## Μη παραμετρική εκτίμηση πυκνότητας.

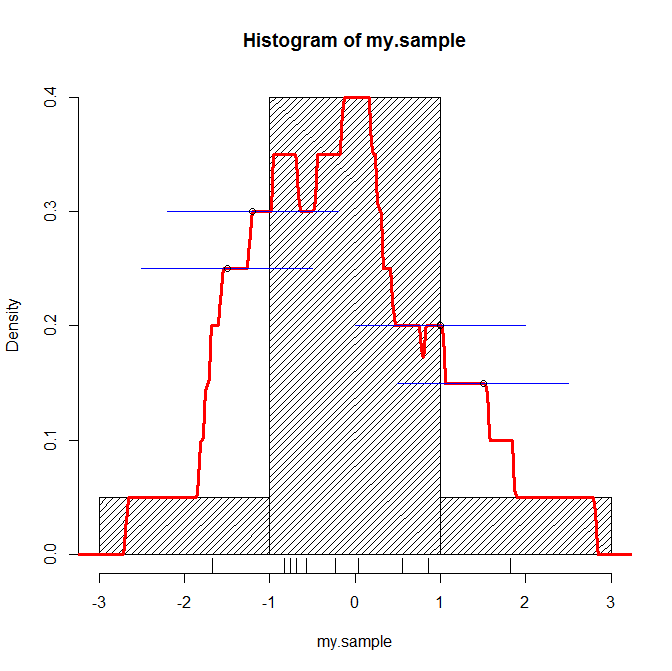
### Καλύτερη εκτιμήτρια της : η εκτιμήτρια βασισμένη σε πυρήνα (kernel estimator)

Έστω ότι θέλουμε να εκτιμήσουμε την  σε κάποιο (θα το κάνουμε για «όλα» τα , η τουλάχιστον για ένα πλέγμα από ).

Ορίζουμε την  **παρόμοια με την εκτιμήτρια του ιστογράμματος, αλλά με διάστημα κεντραρισμένο στο **



Δηλαδή: το ποσοστό των παρατηρήσεων  που απέχουν από το  κατά λιγότερο από  (διαιρεμένο με ).





Παρατηρείστε ότι αν θέταμε , δηλαδή



τότε 



και από



παίρνουμε



**Για να πάρουμε συνεχή συνάρτηση** μπορούμε να χρησιμοποιήσουμε **πυρήνες συνεχείς** όπως του

Epanechnikov ή τον tri-cube 

Βλέπε π.χ.Wassermann, Figure 4.10

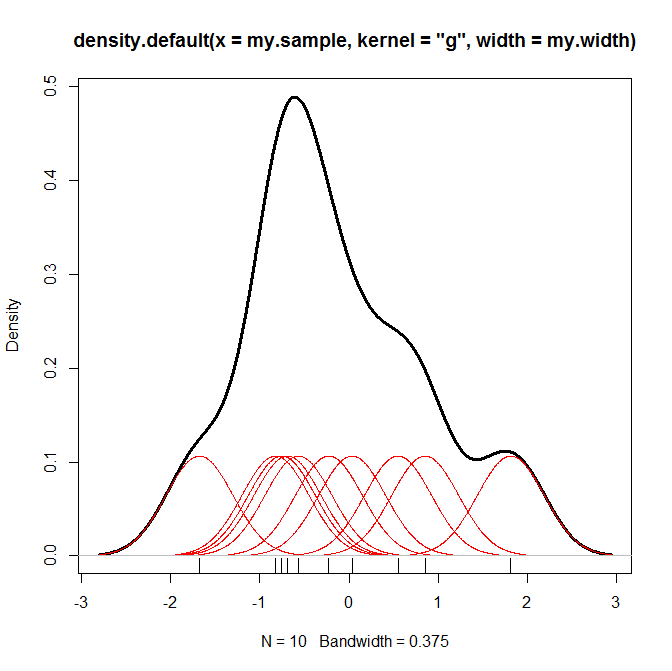


**Οι πυρήνες θέλουμε να έχουν τις εξής ιδιότητες:**

* συμμετρικός γύρω από το 0
* 
* 

Την εκτιμήτρια  μπορούμε να τη φανταστούμε και ως

* άθροισμα των πυρήνων σε κλίμακα που να έχουν φορέα το , διαιρεμένων με  και κεντραρισμένων στις παρατηρήσεις :
  + **κάθε παρατήρηση** «φορτώνεται» μάζα , η οποία απλώνεται γύρω της σε απόσταση h, όπως περιγράφει ο πυρήνας.
  + Αθροίζοντας αυτές τις μάζες παίρνουμε την : Βλέπε Wassermann, Figure 6.4 & 6.5.





Το **h** παίζει πάλι το ρόλο παραμέτρου εξομάλυνσης.

Όσο μεγαλύτερο το h τόσο πιο ομαλή είναι η εκτιμήτρια (μικρότερη ), αλλά μεγαλύτερο ).





#### Μεροληψία της εκτιμήτριας

**Πρόταση** (bias ). Υποθέτουμε ότι

* Η είναι 3 φορές παραγωγίσιμη , με 
* Ο πυρήνας Κ είναι συμμετρικός και έχει την ιδιότητα ότι 

Τότε

 και



Απόδ.









Όμως είναι ,  και 

Άρα: 

#### Διακύμανση της εκτιμήτριας

Παρόμοια μπορεί να δείξει κανείς ότι





#### Συνέπεια και μέσο τετραγωνικό σφάλμα της εκτιμήτριας

**Θεώρημα (συνέπεια):** (όπως για το ιστόγραμμα έχουμε ότι) Υπό τις παραπάνω υποθέσεις

Αν και , **τότε** η  είναι συνεπής: 

**Θεώρημα (IMSE)**: αν η είναι 3-φορές παραγωγίσιμη με και συμμετρικός με  τότε έχουμε



### Θεωρητικά βέλτιστο h

Παραγωγίζοντας την ως προς h και θέτοντας τη παράγωγο ίση με 0 παίρνουμε ότι:

Για το θεωρητικά βέλτιστο  ισχύει ότι , με  , και . Άρα το ελάχιστο IMSE δίνεται από 

Αυτό είναι καλύτερο από το  του ιστογράμματος  αλλά χειρότερο rate από τη παραμετρική 

**Απόδ.**

* 



* 
* 

### Πρακτική επιλογή της h

Στο προηγούμενο τύπο άγνωστο είναι το . Η εκτίμηση της  είναι ακόμα πιο προβληματική απ’ ότι της . Αν όμως πιστεύαμε ότι η  είναι αρκετά ομαλή θα μπορούσαμε εδώ να αντικαταστήσουμε την πυκνότητα  της κανονικής.

Τότε παίρνουμε . Εδώ το  εκτιμάται συνήθως με , όπου  η δειγματική διακύμανση και  η διαφορά πρώτου και τρίτου τεταρτημορίου:  εκτιμάει το  υπό την υπόθεση της κανονικότητας.

Αυτή η επιλογή της h είναι το λεγόμενο **Normal reference rule**

**Καλύτερη επιλογή: Cross-validation (όπως στο ιστόγραμμα)**

Επέλεξε το  που ελαχιστοποιεί την

,

όπου  η εκτιμήτρια που παίρνουμε παραλείποντας την παρατήρηση 

Και εδώ η ** εκτιμάει αμερόληπτα την ** καθώς:



λόγω ανεξαρτησίας της  από την .

Για διευκόλυνση των υπολογισμών , μπορεί να χρησιμοποιήσει κανείς ότι

 ,

όπου  και 

Π.χ. αν 

Υπολογίζει κανείς λοιπόν την  για ένα πλέγμα τιμών  και **διαλέγει ως  αυτήν που δίνει το βέλτιστο .**

Αποδεικνύεται ότι αν  **η με Cross-validation επιλεγείσα bandwidth συγκλίνει στην ιδανική**  **με ταχύτητα **! (πολύ αργή) : 

Αυτή όμως η ταχύτητα είναι και η βέλτιστη δυνατή!

Είναι δε αρκετή για να αποδείξει κανείς την **ασυμπτωτική αποτελεσματικότητα της CV στην επιλογή της h**:



Δηλαδή η επιλογή της h με CV οδηγεί σε ένα IMSE που είναι το ελάχιστο δυνατό!!

### Διαστήματα εμπιστοσύνης για και .

Έστω  

Τότε για σταθερό  έχουμε για  με 



όπου 

Άρα ένα  δ.ε. για το  θα ήταν το

 , με 

Εναλλακτικά, τα ίδια δ.ε γράφονται και ως 

(υπάρχουν και καλύτερες εκτιμήτριες , βλέπε π.χ. Shao & Tu

* Αν θέλαμε αυτά να είναι και διαστήματα εμπιστοσύνης για το ίδιο το 
* θα έπρεπε να έχουμε ,
* που ικανοποιείται όταν  δηλαδή όταν 
* Αυτό σημαίνει : όταν το ** τείνει τόσο γρήγορα στο 0 που η Bias είναι αμελητέα** σε σχέση με τη διακύμανση.
* Όμως αυτό δεν ισχύει για τη βέλτιστη Bandwidth , η οποία «ισορροπεί» Bias και variance.
  + - Γι αυτό όταν  μιλάμε για under smoothing.
  + Εναλλακτική: κατασκευή δ.ε. για το  στη βάση της χρησιμοποίησης εκτίμησης της 
    - Παρ’ ότι υπάρχουν μέθοδοι για την εκτίμηση της μεροληψίας, , ασυμπτωτική θεωρία και προσομοιώσεις δείχνουν ότι η καταλήγει να έχει χειρότερη πιθανότητα κάλυψης από το “undersmoothing”.

### Ζώνες εμπιστοσύνης για και .

Για τη κατασκευή **ζωνών εμπιστοσύνης για την ** (η και για  όταν έχουμε under smoothing) χρειαζόμαστε τα α-ποσοστιαία της κατανομής της



Υπάρχει ασυμπτωτική θεωρία, αρκετή σύνθετη όμως. Εναλλακτική: η χρήση bootstrap για την εκτίμηση των ποσοστιαίων σημείων αυτής της κατανομής:

Στο Joun , M (1988) “Bootstrapping density estimates”, *Comm in Stat. Theory and Methods*, 17, p. 61-78, προτείνεται η εξής προσέγγιση:

* Έστω  bootstrap δείγμα από  και
*  η εκτιμήτρια της  βασισμένη στα .

Τότε, με 

(θέτοντας ) και υπό με  έχουμε:

, για 

Άρα εδώ το bootstrap «δουλεύει»

Άρα με  όπου  η δειγματική κατανομή της  τα δ.ε. για την  θα ήταν τα (αντικαθιστώντας  με  στην )

**ταυτόχρονα για όλα τα **

Προτείνεται και η «καλύτερη» (?) εναλλακτική του δ.ε.



Πάντως τα παραπάνω ισχύουν **για προκαθορισμένο ** και δεν παίρνουν υπόψη τη μεταβλητότητα της επιλογής της  με κάποιο αυτόματο κριτήριο (π.χ. CV).