Υπολογιστική Στατιστική

Κατερίνα Ορφανογιαννάκη

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

2020-2021

Έλεγχος υποθέσεων - ανασκόπηση

Τα βασικά συστατικά του ελέγχου υποθέσεων είναι:

- Η μηδενική και η εναλλακτική υπόθεση
- Μία ελεγχοσυνάρτηση η οποία ισχυριζόμαστε ότι είναι ικανή να κάνει διαχωρισμό ανάμεσα στις 2 υποθέσεις (θέλαμε η συνάρτηση κατανομής της ελεγχοσυνάρτησης να έχει απλή συνάρτηση κατανομής).

Η τυπικά διαδικασία είναι να βρούμε την τιμή της ελεγχοσυνάρτησης από το δείγμα t_{obs} και στη συνέχεια να ελέγξουμε εάν αυτή η τιμή μπορεί να έχει προέλθει από τη μηδενική υπόθεση. Δηλαδή να συγκρίνουμε αυτή την τιμή με την τιμή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση και να βρούμε πόσο ακραία είναι.
Το *p-value* είναι η πιθανότητα μία τόσο ακραία ή περισσότερο ακραία τιμή της ελεγχοσυνάρτησης να έχει προκύψει κάτω από τη μηδενική υπόθεση. Η τυπική διαδικασία είναι ότι χρησιμοποιούμε το *p-value* για να βασίσουμε την συμπερασματολογία μας.

Έλεγχος υποθέσεων - πρόσφατη ανασκόπηση

- Η βασική ιδέα είναι ότι η κατανομή της ελεγχοσυνάρτησης είναι γνωστή κάτω από τη μηδενική υπόθεση. Και αν δεν είναι γνωστή;
- Γιατί να μην χρησιμοποιήσουμε *Bootstrap* για να έχουμε μία εκτίμηση της άγνωστης κατανομής της ελεγχοσυνάρτησης;
- Γιατί να μην χρησιμοποιήσουμε *Bootstrap* για να έχουμε μία εκτίμηση του *p-value*;

Αυτά τα επιχειρήματα μας δείχνουν ότι μπορούμε να χρησιμοποιήσουμε *Bootstrap* για έλεγχο υποθέσεων.
Δεν χρειαζόμαστε πλέον να γνωρίζουμε την κατανομή της ελεγχοσυνάρτησης. Επομένως μπορούμε να χρησιμοποιήσουμε περισσότερες ελεγχοσυναρτήσεις!

Έλεγχος υποθέσεων

Η παραμετρική *Bootstrap* είναι πολύ κατάλληλη για έλεγχο υποθέσεων.
Παράδειγμα: Έστω ότι έχουμε ανεξάρτητες παρατηρήσεις $(x_1, \ldots, x_n)$, και γνωρίζουμε ότι η κατανομή τους είναι η Γάμμα με κάποιες παραμέτρους. Επιθυμούμε να κάνουμε έλεγχο για μία τιμή του μέσου του πληθυσμού μ: H_0: $\mu = 1$ έναντι H_1: $\mu \neq 1$.
Το τυπικό t-τέστ είναι εφαρμόσιμο μόνο μέσω του Κεντρικού Οριακού Θεωρήματος και για μεγάλο μέγεθος δείγματος. Για μικρότερο μέγεθος δείγματος η κατάσταση δεν είναι τόσο εύκολη. Η ιδέα είναι να κατασκευάσουμε την κατανομή της ελεγχοσυνάρτησης χρησιμοποιώντας *bootstrap* (παραμετρική).
Οπότε, ο γενικός αλγόριθμος είναι ο ακόλουθος:

Έλεγχος υποθέσεων: ο Αλγόριθμος

- Όρισε τις 2 υποθέσεις.
- Διάλλεξε μία ελεγχοσυνάρτηση T που μπορεί να κάνει διαχωρισμό ανάμεσα στις 2 υποθέσεις. **Σημαντικό:** Δε μας ενδιαφέρει αν η ελεγχοσυνάρτηση μας έχω γνωστή κατανομή κάτω από τη μηδενική υπόθεση.
- Υπολόγισε της παρατηρούμενη τιμή της ελεγχοσυνάρτησης t_{obs} για το δείγμα.
- Γέννησε B δείγματα από την κατανομή που συνεπάγεται κάτω από τη μηδενική υπόθεση. Για κάθε δείγμα υπολόγισε την τιμή της ελεγχοσυνάρτησης t_i για $i = 1, \ldots, B$.
- Απορρίπτεις ή δεν απορρίπτεις ανάλογα με το επίπεδο στατιστικής σημαντικότητας.

Έλεγχος υποθέσεων: ο Αλγόριθμος (2)

Για να είμαστε ποιο τυπικοί ας υποθέσουμε ότι απορρίπτουμε τη μηδενική υπόθεση στη δεξιά ουρά της κατανομής. Τότε ένα προσεγγιστικό *p-value* δίνεται από τη σχέση:

$$\hat{p} = \frac{\sum_{i=1}^{B} I(t_i \geq t_{obs}) + 1}{B + 1}$$

όπου το $\hat{p}$ αποτελεί μία εκτίμηση του πραγματικού *p-value* και μπορούμε να χτίσουμε διαστήματα εμπιστοσύνης για αυτή την εκτίμηση. Μία καλή στρατηγική είναι να αυξάνουμε το B αν το $\hat{p}$ είναι κοντά στο επίπεδο στατιστικής σημαντικότητας.

Σημειώστε ότι μερικοί ερευνητές χρησιμοποιούν μία απλούστερη εκτίμηση του *p-value*, συγκεκριμένα:

$$\tilde{p} = \frac{\sum_{i=1}^{B} I(t_i \geq t_{obs})}{B}$$

Το $\hat{p}$ έχει καλύτερες ιδιότητες από το $\tilde{p}$.

Έλεγχος υποθέσεων: (Μη παραμετρική *Bootstrap*)

Η μόνη δυσκολία είναι ότι δεν γνωρίζουμε την κατανομή του πληθυσμού. Αυτό σημαίνει ότι πρέπει να προσομοιώσουμε τα δείγματα από την $\hat{F}_n$. Σύμφωνα με τη θεωρία του κλασικού ελέγχου υποθέσεων, χρειαζόμαστε την κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση. Πράγμα που σημαίνει ότι η $\hat{F}_n$ δεν είναι κατάλληλη. Μία λύση είναι να μετασχημαμτίσουμε την $\hat{F}_n$ έτσι ώστε να ικανοποιείται η μηδενική υπόθεση. Τότε μπορούμε να γεννήσουμε *Bootstrap* δείγματα από αυτή την κατανομή και να χτίσουμε την κατανομή της επιλεγμένης ελεγχοσυνάρτησης.

Παράδειγμα

Θεωρείστε τα παρακάτω δεδομένα:

$$\mathbf{x} = (-0.89, -0.47, 0.05, 0.155, 0.279, 0.775, 1.0016, 1.23, 1.89, 1.96).$$

Θέλουμε να ελέγξουμε την $H_0 : \mu = 1$, έναντι της $H_1 : \mu \neq 1$.
Επιλέγουμε σαν ελεγχοσυνάρτηση την $T = |\bar{x} - 1|$. Θα μπορούσαμε να χρησιμοποιήσουμε και άλλες ελεγχοσυναρτήσεις. Αφού $\bar{x} = 0.598$ βρίσκουμε ότι $T_{obs} = 0.402$. Προκειμένου η $\hat{F}_{10}$ να αναπαριστά τη μηδενική υπόθεση μετασχηματίζουμε τα δεδομένα έτσι ώστε να έχουν μέση τιμή ίση με 1. Για να το πετύχουμε αυτό προσθέτουμε σε κάθε παρατήρηση την ποσότητα 0.402. Τότε το καινούριο δείγμα είναι $\mathbf{x}_{null} = \mathbf{x} + 0.402$. Κάνουμε δειγματοληψία με επανάθεση από το $\hat{F}(\mathbf{x}_{null})$.
Για $B = 100$ βρίσκουμε $\hat{p} = 0.18$.
Σημαντικό: Ο μετασχηματισμός δεν είναι προφανής σε όλες τις περιπτώσεις.

Μη παραμετρική *Bootstrap*: Επιλογή ελεγχοσυνάρτησης

Έστω δύο δείγματα $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ και $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. Έστω ότι θέλουμε να ελέγξουμε την υπόθεση ότι έχουν τον ίδιο μέσο δηλαδή $H_0 : \mu_x = \mu_y$ έναντι της $H_1 : \mu_x \neq \mu_y$. Χρησιμοποιούμε την ελεγχοσυνάρτηση $T = |\bar{x} - \bar{y}|$. Κάτω από τη μηδενική υπόθεση μία καλή εκτίμηση της κατανομής του πληθυσμού είναι το δείγμα από ζεύγη τιμών $\mathbf{z} = (x_1, \ldots, x_n, y_1, \ldots, y_m)$. Επομένως, γεννάμε δείγματα από το $\mathbf{z}$. Για κάθε ένα από τα B *bootstrap* δείγματα υπολογίζουμε την τιμή της ελεγχοσυνάρτησης T_i^*, $i = 1, \ldots B$. Εκτιμάμε το *p-value* του ελέγχου ως

$$\hat{p} = \frac{\sum_{i=1}^{B} I(T_i \geq t_{obs}) + 1}{B + 1}$$

Μπορούν να χρησιμοποιηθούν και άλλες ελεγχοσυναρτήσεις όπως για παράδειγμα η γνωστή ελεγχοσυνάρτηση του δίπλευρου ελέγχου t.

Μη παραμετρική *Bootstrap*: Επιλογή ελεγχοσυνάρτησης (2)

Ένα σημείο που θέλει προσοχή όταν επιλέγουμε ελεγχοσυναρτήσεις είναι: Θέλουμε "*pivotal*" ελεγχοσυνάρτηση δηλαδή μία ελεγχοσυνάρτηση για την οποία η κατανομή δεν μεταβάλλεται (η κατανομή της δεν εξαρτάται από τις παραμέτρους του πληθυσμού). Για παράδειγμα η ελεγχοσυνάρτηση που χρησιμοποιείται για τον έλεγχο t έχει αυτή την ιδιότητα καθώς είναι τυποποιημένη.

"*Pivotal*" Ελεγχοσυνάρτηση

- Μία ελεγχοσυνάρτηση λέγεται *pivotal* εάν δεν εξαρτάται από τις παραμέτρους που εκτιμάμε.
- Οι *Pivotal* ελεγχοσυναρτήσεις επιτρέπουν τον διαχωρισμό ανάμεσα στη μηδενική και την εναλλακτική υπόθεση ανεξάρτητα από το ποιες είναι οι τιμές των παραμέτρων.
- Στο πλαίσιο της *bootstrap* ελαχιστοποιούμε την επίδραση των παραμέτρων κάνοντας την διαδικασία ποιο στιβαρή και μειώνοντας το σφάλμα (καθώς χρειάζεται να εκτιμήσουμε λιγότερα πράγματα).
- Δεν υπάρχουν πάντα *Pivotal* ελεγχοσυναρτήσεις.
- Συνήθως τα τεστ που βασίζονται σε *Pivotal* ελεγχοσυναρτήσεις έχουν μεγαλύτερη ισχή.

Παράδειγμα για ελεγχοσυνάρτηση

Έστω ότι προσομοιώνουμε από την εκθετική κατανομή με μέση τιμή θ και στη συνέχεια κάνουμε έλεγχο της μηδενικής υπόθεσης ότι η μέση τιμή είναι θ έναντι της εναλλακτικής ότι είναι διάφορη του θ. Χρησιμοποιώντας δύο ελεγχοσυναρτήσεις $T_1 = |\bar{x} - \theta|$ και $T_2 = \sqrt{n}\frac{\bar{x}-\theta}{s}$ αλλά και το ασυμπτωτικό αποτέλεσμα του κεντρικού οριακού θεωρήματος (δηλαδή το t-τεστ) μετράμε πόσες φορές απορρίψαμε τη μηδενική υπόθεση σε επίπεδο στατιστικής σημαντικότητας 5%. Από τον ορισμό και τον τρόπο που κατασκευάσαμε τους ελέγχους *bootstrap* περιμένουμε πως το ποσοστό θα είναι κοντά στο 5%.

Αποτελέσματα προσομοίωσης

		$B = 100$		$B = 500$		
	n	T_1	T_2	T_1	T_2	t-τεστ
	10	0.14740	0.07100	0.14250	0.06890	0.12880
	20	0.10550	0.06190	0.10680	0.06180	0.09750
$\theta = 1$	50	0.06900	0.05480	0.07430	0.05380	0.07120
	100	0.05830	0.05130	0.06100	0.05040	0.06060
	250	0.05250	0.05070	0.05460	0.05020	0.05320

	n	$B = 100$		$B = 500$		
	10	0.14320	0.06810	0.14380	0.06550	0.12640
	20	0.10790	0.06490	0.10400	0.05880	0.09440
$\theta = 3$	50	0.07750	0.05560	0.07540	0.05670	0.07320
	100	0.06530	0.05440	0.05790	0.05280	0.05650
	250	0.05540	0.05150	0.05080	0.05140	0.05190

Πίνακας: Το ποσοστό των φορών που απορρίψαμε τη μηδενική υπόθεση ότι η μέση τιμή του πληθυσμού είναι θ έναντι δίπλευρης εναλλακτικής υπόθεσης. Τα δεδομένα είχαν προσομοιωθεί από εκθετική κατανομή με μέση τιμή θ.

Έλεγχος καλής προσαρμογής με τη χρήση *Bootstrap*

Η παραμετρική *Bootstrap* είναι κατάλληλη για έλεγχο καλής προσαρμογής. Έστω ότι θέλουμε να ελέγξουμε την υπόθεση της κανονικότητας $H_0 : F = N(\mu, \sigma^2)$ ερσυς $H_1 : F \neq N(\mu, \sigma^2)$. Ένα πολύ γνωστό τεστ για τον έλεγχο κανονικότητα είναι το *Kolmogorov - Smirnov* τεστ $D = max(|\hat{F}_n(x) - F(x)|)$. Αυτό το τεστ έχει ασυμπτωματικά και κάτω από την μηδενική υπόθεση πινακοποιημένη κατανομή. Το τεστ που βασίζεται στην *Bootstrap* δεν χρησιμοποιεί τα ασυμπτωτικά αποτελέσματα αλλά γεννάμε δείγματα κάτω από τη μηδενική υπόθεση της κανονικότητας H_0 (αν οι παράμετροι δεν είναι γνωστοί χρειάζεται να τους εκτιμήσουμε). Για κάθε δείγμα υπολογίζουμε την τιμή του τεστ και κατασκευάζουμε την κατανομή του τεστ. Δεν χρειάζεται να χρησιμοποιήσουμε την D. Άλλες αποστάσεις μπορούν επίσης να χρησιμοποιηθούν προκειμένου να μετρήσουμε αποκλίσεις από την κανονικότητα! Το τεστ μπορεί να χρησιμοποιηθεί για οποιαδήποτε κατανομή!

Πότε η *Bootstrap* αποτυγχάνει

- Σε περιπτώσεις που έχουμε μικρό μέγεθος δείγματος (γιατί η $\hat{F}_n$ δεν είναι καλή προσέγγιση της F)
- Σε περιπτώσεις όπου εκτιμάμε ποσότητες οι οποίες δεν έχουν καλά ορισμένες ροπές. (Για παράδειγμα αν τα δεδομένα προέρχονται από την κατανομή *Cauchy* ξέρουμε πως η αναμενόμενη τιμή της κατανομής δεν υπάρχει. Επομένως αν προσπαθήσουμε να χρησιμοποιήσουμε *bootstrap* για να εκτιμήσουμε το τυπικό σφάλμα της δειγματικής μέσης τιμής είμαστε καταδικασμένοι να αποτύχουμε).
- Σε περιπτώσεις ύπαρξης εξάρτησης (π.χ. χρονοσειρές, χωρικά προβλήματα). Η *Bootstrap* βασίζεται στην υπόθεση της ανεξαρτησίας.

Πότε η *Bootstrap* αποτυγχάνει (συνέχεια)

- Σε περιπτώσεις που θέλουμε να εκτιμήσουμε ακραία ποσοστιαία σημεία (π.χ. το 99.99% ποσοστιαίο σημείο ή το $max(X_i)$).
- Σε περιπτώσεις ακραίων τιμών: Καθώς τότε δεν έχουμε καλή εκτίμηση της κατανομής F και προσθέτουμε μεταβλητότητα στις εκτιμήσεις μας.
- Μη λείες ποσόστητες
- Πολυμεταβλητά δεδμένα: Όταν οι διαστάσεις του προβλήματος είναι μεγάλες τότε η $\hat{F}_n$ γίνεται λιγότερο καλή εκτιμήτρια της F. Το γεγονός αυτό μπορεί να προκαλέσει προβλήματα.

Η επιλογή του B

Η επιλογή του B εξαρτάται από:

- Τη δυνατότητα σε υπολογιστές που υπάρχει
- Τον τύπο του προβλήματος: ενώ $B = 500$ μπορεί να επαρκεί για την εκτίμηση τυπικών σφαλμάτων, ίσως δεν είναι αρκετό για διαστήματα εμπιστοσύνης.
- Την πολυπλοκότητα του προβλήματος.

Παραλαγές της μεθόδου *Bootstrap*

Εαν έχουμε κάποια πληροφορία ή κάποιο λόγο να πιστεύουμε πως η κατανομή του πληθυσμού έχει κάποια συγκεκριμένη μορφή τότε μπορούμε να χρησιμοποιήσουμε παραμετρικό *bootstrap*.

Παραλαγές της μεθόδου *Bootstrap*: *Smoothed Bootstrap*

Χρησιμοποιούμε μια άλλη πιο λεία εκτίμηση της συνάρτησης πυκνότητας πιθανότητας. Ιδίως όταν η εμπειρική κατανομή είναι διακριτή και μπορεί να πάρει μόνο n διαφορετικές τιμές. Τότε αντί να πάρουμε δείγμα από την εμπειρική κατανομή (που είναι διακριτή) παίρνουμε δείγμα από μια συνεχή κατανομή (ποιο λεία) που την εκτιμά.

Παραλαγές της μεθόδου *Bootstrap*: Επαναληπτική μέθοδος *Bootstrap*

Έχοντας πάρει τις τιμές βοοτστραπ $\hat{\theta}_i^*$, τις χρησιμοποιούμε για να ξανακάνουμε *bootstrap*, κάνοντας πάλι τυχαία δειγματοληψία με επανάθεση από αυτές. Με αυτό τον τρόπο κάνουμε ακόμα πιο λεία την εκτίμηση και βελτιώνουμε την ακρίβεια του τυπικού σφάλματος. Η μέθοδος μπορεί να βελτιώσει την εκτίμηση στην περίπτωση μη λείων συναρτήσεων, όπως για παράδειγμα είναι η διάμεσος.

Παραλαγές της μεθόδου *Bootstrap*: *Bayesian Bootstrap*

Μια παραλλαγή της μεθόδου είναι και η χρήση της Μπευζιανής *bootstrap* προσέγγισης. Συγκεκριμένα αντί να δίνουμε σε κάθε παρατήρηση πιθανότητα $1/n$ αντιστοιχούμε σε κάθε παρατήρηση μια πιθανότητα, έστω g_i η οποία έχει διάμεσο την τιμή $1/n$ αλλά μεταβάλλεται από επανάληψη σε επανάληψη.
Προσομοιώνουμε $n-1$ τυχαίες μεταβλητές από την ομοιόμορφη κατανομή στο διάστημα (0,1) και τις διατάσσουμε. Έστω πως $u_{(1)}, u_{(2)}, \ldots, u_{(n-1)}$ οι διατεταγμένες αυτές τιμές. Έστω $u_{(0)} = 0, u_{(n)} = 1$. Υπολόγισε τις ποσότητες $g_i = u_{(i)} - u_{(i-1)}$.

Παραλαγές της μεθόδου *Bootstrap*: *Bayesian Bootstrap* (συνέχεια)

Το *Bootstrap* δείγμα θα δημιουργηθεί με δειγματοληψία με επανάθεση αλλά οι πιθανότητες κάθε παρατήρησης δεν θα είναι το p_i αλλά τα g_i που υπολογίσαμε. Μάλιστα για το επόμενο *bootstrap* δείγμα θα πρέπει να ξαναδημιουργήσουμε το διάνυσμα με τα g_i και αυτό θα επαναλαμβάνεται σε κάθε δείγμα. Μπορεί κανείς εύκολα να δει πως η αναμενόμενη τιμή της πιθανότητας κάθε παρατήρησης είναι ίση με $1/n$. Η μέθοδος χρησιμοποιείται σε Μπευζιανές εφαρμογές καθώς η κατανομή της παραμέτρου θ που δημιουργούμε με τις *bootstrap* επαναλήψεις μπορεί να θεωρηθεί ως εκ των υστέρων κατανομή κατά τη Μπευζιανή προσέγγιση.

Άλλα σχήματα δειγματοληψίας: Η μέθοδος *Jackknife*

Η μέθοδος όπως έχουμε ήδη πει αρχικά παρουσιάστηκε ώς τεχνική μείωσης της μεροληψίας. Με $\hat{\theta}_{(i)}$, συμβολίσαμε την εκτίμηση που πέρναμε όταν η όλες οι παρατηρήσεις εκτός από την i χρησιμοποιούνται για εκτίμηση.
Με τη μέθοδος *Jackknife* αγνοούμε μία παρατήρηση κάθε φορά. Τα δείγματα δημιουργούνται χωρίς επανάθεση. Έτσι έχουμε n δείγματα κάθε φορά

Άλλα σχήματα δειγματοληψίας: *Subsampling*

Μπορούμε να γενικεύσουμε τη λογική της μεθόδου *Jackknife* αγνοόντας $b \geq 1$ παρατηρήσεις κάθε φορά. Θα προκύψουν αντίστοιχες φόρμουλες. Πλήρης απαρίθμηση είναι πλέον δύσκολη: χρησιμοποιείστε μεθόδους *Monte Carlo* πέρνοντας δείγμα από αυτές. Αυτή είναι η ιδέα του *Subsampling*. *Subsamples* είναι δείγματα από την F και όχι από την $\hat{F}_n$. Σε μερικές περιπτώσεις που η *bootstrap* αποτυγχάνει μπορεί να δόσει λύση. Τα *Subsamples* είναι μικρότερου μεγέθους και μπορεί να χρειαστεί να τα κάνουμε *rescale* (θυμηθείτε τον παράγοντα $n-1$ στο *se* στην περίπτωση της μεθόδου *Jackknife*). Χωρίζουμε το αρχικό δείγμα σε μικρότερα δείγματα ίσου μεγέθους.