 4 ,3 ,7 ,6 ,5 ,9. Να υπολογιστούν οι εκτιμήτριες jackknife για τη μέση τιμή και τη διάμεσο του πληθυσμού.

Διάφορες Jackknife εκτιμήτριες: Μέση τιμή

Έστω ότι η εκτιμήτρια η οποία μας ενδιαφέρει είναι η μέση τιμή. Δηλαδή: $\hat{\theta} = \bar{x}$. Οι ψευδοτιμές υπολογίζονται ως:

$$p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} = n\hat{\theta} - (n-1)\frac{\sum_{j=1}^{n} x_j - x_i}{n-1} =$$
$$= n\hat{\theta} - (n-1)\frac{n\hat{\theta} - x_i}{n-1} = x_i$$

Η εκτιμήτρια jackknife επομένως θα είναι: $\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} p_i = \bar{x}$.

Γενίκευση:

Για οποιαδήποτε εκτιμήτρια της μορφής

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

η Jackknife εκτιμήτρια ταυτίζεται με την απλή εκτιμήτρια.

Διάφορες Jackknife εκτιμήτριες: Διάμεσος

Αν το πλήθος των παρατηρήσεων είναι άρτιος αριθμός τότε η Jackknife εκτιμήτρια της διαμέσου ταυτίζεται με τη δειγματική διάμεσο αν όμως είναι περιττός αριθμός κάτι τέτοιο δεν ισχύει.
Για άρτιο μέγεθος δείγματος έχουμε:

$$\hat{\theta}_{(i)} = \begin{cases} X_{n/2} & i \geq \frac{n}{2} + 1 \\ X_{\frac{n}{2}+1} & i < \frac{n}{2} + 1 \end{cases}$$

και $\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\frac{n}{2}\left(X_{n/2} + X_{\frac{n}{2}+1}\right) = \frac{1}{2}\left(X_{n/2} + X_{\frac{n}{2}+1}\right)$

Διάφορες Jackknife εκτιμήτριες: Διάμεσος (συνέχεια)

Για περιπτό μέγεθος δείγματος έχουμε:

$$\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{2n}\left(\frac{n-1}{2}+1\right)\left(X_{\frac{n+1}{2}+1}+X_{\frac{n+1}{2}-1}\right)+\frac{1}{2n}(n-1)X_{\frac{n+1}{2}}$$

και

$$\hat{\theta}_J = \left[n-\frac{1}{2n}(n-1)^2\right]X_{\frac{n+1}{2}}-(n-1)\frac{1}{2n}\left(\frac{n-1}{2}+1\right)\left(X_{\frac{n+1}{2}+1}+X_{\frac{n+1}{2}-1}\right)$$

Διάφορες Jackknife εκτιμήτριες: Διακύμανση

Όταν έχουμε κανονικό πληθυσμό η εκτιμήτρια μεγίστης πιθανοφάνειας της παραμέτρου σ^2 είναι η δειγματική διακύμανση $\hat{\theta} = s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ για την οποία ξέρουμε πως είναι μεροληπτική (για κανονικούς πληθυσμούς). Μπορεί να αποδειχτεί πως :
Η Jackknife εκτιμήτρια της διακύμανσης είναι:

$$\hat{\theta}_J = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

που είναι αμερόληπτη.

Διάφορες Jackknife εκτιμήτριες: Ποσοστά

Έστω ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ δοκιμών *Bernoulli*. Δηλαδή κάθε X_i παίρνει την τιμή 1 με πιθανότητα p και 0 με πιθανότητα $1 - p$. Έστω ότι θέλουμε να εκτιμήσουμε την ποσότητα p^2. Θα δείξουμε ότι η εκτιμήτρια *jackknife* βελτιώνει τη μεροληψία της εκτιμήτριας μεγίστης πιθανοφάνειας.

Διάφορες Jackknife εκτιμήτριες: Ποσοστά (συνέχεια)

Ορίζουμε ως $R = \sum_{i=1}^{n} X_i$ τον αριθμό των επιτυχιών (των μονάδων δηλαδή) στις n δοκιμές. Η εκτιμήτρια μεγίστης πιθανοφάνειας για το p είναι R/n. Επομένως από τις ιδιότητες της μεθόδου μεγίστης πιθανοφάνειας η εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\theta}$ της παραμέτρου $\theta = p^2$ θα είναι

$$\hat{\theta} = \left(\frac{R}{n}\right)^2.$$

θ.δ.ο.

$$E(\hat{\theta}) = \frac{p(1-p)}{n} + p^2$$

Δηλαδή η εκτιμήτρια είναι μεροληπτική.

Διάφορες Jackknife εκτιμήτριες: Ποσοστά (συνέχεια)

θ.δ.ο. η εκτιμήτρια *jackknife* είναι:

$$\hat{\theta}_J = \frac{R(R-1)}{n(n-1)}$$

για την οποία ισχύει:

$$E(\hat{\theta}_J) = p^2$$

και επομένως η *jackknife* εκτιμήτρια είναι αμερόληπτη.

Τυπικά σφάλματα

Από τον ορισμό της εκτιμήτριας *jackknife* με τη χρήση των ψευδοτιμών προκύπτει ότι:

$$Var(\hat{\theta}_J) = \frac{\sum_{i=1}^{n} Var(p_i)}{n^2} = \frac{Var(p_1)}{n},$$

όμως επειδή τα p_i έχουν όλα την ίδια διακύμανση για αυτό μπορούμε να χρησιμοποιούμε, έστω, την τιμή p_1.

Μια αμερόληπτη εκτίμηση για τη διακύμανση των ψευδοτιμών είναι η:

$$s_p^2 = \frac{1}{n-1} \sum_{i=1}^{n} (p_i - \bar{p})^2$$

και επομένως μπορούμε να εκτιμήσουμε τη διακύμανση της *jackknife* εκτιμήτριας ως

$$\begin{aligned} s_{\hat{\theta}_j}^2 &= \frac{1}{n(n-1)} \sum_{i=1}^{n} (p_i - \bar{p})^2 \\ &= \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\bullet)})^2 \end{aligned}$$

Τυπικά σφάλματα (συνέχεια)

Για μια μεγάλη ομάδα εκτιμητριών η εκτιμήτρια *jackknife* ταυτίζεται με την απλή εκτιμήτρια και επομένως μπορούμε να εκτιμήσουμε το τυπικό σφάλμα της απλής εκτιμήτριας χρησιμοποιώντας αυτό της *jackknife* που είναι πιο εύκολο. Η διακύμανση της *jackknife* εκτιμήτριας θα είναι περίπου ίση με τη διακύμανση της αρχικής εκτιμήτριας. Από τον ορισμό της *jackknife* εκτιμήτριας έχουμε ότι:

$$Var(\hat{\theta}_J) = n^2 Var(\hat{\theta}) + (n-1)^2 Var(\bar{\hat{\theta}}_{(\bullet)}) - 2n(n-1)Cov(\hat{\theta}, \bar{\hat{\theta}}_{(\bullet)})$$

Όμως η διακύμανση των $\hat{\theta}$ και $\bar{\hat{\theta}}_{(\bullet)}$ να είναι περίπου ίδια, ενώ η συνδιακύμανση τους θα είναι περίπου ίδια με την κοινή τους διακύμανση. Επομένως προκύπτει ότι:

$$Var(\hat{\theta}_J) \approx Var(\hat{\theta})$$

Εφαρμογές

Η μέθοδος *jackknife* βρίσκει πολλές εφαρμογές στη δειγματοληψία. Δύο ενδεικτικά παραδείγματα είναι τα εξής:

- Η εκτιμήτρια λόγου $Z = \bar{X}/\bar{Y}$ για να εκτιμήσουμε το λόγο των μέσων τιμών στον πληθυσμό. Η εκτιμήτρια αυτή είναι μεροληπτική. Χρησιμοποιώντας *jackknife* εκτιμήτριες καταφέρνουμε όχι μόνο να μειώσουμε τη μεροληψία αλλά να πάρουμε και μια εκτίμηση της διακύμανσης της ποσότητας που μας ενδιαφέρει κάτι που δεν είναι εύκολο με τη χρήση ασυμπτωτικών αποτελεσμάτων.
- Σε ένα δειγματοληπτικό σχέδιο δειγματοληψίας σε ομάδες (*cluster sampling*) μπορούμε να κάνουμε *jackknife* αφαιρώντας κάθε φορά μια ομάδα. Έτσι ακόμα και σε ιδιαίτερα πολύπλοκα δειγματοληπτικά σχέδια (τα οποία όμως είναι αρκετά διαδεδομένα σε μεγάλες έρευνες στην πράξη) η *jackknife* αποτελεί πολύτιμο εργαλείο.

Συμπεράσματα

- Οι εκτιμήτριες *jackknife* μειώνουν την μεροληψία σε σχέση με τις απλές εκτιμήτριες και μπορούμε σχετικά εύκολα να υπολογίσουμε τα τυπικά τους σφάλματα.
- Αν η απλή εκτιμήτρια είναι γραμμική συνάρτηση των δεδομένων, τότε η *jackknife* εκτιμήτρια ταυτίζεται με αυτή και άρα μπορούμε να βρούμε μια εκτίμηση του τυπικού της σφάλματος απλά υπολογίζοντας το τυπικό σφάλμα της *jackknife* εκτιμήτριας
- Αν η μορφή της εκτιμήτριας όμως είναι διαφορετική (π.χ. διάμεσος, μέγιστη τιμή) τότε η μέθοδος *jackknife* δεν είναι ικανοποιητική και πρέπει να χρησιμοποιείται με προσοχή! Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε μια γενίκευση της μεθόδου όπου αφαιρούμε όχι μόνο μια παρατήρηση τη φορά αλλά περισσότερες.
- Η μέθοδος *jackknife* είναι ξεκάθαρα μια μη παραμετρική μέθοδος που δεν βασίζεται σε καμιά παραμετρική υπόθεση σχετικά με τον πληθυσμό. Επομένως τα αποτελέσματα της δεν εξαρτώνται από καμιά υπόθεση.