Προχωρημένες Μέθοδοι Δειγματοληψίας

Εργασία

1. Επαληθεύστε ότι $V(I_i(s)) = \pi_i(1-\pi_i)$, και $V(w_i) = (1-\pi_i)/\pi_i$.

2. Ποιά είναι η ερμηνεία του βάρους $w_i = 1$ για μία μονάδα $i \in U$;

3. Σε απλή τυχαία δειγματοληψία έχουμε $\pi_i = n/N$ για κάθε $i \in U$. Επαληθεύστε ότι $\hat{N} = N$.

4. Εκφράστε το δειγματικό μέγεθος n ως άθροισμα δεικτριών τυχαίων μεταβλητών, και υπολογίζοντας την αναμενόμενη τιμή αυτού του αθροίσματος δείξτε ότι $n = \sum_{i=1}^{N} \pi_i$.

5. Σε ένα πληθυσμό τεσσάρων μονάδων $U = \{1, 2, 3, 4\}$, οι τιμές μιάς μεταβλητής y είναι $y_1 = 10$, $y_2 = 12$, $y_3 = 15$, $y_4 = 22$, και άρα ο ολικός είναι $Y = 59$. Θεωρώντας όλα τα έξι δυνατά δείγματα μεγέθους $n = 2$, όλα με πιθανότητα επιλογής ίση με $1/6$, υπολογίστε τις έξι δυνατές τιμές του εκτιμητή $\hat{Y}$ του ολικού Y καθώς και τις διαφορές αυτών των εκτιμήσεων απο την πραγματική τιμή του Y. Τί παρατηρείτε; Υπολογίστε τον μέσο όρο αυτών των έξι αυτών τιμών. Τί παρατηρείτε;

6. Για τον πληθυσμό $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$, θεωρείστε τα διαφορετικά δείγματα, μεγέθους $n = 4$, $s_1 = \{1, 3, 5, 6\}$, $s_2 = \{2, 3, 7, 8\}$, $s_3 = \{1, 4, 6, 8\}$, $s_4 = \{2, 4, 6, 8\}$, $s_5 = \{4, 5, 7, 8\}$, με τίς αντίστοιχες πιθανότητες $p(s_1) = 1/8$, $p(s_2) = 1/4$, $p(s_3) = 1/8$, $p(s_4) = 3/8$, $p(s_5) = 1/8$. [Τα υπόλοιπα δυνατά δείγματα έχουν $p(s) = 0$.] Βρείτε την πιθανότητα επιλογής π_i για όλες τις μονάδες $i \in U$.

7. Για απλή τυχαία δειγματοληψία (όπου $\pi_i = n/N$) βρείτε τον εκτιμητή μέσου $\hat{\bar{Y}}$. Τί παρατηρείτε ως προς την σχέση δείγματος με πληθυσμό;

8. Το αρχείο "syc" (βρίσκεται μαζί με περιγραφή του στο e-class/έγγραφα) περιέχει επιλεγμένες μεταβλητές της δειγματοληπτικής έρευνας *Survey of Youth in Custody* που συνέλεξε πληροφορίες απο άτομα νεαρής ηλικίας που είναι έγκλειστα σε σωφρονιστικά ιδρύματα Πολιτειών της Αμερικής. Οι πιθανότητες επιλογής των ατόμων του δείγματος είναι άνισες. Χρησιμοποιείστε τα βάρη που δίνονται απο την μεταβλητή *finalwt* για να εκτιμήσετε την μέση ηλικία πρώτης σύλληψης (μεταβλητή *agefirst*). Συγκρίνετε αυτή την εκτίμηση με αυτή που προκύπτει χωρίς την χρήση των βαρών (ή θεωρώντας οτι τα βάρη είναι ίσα).
   Εκτιμήστε τα ποσοστά των νέων που: (α) είναι ηλικίας 14 και κάτω, (β) είναι άντρες, (γ) κρατούνται για βίαιη επίθεση (crimtype 1), (δ) έχουν κάνει χρήση ναρκωτικών (everdrug), (ε) έχουν και τις δύο ιδιότητες (γ) και (δ). Για τους υπολογισμούς να παραληφθούν τα missing values στις μεταβλητές agefirst και everdrug.

9. Για το *ratio estimator*, όπου υποθέτουμε $y_i \approx Bx_i$ και $q_i = x_i$, δείξτε ότι $\hat{B} = \hat{Y}/\hat{X}$ και $\hat{Y}^R = \hat{Y}X/\hat{X}$.

10. Το σύνολο δεδομένων "cherry" (βρίσκεται μαζί με περιγραφή του στο e-class/ έγγραφα) περιέχει μετρήσεις διαμέτρου, ύψους και όγκου για δείγμα τριανταενός ($n = 31$) δέντρων κερασιάς.
(α) Φτιάξτε διάγραμμα σημείων (scatter plot) που να δείχνει την γραμμική συσχέτιση όγκου με διάμετρο (y ο όγκος, x η διάμετρος).
(β) Υποθέστε ότι τα 31 δέντρα έχουν επιλεγεί με απλή τυχαία δειγματοληψία απο δάσος με N=2967 κερασιές ($\pi_i = n/N$), και ότι το άθροισμα των διαμέτρων όλων των κερασιών του δάσους είναι $X = 41835$. Χρησιμοποιείστε τον εκτιμητή λόγου (ratio estimator), με βοηθητική μεταβλητή x την διάμετρο, για να εκτιμήσετε τον ολικό όγκο όλων των κερασιών του δάσους.

11. Στην ειδική *περίπτωση* 1 του calibration, επαληθεύστε ότι $g_i = N_j/\hat{N}_j$ για όλα τα $i \in U_j$. (Υποθέστε ότι $q_i = 1$ για όλα τα $i \in U$)

12. Για απλή τυχαία δειγματοληψία (όπου $\pi_i = n/N$) και για κάποιον υποπληθυσμό $U_d \subset U$, βρείτε τα $\hat{N}_d$, $\hat{Y}_d$, $\hat{\bar{Y}}_d$ και $\hat{P}_d$. Τί παρατηρείτε για τα $\hat{\bar{Y}}_d$ και $\hat{P}_d$;

13. Αν έχει γίνει calibration της ειδικής *περίπτωσης* 1, ποιά θα είναι η μορφή του $\hat{Y}_d^C$ για κάποιο $U_d \subset U$;

14. Για τον υπολογισθέντα εκτιμητή $\hat{Y}$ της άσκησης 10, υπολογίστε την διακύμανση Jackknife $\hat{V}_{JK}(\hat{Y})$ με $k = n$, $m = 1$ (αναζητήστε τον τύπο στις διαφάνειες). Έπειτα υπολογίστε την διακύμανση Jackknife $\hat{V}_{JK}(\hat{Y}^R)$, του επίσης υπολογισθέντος εκτιμητή λόγου $\hat{Y}^R$, χρησιμοποιώντας τον τύπο $[(n-1)/n]\sum_{k=1}^{n}(\hat{\theta}_{(k)} - \hat{\bar{\theta}})^2$, με τα καταλλήλως ορισμένα $\hat{\theta}_{(k)}$, και όπου $\hat{\bar{\theta}} = (1/n)\sum_{k=1}^{n}\hat{\theta}_{(k)}$. Συγκρίνετε τις διακυμάνσεις $\hat{V}_{JK}(\hat{Y})$ και $\hat{V}_{JK}(\hat{Y}^R)$.

15. Στον τύπο της $\hat{V}_{JK}(\hat{Y})$ στην περίπτωση στρωματικής πολυσταδιακής δειγματοληψίας, χρησιμοποιείστε έκφραση του $\hat{Y}$ ("σπάζοντας" το s κατάλληλα), ανάλογη της έκφρασης του $\hat{Y}_{(hk)}$ που υπάρχει στις διαφάνειες, και δείξτε οτι

$$\hat{V}_{JK}(\hat{Y}) = \sum_{h=1}^{H}\frac{K_h-1}{K_h}\sum_{k=1}^{K_h}\left(\hat{Y}_{(hk)} - \hat{Y}\right)^2 = \sum_{h=1}^{H}\frac{K_h-1}{K_h}\sum_{k=1}^{K_h}\left(\frac{1}{K_h-1}\hat{Y}_{s_h - s_{hk}} - \hat{Y}_{s_{hk}}\right)^2,$$

όπου $\hat{Y}_{s_h - s_{hk}} = \sum_{s_h - s_{hk}} w_i y_i$ και $\hat{Y}_{s_{hk}} = \sum_{s_{hk}} w_i y_i$. Ερμηνεύστε τον όρο $\frac{1}{K_h-1}\hat{Y}_{s_h - s_{hk}} - \hat{Y}_{s_{hk}}$ για το στρώμα h.

16. Στο αρχείο "syc" (βλέπε άσκηση 8), η μεταβλητή everdrug έχει 6 missing values (λείπει η απάντηση σε 6 μονάδες). Να γίνει imputation με τις τεχνικές:
(α) Τυχαίο hot-deck imputation κατά τάξεις, με τις 8 τάξεις που ορίζονται απο τις μεταβλητές sex και educ (4 κατηγορίες, βλέπε περιγραφή του αρχείου στο e-class). Για να επιλέξετε τυχαία μία μονάδα απο ένα σύνολο s με m μονάδες, χρησιμοποιείστε στην R την εντολή s[sample(m,1)].
(b) Nearest-neighbor hot-deck imputation, με κριτήριο απόστασης που ορίζεται απο τις μεταβλητές sex και educ. (Αυτή η τεχνική να χρησιμοποιηθεί για τις μονάδες με τιμή της μεταβλητής educ στο διάστημα [01,12]).