Υπολογιστική Στατιστική

Κατερίνα Ορφανογιαννάκη

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

2020-2021

Κίνητρο

- Για κάθε σημειακή εκτίμηση που λαμβάνουμε στη Στατιστική είναι σημαντικό να συνοδεύεται και από μία εκτίμηση μεταβλητότητας. Κάποιο τυπικό σφάλμα που να μας δίνει μία ένδειξη της μεταβλητότητας γύρω από αυτή την εκτίμηση.
- Παραδείγματα: Ο μέσος του δείγματος, το ποσοστό των ψήφων ενός κόμματος σε μία δημοσκόπηση κ.α.
- Πώς μπορούμε να εκτιμήσουμε τα τυπικά σφάλματα;
- Δεν είναι πάντα εύκολο. Π.χ. ενώ γνωρίζουμε το τυπικό σφάλμα για τον μέσο είναι δύσκολο να το υπολογίσουμε για τη διάμεσο.

Προκειμένου να λοιπόν να είμαστε σε θέση να υπολογίζουμε το τυπικού σφάλμα και σε περιπτώσεις όπως η διάμεσος χρειαζόμαστε κάποια εργαλεία για αυτό.

Παράδειγμα

16 ποντίκια συμμετείχαν σε ένα πείραμα, σε 7 από αυτά δόθηκε ένα καινούριο φάρμακο ενώ στα υπόλοιπα 9 δόθηκε ένα άλλο ήδη υπάρχον φάρμακο. Σκοπός ήταν να μελετηθεί η επιβίωση σε μέρες και αν το καινούριο φάρμακο μεγαλώνει την επιβίωση των ποντικιών. Οι τιμές εμφανίζονται στον παρακάτω πίνακα:

Φάρμακο	Παρατηρήσεις	Μέση τιμή	Τυπική απόκλιση της μέσης τιμής
Νέο	94, 197, 16, 38, 99, 141, 23	86.86	25.24
Παλιό	52, 104, 146, 10, 50, 31, 40, 27, 46	56.22	14.14
	Διαφορά	30.63	28.93

Η βασική ιδέα

- Ακόμα και αν ο υπολογισμός των τυπικών σφαλμάτων είναι δύσκολος να γίνει θεωρητικά μπορούμε να χρησιμοποιήσουμε την ισχή των υπολογιστών για να εκτιμήσουμε τυπικά σφάλματα.
- Το μόνο που χρειαζόμαστε είναι να είμαστε σε θέση να αναπαράγουμε τον μηχανισμό που γέννησε τα δεδομένα και κατέπέκταση την πηγή που που παρήγαγε την μεταβλητότητα/αβεβαιότητα.
- Παράδειγμα: Έστω ότι χρειαζόμαστε να αναπαραστήσουμε το τυπικό σφάλμα του μέσου από ένα δείγμα μεγέθους n από έναν συγκεκριμένο πληθυσμό. Λύση: Πάρε όσα περισσότερα δείγματα μεγέθους n μπορείς και για κάθε ένα εκτίμησε τον μέσο. Τότε έχεις αρκετές τιμές που να αντανακλούν την αβεβαιότητα γύρω από την εκτίμηση.
- Η μέθοδος βασίζεται στην προσωμοίωση. Χρειάζεται να προσωμοιώσουμε από την κατανομή του πληθυσμού.

Αλλά στην πραγματικότητα ...

- Αλλά όταν μας δίνεται ένα δείγμα πραγματικών δεδομένων γνωρίζουμε την F;
- Σπάνια η απάντηση είναι ναι σάυτή την ερώτηση. Γενικά δεν ισχύει.
- Οπότε δεν γνωρίζουμε από ποια F να προσωμοιώσουμε.
- Εδώ είναι που μπαίνει η *bootstrap* προκειμένου να δώσει λύση στο συγκεκριμένο ζήτημα.

Περίγραμμα

Τι είναι η *bootstrap*

- Η *Bootstrap* είναι μία υπολογιστική μέθοδος που χρησιμοποιείται ευρέως στις μέρες μας παρέχοντας μία γενική διαδικασία για την εκτίμηση τυπικών σφαλμάτων, τη δημιουργία διαστημάτων εμπιστοσύνης και για την εξαγωγή στατιστικών συμπερασμάτων βασιζόμενη σε δειγματοληψία από τα δεδομένα.
- Στις περισσότερες περιπτώσεις βασίζεται σε υπολογιστική ισχή.
- Βρίσκει εφαρμογή με μία ευρεία γκάμα προβλημάτων με ελάχιστο αριθμό υποθέσεων.

Προσέγγιση *Monte – Carlo*

Έστω ότι η συνάρτηση κατανομής F του πληθυσμού είναι γνωστή.
Θέλουμε να υπολογίσουμε το:

$$\mu(F) = \int \phi(y) dF(y)$$

μπορούμε να το προσεγγίσουμε χρησιμοποιόντας το:

$$\hat{\mu}(F) = \frac{1}{M} \sum_{i=1}^{M} \phi(y_i)$$

όπου y_i, $i = 1, \ldots, M$ τυχαίες μεταβλητές προσομοιωμένες από την F (ή απλά ένα δείγμα από την F).
Ξέρουμε ότι όταν $M \to \infty$ τότε $\hat{\mu}(F) \to \mu(F)$.

Τί συμβαίνει όταν η F δεν είναι γνωστή;

Κίνητρο

Θεραπεία: Γιατί να μην χρησιμοποιήσουμε μία εκτίμηση της F βασιζόμενοι στο δείγμα $(x_1, \ldots, x_n)$ που έχουμε στα χέρια μας;
Η ποιο γνωστή εκτίμηση της F είναι η εμπειρική συνάρτηση κατανομής

$$\hat{F}_n(x) = \frac{\#\ \text{παρατηρήσεων} \leq x}{n}$$

ή

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

όπου $I(A)$ είναι η δείκτρια συνάρτηση και ο δείκτης n μας υπενθυμίζει ότι βασίζεται στο δείγμα μεγέθους n.

Η βασική ιδέα της *Bootstrap*

Χρησιμοποίησε μία εκτίμηση της ποσότητας $\mu(\hat{F}_n)$ αντί για $\mu(F)$. Αφού η $\hat{F}_n$ είναι μία συνεπής εκτιμήτρια της F (π.χ. $\hat{F}_n \to F$ αν $n \to \infty$) τότε $\mu(\hat{F}_n) \to \mu(F)$.

Σημαντικό: $\mu(\hat{F}_n)$ είναι ένα ακριβές αποτέλεσμα. Στην πράξη δεν είναι εύκολο να το βρούμε οπότε χρησιμοποιούμε την *Monte Carlo* προσέγγιση του.

Οπότε η ιδέα της *bootstrap* που χρησιμοποιείται στην πράξη είναι η ακόλουθη:

Γέννησε δείγματα από την $\hat{F}_n$ και χρησιμοποίησε σαν εκτίμηση της $\mu(F)$ την ποσότητα:

$$\hat{\mu}(F) = \frac{1}{M} \sum_{i=1}^{M} \phi(y_i^*)$$

όπου y_i^*, $i = 1, \ldots, M$ τυχαίες μεταβλητές προσωμοιωμένες από την $\hat{F}_n$.

Προσωμοίωση από την $\hat{F}_n$

- Η προσομοίωση από την $\hat{F}_n$ είναι μία σχετικά εύκολη και σαφής διαδικασία. Η συνάρτηση πυκνότητας $\hat{f}_n$ που συνδέεται με την $\hat{F}_n$ θα είναι αυτή που δίνει πιθανότητα $1/n$ σε όλα τα παρατηρούμενα σημεία x_i, $i = 1, \ldots, n$ και 0 αλλού.
- Σημείωση: αν κάποια τιμή εμφανίζεται περισσότερες από μία φορές τότε στην τιμή αυτή αντιστοιχεί πιθανότητα μεγαλύτερη από $1/n$.
- Οπότε επιλέγουμε τυχαία δείγματα από τις παρατηρήσεις με δειγματοληψία με **επανάθεση** από το αρχικό δείγμα.
- Μία τιμή μπορεί να εμφανίζεται στο *bootstrap* δείγμα περισσότερες από μία φορές και κάποιες άλλες τιμές να μην εμφανίζονται καμία φορά στο δείγμα *bootstrap*.

Παράδειγμα

Έστω 10 τιμές: 1,3,6,8,9,11,14,16,19,18.

Ένα δείγμα που μπορεί να προκύψει με δειγματοληψία με επανάθεση είναι:

1,1,3,3,3,9,11,16,18,19.

Τι παρατηρείτε;

Μία γρήγορη αναδρομή στην ιστορία της *Bootstrap* (1)

- Εμφανίστηκε το 1979 σε μία δημοσίευση του Εφρον. Προκάτοχοι υπήρχαν για πολύ καιρό
- Έγινε ιδιαίτερα δημοφιλής τη δεκαετία του 80 εξαιτίας της εισαγωγής των υπολογιστών σε Στατιστικές εφαρμογές.
- Έχει δυνατό Μαθηματικό υπόβαθρο (με το οποίο δεν θα ασχοληθούμε εδώ).
- Στην πράξη βασίζεται στην προσωμοίωση αλλά σε μερικά παραδείγματα υπάρχουν ακριβή αποτελέσματα για τα οποία δεν χρειάζεται προσωμοίωση.
- Παρόλο που είναι μέθοδος για βελτίωση εκτιμητών, είναι ιδιαίτερα γνωστή ως μέθοδος εκτίμησης τυπικών σφαλμάτων, μεροληψίας και κατασκευής διαστημάτων εμπιστοσύνης.

Μία γρήγορη αναδρομή στην ιστορία της *Bootstrap*

- Απαιτεί ελάχιστες υποθέσεις. Κυρίως βασίζεται στην υπόθεση ότι το δείγμα είναι μία καλή αναπαράσταση του άγνωστου πληθυσμού.
- Δεν είναι μαύρο κουτί. Δουλεύει για την πλειοψηφία των προβλημάτων αλλά μπορεί να είναι προβληματική για άλλα.
- Στην πράξη είναι υπολογιστικά απαιτητική. Αλλά η πρόοδος στους υπολογιστές την κάνει εύκολα διαθέσιμη σε καθημερινές πρακτικές.

Είδη *Bootstrap*

- Παραμετρική *Bootstrap*: Γνωρίζουμε ότι η F ανήκει σε μία παραμετρική οικογένεια από κατανομές και απλά εκτιμάμε τις παραμέτρους από το δείγμα. Γεννάμε δείγματα από την F χρησιμοποιόντας τις παραμέτρους που έχουμε εκτιμήσει.
- Μη παραμετρική Βοοτστραπ: Δεν γνωρίζουμε τι μορφή έχει η F και την εκτιμάμε από την $\hat{F}$ την εμπειρική κατανομή που παίρνουμε από τα δεδομένα.

Ο γενικός αλγόριθμος *bootstrap*

1. Γέννησε ένα δείγμα $\mathbf{x}^*$ μεγέθους n από την $\hat{F}_n$.
2. Υπολόγισε την $\hat{\theta}^*$ γιάυτό το δείγμα *bootstrap*.
3. Επανέλαβε τα βήματα 1 και 2, B φορές.

Από αυτή τη διαδικασία καταλήγουμε με τις *bootstrap* τιμές $\hat{\theta}^* = (\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B)$. Θα χρησιμοποιήσουμε αυτές τις *bootstrap* τιμές για να υπολογίσουμε τις ποσότητες που μας ενδιαφέρουν.
Παρατηρείστε ότι $\hat{\theta}^*$ είναι ένα δείγμα από την άγνωστη κατανομή της $\hat{\theta}$ και άρα περιλαμβάνει όλη την πληροφορία που αφορά την $\hat{\theta}$!

Ένα παράδειγμα: Η διάμεσος

Θεωρείστε τα δεδομένα $\mathbf{x} = (x_1, \ldots, x_n)$. Έστω ότι θέλουμε να βρούμε το τυπικό σφάλμα της δειγματικής διαμέσου. Υπάρχουν ασυμπτωτικά αποτελέσματα αλλά ανεφέρονται σε μεγάλα μεγέθη δείγματος και δεν είναι εφαρμόσιμα στην περίπτωση μας αν το n μικρό. Χρησιμοποιούμε *bootstrap*.

- Γεννάμε ένα δείγμα $\mathbf{x_1}^*$ με δειγματοληψία με επανάθεση από το $\mathbf{x}$. Αυτό είναι το πρώτο μας *bootstrap* δείγμα.
- Γιάυτό το δείγμα υπολογίζουμε τη δειγματική διάμεσο έστω ότι την συμβολίζουμε με: $\hat{\theta}_1^*$
- Επανέλαβε τα βήματα 1 και 2, B φορές.

Στο τέλος έχουμε B τιμές $\hat{\theta}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*)$. Αυτό είναι ένα τυχαίο δείγμα από την κατανομή της δειγματικής διαμέσου και άρα μπορούμε να το χρησιμοποιήσουμε για να προσεγγίσουμε όποια ποσότητα μας ενδιαφέρει (μέσο, τυπική απόκλιση κ.α.) Επιπλέον ένα ιστόγραμμα είναι μία εκτίμηση της άγνωστης πικνότητας της δειγματικής διαμέσου. Μπορούμε να μελετήσουμε την λοξότητα κλπ.

Bootstrap τυπικά σφάλματα

Έστω θ_i^* η *Bootstrap* τιμή από το i δείγμα, $i = 1, \ldots, B$. Η *Bootstrap* εκτίμηση για το τυπικό σφάλμα της $\hat{\theta}$ υπολογίζεται ως:

$$se_B(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \hat{\theta}^*\right)^2}$$

όπου

$$\hat{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

Αυτό δεν είναι τίποτα άλλο από την τυπική απόκλιση των *Bootstrap* τιμών.

Bootstrap εκτίμηση της μεροληψίας

Όμοια μία εκτίμηση της μεροληψίας της $\hat{\theta}$ υπολογίζεται ως

$$\widehat{Bias}(\hat{\theta}) = \hat{\theta}^* - \hat{\theta}$$

Σημείωση: Ακόμα και αν η $\hat{\theta}$ είναι αμερόληπτη εκτιμήτρια, αφού είναι μόνο μία εκτίμηση μπορεί να μην είναι μηδέν. Άρα αυτή η εκτίμηση πρέπει να αξιολογείται συνδιαστικά με τα τυπικά σφάλματα.

Bootstrap εκτίμηση της Συνδιακύμανσης

Σε όμοιο τρόπο με τα τυπικά σφάλματα για μία παράμετρο μπορούμε να αποκτήσουμε εκτιμήσεις *Bootstrap* για τη συνδιακύμανση δύο παραμέτρων. Έστω ότι $\hat{\theta}_1$ και $\hat{\theta}_2$ είναι δύο εκτιμήτριες ενδιαφέροντος (π.χ. στην κανονική κατανομή μπορεί να είναι ο μέσος και η διακύμανση ή στην παλινδρόμηση δύο συντελεστές παλινδρόμησης). Τότε η *Bootstrap* εκτίμηση της συνδιακύμανση προκύπτει από:

$$Cov_B(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\theta}^*_{1i} - \hat{\theta}^*_1 \right) \left(\hat{\theta}^*_{2i} - \hat{\theta}^*_2 \right)$$

όπου $\left(\hat{\theta}^*_{1i}, \hat{\theta}^*_{2i} \right)$ είναι οι *Bootstrap* τιμές των δύο παραμέτρων που προκύπτουν από το *i Bootstrap* δείγμα.

Παράδειγμα: Συνδιακύμανση ανάμεσα στο μέσο και τη διακύμανση

Παραμετρική *Bootstrap*. Δείγματα μεγέθους $n = 20, 200$ γεννιούνται από $N(1, 1)$ και $Gamma(1, 1)$ κατανομές. Και οι δύο έχουν $\mu = 1$ και $\sigma^2 = 1$. Υποθέστε ότι $\hat{\theta}_1 = \bar{x}$ και $\hat{\theta}_2 = s^2$. Βασιζόμενοι σε $B = 1000$ επαναλήψεις εκτιμείστε την συνδιακύμανση ανάμεσα στην $\hat{\theta}_1$ και την $\hat{\theta}_2$.

	Κατανομή	
	Κανονική	Γάμμα
v=20	0.00031	0.0998
v=200	0.0007	0.0104

Πίνακας: Οι εκτιμήσεις της συνδιακύμανσης για το μέσο και τη διακύμανση βασοζόμενοι σε παραμετρική *Bootstrap* ($B = 1000$). Από τη θεωρία για την Κανονική κατανομή γνωρίζουμε ότι η συνδιακύμανση είναι 0.

Απλό *Bootstrap* δ.ε.

Χρησιμοποιούμε την εκτίμηση του τυπικού σφάλματος με *Bootstrap* και την υπόθεση της κανονικότητας (η οποία είναι αυθαίρετη σε αρκετές περιπτώσεις) για να κατεσκευάσουμε ένα διάστημα της μορφής:

$$(\hat{\theta} - Z_{1-a/2} se_B(\hat{\theta}), \hat{\theta} - Z_{a/2} se_B(\hat{\theta}))$$

όπου Z_a συμβολίζει το a ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής. Αυτό είναι ένα $(1 - a)$ διάστημα εμπιστοσύνης για τη θ. Υπονοεί ότι έχουμε υποθέσει ότι η $\hat{\theta}$ ακολουθεί την κανονική κατανομή.

Ποσοστιαία δ.ε.

Το διάστημα εμπιστοσύνης υπολογίζεται ως:

$$\left(\kappa_{a/2}, \kappa_{1-a/2}\right)$$

όπου κ_a συμβολίζει το a εμπειρικό ποσοστιαίο σημείο των $\hat{\theta}_i^*$.
Ξεκάθαρο αυτό είναι μη συμμετρικό και λαμβάνη υπόψη την μορφή της κατανομής των $\hat{\theta}_i^*$.

$\hat{\theta}_i^*$-t δ.ε.

Βελτιώνει το απλό *bootstrap* δ.ε. με την έννοια ότι δεν χρειάζεται την υπόθεση της κανονικότητας. Το διάστημα έχει τη μορφή:

$$(\hat{\theta} - \zeta_{1-a/2} se_B(\hat{\theta}), \hat{\theta} - \zeta_{a/2} se_B(\hat{\theta}))$$

όπου ζ_a είναι το a ποσοστιαίο σημείο των τιμών ξ_i, όπου

$$\xi_i = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)}$$

Παρατηρείστε ότι για τις τιμές ξ_i χρειαζόμαστε τις ποσότητες $se(\hat{\theta}_i^*)$. Αν δεν αυτές δεν δίνονται από κλειστού τύπου εξισώσεις (π.χ. ασυμπτωτικά τυπικά σφάλματα) μπορούμε να χρησιμοποιήσουμε $\hat{\theta}_i^*$ να τα εκτιμήσουμε. Το υπολογιστικό βάρος είναι διπλάσιο!
Παρατηρείστε ότι τα ξ_i είναι τυποποιημένες τιμές των $\hat{\theta}_i^*$.

BCa δ.ε.

Βελτιώνει το Ποσοστιαία *Bootsrap* δ.ε. με την έννοια ότι λαμβάνει υπόψη τη μεροληψία.
Το *Bootsrap BCa* δ.ε. είναι:

$$\left(\kappa_{p_1}, \kappa_{p_2}\right)$$

όπου $p_1 = \Phi\left(z_{\alpha/2} + 2b_0\right)$ και $p_2 = \Phi\left(z_{1-\alpha/2} + 2b_0\right)$ με $\Phi(\cdot)$ η συνάρτηση της τυποποιημένης κανονικής κατανομής.
κ_a είναι το $a-$ ποσοστιαίο σημείο της κατανομής των *Bootsrap* τιμών (όμοια με το συμβολισμό για το ποσοστιαίο δ.ε.).

$$b_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{i=1}^{B} I(\hat{\theta}_i^* \leq \hat{\theta})\right)$$

Αν η κατανομή των $\hat{\theta}_i^*$ είναι συμμετρική, τότε $b_0 = 0$ και $p_1 = a/2$ και $p_2 = 1 - a/2$, και επομένως καταλήγουμε στα απλά ποσοστιαία δ.ε.

Σύκριση των δ.ε.

Ποιό δ.ε. να χρησιμοποιήσω; Προκειμένου να καταλήξω σε μία απόφαση λαμβάνω υπόψη ότι:

- Τα Ποσοστιαία δ.ε. είναι εύκολα εφαρμόσιμα σε αρκετές περιπτώσεις.
- Για τα ποσοστιαία -*t* δ.ε. χρειάζεται να γνωρίζουμε τα: $se(\hat{\theta})$.
- Προκειμένου να μπορέσουμε να εκτιμήσουμε επαρκώς δ.ε. χρειάζεται να αυξήσουμε το *B*.
- Καταλήγουμε σε μη συμμετρικά δ.ε.(εκτός από την περίπτωση των απλών δ.ε., το οποίο μας δίνει μικρότερη διαίσθηση)
- Παρατηρείστε τη σύνδεση ανάμεσα στα δ.ε. και τον έλεγχο υποθέσεων!

Παράδειγμα 1: Συντελεστής συσχέτισης

Θεωρείστε τα ζεύγη τιμών (x_i, y_i), $i = 1, \ldots, n$. Ο συντελεστής συσχέτισης του *Pearson* δίνεται από τη σχέση:

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Η συμπερασματολογία δεν είναι εύκολη. Αποτελέσματα υπάρχουν μόνο κάτω από την υπόθεση της διδιάστατης κανονικής κατανομής. Η *Bootstrap* δίνει μια πιθανή λύση:

Αμερικανικές εκλογές

Τα δεδομένα αναφέρονται σε $n = 24$ πολιτείες στην Αμερική. Σχετίζονται με τις Αμερικανικές εκλογές του 1844. Οι δύο μεταβλητές είναι: το ποσοστό συμμετοχής στις εκλογές για την κάθε πολιτεία και η διαφορά ανάμεσα στους 2 υποψήφιους. Η ερώτηση είναι κατά πόσο υπάρχει συσχέτιση ανάμεσα στις 2 μεταβλητές. Ο παρατηρούμενος συντελεστής συσχέτισης είναι: $\hat{\theta} = -0.359$.

Αμερικανικές εκλογές: Τα δεδομένα

Πολιτεία	Συμμετοχή %	Απόλυτη διαφορά %
Maine	67.5	13
New Hampshire	65.6	19
Vermont	65.7	18
Masschusetts	59.3	12
Rhode Island	39.8	20
Connecticut	76.1	5
New York	73.6	1
New Jersey	81.6	1
Pensylvania	75.5	2
Maryland	80.3	3
Virginia	54.5	6
North Carolina	79.1	5

Πίνακας: Τα αποτελέσματα των αμερικάνικων προεδρικών εκλογών του 1844 σε 24 πολιτείες. Στον πίνακα βλέπετε το ποσοστό της συμμετοχής και την απόλυτη διαφορά ανάμεσα στους δύο υποψηφίους για τις 12 πρώτες πολιτείες.

Αμερικανικές εκλογές: Τα δεδομένα (συνέχεια)

Πολιτεία	Συμμετοχή %	Απόλυτη διαφορά %
Georgia	94.0	4
Kentucky	80.3	8
Tennessee	89.6	1
Louisianna	44.7	3
Alabama	82.7	18
Missisipi	89.7	13
Ohio	83.6	2
Indiana	84.9	2
Illinois	76.3	12
Missouri	74.7	17
Arkansas	68.8	26
Michigan	79.3	6

Πίνακας: Τα αποτελέσματα των αμερικάνικων προεδρικών εκλογών του 1844 σε 24 πολιτείες. Στον πίνακα βλέπετε το ποσοστό της συμμετοχής και την απόλυτη διαφορά ανάμεσα στους δύο υποψηφίους για τις επόμενες 12 πολιτείες.

Αμερικανικές εκλογές: Αποτελέσματα

Για $B = 1000$ επαναλήψεις βρήκαμε $\hat{\theta}^* = -0.3656$, $se_B(\hat{\theta}) = 0.1801$.
Το ασυμπτωτικό τυπικό σφάλμα που δίνεται από την θεωρία είναι
$se_N(\hat{\theta}) = \frac{1-\hat{\theta}^2}{\sqrt{n-3}} = 0.1881$. $Bias(\hat{\theta}) = 0.0042$

απλό δ.ε.	(-0.7228, -0.0167)
Ποσοστιαίο	(-0.6901, 0.0019)
Ποσοστιαίο-*t*	(-0.6731, 0.1420)
BCa	(-0.6806, 0.0175)

Πίνακας: 95% *bootstrap* δ.ε.

Περισσότερες λεπτομέρειες για το Ποσοστιαίο-*t* δ.ε.

Σε περίπτωση σαν αυτή στην όποια έχουμε μία εκτίμηση του τυπικού σφάλματος για τον συντελεστή συσχέτισης (παρόλο που είναι ασυμπτωτική εκτίμηση) μπορούμε να χρησιμοποιήσουμε το Ποσοστιαίο-*t* δ.ε. χωρίς να χρειάζεται να επαναλάβουμε τις επαναλήψεις *bootstrap*. Για να το κάνουμε αυτό, από τις *bootstrap* τιμές $\hat{\theta}_i^*$, $i = 1, \ldots, B$, υπολογίζουμε

$$\xi_i = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)}$$

όπου

$$se(\hat{\theta}_i^*) = \frac{1 - \hat{\theta}_i^{*2}}{\sqrt{n-3}}$$

Μετά βρίσκουμε τα ποσοστιαία σημεία των ξ. Παρατηρείστε ότι η κατανομή των ξ είναι ασύμμετρη, αυτό εξηγεί το διαφορετικό αριστερό όριο στο Ποσοστιαίο-*t* δ.ε.

Παράδειγμα 2: Δείκτης διασποράς

Για *count data* μία ποσότητα ενδιαφέροντος είναι ο λόγος $I = s^2/\bar{x}$. Για δεδομένα από την κατανομή *Poisson* η ποσότητα αυτή είναι θεωρητικά ίση με, 1. Έστω τα δεμένα ότι είναι ατυχήματα σε $n = 20$ διασταυρώσεις κατά τη διάρκεια ενός έτους. Θέλουμε να κατασκευάσουμε δ.ε. για το I. Τα δεδομένα είναι:
(1,2,5,0,3,1,0,1,1,2,0,1,8,0,5,0,2,1,2,3).
Χρησιμοιούμε *bootstrap* με επαναδειγματοληψία από τις παρατηρούμενες τιμές. Χρησιμοποιόντας $\hat{\theta} = I$ βρίσκουμε ($B = 1000$): $\hat{\theta} = 2.2659$, $\hat{\theta}^* = 2.105$, $\widehat{Bias}(\hat{\theta}) = 0.206$, $se_B(\hat{\theta}) = 0.6929$.

απλό *bootstrap* δ.ε.	(0.9077, 3.6241)
Ποσοστιαίο	(0.8981, 3.4961)
Ποσοστιαίο-*t*	(0.8762, 3.4742)
BCa	(1.0456, 3.7858)

Πίνακας: 95% *bootstrap* δ.ε. για το I

Μερικά σχόλια

- Ασυμπτωτική θεωρία για το I δεν είναι διαθέσιμη· η *bootstrap* είναι μία εύκολη μεθοδολογία για να εκτιμήσουμε τυπικά σφάλματα.
- Τα δ.ε. δεν είναι ίδια. Αυτό οφείλεται στην ύπαρξη μικρής ασσυμετρίας των *bootstrap* τιμών.
- Η δειγματική εκτίμηση είναι μεροληπτική.
- Για τα Ποσοστιαία-t δ.ε. χρησιμοποιούμε *bootstrap* να εκτιμήσουμε τυπικά σφάλματα για κάθε *bootstrap* δείγμα.