## Το Jackknife και το Βootstrap

### Βασικά (basic ή pivotal) bootstrap δ.ε.:

* Εδώ παίρνουμε ως .
* Η κατανομή της , , αντί να προσεγγιστεί με κάποια , εδώ **εκτιμάται από την αντίστοιχη bootstrap-εμπειρική** συνάρτηση κατανομής:

,

* + - Όπου  και  με 
* Τότε το  δ.ε. για το  θα ήταν το

,

* + - όπου  και ,
		- καθώς  είναι το  ποσοστιαία σημείο της δειγματικής κατανομής της , άρα ισούται με την  μικρότερη παρατήρηση , την 
* Άρα τελικά τα δ.ε. για το  θα είναι το

.

***Bias-correction.*** Αυτό δίνει διορθωμένα δε.:

,

Με , όπου , 

Ή ισοδύναμα:



Με (υπενθυμ.)  και 

**Θεώρημα: *για να έχει το παραπάνω δ.ε. για την ονομαστική πιθανότητα κάλυψης  αρκεί η  να είναι παραγωγίσιμη κατά Hadamard. (την ίδια υπόθεση χρειαζόμαστε για τη θεωρία των συναρτήσεων επιρροής)****.*



* **Asymptotic normal:** Dashed, red = cdf of normal with
	+ mean= mean(sample)=56.2
	+ stdev= se of sample /sqrt(n) =14.1
	+ confidence intervals = 



* **Normal bootstrap (vboot)** dotted, blue = cdf of normal with
	+ mean= mean(sample)=56.2
	+ stdev= **se of bootstrap sample of means** = sqrt(v.boot) = 13.17
	+ confidence intervals = 



* **Basic Bootstrap** continuous = cdf of centered bootstrap sample
	+ confidence intervals = 



### Student Pivotal bootstrap (ή t- bootstrap) διαστήματα εμπιστοσύνης.

* **Ακριβέστερη κάλυψη** (**καλύτερη προσέγγιση στην ονομαστική**) παίρνουμε για πεπερασμένο n ή αν χρησιμοποιήσουμε ως στατιστικό το , όπου  μια εκτίμηση της τυπικής απόκλισης της , π.χ. η .
* *Προσεγγίζουμε* την κατανομή 

με την δειγματική bootstrap-κατανομή  της ,

όπου  είναι μια εκτίμηση της τυπικής απόκλισης της  (και όχι της ).

* Αν  τότε το δ.ε. είναι :

,

 όπου είναι το  δειγματικό ποσοστιαία σημείο του , δηλαδή η  μικρότερη από τις  δηλαδή τέτοια ώστε .

***Bias-correction.*** Αυτό δίνει διορθωμένα δε.:

,

Με , oπου , 

Υπενθυμ.  και 

* Το πρακτικό μειονέκτημα αυτής της μεθόδου είναι ότι εκτός από την  χρειαζόμαστε να υπολογίσουμε και **τις  δηλαδή εκτιμήτριες της διακύμανσης της  για κάθε bootstrap δείγμα**.
	+ - Αυτό είναι εύκολο αν υπάρχει θεωρητικός τύπος για τη διακύμανση αυτή π.χ. με βάση τη θεωρία για τις συναρτήσεις επιρροής.
		- Αν όχι πρέπει να κάνουμε δεύτερο bootstrap (ή jackknife) για κάθε  bootstrap δείγμα ώστε να προσεγγίσουμε την  με bootstrap. Αν αυτό το δεύτερο bootstrap χρησιμοποιούσε  δείγματα θα χρειαζόμασταν συνολικά bootstrapδείγματα. [double bootstrap].
* Ένας τρόπος να δει κανείς γιατί αυτή η μέθοδος είναι **καλύτερη από το βασικό bootstrap** είναι η σύγκριση με την ασυμπτωτική θεωρία: για μικρό δείγμα η χρησιμοποίηση ποσοστιαίων σημείων της  κατανομής αντί για εκείνα της κανονικής στη κατασκευή δ.ε. είναι ακριβέστερη διότι η  παίρνει υπόψη της καλύτερα το γεγονός ότι εκτιμήθηκε η διακύμανση: έτσι δίνει λίγο μεγαλύτερα δ.ε. για μικρά n. Και αυτά όμως υποθέτουν εγγύτητα της κατανομής της  με τη κανονικότητα.
* *Τα bootstrap-student δ.ε. έμμεσα προσπαθούν να πάρουν υπόψη και το πεπερασμένο n και αποκλίσεις από τη κανονικότητα εκτιμώντας τα κατάλληλα ποσοστιαία σημεία της κατανομής του στατιστικού μας για τη συγκεκριμένη κατανομή από την οποία προήλθαν οι παρατηρήσεις μας.*

**Παράδειγμα**. Για τα δεδομένα του control group των δεδομένων χρόνου επιβίωσης των ποντικών και



Μέρες διάρκειας ζωής μετά από χειρουργική επέμβαση.

Με μέσος παίρνουμε  και 

**Τα ποσοστιαία σημεία** είναι 

που **δίνουν τα ακόλουθα δ.ε.** (95%).

Σημειώστε ότι τα **bootstrap-t ποσοστιαία σημεία δεν είναι συμμετρικά** (όπως αυτά της t και της N κατανομής), καθώς «προσπαθούν» να αναπαράγουν ασυμμετρίες στη κατανομή του πληθυσμού μας. Επομένως ούτε τα δ.ε. είναι συμμετρικά γύρω από το .



* **Asymptotic Student**  Dashed = cdf of normal with
	+ mean= mean(sample)=56.2
	+ stdev= se of sample /sqrt(n) =14.1
	+ confidence intervals = 



* **Student Bootstrap** continuous = cdf of centered bootstrap sample
	+ confidence intervals = 



Tα t-bootstrap δ.ε. είναι ακριβέστερα από τα προηγούμενα (βλέπε παρακάτω) αλλά έχουν δύο ζητήματα:

* Χρειαζόμαστε **τύπο για την se της εκτιμήτριας**. **Αν αυτός δεν υπάρχει** πρέπει να εκτιμήσουμε την  για κάθε bootstrap-δείγμα π.χ. με **double bootstrap**.
	+ -





* + - Έτσι π.χ. σε παράδειγμα  δεδομένων , όπου  είναι ο συντελεστή συσχέτισης ρ, και όπου πήραμε .



Για κάθε ένα από τα B=1000 bootstrap-δείγματα χρησιμοποιήθηκαν 25 νέα bootstrap-δείγματα για την εκτίμηση της . Αυτό έδωσε 90% δ.ε.  για το .

* Ένα δεύτερο ζήτημα είναι ότι για **πολύ μικρά και μη κανονικά δείγματα** τα t-bootstrap δ.ε. **μπορεί να έχουν μεγάλη διακυμαντικότητα**.
	+ - Εδώ ***βολεύει αν γνωρίζουμε μετασχηματισμό που κανονικοποιεί την εκτιμήτρια μας και ταυτόχρονα σταθεροποιεί τη διακύμανση***.
		- Π.χ. για το συντελεστή συσχέτισης είναι γνωστό ότι ανκαι τότε . (Σημείωση: ο μετασχηματισμός είναι μονότονος).



* + - Μπορεί λοιπόν κανείς να κατασκευάσει δ.ε. για το  και ακολούθως να τα γυρίσει με τον αντίστροφο μετασχηματισμό σε δ.ε. για το .

Στα δεδομένα μας αυτό θα έδινε το 90% δ.ε.  για το . Σημειώστε ότι τα δ.ε. δεν είναι τα ίδια!

**Παρένθεση: Vboot and theoretical variance**

****

**Τα t-bootstrapδιαστήματα εμπιστοσύνης εξαρτώνται από τη κλίμακα (μετασχηματισμού) που χρησιμοποιεί κανείς.**

**Δυστυχώς ένας «καλός» μετασχηματισμός δεν είναι πάντα διαθέσιμος.**