##  Μη παραμετρική εκτίμηση πυκνότητας.

* Έστω  ανεξάρτητες και ισόνομες από cdf , με πυκνότητα . Γράφουμε και 
* Θα υποθέσουμε ότι η ως καμπύλη είναι κατά κάποιο τρόπο (?) «**ομαλή**» ή «**λεία**». Έτσι η εκτιμήτρια της θα είναι και αυτή «ομαλή».
* Οι μέθοδοι κατασκευής τέτοιων ομαλών εκτιμητριών καμπυλών, γνωστές και ως **«τεχνικές εξομάλυνσης» (smoothing techniques)** είναι κοινές με άλλα προβλήματα στη μη–παραμετρική στατιστική, όπως στη μη παραμετρική παλινδρόμηση. Κοινά είναι και τα προβλήματα που ανακύπτουν: συνήθως ο χρήστης επιλέγει τη τιμή μιας **παραμέτρου που ορίζει την «ομαλότητα»** της εκτιμήτριας έτσι ώστε να εξισορροπεί μεροληψία και διακύμανση και οι ταχύτητες σύγκλισης (rates) που επιτυγχάνονται είναι χειρότερες από τις παραμετρικές.
* Παράδειγμα από **Wasserman Example 4.3**: Sloan digital sky survey (SDSS).
	+  η απόσταση του  γαλαξία που συναντάμε σε ορισμένη κατεύθυνση από τη Γη ,
		- μετρούμενη από τη μετατόπιση του χρώματός του προς το κόκκινο (“redshift”),
		- που είναι ανάλογη της ταχύτητα απομάκρυνσής του από τη Γή. (Βλέπε Wasserman Figure 4.2.)
	+ Στόχος η κατανόηση της κατανομής των γαλαξιών και πως εξελίσσονται στο χρόνο.



* Πρώτη εκτιμήτρια: **Ιστογράμματα**. Βλέπε Wasserman Figure 4.1.



* + **Κρίσιμη η επιλογή της παραμέτρου εξομάλυνσης** (smoothing parameter) που **είναι εδώ ο αριθμός των διαστημάτων του ιστογράμματος ή ισοδύναμα το εύρος τους.**
	+ Πολύ **λίγα διαστήματα**: κρύβονται υπάρχουσες δομές, και άρα *σημαντικό Bias*. (**Over smoothing**)
	+ Πάρα **πολλά διαστήματα**: εμφανίζονται τυχαίες αποκλίσεις ως «υπάρχουσες» δομές, που οφείλεται στη *μεγάλη διακύμανση*. (**Under smoothing**).
	+ **Στόχος**: η επιλογή της παραμέτρου εξομάλυνσης h (ή του αριθμού των διαστημάτων) έτσι **ώστε να ελαχιστοποιείται το μέγιστο τετραγωνικό σφάλμα**

.

Μέθοδος για να το πετύχουμε Cross-validation (Βλέπε παρακάτω).

* Καλύτερη εκτιμήτρια: **Εκτιμήτρια βασισμένη σε πυρήνες**. (kernel estimator, αργότερα…) Βλέπε Wasserman Figure 4.2.



### Ιστόγραμμα

* Έστω να παίρνουν τιμές στο .
* Για έναν αριθμό διαστημάτων **m** (που επιλέγουμε εμείς ) θέτουμε  το εύρος των διαστημάτων και σπάμε το  σε m διαστήματα  εύρους . 
* Έστω  και 
* Εκτιμάμε την για με μια συνάρτηση  σταθερή στο, που θα έδινε .
* Άρα 

Τότε:

*  αυξάνεται με το h
* μικραίνει με το h

From Patrick Breheny’s course at the university of Kenducky



#### Το θεωρητικά βέλτιστο h:

Το h ή το m παίζει το ρόλο της παραμέτρου εξομάλυνσης. Πως επιλέγεται το «βέλτιστο»; Χρειάζεται να εισάγουμε **κριτήριο «σφάλματος»**

* Έστω



η (ολοκληρωμένη) τετραγωνική συνάρτηση σφάλματος , αν συμφωνήσουμε ότι μετράμε το σφάλμα με τετραγωνική απόσταση.

* Και το μέσο (ολοκληρωμένο) τετραγωνικό σφάλμα (integrated mean squared error, IMSE)





**«Kεντρικό» πρόβλημα σε πολλά παραμετρικά προβλήματα**

* Αν πολύ μεγάλο over smoothing: μεγάλα bias
* Αν  πολύ μικρόunder smoothing: μεγάλη διακύμανση

Βλέπε Wassermann, Figure 4.9 για προσδιορισμό του θεωρητικά βέλτιστου .



### Το IMSE της

#### Η μεροληψία του Ιστογράμματος

#### Πρόταση. Αν  τότε

Ακριβέστερα μπορεί να δειχτεί ότι αν  και  τότε



Απόδ. Για 





 με

άρα 

#### Η διακύμανση του Ιστογράμματος

Αν 



Αλλά μόλις είδαμε ότι . Άρα:





**Θεώρημα:** αν  και . Τότε





Αυτό ελαχιστοποιείται για . Για αυτό το  θα έχουμε 

**Απόδ.**

* 



* 
* 

**Πρόταση (συνέπεια):** αν  και , , τότε η  (είναι στοχαστικά συνεπής)

σημείωση:

-εξασφαλίζει ότι ,ενώ

-εξασφαλίζει ότι 

### Πρακτική επιλογή της βέλτιστης h: Cross-Validation

Στόχος (ιδανικός) η επιλογή του h που ελαχιστοποιεί την



Αυτό ισούται με



και η ελαχιστοποίηση αυτής της ποσότητας ως προς h ισοδυναμεί με ελαχιστοποίηση της

,

καθώς ο τελευταίος όρος δεν εξαρτάται από h.

* Καθώς και αυτή η ποσότητα περιέχει το άγνωστο , αντ’ αυτής **ελαχιστοποιούμε μια αμερόληπτη εκτιμήτρια της**: την

,

όπου  η εκτιμήτρια στη κατασκευή της οποίας έχουμε παραλείψει την i-παρατήρηση .

**Αιτία:**

 Η είναι αμερόληπτη για :.

Διότι, με 

,

 λόγω ανεξαρτησίας της  από την .

Για το ιστόγραμμα ο υπολογισμός αυτής της ποσότητας είναι ιδιαίτερα εύκολος, καθώς ισούται με:

.

Άρα υπολογίζουμε την  για ένα κατάλληλο «πλέγμα» σημείων , π.χ.  και βρίσκουμε εκείνο το h που ελαχιστοποιεί την .

### Διαστήματα εμπιστοσύνης για f

* Θα κατασκευάσουμε δ.ε. για την ,

τη προσέγγιση στην  που είναι σταθερή στα διαστήματα , ώστε να αποφύγουμε το ζήτημα της μεροληψίας .

Κατ’ αρχάς για σταθερό και , έχουμε



και άρα ένα  δ.ε. για το θα ήταν το

.

Για να έχουμε ένα **ζώνη εμπιστοσύνης ταυτόχρονα για όλα τα ** θα μπορούσαμε να κατασκευάσουμε τα διαστήματα **σε κάθε  σε επίπεδο ** δηλαδή ως:

.

Τότε η πιθανότητα να ανήκει η  στη παραπάνω ζώνη **για όλα τα  ταυτόχρονα** είναι .

**Για μικρά ** θα μπορούσαμε να κατασκευάσουμε δ.ε. για την  στηριζόμενοι στη διωνυμική κατανομή, αντί για τη κανονική προσέγγιση:

Αν  και  τότε .

Μπορουν λοιπόν να κατασκευαστούν είτε τα ακριβή δ.ε. οιυ στηρίζονατι στη διωνυμική κατανομή (Clopper-Pearson”) είτε π.χ. αυτά που στηρίζοαντα στο μετασχηματισμό, για τον οποίο ισχύει (βλ. Κεφ. 2):



Έτσι το  δ.ε. για το  θα ήταν  και άρα το δ.ε. για το  θα ήταν 

Άρα δ.ε. για το  για  θα ήταν



Βλέπε Figures από slides του Breheny

#### Κατασκευή ζωνών εμπιστοσύνης με bootstrap

Για τη κατασκευή ζωνών εμπιστοσύνης για την  (ταυτόχρονα για όλα τα ) χρειαζόμαστε τα α-ποσοστιαία της κατανομής των (με ):

 και 

Ta ποσοστιαία σημεία αυτών των κατανομών μπορούν να εκτιμηθούν με τη χρήση bootstrap:

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

 και 

και υπό με  έχουμε:

 , και αντίστοιχα για , για 

**Άρα εδώ το bootstrap «δουλεύει».**

Άρα με

*  το  ποσοστιαίο σημείο της (bootstrap) κατανομής της  και
* με  το  ποσοστιαίο σημείο της (bootstrap) κατανομής της  θα είχαμε:



και άρα θα παίρναμε  διαστήματα εμπιστοσύνης ταυτόχρονα για όλα τα από



