## Εκτίμηση της αθροιστικής συνάρτησης κατανομής (cdf)

* Έχω ανεξάρτητα, όπου συνεχής
* **Θέλω  εκτιμήτρια της **
* Πρώτα σταθερό. Έστω  και 

Τότε:

και

 ανεξάρτητες Bernoulli

Το 

είναι επομένως συνεπής εκτιμήτρια της :

 αν 

* Κάνοντας το ίδιο για όλα τα στη θέση τουπαίρνουμε την **εμπειρική συνάρτηση κατανομής (empirical cumulative distribution function, ecdf)** 



Για δοσμένο δείγμα  **η  είναι η αθροιστική (cdf) μιας υποθετικής τυχαίας μεταβλητής που βάζει πιθανότητα  σε κάθε παρατήρηση**: πηδάει κατά  σε κάθε παρατήρηση:



******



##### Σημείωση

**Η ecdf  είναι μη-παραμετρική εκτιμήτρια μέγιστης πιθανοφάνειας της .**

Εξήγηση:

* Έστω  η **μη παραμετρική πιθανοφάνεια του δείγματος μας**.(η θεωρείται «παράμετρος» παρ’ ότι είναι μια συνάρτηση, εξ ου και «μη-παραμετρική» πιθανοφάνεια)
* Έχουμε , όπου  η πιθανότητα μια τυχαία μεταβλητή Τ με cdf την να «φέρει» την τιμή .

Άρα θέτοντας παίρνουμε .

* Τότε μπορεί να δειχτεί ότι η μεγιστοποιείται όταν  δηλαδή για , δηλαδή ότι η ecdf είναι μη-παραμετρική εκτιμήτρια μέγιστης πιθανοφάνειας της .

Θεμελιώδες θεώρημα της μη-παραμετρικής στατιστικής.

**Glivenko-Cantelli:**

**Αν συνεχής, μπορούμε να την εκτιμήσουμε με συνέπεια (την ) ομοιόμορφα σε όλα τα :**

Αν ,  συνεχής, τότε 



**Κλειδί της απόδειξης:**

**Η κατανομή του  δεν εξαρτάται από την (αν  συνεχής).**

Αυτό οφείλεται στο ότι:

* Αν τότε  καθώς:



* Και

 (θέτοντας)







,

όπου η ecdf των τυχαίων μεταβλητών 

Για το στατιστικό μπορούμε μάλιστα να έχουμε ένα φράγμα για όλα τα n (όχι απλώς για )

Είναι η **ανισότητα Dvoretzky-Kiefer-Wolfowitz**:

 

Πρόκειται για επέκταση της ανισότητας Hoeffding ομοιόμορφα σε όλα τα . Η επέκταση αυτή, το να ισχύει δηλαδή η ανισότητα ταυτόχρονα για όλα τα x είναι πολύ πιο απαιτητικό να δειχτεί. Σκεφτείτε το «πεπερασμένο» αντίστοιχο. Αν έχω τα 



Αυτό μας επιτρέπει να κατασκευάσουμε **σύνολα εμπιστοσύνης πεπερασμένου μεγέθους δείγματος (DKW)** για την :

* Για επίπεδο  και θέτοντας πάλι , ένας «φάκελος» εμπιστοσύνης για την  κατασκευάζεται ως εξής:
	+ Θέτουμε  και .
	+ Τότε 

Τα παραπάνω διαστήματα στηρίζονται σε ένα άνω φράγμα της 

και θα μπορούσαν να είναι ενδεχομένως πολύ συντηρητικά.

Μπορούν να κατασκευαστούν και κάπως στενότερα **ασυμπτωτικά (δηλαδή για ) σύνολα εμπιστοσύνης για την .**

Αυτά θα στηρίζονται στην **ασυμπτωτική κατανομή** της  που ονομάζεται και στατιστική Kolmogorov**-**Smirnov



Όπου θα έπαιρνε κανείς ως κρίσιμη τιμή k το 1-α ποσοστιαίο σημείο αυτής της κατανομής. Έτσι

Για παράδειγμα

Τέλος **σύνολα εμπιστοσύνης πεπερασμένου μεγέθους δείγματος** μπορουν επίσης να κατασκευαστούν καθωςείναι επίσης **γνωστή και** **η κατανομή της**  **για πεπερασμένο μέγεθος δείγματος**, βλ. π.χ. Gibbons, theorem 3.2 και Table F στο Appendix για . Βλέπε και Table 3.1 για σύγκριση των κρίσιμων τιμών πεπερασμένου μεγέθους με τα ασυμπτωτικά .

Πάλι θα έπαιρνε κανείς ως κρίσιμη τιμή k το 1-α ποσοστιαίο σημείο της κατανομής για πεπερασμένο μέγεθος δείγματος:

Π.χ. για κρίσιμες τιμές για :



******

******

O Kolmogorov-Smirnov έλεγχος καλής προσαρμογής (goodness of fit test)

Έστω

**Δίπλευρος έλεγχος: Θέλουμε να ελέγξουμε αν η cdf των,η , ταυτίζεται *με κάποια δοσμένη*  που μας ενδιαφέρει, έναντι της εναλλακτικής να διαφέρει:**



O Kolmogorov**-**Smirnov έλεγχος ***βασίζεται στο στατιστικό , το οποίο υπό την  έχει γνωστή κατανομή*.**

Προσοχή! Αρκεί το στατιστικό να υπολογιστεί ως maximum επί των  και των , για μικρό ε.

Ισοδύναμα, **θα απορρίπταμε αν  δεν άνηκε στο φάκελο εμπιστοσύνης** που φτιάξαμε παραπάνω. Το εύρος των δ.ε. , και αντίστοιχα η κρίσιμη τιμή του ελέγχου, θα εξαρτιόταν από την κατανομή του στατιστικού KS που θα διαλέγαμε (ακριβής, ασυμπτωτική)

**Μονόπλευρος έλεγχος: Θέλουμε να ελέγξουμε αν η cdf των, η , ταυτίζεται *με κάποια δοσμένη*  που μας ενδιαφέρει, έναντι της εναλλακτικής η  να είναι στοχαστικά μεγαλύτερη της :**

΄



O Kolmogorov**-**Smirnov έλεγχος ***βασίζεται στο στατιστικό , το οποίο υπό την  έχει γνωστή κατανομή και απορρίπτουμε για μεγάλες τιμές του στατιστικού μας*.** (η κρίσιμη τιμή είναι περίπου αυτή που αντιστοιχεί στο διπλάσιο α του δίπλευρου ελέγχου.)

Για 

O Kolmogorov**-**Smirnov έλεγχος ***βασίζεται στο στατιστικό , το οποίο υπό την  έχει γνωστή κατανομή και απορρίπτουμε για μεγάλες τιμές του στατιστικού μας*.** (η κρίσιμη τιμή είναι περίπου αυτή που αντιστοιχεί στο διπλάσιο α του δίπλευρου ελέγχου.)

### Έλεγχος Κανονικότητας του Lillierfors

* **Ο έλεγχος Kolmogorov-Smirnov μπορεί να χρησιμοποιηθεί για να γίνει έλεγχος κανονικότητας .**
* Καθώς όμως τα  δεν είναι δοσμένα, δεν είναι δοσμένη και μια : άρα **η μηδενική αποτελείται από «πολλές»**.
* **Αντικαθιστώντας την  με την ** παίρνουμε ένα στατιστικό που έχει νόημα, διότι μετράει την απόσταση της από την «πλησιέστερη» cdfκανονικής κατανομής.
* **Όμως:** η κατανομή της  δεν είναι ταυτόσημη με την, ακριβώς για διαλέξαμε την «πλησιέστερη» στην cdf.

Κατάλληλες κρίσιμες τιμές για αυτόν τον έλεγχο υπολογίστηκαν με προσομοιώσεις (πώς;) , βλέπε π.χ. Table 0 στο Appendix του Gibbonns.



(Αν χρησιμοποιήσουμε τις τιμές για την KS, αδιαφορώντας για το γεγονός ότι εκτιμήσαμε παραμέτρους, ο έλεγχος θα ήταν πολύ συντηρητικός: θα τηρούσε το ονομαστικό λάθος τύπου Ι, αλλά θα είχε χαμηλή ισχύ).

**Άλλος έλεγχος καλής προσαρμογής** (και κανονικότητας) που βασίζεται στην : ο **«Anderson – Darling» έλεγχος** βασίζεται σε *σταθμισμένη τετραγωνική απόσταση μεταξύ και :*

,

 ο οποίος χρησιμοποιεί στάθμιση της  ανάλογη της διακύμανσής της και άρα κοιτάζει καλύτερα στης ουρές η .