## Το Jackknife και το Βootstrap

### Το bootstrap

* Το bootstrap είναι μια μέθοδος που **σκοπό έχει να εκτιμήσει τη διακύμανση ή και γενικότερα τη κατανομή ενός στατιστικού** , όπως π.χ. μια εκτιμήτρια . Τον ίδιο στόχο είχαν άλλωστε και το Jackknife, καθώς και η ασυμπτωτική θεωρία με τις συναρτήσεις επιρροής (influence functions).
* Η ιδέα και στο bootstrap όπως και στο Jackknife είναι: **να μελετήσουμε τη μεταβλητότητα μιας εκτιμήτριας όταν γίνονται «μικρές αλλαγές» στο δείγμα μας.**

Οι Casella & Berger (2002) σημείωσαν για το bootstrap ότι «ίσως είναι η μοναδική πιο σημαντική εξέλιξη στη στατιστική μεθοδολογία των τελευταίων ετών»

*Δες υλικό Παπασταμούλη σελ. 4*

Έστω λοιπόν  και . Υποθέτουμε iid.

* Η διακύμανση και η αναμενόμενη τιμή της  εξαρτώνται από την 



όπου 

* H ***ιδέα του bootstrap*** είναι να εκτιμήσουμε την  και την  αντικαθιστώντας στη παραπάνω σχέση την  με την 

 .

όπου 

Αυτό ονομάζεται το «**ιδανικό bootstrap**»

Τι είναι δηλαδή το «**ιδανικό bootstrap**»;

Υπενθύμ. 

όπου 

* Καθώς η  είναι η cdf διακριτής τυχαίας μεταβλητής που παίρνει τις τιμές  (το δείγμα των που παρατηρήσαμε), τη κάθε μια με πιθανότητα ,
* **η αναμενόμενη τιμή ως προς ** είναι σα να λέμε ότι παίρνουμε δείγμα  από την **** , δηλαδή να επιλέγονται τα , ανεξάρτητα μεταξύ τους, από τα , κάθε ένα τιμή με πιθανότητα , **με** επανατοποθέτηση.
* Ακολούθως παίρνουμε το μέσο όρο ως προς όλες τις πιθανές εκβάσεις των 

,

όπου  όλες οι δυνατές (επαναληπτικές) διατάξεις του αρχικού μας δείγματος .

Υπολογιστικά είναι αδύνατο να επιτευχτεί αυτό καθώς το  είναι πολύ μεγάλο.

Υπολογίζεται όμως **μια εκτίμηση του με ολοκλήρωση κατά “Monte Carlo”:**

* Πάρε  δείγματα  μεγέθους  από την :
	+ κάθε τέτοιο δείγμα είναι μια τυχαία (επαναληπτική) διάταξη του αρχικού μας δείγματος 
	+ στην οποία κάθε  συμπεριλαμβάνεται (με επανατοποθέτηση ) με πιθανότητα 
	+ είναι λοιπόν ένα δείγμα που λαμβάνεται με απλή τυχαία δειγματοληψία με επανατοποθέτηση από το αρχικό μας δείγμα , εξ ου και ο χαρακτηρισμός “**resampling**”).
* Υπολόγισε τα  τις bootstrap επαναλήψεις (replications)
* Υπολόγισε την 

Σύμφωνα με το νόμο των μεγάλων αριθμών  όταν 

*Δες υλικό Παπασταμούλη σελ. 14*

* **Σχηματικά**:



* bootstrap κόσμος μιμείται τον πραγματικό, με την ελπίδα
* η *μεταβλητότητα* της  γύρω από την  **στον bootstrap κόσμο να μοιάζει με την**
* *μεταβλητότητα* της  γύρω από την  **στον πραγματικό κόσμο.**

*Δες υλικό Παπασταμούλη σελ. 10*

* Η «απόσταση» πραγματικού και bootstrap κόσμου είναι τάξης 

Δηλαδή οι ποσότητες όπως π.χ. διακύμανση  από τις αντίστοιχες του δειγματικού κόσμου  εκτιμώνται με δ.ε. που έχουν μήκος ένα πολλαπλάσιο το 

* Όμως η απόσταση μεταξύ «ιδανικής » και (προσομοιωμένης από B bootstrap δείγματα» είναι δείγματα τάξης : μπορούμε να τη κάνουμε όσο μικρή θέλουμε δημιουργώντας περισσότερα bootstrap δείγματα.

Δείγμα το controlgroup των ποντικών από το παράδειγμα των Efron & Tibshirani.

Ημέρες επιβίωσης =10,27,30,40,46,51,52,104,146

**Εκτιμώμενη πυκνότητα της**

**bootstrap-κατανομής του δειγματικού μέσου (μαύρη) και**

**της θεωρητικής κατανομής του, με δειγματική τυπ,. απόκλ. (κόκκινη)**



#### Παραμετρικό bootstrap:

* Στο bootstrap που είδαμε παραπάνω, εκτιμήσαμε την  μη-παραμετρικά με την  (την ecdf).
* Αν όμως θέλουμε να κάνουμε παραμετρική υπόθεση, ότι δηλαδή το  που γέννησε τα δεδομένα περιγράφεται παραμετρικά ως  τότε μπορούμε να κάνουμε **«παραμετρικό bootstrap»:**
	+ - Προκειμένου να εκτιμήσουμε την  θα γεννούσαμε παρατηρήσεις από την , όπου  εκτιμήτρια του  , αντί να γεννήσουμε από την  .
		- Προφανώς, σε αυτή τη περίπτωση οι όποιες εκτιμήτριες, διαστήματα εμπιστοσύνης κλπ. έχουν αξία μόνο αν πληρούνται οι παραμετρικές υποθέσεις μας.

### Διαστήματα εμπιστοσύνης με (μη παραμετρικό) Bootstrap

Επειδή υπάρχουν πολλοί διαφορετικοί τρόποι κατασκευής δ.ε. με bootstrap, ας πιάσουμε το θέμα από τα βασικά, ώστε να εξηγήσουμε τις διαφορές μεταξύ αυτών των τρόπων:

Η κατασκευή διαστημάτων εμπιστοσύνης π.χ. για μια παράμετρο , τα οποία να έχουν επιθυμητή (ή ονομαστική) πιθανότητα κάλυψης , **στηρίζεται συνήθως σ’ ένα στατιστικό** , το οποίο μπορεί να περιλαμβάνει π.χ. την  καθώς και μια εκτιμήτρια της π.χ. την και **του οποίου στατιστικού είναι γνωστή** (και άρα δεν εξαρτάται από άγνωστες παραμέτρους) **η κατανομή**



* Για παράδειγμα θα μπορούσε το στατιστικό να είναι το ,

όπου  μια εκτιμήτρια της .

* τότε από έπεται ότι:









Άρα: 

είναι ένα δ.ε. στο οποίο το  θα ανήκει με πιθανότητα ακριβώς .

Συνήθως είναι **δύσκολο όμως να γνωρίζουμε** σε ένα δοσμένο πρόβλημα ένα **στατιστικό , του οποίου είναι γνωστή η κατανομή**  και χωρίς να εξαρτάται από άγνωστες παραμέτρους.

Γι αυτό το λόγο, αντί για το παραπάνω «ακριβές» δ.ε.

θα μπορούσαμε π.χ. **να προσεγγίσουμε την  με την , την ασυμπτωτική  κατανομή του στατιστικού μας,**

αν αυτή είναι γνωστή και αν δεν εξαρτάται από άλλες (άγνωστες) παραμέτρους. Συχνά αυτή η  είναι κανονική κατανομή και έτσι κατασκευάζονται τα:



### Κανονικά bootstrap διαστήματα εμπιστοσύνης:

Αν π.χ. , τότε λαμβάνουμε (ασυμπτωτικά) bootstrap δ.ε. ως

,

όπου και ,  η bootstrap–εκτιμήτρια της , 

**Πλεονέκτημα έναντι της**  που στηρίζεται στις συναρτήσεις επιρροής: **δεν χρειάζεται ο υπολογισμός τύπου για την .** Η  εκτιμάται από το bootstrap.

***Και στις δύο περιπτώσεις όμως χρησιμοποιείται η υπόθεση ότι η κατανομή του  μπορεί να προσεγγιστεί από τη κανονική***, ***που για μικρά δείγματα μπορεί να μην ευσταθεί.***

***Bias-correction.*** Όπως και στο Jackknife και το bootstrap μας επιτρέπει να υπολογίσουμε μια διόρθωση μεροληψίας για την εκτιμήτρια μας. Αυτή υπολογίζεται ως

 και 

Αυτό δίνει διορθωμένα δε.:





* **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 = 

