



ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ



Ιωάννης Καππάς

Μία μη παραμετρική μέθοδος για ελέγχους καλής
προσαρμογής συναρτήσεων πυκνότητας πιθανότητας
με τη μέθοδο των πυρήνων και η διερεύνηση της
επίδρασης της παραμέτρου εύρους ζώνης του
ιστογράμματος στην ισχύ του στατιστικού ελέγχου

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννινα, 2025



UNIVERSITY OF IOANNINA
Department of Mathematics



John Kappas

A non-parametric method for goodness-of-fit tests of probability density functions using the kernel method and an investigation into the effect of the histogram bandwidth parameter on the power of the statistical test

Master's Thesis

Ioannina, 2025

Αφιερώνεται στην οικογένεια μου.

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική και Επιχειρησιακή Έρευνα, που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 25/06/2025 από την εξεταστική επιτροπή:

Ονοματεπώνυμο	Βαθμίδα
Δημήτριος Μπάγκαβος	Αναπληρωτής Καθηγητής
Απόστολος Μπατσίδης	Αναπληρωτής Καθηγητής
Σταύρος Νικολακόπουλος	Αναπληρωτής Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

“Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή.”

Ιωάννης Καππάς

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη στον επιβλέποντα καθηγητή μου, κύριο Μπάγκαβο Δημήτριο, για την πολύτιμη καθοδήγηση, τη συνεχή υποστήριξη και την εμπιστοσύνη που μου έδειξε καθ' όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας. Ευχαριστώ επίσης τους καθηγητές και το προσωπικό του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων για τις γνώσεις και τα εφόδια που μου προσέφεραν κατά τη διάρκεια των σπουδών μου. Ιδιαίτερες ευχαριστίες απευθύνω στην οικογένειά μου για την αμέριστη ηθική και υλική στήριξη σε όλη τη διάρκεια των σπουδών μου. Η υπομονή, η ενθάρρυνση και η αγάπη τους υπήρξαν για μένα πολύτιμη δύναμη. Τέλος, ευχαριστώ τους φίλους και συμφοιτητές μου για τη συνεργασία, τη βοήθεια και τις όμορφες στιγμές που μοιραστήκαμε αυτά τα χρόνια.

ΠΕΡΙΛΗΨΗ

Σε αυτήν τη διατριβή παρουσιάζεται η εργασία των Bagkavos et al. (2023) για την περίπτωση του μονοδιάστατου δείγματος. Συγκεκριμένα αναλύεται ένας μη παραμετρικός έλεγχος καλής προσαρμογής για συνεχείς παραμετρικά εκτιμώμενες συναρτήσεις πυκνότητας πιθανότητας στη μονοδιάστατη περίπτωση. Ο μηχανισμός του ελέγχου βασίζεται στην ανίχνευση τοπικών διαφορών μεταξύ μίας μη παραμετρικής εκτίμησης της άγνωστης συνάρτησης πυκνότητας πιθανότητας, με τη μέθοδο του ιστογράμματος, και της δοθείσας υπό την μηδενική υπόθεση παραμετρικά εκτιμημένης συνάρτησης πυκνότητας πιθανότητας. Η μελέτη παρέχει θεωρητικά αποτελέσματα, συμπεριλαμβανομένης της αναλυτικής περιγραφής της ασυμπτωτικής κατανομής του στατιστικού ελέγχου τόσο υπό τη μηδενική όσο και υπό την εναλλακτική υπόθεση όταν αυτή είναι της μορφής Pitman. Για αυτές τις μορφές εναλλακτικών υποθέσεων περιλαμβάνονται κλειστοί τύποι για την ασυμπτωτική ισχύ της στατιστικής συνάρτησης. Για τη βελτίωση του ρυθμού της σύγκλισης του στατιστικού ελέγχου στην κανονική κατανομή, αναπτύσσεται στο άρθρο των Bagkavos et al. (2023), μία μέθοδος επιλογής εύρους ζώνης του πυρήνα, ο οποίος χρησιμοποιείται ως συνάρτηση βάρους στην κατασκευή του ελέγχου, βάση μίας ανισότητας τύπου Berry-Esseen. Επιπλέον, χρησιμοποιείται μία προσεγγιστική bootstrap μέθοδος για τον προσδιορισμό των κατάλληλων κρίσιμων σημείων για εφαρμογές σε πεπερασμένα δείγματα. Μία εκτενής προσομοίωση, που διεξάγεται υπό εναλλακτικές υποθέσεις τύπου Pitman, συγκρίνει την απόδοση του προτεινόμενου ελέγχου με καθιερωμένες μεθόδους από τη βιβλιογραφία, αναδεικνύοντας την ισχυρή απόδοση και την ανταγωνιστική αποτελεσματικότητά του σε σύγκριση με γνωστούς μη παραμετρικούς ελέγχους καλής προσαρμογής. Τέλος, παρουσιάζεται ο κύριος σκοπός της διατριβής που είναι η διερεύνηση της ανεξαρτησίας του ελέγχου από μεταβολές της παραμέτρου του εύρους ζώνης του ιστογράμματος, σε δείγματα πεπερασμένου μεγέθους.

ABSTRACT

This dissertation presents the work of Bagkavos et al. (2023) for the case of a univariate sample. Specifically, it analyzes a non-parametric goodness-of-fit test for continuous, parametrically estimated probability density functions in the one-dimensional case. The test's mechanism is based on detecting local differences between a non-parametric estimate of the unknown probability density function, using the histogram method, and the parametrically estimated probability density function given under the null hypothesis. The study provides theoretical results, including an analytical description of the asymptotic distribution of the test statistic under both the null and the alternative hypothesis, when the latter is of the Pitman type. For these forms of alternative hypotheses, closed-form formulas for the asymptotic power of the test function are included. To improve the rate of convergence of the test statistic to the normal distribution, a method for selecting the kernel bandwidth—which is used as a weight function in the test's construction—is developed in the article by Bagkavos et al. (2023), based on a Berry-Esseen type inequality. Furthermore, an approximate bootstrap method is used to determine the appropriate critical values for applications in finite samples. An extensive simulation study, conducted under Pitman-type alternative hypotheses, compares the performance of the proposed test with established methods from the literature, highlighting its strong performance and competitive efficiency in comparison with well-known non-parametric goodness-of-fit tests. Finally, the main purpose of this dissertation is presented, which is to investigate the test's independence from variations in the histogram bandwidth parameter in finite-sized samples.

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη

Abstract

1	Μη παραμετρικοί μεθόδοι εκτίμησης της σ.π.π. και μοντέλα ελέγχων καλής προσαρμογής	3
1.1	Εκτιμητής Μέγιστης Πιθανοφάνειας και τάξεις σύγκλισης	4
1.2	Μη παραμετρική εκτίμηση της σ.π.π.	6
1.3	Έλεγχοι καλής προσαρμογής βασισμένοι στην α.σ.χ.	20
1.4	Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων	26
2	Μη παραμετρικός έλεγχος καλής προσαρμογής έναντι εναλλακτικών Pitman υποθέσεων	37
2.1	Περιγραφή του ελέγχου υποθέσεων	37
2.2	Κατασκευή του ελέγχου καλής προσαρμογής	38
2.3	Υποθέσεις	39
2.4	Ασυμπτωτική κατανομή	40
2.5	Ασυμπτωτική ισχύς του ελέγχου	41
2.6	Επιλογή εύρους ζώνης h	46
3	Προσομοιώσεις της ισχύος	53
3.1	Σχεδιασμός προσομοιώσεων	53
3.2	Αποδοτικότητα των ελέγχων υπό Pitman εναλλακτικές	59
3.3	Μεταβολή της παραμέτρου Δ	63

4 Παράρτημα	75
Βιβλιογραφία	94

ΚΕΦΑΛΑΙΟ 1

ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΙ ΜΕΘΟΔΟΙ ΕΚΤΙΜΗΣΗΣ ΤΗΣ Σ.Π.Π. ΚΑΙ ΜΟΝΤΕΛΑ ΕΛΕΓΧΩΝ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ

Το πρόβλημα της εκτίμησης της συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) ενός διαθέσιμου δείγματος, είναι κεντρικό πρόβλημα στη στατιστική. Έστω ότι μας δίνεται ένα δείγμα ισόνομων και ανεξάρτητων παρατηρήσεων $X_1, X_2, \dots, X_n$. Προκειμένου ο στατιστικός να εξετάσει τη συμπεριφορά των παρατηρήσεων και να βγάλει περαιτέρω συμπεράσματα, πρέπει να γνωρίζει τη συνάρτηση που περιγράφει την κατανομή τους στην ευθεία των πραγματικών αριθμών. Ο ένας τρόπος είναι ο παραμετρικός, όπου υποθέτουμε μια συγκεκριμένη συναρτησιακή μορφή της σ.π.π., που ανάγεται σε εκτίμηση των παραμέτρων της με βάση το δείγμα. Ο άλλος τρόπος είναι ο μη παραμετρικός, στον οποίο η άγνωστη σ.π.π. εκτιμάται χρησιμοποιώντας το δείγμα, με διάφορες μεθόδους, χωρίς να υιοθετείται κάποιος συναρτησιακός τύπος για αυτή. Στο κεφάλαιο αυτό θα γίνει αρχικά μία μικρή ανασκόπηση, για τον εκτιμητή μέγιστης πιθανοφάνειας (Ε.Μ.Π.) και τις ιδιότητες του, καθώς και μερικούς ορισμούς για τις τάξεις σύγκλισης ακολουθιών, τα οποία θα μας χρησιμεύσουν έτσι ώστε να επεξηγήσουμε αργότερα τα συμπεράσματα μας. Έπειτα θα παρουσιάσουμε δύο τρόπους μη παραμετρικής εκτίμησης της σ.π.π., το ιστόγραμμα και τη μέθοδο των πυρήνων. Κατόπιν θα γίνει μία σύντομη αναφορά σε μη παραμετρικά τεστ καλής προσαρμογής βασισμένα στην αρθροιστική συνάρτηση κατανομής (α.σ.κ.), που θα χρησιμοποιήσουμε για να συγκρίνουμε τα αποτελέσματα μας, τα οποία θα παρουσιάσουμε στο Κεφάλαιο 3. Στο τέλος του κεφαλαίου, θα αναλυθεί ένα μοντέλο ελέγχου υποθέσεων, για πυκνότητες, με χρήση της εκτιμήτριας πυρήνων.

1.1 Εκτιμητής Μέγιστης Πιθανοφάνειας και τάξεις σύγκλισης

Έστω δείγμα $X_1, \dots, X_n$ ισόνομων και ανεξάρτητων (συνεχών ή διακριτών) παρατηρήσεων, διάστασης $d = 1$. Αν οι παρατηρήσεις προέρχονται από μία σ.π.π. $f(x; \theta)$, $\theta \in \Theta$, τότε ορίζουμε τη συνάρτηση πιθανοφάνειας να είναι η από κοινού σ.π.π. των παρατηρήσεων υπολογισμένη στην παρατηρηθείσα τιμή $\mathbf{x} = (x_1, \dots, x_n)$, αλλά ως συνάρτηση της παραμέτρου θ .

$$L(\theta) = L(\theta; \mathbf{x}) = \prod_{i=1}^n f(x_i; \theta), \theta \in \Theta.$$

Ορισμός 1.1. Ο εκτιμητής $\hat{\theta}(\mathbf{X})$ ονομάζεται εκτιμητής μέγιστης πιθανοφάνειας (E.M.Π.) του θ , αν $\forall \mathbf{X} = (X_1, \dots, X_n)$ ισχύει:

$$L(\hat{\theta}(\mathbf{X})) = \max_{\theta \in \Theta} L(\theta; \mathbf{X}).$$

Παρακάτω δίνουμε τον ορισμό για την ασθενή και ισχυρή συνέπεια εκτιμητών, αντίστοιχα.

Ορισμός 1.2. Ένας εκτιμητής $T_n(X_1, \dots, X_n)$ του $g(\theta)$ (όπου $g : \Theta \rightarrow \mathbb{R}$ είναι συνάρτηση της παραμέτρου θ) ονομάζεται ασθενώς συνεπής, αν συγκλίνει κατά πιθανότητα στο $g(\theta)$, δηλαδή αν

$$T_n(X_1, \dots, X_n) \xrightarrow{P} g(\theta), \text{ καθώς } n \rightarrow \infty, \forall \theta \in \Theta,$$

ή ισοδύναμα αν

$$P(|T_n(X_1, \dots, X_n) - g(\theta)| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0, \forall \varepsilon > 0, \forall \theta \in \Theta.$$

Ο εκτιμητής $T_n(X_1, \dots, X_n)$ ονομάζεται ισχυρά συνεπής, αν συγκλίνει σχεδόν βέβαια (*almost surely*) στο $g(\theta)$, δηλαδή:

$$T_n(X_1, \dots, X_n) \xrightarrow{a.s.} g(\theta), \text{ καθώς } n \rightarrow \infty, \forall \theta \in \Theta,$$

ή ισοδύναμα

$$P(\lim_{n \rightarrow \infty} T_n(X_1, \dots, X_n) = g(\theta)) = 1, \forall \theta \in \Theta.$$

Παραθέτουμε τις παρακάτω υποθέσεις:

ΚΕΦΑΛΑΙΟ 1 1.1. Εκτιμητής Μέγιστης Πιθανοφάνειας και τάξεις σύγκλισης

Σ0. Ο παραμετρικός χώρος Θ είναι ανοικτό υποσύνολο του $\mathbb{R}$.

Σ1. Το κοινό σύνολο τιμών των X_i , $S = \{x \mid f(x; \theta) > 0\}$, δεν εξαρτάται από την παράμετρο θ .

Σ2. $\forall \theta_1 \neq \theta_2$, με $\theta_1, \theta_2 \in \Theta$, ισχύει $f(x; \theta_1) \neq f(x; \theta_2), \forall x \in S_2 = \{x \mid f(x; \theta_2) > 0\} \subset S_1 = \{x \mid f(x; \theta_1) > 0\}$, με $P(S_2) > 0$. Δηλαδή σε κάθε σημείο του Θ αντιστοιχεί διαφορετική κατανομή πιθανότητας.

Σ3. $\forall x \in S$ και $\forall \theta \in \Theta$, υπάρχει η παράγωγος $\frac{d}{d\theta} f(x; \theta)$ και είναι πεπερασμένη.

Πρόταση 1.1. Έστω $X_1, \dots, X_n$ ένα τυχαίο δείγμα ανεξάρτητων και ισόνομων παρατηρήσεων που προέρχονται από μία κοινή σ.π.π. $f(x; \theta)$ και έστω ότι ισχύουν οι συνθήκες Σ0, Σ1, Σ2, Σ3. Τότε:

Αν υπάρχει ο ε.μ.π. του θ και η εξίσωση πιθανοφάνειας $\frac{d}{d\theta} L_n(\theta | \mathbf{X}) = 0$ έχει μοναδική λύση, τότε ο Ε.Μ.Π. του θ είναι ισχυρά συνεπής εκτιμητής του θ . Δηλαδή:

$$\hat{\theta}_n(\mathbf{X}) \xrightarrow{a.s.} \theta_0,$$

όπου θ_0 είναι η πραγματική τιμή της παραμέτρου θ .

Η απόδειξη της παραπάνω πρότασης παρουσιάζεται αναλυτικά στο Κεφάλαιο 7 του συγγράμματος των Κουρούκλης, Σ. et al. (2015).

Παρακάτω παραθέτουμε κάποιους ορισμούς για τάξεις σύγκλισης ακολουθιών.

Έστω $(a_n)_{n \in \mathbb{N}}$ και $(b_n)_{n \in \mathbb{N}}$ ακολουθίες πραγματικών αριθμών.

Τότε λέμε, ότι $(a_n)_{n \in \mathbb{N}}$ είναι τάξης $(b_n)_{n \in \mathbb{N}}$, καθώς $n \rightarrow \infty$, και γράφουμε:

$$a_n = O(b_n), \text{ καθώς } n \rightarrow \infty \quad \text{αν-ν} \quad \limsup_{n \rightarrow \infty} \left| \frac{a_n}{b_n} \right| < \infty.$$

Επίσης λέμε ότι $(a_n)_{n \in \mathbb{N}}$ είναι μικρής τάξης $(b_n)_{n \in \mathbb{N}}$ και γράφουμε:

$$a_n = o(b_n), \text{ καθώς } n \rightarrow \infty \quad \text{αν-ν} \quad \lim_{n \rightarrow \infty} \left| \frac{a_n}{b_n} \right| = 0.$$

Επιπλέον συμβολίζουμε:

$$a_n \sim b_n \quad \text{αν-ν} \quad \lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1.$$

Από την άλλη, για ακολουθίες $(a_n)_{n \in \mathbb{N}}$ και $(b_n)_{n \in \mathbb{N}}$ θετικών αριθμών, ο συμβολισμός $a_n \asymp b_n$ σημαίνει $a_n = O(b_n)$ και $b_n = O(a_n)$.

Τέλος για ακολουθίες $(A_n)_{n \in \mathbb{N}}$ και $(B_n)_{n \in \mathbb{N}}$ τυχαίων θετικών μεταβλητών, ο συμβολισμός $A_n \prec_p B_n$ σημαίνει $A_n = O_p(B_n)$ και $B_n = O_p(A_n)$, όπου οι τάξεις σύγκλισης κατα πιθανότητα ορίζονται ως εξής:

- O_p κατά πιθανότητα:

$$A_n = O_p(B_n),$$

σημαίνει ότι για κάθε $\varepsilon > 0$, υπάρχει σταθερά $C > 0$ τέτοια ώστε:

$$\sup_{n \in \mathbb{N}} \mathbb{P}(|A_n| > C|B_n|) < \varepsilon,$$

το οποίο σημαίνει ότι το πηλίκο $\frac{A_n}{B_n}$ είναι φραγμένο κατά πιθανότητα.

- o_p κατά πιθανότητα:

$$A_n = o_p(B_n),$$

σημαίνει ότι:

$$\frac{A_n}{B_n} \xrightarrow{P} 0,$$

δηλαδή το παραπάνω πηλίκο συγκλίνει στο μηδέν κατά πιθανότητα.

1.2 Μη παραμετρική εκτίμηση της σ.π.π.

Ιστόγραμμα

Ένας από τους πιο δημοφιλείς και απλούς τρόπους εκτίμησης της σ.π.π. είναι το ιστόγραμμα. Για την περιγραφή της μεθόδου υποθέτουμε ότι μας παρέχεται ένα δείγμα ισόνομων και ανεξάρτητων συνεχών παρατηρήσεων $X_1, X_2, \dots, X_n$. Η κεντρική ιδέα του ιστογράμματος βασίζεται στην άθροιση του πλήθους των παρατηρήσεων στα διαστήματα της μορφής $[x_0, x_0 + h)$ και έπειτα τη χρήση της σχετικής συχνότητας (ή πυκνότητας) στο σημείο $x \in [x_0, x_0 + h)$, εκτιμώντας έτσι την ποσότητα:

$$f(x_0) = F'(x_0) = \lim_{h \rightarrow 0} \frac{F(x_0 + h) - F(x_0)}{h} = \lim_{h \rightarrow 0^+} \frac{P(x_0 < X < x_0 + h)}{h}.$$

Συγκεκριμένα αν μας δίνεται ένα αρχικό σημείο t_0 και ένα εύρος (bandwidth) $h > 0$, τότε μπορούμε να κατασκευάσουμε τα διαστήματα:

$$\{B_k = [t_k, t_{k+1}) \mid t_k = t_0 + hk, k \in \mathbb{Z}\},$$

στα οποία μετράμε το πλήθος των παρατηρήσεων που βρίσκονται μέσα σε αυτά. Το γεγονός ότι τα διαστήματα αυτά έχουν σταθερό μήκος h είναι σημαντικό καθώς μας επιτρέπει να κανονικοποιήσουμε με h , έτσι ώστε να έχουμε σχετικές συχνότητες ανά μήκος κάθε διαστήματος.

Το ιστόγραμμα στο σημείο x ορίζεται ως εξής:

$$\hat{f}_H(x; t_0, h) = \frac{1}{nh} \sum_{i=1}^n I_{\{X_i \in B_k | x \in B_k\}},$$

Σημειώνουμε εδώ ότι δείτρια τυχαία μεταβλητή ενός ενδεχομένου $A \subseteq \Omega$ του δειγματικού χώρου Ω , είναι η τυχαία μεταβλητή $I_{\{X \in A\}} : \Omega \rightarrow \mathbb{R}$, η οποία ορίζεται ως

$$I_{\{X \in A\}} = \begin{cases} 1, & \text{αν } X \in A \\ 0, & \text{αλλού.} \end{cases}$$

Μέσω του ορισμού διαπιστώνουμε ότι το ιστόγραμμα εξαρτάται από:

- την επιλογή του αρχικού σημείου t_0 στο οποίο βασίζεται η κατασκευή των διαστημάτων,
- το εύρος h των διαστημάτων.

Παρατηρούμε ότι κρατώντας σταθερό το εύρος h και αλλάζοντας το αρχικό σημείο t_0 , αλλάζει σημαντικά η ακρίβεια της εκτίμησης της άγνωστης σ.π.π. f χρησιμοποιώντας τα ίδια δεδομένα. Για να απαλλαγούμε από την εξάρτηση του αρχικού σημείου t_0 , χρησιμοποιούμε μια εναλλακτική προσέγγιση του ιστογράμματος. Η ιδέα είναι ότι μετράμε τις παρατηρήσεις στα διαστήματα της μορφής $(x - h, x + h)$ και μετά χρησιμοποιούμε τη σχετική συχνότητα στο διάστημα $(x - h, x + h)$ για να προσεγγίσουμε τη σ.π.π. στο σημείο x , δηλαδή την ποσότητα:

$$\begin{aligned} f(x) = F'(x) &= \lim_{h \rightarrow 0} \frac{F(x + h) - F(x - h)}{2h} = \lim_{h \rightarrow 0^+} \frac{P(x - h < X < x + h)}{2h} \\ &= \lim_{h \rightarrow 0^+} \frac{P\left(x - \frac{h}{2} < X < x + \frac{h}{2}\right)}{h}. \end{aligned}$$

Συνεπώς έχοντας ένα δεδομένο εύρος $h > 0$, η μη παραμετρική εκτίμηση της σ.π.π. σε ένα σημείο x είναι:

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^n I_{\{x - \frac{h}{2} < X_i < x + \frac{h}{2}\}} = \frac{1}{nh} \sum_{i=1}^n I_{\{|X_i - x| < \frac{h}{2}\}}.$$

Παρατηρούμε ότι η ποσότητα $\hat{f}(x; h)$ αποτελεί τυχαία μεταβλητή καθώς:

$$I_{\{|X_i - x| < \frac{h}{2}\}} \sim B(1, p_{x,h}),$$

$$\text{όπου } p_{x,h} = P\left(x - \frac{h}{2} < X < x + \frac{h}{2}\right) = F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right).$$

Σημειώνουμε εδώ ότι $B(n, p)$ αποτελεί την Διωνυμική κατανομή η οποία περιγράφει το πλήθος των επιτυχιών σε n ανεξάρτητες επαναλήψεις ενός τυχαίου πειράματος με δυο πιθανά αποτελέσματα (επιτυχία - αποτυχία) και πιθανότητα επιτυχίας p . Επιπλέον η αναμενόμενη τιμή και η διακύμανση μίας τυχαίας μεταβλητής X που ακολουθεί τη Διωνυμική κατανομή είναι $E[X] = np$ και $Var[X] = np(1 - p)$, αντίστοιχα.

Καθώς λοιπόν $X_1, \dots, X_n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, από της ιδιότητες της Διωνυμικής κατανομής προκύπτει ότι:

$$\sum_{i=1}^n I_{\{|X_i - x| < \frac{h}{2}\}} \sim B(n, p_{x,h}).$$

Επομένως βλέποντας την ποσότητα $\hat{f}(x; h)$ ως τυχαία μεταβλητή, μπορούμε να υπολογίσουμε τη μέση τιμή και τη διακύμανση της, βρίσκοντας έτσι την μεροληψία του ιστογράμματος, άμεσα από τις ιδιότητες της Διωνυμικής κατανομής. Συνεπώς προκύπτει ότι:

$$\begin{aligned} E[\hat{f}(x; h)] &= E\left[\frac{1}{nh} \sum_{i=1}^n I_{\{|X_i - x| < \frac{h}{2}\}}\right] \\ &= \frac{1}{nh} E\left[\sum_{i=1}^n I_{\{|X_i - x| < \frac{h}{2}\}}\right] \\ &= \frac{1}{nh} np_{x,h} \\ &= \frac{p_{x,h}}{h} \\ &= \frac{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)}{h}. \end{aligned}$$

$$\begin{aligned}
\text{Var}[\hat{f}(x; h)] &= \text{Var} \left[\frac{1}{nh} \sum_{i=1}^n I_{\{|X_i - x| < \frac{h}{2}\}} \right] \\
&= \frac{1}{n^2 h^2} \text{Var} \left[\sum_{i=1}^n I_{\{|X_i - x| < \frac{h}{2}\}} \right] \\
&= \frac{1}{n^2 h^2} n p_{x,h} (1 - p_{x,h}) \\
&= \frac{p_{x,h}}{nh^2} - \frac{p_{x,h}^2}{nh^2} \\
&= \frac{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)}{nh^2} - \frac{\left\{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)\right\}^2}{nh^2}.
\end{aligned}$$

Η ακρίβεια του ιστογράμματος ως εκτιμητή της άγνωστης σ.π.π. $f(x)$ καθορίζεται από το μέσο τετραγωνικό σφάλμα (Mean Squared Error = MSE) το οποίο ορίζεται ως εξής:

$$MSE[\hat{f}(x; h)] = E \left[\left\{ \hat{f}(x; h) - f(x) \right\}^2 \right] = \text{Bias}^2[\hat{f}(x; h)] + \text{Var}[\hat{f}(x; h)].$$

όπου $\text{Bias}[\hat{f}(x; h)] = E[\hat{f}(x; h)] - f(x)$ ορίζεται ως η μεροληψία του εκτιμητή $\hat{f}(x; h)$.

Επίσης, παρατηρούμε ότι:

$$\begin{aligned}
E[\hat{f}(x; h)] &= \frac{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)}{h} = \frac{\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y) dy}{h} \xrightarrow{h \rightarrow 0} f(x), \\
\text{Var}[\hat{f}(x; h)] &= (nh)^{-1} \frac{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)}{h} \\
&\quad - n^{-1} \left\{ \frac{F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)}{h} \right\}^2 \\
&\xrightarrow{h \rightarrow 0} (nh)^{-1} f(x) - n^{-1} f^2(x).
\end{aligned}$$

Από τα παραπάνω λοιπόν συμπεραίνουμε ότι όσο το h μικραίνει η μεροληψία μειώνεται αλλά η διακύμανση αυξάνεται, ενώ όσο μεγαλώνει το h συμβαίνει το αντίστροφο. Αυτό μας υποδηλώνει πόσο σημαντική είναι η επιλογή του h προκειμένου να έχουμε όσο το δυνατόν μικρότερο MSE, δηλαδή όσο το δυνατόν καλύτερη εκτίμηση της άγνωστης σ.π.π. $f(x)$.

Με παρόμοιο σκεπτικό μπορούμε να προσεγγίσουμε και την αθροιστική συνάρτηση κατανομής (α.σ.κ.) εκτιμώντας την ποσότητα $F(x) = P(X_i \leq x)$.

$$\hat{F}(x) = n^{-1} \sum_{i=1}^n I_{\{X_i \leq x\}} = \frac{\#\{X_i \leq x\}}{n},$$

όπου $\#\{X_i \leq x\}$ συμβολίζει το πλήθος των παρατηρήσεων X_i οι οποίες είναι μικρότερες ή ίσες από την τιμή x .

Σε περίπτωση που το δείγμα είναι διατεταγμένο, τότε η εμπειρική α.σ.κ. παίρνει τη μορφή:

$$\hat{F}(x) = n^{-1} \sum_{i=1}^n I_{\{X_{(i)} \leq x\}} = \begin{cases} 0, & \text{αν } x < X_{(1)} \\ \frac{i}{n}, & \text{αν } X_{(i)} \leq x < X_{(i+1)}, i = 1, \dots, n-1, \\ 1, & \text{αν } X_{(n)} \leq x \end{cases}$$

όπου $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ αποτελούν τις διατεταγμένες παρατηρήσεις σε αύξοντα αριθμό.

Επίσης, διαπιστώνουμε ότι η εμπειρική α.σ.κ. σχετίζεται με το ιστόγραμμα καθώς $\forall x_j \in B_j = [x_j - \frac{h}{2}, x_j + \frac{h}{2}]$ θα ισχύει:

$$\hat{F}(x_j) = \frac{\#\{X_i \leq x_j\}}{n} = \sum_{k=1}^j \frac{\#\{X_i \leq x_k \mid x_k \in B_k\}}{n} = h \sum_{k=1}^j \hat{f}(x_k).$$

Επιπλέον η εμπειρική α.σ.κ. συγκλίνει στην πραγματική α.σ.κ. μέσω του κεντρικού οριακού θεωρήματος (Κ.Ο.Θ.). Έστω ότι μας δίνεται ένα δείγμα ανεξάρτητων και ισόνομων παρατηρήσεων $X_1, \dots, X_n$. Από όσα έχουμε αναφέρει παραπάνω, θα ισχύουν:

$$\hat{F}(x) = \frac{\sum_{i=1}^n I_{\{X_i \leq x\}}}{n}, \text{ με } I_{\{X_i \leq x\}} \sim B(1, p_x), \forall i \in \{1, \dots, n\},$$

όπου $p_x = P(X \leq x) = E[I_{\{X \leq x\}}] = F(x)$.

Συνεπώς εφαρμόζοντας το Κ.Ο.Θ. στις ανεξάρτητες και ισόνομες τυχαίες μεταβλητές $I_{\{X_1 \leq x\}}, \dots, I_{\{X_n \leq x\}}$ προκύπτει ότι:

$$\hat{F}(x) \stackrel{d}{\sim} \mathcal{N} \left(E[I_{\{X \leq x\}}], \frac{\text{Var}[I_{\{X \leq x\}}]}{n} \right),$$

ισοδύναμα:

$$\hat{F}(x) \stackrel{d}{\sim} \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right).$$

Επομένως το ιστόγραμμα είναι μια καλή αρχική προσέγγιση της πραγματικής σ.π.π. αλλά με δύο βασικά μειονεκτήματα:

- η εκτίμηση $\hat{f}(x)$ είναι μη συνεχή,
- δίνεται το ίδιο βάρος στις παρατηρήσεις του διαστήματος $(x-h, x+h)$, που είναι κοντά στο x με τις αντίστοιχες που βρίσκονται πιο μακριά.

Για αυτόν τον λόγο θα αναφερθεί παρακάτω μία άλλη μη παραμετρική μέθοδος εκτίμησης της σ.π.π., η οποία διορθώνει αυτά τα δύο βασικά προβλήματα.

Μέθοδος των πυρήνων

Σύμφωνα με τα παραπάνω έχουμε ότι η εκτίμηση της σ.π.π. με τη μέθοδο του ιστογράμματος είναι:

$$\hat{f}(x; h) = \frac{1}{2nh} \sum_{i=1}^n I_{\{x-h < X_i < x+h\}}, \quad (1.1)$$

η οποία ισοδύναμα, γράφεται:

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} I_{\{-1 < \frac{x-X_i}{h} < 1\}} = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right),$$

όπου $K(z) = \frac{1}{2} \cdot I_{\{-1 < z < 1\}}$ είναι η σ.π.π. της ομοιόμορφης κατανομής στο $(-1, 1)$.

Αυτό μας δείχνει ότι όταν προσεγγίζουμε την πιθανότητα:

$$P(x-h < X < x+h) = P\left(-1 < \frac{x-X}{h} < 1\right),$$

με βάση τη σχέση (1.1), δίνουμε το ίδιο βάρος σε όλες τις παρατηρήσεις $X_1, \dots, X_n$.

Από τα παραπάνω μπορούμε να ισχυριστούμε ότι η γενίκευση της (1.1) είναι προφανής. Πιο συγκεκριμένα αντικαθιστούμε τη συνάρτηση K με μια αυθαίρετη πυκνότητα. Τότε η συνάρτηση K ονομάζεται πυρήνας (kernel) και η παράμετρος h καλείται εύρος του πυρήνα (bandwidth).

Συγκεκριμένα η συνάρτηση K επιλέγεται να είναι:

- συμμετρική στο 0 διότι εξασφαλίζει ότι η εκτιμώμενη πυκνότητα σε κάθε σημείο βασίζεται ισότιμα σε σημεία δεξιά και αριστερά από αυτό. Αν ο πυρήνας ήταν ασύμμετρος, θα μετατόπιζε συστηματικά την εκτίμηση της πυκνότητας προς μία πλευρά,
- μονοκόρυφη στο 0 γιατί κάθε σημείο x λαμβάνει με αυτόν τον τρόπο τη μέγιστη επιρροή από παρατηρήσεις που βρίσκονται κοντά του. Καθώς απομακρυνόμαστε από το x , η επιρροή μειώνεται ομαλά,
- να ικανοποιεί $\int K(x) dx = 1$. Αυτό μας εξασφαλίζει το γεγονός ότι η εκτίμηση $\hat{f}(x; h)$, που θα ορίσουμε στη συνέχεια, θα αποτελεί συνάρτηση πυκνότητας πιθανότητας.

Η γενίκευση του ορισμού του ιστογράμματος μας προσφέρει τη δυνατότητα να ορίσουμε τον εκτιμητή της σ.π.π. με τη μέθοδο των πυρήνων ως εξής:

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right). \quad (1.2)$$

Θέτουμε $K_h(z) := \frac{1}{h} \cdot K\left(\frac{z}{h}\right)$.

Τότε η σχέση (1.2) ισοδύναμα γράφεται:

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^n K_h(x - X_i).$$

Αν επιλέξουμε η πυκνότητα K να είναι η σ.π.π. της κανονικής κατανομής τότε θα έχουμε:

$$K_h(x - X_i) = \varphi(x; X_i, h^2) = \varphi(x - X_i; 0, h^2),$$

όπου $\varphi(x; X_i, h^2)$ είναι η σ.π.π. της κανονικής κατανομής με μέση τιμή X_i και τυπική απόκλιση h .

Η εκτίμηση με τη μέθοδο των πυρήνων έχει κατασκευαστεί κεντράροντας έναν κανονικοποιημένο πυρήνα σε κάθε παρατήρηση. Η τιμή της εκτίμησης $\hat{f}(x; h)$ σε ένα σημείο x είναι απλά ο μέσος όρος n το πλήθος τιμών που παίρνει ο πυρήνας στο σημείο x , το οποίο είναι κεντραρισμένο κάθε φορά με μία παρατήρηση. Ουσιαστικά αυτή η μέθοδος συνδιάζει τη συνεισφορά όλων των παρατηρήσεων στην εκτίμηση της άγνωστης σ.π.π. $f(x)$ σε ένα σημείο x . Αυτό σημαίνει ότι σε περιοχές (διαστήματα) που υπάρχουν πολλές παρατηρήσεις και αναμένουμε μεγάλη τιμή

της πραγματικής σ.π.π. $f(x)$, ο πυρήνας K θα πρέπει και αυτός να δίνει σχετικά μεγάλη τιμή. Το αντίθετο θα συμβαίνει σε περιοχές όπου υπάρχουν λιγότερες παρατηρήσεις. Αν και η επιλογή του πυρήνα επηρεάζει την εκτίμηση, έρευνες δείχνουν ότι ο ρόλος του h είναι κρίσιμος για την ισορροπία μεταξύ μεροληψίας και διακύμανσης (βλέπε Wand and Jones (1995)). Προκειμένου όμως να αντιληφθούμε καλύτερα τον τρόπο με τον οποίο η μέθοδος των πυρήνων εκτιμάει την άγνωστη σ.π.π., θα υπολογίσουμε τη μέση τιμή και τη διακύμανση της $\hat{f}(x; h)$.

Αρχικά υπολογίζουμε τη μέση τιμή:

$$\begin{aligned} E[\hat{f}(x; h)] &= E \left[\frac{1}{nh} \sum_{i=1}^n K \left(\frac{x - X_i}{h} \right) \right] = \frac{1}{nh} \sum_{i=1}^n E \left[K \left(\frac{x - X_i}{h} \right) \right] \\ &= \frac{1}{nh} n E \left[K \left(\frac{x - X}{h} \right) \right] = h^{-1} \int K \left(\frac{x - y}{h} \right) f(y) dy \\ &= h^{-1} \int K \left(\frac{y - x}{h} \right) f(y) dy \quad (K \text{ συμμετρική}) \\ &= \int K(u) f(hu + x) du \quad \left(u = \frac{y - x}{h} \right). \end{aligned}$$

Άρα η μέση τιμή της $\hat{f}(x; h)$ είναι:

$$E[\hat{f}(x; h)] = \int K(y) \cdot f(hy + x) dy.$$

Συνεχίζουμε με τη διακύμανση:

$$\begin{aligned} Var[\hat{f}(x; h)] &= Var \left[\frac{1}{nh} \sum_{i=1}^n K \left(\frac{x - X_i}{h} \right) \right] \quad (X_1, \dots, X_n \text{ ανεξάρτητες}) \\ &= (n^2 h^2)^{-1} \sum_{i=1}^n Var \left[K \left(\frac{x - X_i}{h} \right) \right] \quad (X_1, \dots, X_n \text{ ισόνομες}) \\ &= (n^2 h^2)^{-1} n Var \left[K \left(\frac{x - X}{h} \right) \right] \\ &= (nh^2)^{-1} Var \left[K \left(\frac{x - X}{h} \right) \right] \\ &= (nh^2)^{-1} \left\{ E \left[K^2 \left(\frac{x - X}{h} \right) \right] - E \left[K \left(\frac{x - X}{h} \right) \right]^2 \right\}. \end{aligned}$$

Υπολογίζουμε την ποσότητα $E \left[K^2 \left(\frac{x-X}{h} \right) \right]$:

$$\begin{aligned} E \left[K^2 \left(\frac{x-X}{h} \right) \right] &= \int K^2 \left(\frac{x-y}{h} \right) f(y) dy \\ &= \int K^2 \left(\frac{y-x}{h} \right) f(y) dy \quad (\text{K συμμετρική}) \\ &= h \int K^2(u) f(hu+x) du \quad \left(u = \frac{y-x}{h} \right) \\ &= h \int K^2(y) f(hy+x) dy. \end{aligned}$$

Συνδυάζοντας τα παραπάνω έχουμε:

$$\begin{aligned} \text{Var}[\hat{f}(x; h)] &= (nh^2)^{-1} \left\{ h \int K^2(y) f(hy+x) dy \right. \\ &\quad \left. - h^2 \left(\int K(y) f(hy+x) dy \right)^2 \right\}, \end{aligned}$$

από όπου προκύπτει ότι η διακύμανση της $\hat{f}(x; h)$ είναι:

$$\text{Var}[\hat{f}(x; h)] = (nh)^{-1} \int K^2(y) f(hy+x) dy - n^{-1} \left\{ \int K(y) f(hy+x) dy \right\}^2.$$

Έχοντας λοιπόν υπολογίσει τη μέση τιμή και τη διακύμανση του εκτιμητή της μεθόδου των πυρήνων, μπορούμε να μελετήσουμε την ασυμπτωτική του συμπεριφορά με σκοπό να αποκτήσουμε μια πιο καθαρή εικόνα για την επιρροή του εύρους h στην εκτίμηση της άγνωστης σ.π.π. f .

Πριν προχωρήσουμε σε ασυμπτωτική προσέγγιση της μέσης τιμής και της διακύμανσης της $\hat{f}(x; h)$, υιοθετούμε τους παρακάτω ισχυρισμούς:

- Η σ.π.π. f έχει παράγωγο δεύτερης τάξης, έστω f'' η οποία είναι συνεχής, ολοκληρώσιμη και γνησίως μονότονη.
- Το εύρος του πυρήνα $h = h_n$ είναι μια μη τυχαία ακολουθία πραγματικών αριθμών. Η εξάρτηση του h από το n θα περιοριστεί αργότερα στους υπολογισμούς για λόγους απλότητας. Επίσης το h είναι τέτοιο ώστε $h \xrightarrow{n \rightarrow \infty} 0$ και $nh \xrightarrow{n \rightarrow \infty} \infty$.

- Ο πυρήνας K είναι μια φραγμένη σ.π.π. που έχει τέταρτη τάξης ροπή και συμμετρία γύρω από το μηδέν.

Αρχικά εφαρμόζουμε το Θεώρημα Taylor για την σ.π.π. f γύρω από το σημείο x , στο σημείο $hy + x$ και για μικρό h .

$$f(hy + x) = f(x) + hyf'(x) + \frac{h^2}{2}y^2f''(x) + o(h^2).$$

Πολλαπλασιάζουμε και τα δύο μέλη της ισότητας με $K(y)$:

$$K(y)f(hy + x) = K(y)f(x) + hyf'(x)K(y) + \frac{h^2}{2}y^2f''(x)K(y) + o(h^2).$$

Στη συνέχεια ολοκληρώνουμε ως προς y και τα δύο μέλη της ισότητας:

$$\begin{aligned} \int K(y)f(hy + x) dy &= f(x) \int K(y) dy + hf'(x) \int yK(y) dy \\ &\quad + \frac{h^2}{2}f''(x) \int y^2K(y) dy + o(h^2). \end{aligned}$$

Η παραπάνω ισότητα, με βάση τις υποθέσεις που έχουμε κάνει, γίνεται:

$$\int K(y)f(hy + x) dy = f(x) + \frac{h^2}{2}f''(x) \int y^2K(y) dy + o(h^2).$$

Έτσι λοιπόν προκύπτει η ασυμπτωτική μέση τιμή του εκτιμητή $\hat{f}(x; h)$:

$$E[\hat{f}(x; h)] = f(x) + \frac{h^2}{2}f''(x) \int y^2K(y) dy + o(h^2).$$

Επομένως η μεροληψία του $\hat{f}(x; h)$ θα είναι:

$$Bias[\hat{f}(x; h)] = E[\hat{f}(x; h)] - f(x) = \frac{h^2}{2}f''(x) \int y^2K(y) dy + o(h^2).$$

Συμπεραίνουμε ότι η μεροληψία εξαρτάται από το h^2 , που μας δείχνει ότι η εκτίμηση $\hat{f}(x; h)$ είναι ασυμπτωτικά αμερόληπτη, καθώς και από την άγνωστη f . Συγκεκριμένα σε σημεία όπου η f παρουσιάζει μεγάλη καμπυλότητα, ισοδύναμα η f'' παίρνει μεγάλες τιμές κατά απόλυτη τιμή, η μεροληψία αποκλίνει από το μηδέν προσθέτοντας έτσι μεροληψία στην εκτίμηση μας. Συνεπώς ο εκτιμητής $\hat{f}(x; h)$

έχει μία τάση να επερεκτιμά ή να υποεκτιμά την πραγματική σ.π.π. f στα σημεία αυτά.

Για την ασυμπτωτική διακύμανση θα έχουμε:

$$\begin{aligned} Var[\hat{f}(x; h)] &= (nh)^{-1} \int K^2(y) f(hy + x) dy - n^{-1} \left\{ \int K(y) f(hy + x) dy \right\}^2 \\ &= (nh)^{-1} \int K^2(y) (f(x) + o(1)) dy - n^{-1} E[\hat{f}(x; h)]^2 \\ &= (nh)^{-1} \int K^2(y) dy f(x) - n^{-1} \{f(x) + o(1)\}^2 \\ &= (nh)^{-1} f(x) \int K^2(y) dy + o((nh)^{-1}). \end{aligned}$$

Οπότε η ασυμπτωτική διακύμανση του $\hat{f}(x; h)$ θα είναι:

$$Var[\hat{f}(x; h)] = (nh)^{-1} f(x) \int K^2(y) dy + o((nh)^{-1}).$$

Με βάση την ασυμπτωτική μεροληψία και διακύμανση, μπορούμε να υπολογίσουμε το μέσο τετραγωνικό σφάλμα του $\hat{f}(x; h)$ και να δούμε πως συμπεριφέρεται η εκτίμηση $\hat{f}$ της άγνωστης σ.π.π. f ως προς το εύρος του πυρήνα h . Είναι:

$$MSE[\hat{f}(x; h)] = E \left[\left\{ \hat{f}(x; h) - f(x) \right\}^2 \right] = Var[\hat{f}(x; h)] + E \left[\hat{f}(x; h) - f(x) \right]^2.$$

Οπότε αντικαθιστώντας προκύπτει το ασυμπτωτικό μέσο τετραγωνικό σφάλμα του εκτιμητή (Asymptotic MSE = AMSE) $\hat{f}(x; h)$:

$$MSE[\hat{f}(x; h)] = AMSE[\hat{f}(x; h)] + o((nh)^{-1} + h^4),$$

$$AMSE[\hat{f}(x; h)] = (nh)^{-1} f(x) \int K^2(y) dy + \frac{h^4}{4} f''(x)^2 \left\{ \int y^2 K(y) dy \right\}^2.$$

Θέτουμε:

$$\mu_m(g) = \int z^m g(z) dz, \quad \forall m \in \mathbb{N} \text{ και για κάθε ολοκληρώσιμη συνάρτηση } g,$$

$$R(g) = \int g(x)^2 dx, \text{ για κάθε τετραγωνικά ολοκληρώσιμη συνάρτηση } g,$$

οπότε το AMSE γράφεται:

$$AMSE[\hat{f}(x; h)] = (nh)^{-1} f(x) R(K) + \frac{h^4}{4} f''(x)^2 \mu_2(K)^2.$$

Γενικά προτιμάται η εκτίμηση να γίνεται σε όλη την ευθεία των πραγματικών αριθμών και όχι σε ένα μόνο σημείο. Για αυτό ολοκληρώνοντας ως προς x προκύπτει το ασυμπτωτικό ολοκλήρωμα του μέσου τετραγωνικού σφάλματος (Asymptotic Mean Integrated Squared Error = AMISE):

$$AMISE[\hat{f}(x; h)] = (nh)^{-1} R(K) + \frac{h^4}{4} R(f''(x)) \mu_2(K)^2.$$

Βλέπουμε ότι το ολοκλήρωμα του τετραγώνου της μεροληψίας είναι ασυμπτωτικά ανάλογο του h^4 . Σκοπός είναι η μείωση αυτής της ποσότητας και άρα η μείωση του h . Όμως αν μειώσουμε το h , θα αυξηθεί το ολοκλήρωμα της διακύμανσης το οποίο είναι ασυμπτωτικά αντιστρόφως ανάλογο της ποσότητας nh . Συγκεντρωτικά λοιπόν, σύμφωνα με τους Marron and Wand (1992) η μεταβολή του h έχει τις εξής συνέπειες:

- Μικραίνοντας το h : Μεγαλώνει η διακύμανση και μικραίνει η μεροληψία. Αυτό έχει ως αποτέλεσμα η εκτίμηση $\hat{f}(x; h)$ να γίνεται αρκετά ακανόνιστη με αιχμηρές κορυφές και συνεπώς να μεταβάλλεται σε επαναλαμβανόμενες δειγματοληψίες από την f . Άρα οι κορυφές αυτές ή τα spikes θα εμφανίζονται κάθε φορά σε διαφορετικά σημεία.
- Μεγαλώνοντας το h : Μικραίνει η διακύμανση και μεγαλώνει η μεροληψία. Αυτό σημαίνει ότι η μεταβλητότητα της $\hat{f}(x; h)$ μειώνεται αλλά προστίθεται μεροληψία. Έχοντας λοιπόν μεγάλη μεροληψία, ομαλοποιούνται όλα τα χαρακτηριστικά της εκτίμησης $\hat{f}(x; h)$, αλλά με μικρή διακύμανση, διότι τα δεδομένα σχεδόν αγνοούνται.

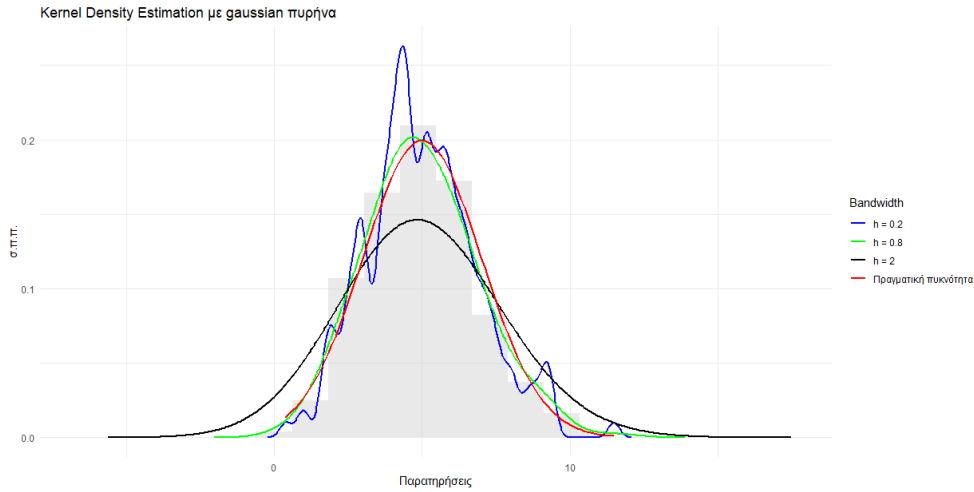
Για αυτόν το λόγο το h πρέπει να κυμαίνεται σε ένα τέτοιο εύρος έτσι ώστε η διακύμανση και η μεροληψία να μειώνονται ταυτόχρονα. Αν λοιπόν παραγωγίσουμε ως προς h το AMISE και εξισώσουμε την παράγωγό του με το μηδέν, προκύπτει το βέλτιστο h για το οποίο ελαχιστοποιείται το AMISE, βελτιστοποιώντας έτσι την εκτίμηση της f από την $\hat{f}$. Η σχέση που δίνει το βέλτιστο εύρος ζώνης του πυρήνα σ' αυτήν την περίπτωση είναι η:

$$h_{AMISE} = \left\{ \frac{R(K)}{n \mu_2(K)^2 R(f'')} \right\}^{\frac{1}{5}}.$$

Σημειώνουμε εδώ ότι η ποσότητα $R(f'')$ μετράει την ολική καμπυλότητα της f . Βλέπουμε ότι το h_{AMISE} εξαρτάται από τον πυρήνα K και το μέγεθος του δείγματος n , που είναι γνωστές ποσότητες, ενώ είναι αντιστρόφως ανάλογο της ποσότητας $R(f'')^{\frac{1}{5}}$, η οποία είναι άγνωστη καθώς δεν γνωρίζουμε την πραγματική σ.π.π. f . Συμπεραίνουμε λοιπόν ότι όταν η f παρουσιάζει μεγάλη καμπυλότητα ή ισοδύναμα η ποσότητα $R(f'')$ παίρνει μεγάλη τιμή, χρειάζεται να επιλέξουμε μικρό h , ενώ σε αντίθετη περίπτωση πρέπει να επιλέγεται μεγάλο h .

Αντικαθιστώντας το βέλτιστο εύρος h_{AMISE} στην προσέγγιση AMISE, προκύπτει το ελάχιστο δυνατό AMISE για την εκτίμηση της f χρησιμοποιώντας τον πυρήνα K , το οποίο είναι το:

$$\inf_{h>0} AMISE[\hat{f}(\cdot, h)] = \frac{5}{4} \{ \mu_2(K)^2 R(K)^4 R(f'') \}^{\frac{1}{5}} n^{-\frac{4}{5}}.$$



Σχήμα 1.1: 200 παρατηρήσεις από την κανονική κατανομή με μέση τιμή 5 και διακύμανση 2. Στο παραπάνω σχήμα απεικονίζεται η πραγματική πυκνότητα από την οποία προήλθαν τα δεδομένα, $N(5,2)$, με κόκκινη γραμμή. Επιπλέον με μπλέ γραμμή παρουσιάζεται η εκτίμηση της σ.π.π. με τη μέθοδο των πυρήνων για $h = 0.2$, με πράσινη για $h = 0.8$ και με μαύρη για $h = 2$. Ταυτόχρονα στο ίδιο σχήμα απεικονίζεται και η εκτίμηση του ιστογράμματος.

Επιπρόσθετα διαπιστώνουμε ότι καθώς $n \rightarrow \infty$:

$$E[\hat{f}(x; h)] \xrightarrow{n \rightarrow \infty} f(x) \quad \text{και} \quad Var[\hat{f}(x; h)] \xrightarrow{n \rightarrow \infty} 0,$$

το οποίο συνεπάγεται ότι η ποσότητα $\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$ είναι ασθενά συνεπής εκτιμητής της πραγματικής σ.π.π. $f(x)$.

Αν επιπλέον ισχύουν οι υποθέσεις:

- $\sum_{i=1}^{\infty} \exp(-\alpha nh) < \infty$, $\forall \alpha > 0$,
- $nh\{\ln(n)\}^{-1} \xrightarrow{n \rightarrow \infty} \infty$,

τότε η ποσότητα $\hat{f}(x; h)$ είναι ισχυρά συνεπής εκτιμητής της $f(x)$. Δηλαδή:

$$\hat{f}(x; h) \xrightarrow{a.s.} f(x).$$

Τέλος, θα παρουσιάσουμε την εκτίμηση της α.σ.κ. με τη μέθοδο των πυρήνων. Γνωρίζουμε ότι η σ.π.π. δεν είναι τίποτα άλλο παρά ο ρυθμός μεταβολής της α.σ.κ.. Έχοντας λοιπόν ένα δείγμα $X_1, \dots, X_n$ ανεξάρτητων και ισόνομων παρατηρήσεων από έναν πληθυσμό με σ.π.π. f , θα έχουμε:

$$f(x) = \frac{d}{dx} F(x) \Rightarrow F(x) = \int_{-\infty}^x f(y) dy.$$

Αντικαθιστούμε στον τύπο της α.σ.κ. τον εκτιμητή πυρήνων της σ.π.π. f :

$$\begin{aligned} \hat{F}(x; h) &= \int_{-\infty}^x \hat{f}(y; h) dy \\ &= \int_{-\infty}^x \left\{ \frac{1}{nh} \sum_{i=1}^n K\left(\frac{y-X_i}{h}\right) \right\} dy \\ &= \frac{1}{nh} \sum_{i=1}^n \int_{-\infty}^x K\left(\frac{y-X_i}{h}\right) dy \\ &= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\frac{x-X_i}{h}} K(u) du. \end{aligned}$$

Συνεπώς ο εκτιμητής της α.σ.κ. με τη μέθοδο των πυρήνων είναι:

$$\hat{F}(x; h) = \frac{1}{n} \sum_{i=1}^n I\left(\frac{x-X_i}{h}\right),$$

όπου $I(x) = \int_{-\infty}^x K(u) du$.

1.3 Έλεγχοι καλής προσαρμογής βασισμένοι στην α.σ.κ.

Έστω ότι μας δίνεται ένα δείγμα $X_1, \dots, X_n$ από κάποια άγνωστη κατανομή με α.σ.κ. F στο σύνολο των πραγματικών αριθμών $\mathbb{R}$. Αυτό που θέλουμε να ελέγξουμε είναι αν το δείγμα προέρχεται από δοθείσα κατανομή με α.σ.κ. $F_0(x; \theta)$. Συνεπώς το ερώτημα αυτό θα απαντηθεί μέσω του ελέγχου της υπόθεσης:

$$H_0: F(x) = F_0(x; \theta) \text{ έναντι } H_1: F(x) \neq F_0(x; \theta), \text{ για κάποιο } x \in \mathbb{R}. \quad (1.3)$$

Για τον παραπάνω έλεγχο διακρίνουμε δύο περιπτώσεις:

- i) Η $F_0(x; \theta) = F_0(x)$ είναι πλήρως προσδιορισμένη, δηλαδή οι παράμετροι της είναι πλήρως γνωστές. Σε αυτή την περίπτωση λέμε ότι έχουμε ένα απλό πρόβλημα ελέγχου υποθέσεων (simple goodness of fit problem).
- ii) Οι παράμετροι της $F_0(x; \theta)$ δεν είναι γνωστές και ανήκει σε κάποια παραμετρική οικογένεια κατανομών $\mathcal{F}$ ($F_0 \in \mathcal{F}$). Επομένως οι άγνωστες παράμετροι εκτιμούνται από τα δεδομένα και τότε το πρόβλημα ελέγχου υποθέσεων καλείται σύνθετο (composite goodness of fit problem).

Σε αυτήν την υποενότητα θα αναφέρουμε τρία αρκετά γνωστά και ισχυρά τεστ που χρησιμοποιούνται στον ευρύ επιστημονικό κύκλο με σκοπό την επίλυση του απλού προβλήματος ελέγχου υποθέσεων. Το κοινό χαρακτηριστικό των ελέγχων αυτών που θα δούμε είναι ότι χρησιμοποιούν την εμπειρική α.σ.κ. $\hat{F}$ ως εκτιμήτρια της άγνωστης α.σ.κ. F για τον έλεγχο της υπόθεσης που διατυπώθηκε στη σχέση (1.3). Επομένως όταν ισχύει η $H_0: F = F_0$, ισοδύναμα μπορούμε να ελέγξουμε την H_0 εξετάζοντας την απόκλιση μεταξύ της εμπειρικής α.σ.κ. $\hat{F}$ και της α.σ.κ. F , υπό την H_0 . Άρα διαφορετικές επιλογές συνάρτησης απόκλισης, θα οδηγήσουν σε διαφορετικά τεστ ελέγχου υποθέσεων. Σκοπός λοιπόν είναι:

- i) να εφαρμοστεί ένα κριτήριο απόκλισης (μία μετρική) με το οποίο θα εξετάζεται η εγγύτητα της $F_0(x; \theta)$ στην πραγματική $F(x)$, και
- ii) να μπορέσουμε να καθορίσουμε την κατανομή της στατιστικής συνάρτησης, που προκύπτει από το παραπάνω βήμα.

Υπάρχουν πολλές μετρικές για να μετρήσουμε την “απόσταση” μεταξύ $F(x)$ και $F_0(x; \theta) = F_0(x)$ (όταν θ γνωστή). Εμείς θα αναφερθούμε στις τρεις αυτές

μετρικές που χρησιμοποιούν τα τεστ ελέγχου υποθέσεων των Kolmogorov - Smirnov, Cramér - Von Mises και Anderson - Darling αντίστοιχα.

$$D_n = \sup_{x \in \mathbb{R}} |F(x) - F_0(x)| \quad : \text{supremum norm}$$

$$C_n = \int \{F(x) - F_0(x)\}^2 dF_0(x) \quad : L_2 \text{ norm}$$

$$A_n = \int \{F(x) - F_0(x)\}^2 w(x) dF_0(x) \quad : \text{weighted } L_2 \text{ norm}$$

Γενικά για τέτοιου τύπου ελέγχους όπως αυτός που δίνεται στη σχέση (1.3), περιμένουμε ότι όταν ισχύει η μηδενική υπόθεση H_0 , η τιμή του στατιστικού να είναι κοντά στο μηδέν, ενώ σε αντίθετη περίπτωση οι ακραίες τιμές της στατιστικής συνάρτησης σημαίνουν σημαντική απόκλιση μεταξύ των $F(x)$ και $F_0(x; \theta)$. Βασική παρατήρηση είναι ότι η στατιστική συνάρτηση είναι τυχαία μεταβλητή, καθώς εξαρτάται πάντα από τα δεδομένα. Συνεπώς έχει κατανομή η οποία είναι διαφορετική αναλόγως τον συνδυασμό μετρικής και εκτιμήτριας της $F(x)$.

Kolmogorov - Smirnov

Ίσως το πιο συνηθισμένο και συχνά χρησιμοποιούμενο τεστ στην πράξη για τον έλεγχο της (1.3) είναι το τεστ των Kolmogorov - Smirnov (K-S) που βασίζεται στη μέτρηση της απόκλισης μεταξύ της άγνωστης α.σ.κ. $F(x)$ και της α.σ.κ. $F_0(x; \theta)$, υπό την H_0 , χρησιμοποιώντας τη sup μετρική, μέσω της στατιστικής συνάρτησης:

$$D = \sup_{x \in \mathbb{R}} |F(x) - F_0(x; \theta)|.$$

Βασικά χαρακτηριστικά του τεστ είναι ότι:

- η $F_0(x; \theta)$ είναι πλήρως προσδιορισμένη, δηλαδή η τιμή της παραμέτρου δεν έχει προκύψει από εκτίμηση μέσω του δείγματος και συνεπώς δεν εξαρτάται από αυτό,
- για να υλοποιηθεί ο έλεγχος, δηλαδή προκειμένου να χρησιμοποιηθεί ο παραπάνω τύπος, η πραγματική α.σ.κ. $F(x)$ εκτιμάται από την εμπειρική α.σ.κ. $\hat{F}(x)$.

Η στατιστική συνάρτηση στην πράξη υπολογίζεται ως εξής:

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}(x) - F_0(x; \theta)| = \max_{1 \leq i \leq n} \begin{cases} \hat{F}(X_i - \varepsilon) - F_0(X_i) \\ \hat{F}(X_i) - F_0(X_i) \end{cases}.$$

Για να μπορούμε να χρησιμοποιήσουμε την D_n έτσι ώστε να έχουμε τη δυνατότητα να ελέγξουμε υποθέσεις και να βγάλουμε συμπεράσματα, πρέπει να προσδιορίσουμε την κατανομή της.

Ασυμπτωτικά η κατανομή του τεστ προσεγγίζεται από την πιθανότητα:

$$P\left(\sqrt{n}\sup_{x \in \mathbb{R}}|\hat{F}(x) - F_0(x; \theta)| \leq t\right) \xrightarrow{n \rightarrow \infty} Q(t) = 1 - 2 \sum_{k=0}^{\infty} (-1)^k e^{-2k^2 t^2}, \quad t > 0,$$

όπου παρατηρούμε ότι δεν εξαρτάται από την άγνωστη α.σ.κ. $F(x)$.

Θεώρημα Glivenko - Cantelli. Έστω τυχαίο δείγμα $X_1, \dots, X_n$ με κοινή α.σ.κ. F . Τότε θα ισχύει:

$$\|\hat{F} - F\|_{\infty} = \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \xrightarrow{a.s.} 0,$$

όπου $\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I_{[X_i, \infty)}$ είναι η εμπειρική α.σ.κ.

Απόδειξη. Η σχεδόν βέβαιη σύγκλιση για την εμπειρική α.σ.κ. $\hat{F}(x)$ σημαίνει ότι

$$P\left(\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| = 0\right) = 1.$$

Επιπλέον από τον ισχυρό νόμο των μεγάλων αριθμών έχουμε ότι $\forall x \in \mathbb{R}$ θα ισχύει

$$P\left(\lim_{n \rightarrow \infty} \hat{F}(x) = F(x)\right) = 1.$$

Ουσιαστικά το supremum της διαφοράς της εμπειρικής α.σ.κ. και της πραγματικής α.σ.κ. στην εκφώνηση του θεωρήματος είναι στην ουσία το μέγιστο των διαφορών $\hat{F}(x_i) - F(x_i)$ για πεπερασμένο ακέραιο αριθμό $i \in \{1, \dots, m\}$. Εφόσον η F είναι συνεχής συνάρτηση, μπορούμε να βρούμε m σημεία τέτοια ώστε $-\infty < x_0 < x_1 < \dots < x_m = \infty$ και $\hat{F}(x_j) - F(x_j) \leq \epsilon$, με $j \in \{1, \dots, m\}$ για δεδομένο $\epsilon > 0$. Έστω λοιπόν ένα αυθαίρετο σημείο $x \in \mathbb{R}$. Τότε υπάρχει ακέραιος αριθμός $j \in \{1, \dots, m\}$ τέτοιος ώστε $x_{j-1} \leq x \leq x_j$. Κατά συνέπεια, αφού η α.σ.κ. F είναι αύξουσα συνάρτηση θα έχουμε ότι

$$\begin{aligned} \hat{F}(x) - F(x) &\leq \hat{F}(x_j) - F(x_{j-1}) = \hat{F}(x_j) - F(x_j) + F(x_j) - F(x_{j-1}) \\ &\leq \hat{F}(x_j) - F(x_j) + \epsilon \\ &\leq \max_{j \in \{0, \dots, m\}} |\hat{F}(x_j) - F(x_j)| + \epsilon. \end{aligned}$$

Ομοίως έχουμε

$$\begin{aligned}\hat{F}(x) - F(x) &\geq \hat{F}(x_{j-1}) - F(x_j) = \hat{F}(x_{j-1}) - F(x_{j-1}) + F(x_{j-1}) - F(x_j) \\ &\geq \hat{F}(x_{j-1}) - F(x_{j-1}) - \epsilon \\ &\geq - \max_{j \in \{1, \dots, m\}} |\hat{F}(x_{j-1}) - F(x_{j-1})| - \epsilon.\end{aligned}$$

Συνδυάζοντας τις δύο παραπάνω ανισότητες προκύπτει

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \leq \epsilon + \max_{j \in \{0, \dots, m\}} |\hat{F}(x_j) - F(x_j)|. \quad (1.4)$$

Στη συνέχεια θεωρούμε τα σύνολα

$$A_j = \left\{ \lim_{n \rightarrow \infty} |\hat{F}(x_j) - F(x_j)| \neq 0 \right\}, j \in \{0, \dots, m\},$$

και

$$A = \left\{ \lim_{n \rightarrow \infty} \max_{j \in \{0, \dots, m\}} |\hat{F}(x_j) - F(x_j)| \neq 0 \right\}.$$

Ξέρουμε ήδη ότι $P(A_j) = 0$ και επιπλέον παρατηρούμε ότι $A = \cup_{j=0}^m A_j$. Συνεπώς θα έχουμε ότι

$$P(A) = P\left(\cup_{j=0}^m A_j\right) \leq \sum_{j=0}^m P(A_j) = 0.$$

Οπότε θα ισχύει ότι $P(A) = 0$. Αυτό σημαίνει ότι

$$\max_{j \in \{0, \dots, m\}} |\hat{F}(x_j) - F(x_j)| \xrightarrow{a.s.} 0. \quad (1.5)$$

Αντικαθιστώντας τη σχέση (1.5) στη (1.4) παίρνουμε ότι

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \leq \epsilon, \forall \epsilon > 0.$$

Το ζητούμενο προκύπτει άμεσα αν θεωρήσουμε ότι $\epsilon \rightarrow 0$. □

Σύμφωνα με το παραπάνω Θεώρημα, καθώς $n \rightarrow \infty$, θα ισχύει ότι $D_n \xrightarrow{a.s.} 0$ σχεδόν βέβαια, που σημαίνει ότι το τεστ των K-S είναι συνεπές, δηλαδή καθώς $n \rightarrow \infty$, η ισχύς του τείνει στο 1, για όλες τις γενικές εναλλακτικές της υπόθεσης (1.3). Παρατηρούμε ότι στην περίπτωση που $F(x) \neq F_0(x; \theta)$, η τιμή της D_n θα είναι η $\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| > 0$, η οποία θα τείνει στο άπειρο καθώς μεγαλώνει το δείγμα με αποτέλεσμα το τεστ να απορρίπτει σωστά τη μηδενική

υπόθεση. Σε αυτό το σημείο βλέπουμε ότι αν η παράμετρος $\theta \in \mathbb{R}^p$, $p \geq 1$ ήταν τυχαία μεταβλητή, τότε η προσέγγιση της κατανομής της $\sqrt{n}D_n$ από την $Q(t)$ δεν θα ίσχυε και θα οδηγούμασταν σε λάθος συμπεράσματα.

Cramér - von Mises

Αν αντί της $\sup$ μετρικής που χρησιμοποιείται στο τεστ των K-S, χρησιμοποιήσουμε την L_2 μετρική, τότε η απόκλιση μεταξύ $F(x)$ και $F_0(x; \theta)$ μετράται από τη σχέση:

$$W = n \int \{F(x) - F_0(x; \theta)\}^2 dF(x).$$

Η προτεινόμενη στατιστική συνάρτηση ελέγχου είναι η:

$$\hat{W}_n = n \int \{\hat{F}(x) - F_0(x; \theta)\}^2 dF_0(x),$$

η οποία εξαρτάται από το δείγμα και τη δοθείσα κατανομή υπό τη μηδενική υπόθεση.

Μία ισοδύναμη και υπολογιστικά πιο απλή έκφραση είναι η:

$$\hat{W}_n = n \int \{\hat{F}(x) - F_0(x; \theta)\}^2 dF_0(x) = \sum_{i=1}^n \left\{ \frac{2i-1}{2n} - F_0(X_{(i)}; \theta) \right\}^2 + \frac{1}{12n},$$

όπου $X_{(i)}$ είναι η i -οστή διατεταγμένη παρατήρηση με $i = 1, \dots, n$ και $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(i-1)} \leq X_{(i)} \leq X_{(i+1)} \leq \dots \leq X_{(n)}$.

Η ασυμπτωτική κατανομή του τεστ είναι:

$$\begin{aligned} \lim_{n \rightarrow \infty} F_{W_n}(x) &= \lim_{n \rightarrow \infty} P(W_n \leq x) \\ &= \begin{cases} 0, & \text{αν } x \leq 0 \\ 1 - \frac{2}{\pi} \sum_{k=1}^{\infty} (-1)^{k+1} \int_{(2k-1)\pi}^{2k\pi} \frac{e^{-\frac{u^2 x}{2}}}{\sqrt{-u \sin u}} du, & \text{αν } x > 0, \end{cases} \end{aligned}$$

Anderson - Darling

Το τεστ των Anderson - Darling βασίζεται, όπως και το τεστ των Cramér - von Mises, σε μία παραλλαγή της L_2 μετρικής. Ειδικότερα είναι:

$$A = n \int \{F(x) - F_0(x; \theta)\}^2 w(x) dF(x).$$

όπου $w(x)$ είναι μία συνάρτηση βάρους. Παρατηρούμε ότι θέτοντας $w(x) = 1$ παίρνουμε το τεστ των Cramér - von Mises.

Η απόκλιση που χρησιμοποιεί το τεστ των Anderson - Darling προκύπτει από τον παραπάνω τύπο, θέτοντας $w(x) = \{F(x)(1 - F(x))\}^{-1}$, οπότε είναι:

$$A = n \int \frac{\{F(x) - F_0(x; \theta)\}^2}{F(x)(1 - F(x))} dF(x).$$

Η επιλογή αυτής της συνάρτησης βάρους $w(x)$ δεν είναι τυχαία καθώς για να μπορεί να χρησιμοποιηθεί η παραπάνω στατιστική συνάρτηση, εκτιμούμε την άγνωστη α.σ.κ. F με την εμπειρική α.σ.κ. $\hat{F}$, η οποία έχει ασυμπτωτική μέση τιμή και διακύμανση $E[\hat{F}] \xrightarrow{n \rightarrow \infty} F(x)$ και $Var[\hat{F}] \xrightarrow{n \rightarrow \infty} F(x)(1 - F(x))$ αντίστοιχα. Συνεπώς υπό την H_0 :

$$\begin{aligned} E[\hat{F}] &\xrightarrow{n \rightarrow \infty} F_0(x), \\ Var[\hat{F}] &\xrightarrow{n \rightarrow \infty} F_0(x)(1 - F_0(x)). \end{aligned}$$

Επομένως η στατιστική συνάρτηση του τεστ των Anderson - Darling, υπό την H_0 είναι η:

$$\hat{A}_n = n \int \frac{\{\hat{F}(x) - F_0(x; \theta)\}^2}{F_0(x)(1 - F_0(x))} dF_0(x).$$

Κανονικοποιώντας λοιπόν με αυτόν τον τρόπο την εμπειρική α.σ.κ. $\hat{F}$, δίνεται μεγαλύτερο βάρος στις παρατηρήσεις στα άκρα της κατανομής και λιγότερο στη μέση.

Μία πιο εύχρηστη και ισοδύναμη μορφή της $\hat{A}_n$ είναι:

$$\hat{A}_n = -n - n^{-1} \sum_{i=1}^n (2i - 1) [\ln F_0(X_{(i)}; \theta) + \ln(1 - F_0(X_{(n+1-i)}; \theta))].$$

Τέλος η ασυμπτωτική κατανομή του τεστ Anderson - Darling προσδιορίζεται από τη σχέση:

$$\begin{aligned} \lim_{n \rightarrow \infty} F_{A_n}(x) &= \lim_{n \rightarrow \infty} P(A_n \leq x) \\ &= \frac{\sqrt{2\pi}}{x} \sum_{i=0}^{\infty} \binom{-1/2}{i} (4i + 1) e^{-\frac{(4i+1)^2 \pi^2}{8x}} \int_0^{\infty} e^{\frac{x}{8(1+u)^2} - \frac{u^2(4i+1)^2 \pi^2}{8x}} du, \end{aligned}$$

όπου παρατηρούμε ότι είναι ανεξάρτητη από την αληθή κατανομή $F_0(x)$, χαρακτηριστικό που την καθιστά χρήσιμη για γενικό έλεγχο καλής προσαρμογής.

1.4 Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

Έστω ότι μας δίνεται ένα δείγμα $X_1, \dots, X_n$ ανεξάρτητων και ισόνομων παρατηρήσεων και θέλουμε να ελέγξουμε αν η σ.π.π. από την οποία προέρχεται το δείγμα, ταυτίζεται με την πυκνότητα μιας κατανομής, που ανήκει στην παραμετρική οικογένεια κατανομών. Δηλαδή θέλουμε να ελέγξουμε:

$$H_0: f = f_0 \in P_\Theta \quad \text{έναντι} \quad H_1: f = f_1 \notin P_\Theta, \quad (1.6)$$

όπου:

- $P_\Theta = \{f(\cdot; \theta) \mid \theta \in \Theta\}$ ορίζει το μοντέλο που απαρτίζεται από μία παραμετρική οικογένεια συναρτήσεων πυκνότητας πιθανότητας ορισμένες στο $\mathbb{R}$ και με πεδίο ορισμού ένα υποσύνολο του $\mathbb{R}$,
- $\Theta \subseteq \mathbb{R}^p$ είναι ο παραμετρικός χώρος με $p \in \mathbb{N}$,
- $\theta \in \Theta$ είναι το διάνυσμα των παραμέτρων της οικογένειας κατανομών.

Όπως παρουσιάσαμε στην υποενότητα 1.3, υπάρχουν αρκετές στατιστικές συναρτήσεις που βασίζονται σε μετρικές που ανιχνεύουν διαφορές μεταξύ της πραγματικής σ.π.π. και αυτής υπό τη μηδενική υπόθεση. Όμως δεν μπορούν να ανιχνεύσουν τέτοιες διαφορές όταν ο εκτιμητής της άγνωστης σ.π.π. f δεν είναι συνεπής. Αυτό έχει ως αποτέλεσμα αρκετά τεστ να μην είναι και αυτά με τη σειρά τους συνεπή έναντι γενικών εναλλακτικών υποθέσεων H_1 , δηλαδή η ισχύς τους να μην συγκλίνει στο 1 καθώς $n \rightarrow \infty$, όπου n το μέγεθος του δείγματος. Ένας τέτοιος εκτιμητής είναι για παράδειγμα το ιστόγραμμα. Όταν το πλάτος των διαστημάτων του ιστογράμματος παραμένει σταθερό και δεν μειώνεται καθώς το μέγεθος του δείγματος μεγαλώνει, τότε η εκτίμηση παρουσιάζει αρκετά διακριτά άλματα και δεν εξομαλύνεται έχοντας ως αποτέλεσμα ο εκτιμητής του ιστογράμματος να μην προσεγγίζει την πραγματική σ.π.π. $f(x)$. Στη συγκεκριμένη περίπτωση για να είναι συνεπής ο εκτιμητής του ιστογράμματος αρκεί να υποθέσουμε ότι $h \rightarrow 0$ και $nh \rightarrow \infty$ καθώς $n \rightarrow \infty$. Για να αποφευχθεί αυτό το πρόβλημα, μπορεί να χρησιμοποιηθεί ως μετρική το ολοκλήρωμα του τετραγωνικού σφάλματος:

$$I = \int \{f(x) - f_0(x; \theta)\}^2 dx.$$

Αν εκτιμήσουμε την f με κάποια εκτιμητήρια, τότε η παραπάνω μετρική γίνεται:

$$\hat{I} = \int \{\hat{f}(x) - f_0(x; \theta)\}^2 dx.$$

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

Αντί να μετρήσουμε τη διαφορά των $f(x)$ και $f_0(x; \theta)$ σε ένα φιξαρισμένο σημείο, υπολογίζουμε τη διαφορά χρησιμοποιώντας όλη την ευθεία των πραγματικών αριθμών με την εφαρμογή του ολοκληρώματος. Από την άλλη η διαφορά υψώνεται στο τετράγωνο, για να ανιχνεύονται οι διαφορές της εκτιμήτριας της πραγματικής σ.π.π. από αυτής υπό τη μηδενική υπόθεση. Αυτό επιτυγχάνεται, διότι προσεγγιστικά οι διαφορές αθροίζονται και επειδή είναι υψωμένες στο τετράγωνο αποτρέπεται το ενδεχόμενο αναίρεσης θετικών διαφορών από αρνητικές.

Προκειμένου η μετρική να είναι υλοποιήσιμη στην πράξη, εκτιμούμε την άγνωστη σ.π.π. με τη μέθοδο των πυρήνων, καθώς έτσι μας δίνεται μία συνεχής εκτίμηση της την οποία μπορούμε και να ολοκληρώσουμε. Επομένως η μετρική αυτή θα πάρει τη μορφή:

$$\hat{I}_n = \int \{\hat{f}(x; h) - f_0(x; \theta)\}^2 dx,$$

όπου $\hat{f}(x) = \hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$ είναι ο εκτιμητής πυρήνων της άγνωστης f .

Το ολοκλήρωμα $\hat{I}_n$ αποτιμά το τετραγωνικό σφάλμα μεταξύ της εκτιμώμενης πυκνότητας και της θεωρούμενης πυκνότητας υπό τη μηδενική υπόθεση. Μικρές τιμές της στατιστικής συνάρτησης $\hat{I}_n$ υποδηλώνουν καλή προσαρμογή, ενώ μεγάλες τιμές υποδηλώνουν απόκλιση. Για λόγους απλότητας υποθέτουμε ότι η σ.π.π., υπό τη μηδενική υπόθεση, είναι πλήρως προσδιορισμένη και ότι η διάσταση της τυχαίας μεταβλητής X είναι $d = 1$. Σε όσα ακολουθούν, για λόγους ευκολίας, θα γράφουμε $\hat{f}(x)$ ή $\hat{f}$ αντί $\hat{f}(x; h)$ και f_0 αντί $f_0(x; \theta)$.

Το $\hat{I}_n$ ισοδύναμα, γράφεται:

$$\begin{aligned} \hat{I}_n &= \hat{I}_n(\hat{f}, f_0) = \int \{\hat{f}(x) - f_0(x)\}^2 dx \\ &= \int (\hat{f} - E\hat{f} + E\hat{f} - f_0)^2 dx \\ &= \int (\hat{f} - E\hat{f})^2 dx + 2 \int (\hat{f} - E\hat{f})(E\hat{f} - f_0) dx + \int (E\hat{f} - f_0)^2 dx. \end{aligned}$$

Αντικαθιστώντας τον τύπο της $\hat{f}$ στις ποσότητες $\hat{f} - E\hat{f}$ και $(\hat{f} - E\hat{f})^2$ προκύπτει

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

ΤΟΥΝ:

$$\begin{aligned}\hat{f} - E\hat{f} &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) - \sum_{i=1}^n \frac{1}{nh} E\left[K\left(\frac{x - X_i}{h}\right)\right] \\ &= \frac{1}{nh} \sum_{i=1}^n \left\{ K\left(\frac{x - X_i}{h}\right) - E\left[K\left(\frac{x - X_i}{h}\right)\right] \right\} \\ &= \frac{1}{nh} \cdot \sum_{i=1}^n \Delta_i(x),\end{aligned}$$

$$\text{όπου } \Delta_i(x) = K\left(\frac{x - X_i}{h}\right) - E\left[K\left(\frac{x - X_i}{h}\right)\right].$$

Οπότε:

$$\begin{aligned}(\hat{f} - E\hat{f})^2 &= (nh)^{-2} \left\{ \sum_{i=1}^n \Delta_i(x) \right\}^2 \\ &= (nh)^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i(x) \Delta_j(x).\end{aligned}$$

Θέτουμε $H_n(X_i, X_j) = \int \Delta_i(x) \Delta_j(x) dx$ και προκύπτει:

$$\int (\hat{f} - E\hat{f})^2 dx = (nh)^{-2} \sum_{i=1}^n \sum_{j=1}^n H_n(X_i, X_j).$$

Επομένως για να μελετήσουμε τη συνεισφορά κάθε όρου στη συνολική διακύμανση του ολοκληρώματος $\hat{I}_n$, διαχωρίζουμε το ολοκλήρωμα σε τέσσερα επιμέρους μέλη:

$$\begin{aligned}\hat{I}_n &= (n^2 h^2)^{-1} \sum_{i=1}^n \sum_{j=1}^n H_n(X_i, X_j) + 2 \int (\hat{f} - E\hat{f})(E\hat{f} - f_0) dx + \int (E\hat{f} - f_0)^2 dx \\ &= (n^2 h^2)^{-1} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n H_n(X_i, X_j) + (n^2 h^2)^{-1} \sum_{i=1}^n H_n(X_i, X_i) \\ &\quad + 2 \int (\hat{f} - E\hat{f})(E\hat{f} - f_0) dx + \int (E\hat{f} - f_0)^2 dx \\ &= \hat{I}_1 + \hat{I}_2 + \hat{I}_3 + \hat{I}_4,\end{aligned}$$

όπου $\hat{I}_1, \hat{I}_2, \hat{I}_3$ και $\hat{I}_4$ είναι αντίστοιχα οι όροι του παραπάνω αθροίσματος (βλέπε Pagan and Ullah (2005)).

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

Η εύρεση της ασυμπτωτικής κατανομής αυτού του αθροίσματος με το οποίο ισούται η συνάρτηση $\hat{I}_n$, βασίζεται στους ισχυρισμούς που παρουσιάσαμε για τη συνάρτηση του πυρήνα στις σελίδες 11 και 14, καθώς και στις παρακάτω υποθέσεις:

- i) Οι παρατηρήσεις $X_1, \dots, X_n$ είναι ανεξάρτητες και ισόνομα κατανεμημένες.
- ii) Ο πυρήνας K είναι συμμετρική συνάρτηση γύρω από το μηδέν και ικανοποιεί:

$$- \int y^2 K(y) dy \neq 0$$

$$- \int K^2(y) dy < \infty.$$

- iii) Οι παράγωγοι δεύτερης τάξης της f είναι συνεχείς, ολοκληρώσιμες και φραγμένες σε μία περιοχή του x .

$$\text{iv) } h = h_n \xrightarrow{n \rightarrow \infty} 0.$$

$$\text{v) } nh_n \xrightarrow{n \rightarrow \infty} \infty.$$

$$\text{vi) } n(h_n)^5 \xrightarrow{n \rightarrow \infty} 0.$$

Με βάση αυτές τις υποθέσεις λοιπόν, αποδεικνύεται, από τους Pagan και Ullah, ότι υπό την H_0 ισχύει ότι:

$$nh^{\frac{1}{2}}(\hat{I}_n - c^*(n)) \xrightarrow{d} N(0, \sigma^2),$$

όπου:

- $c^*(n) = c(n) + \int (Ef - f_0)^2 dx,$
- $c(n) = (nh)^{-1} \int K^2(y) dy,$
- $\sigma^2 = 2\{\int f_0^2(x) dx\} \left[\int \left\{ \int K(y)K(y+t) dy \right\}^2 dt \right].$

Σύμφωνα με την Ενότητα 2.9 του βιβλίου των Pagan and Ullah (2005), οι ποσότητες $\hat{I}_2, \hat{I}_3$ και $\hat{I}_4$ ασυμπτωτικά συγκλίνουν στο μηδέν και επομένως η κατανομή της στατιστικής συνάρτησης $nh^{\frac{1}{2}}(\hat{I}_n - c^*(n))$ είναι η ίδια με αυτήν της $nh^{\frac{1}{2}}\hat{I}_1$. Το παραπάνω αποτέλεσμα βασίζεται στο κεντρικό οριακό Θεώρημα του Hall (1984), για U - στατιστικά, το οποίο διατυπώνεται στη συνέχεια.

Θεώρημα Hall (1984). Έστω ένα δείγμα $X_1, \dots, X_n$ ανεξάρτητων και ισόνομων τυχαίων παρατηρήσεων.

Ορίζουμε U - στατιστικό την παρακάτω στατιστική συνάρτηση:

$$U_n = \sum_{1 \leq i < j \leq n} H_n(X_i, X_j)$$

όπου:

- H_n είναι συμμετρική συνάρτηση,
- $E[H_n(X_1, X_2)|X_1] \xrightarrow{a.s.} 0$,
- $E[H_n(X_1, X_2)^2] < \infty$, $\forall n \in \mathbb{N}$.

Αν ισχύει:

$$\frac{E[G_n(X_1, X_2)^2] + n^{-1}E[H_n(X_1, X_2)^4]}{\{E[H_n(X_1, X_2)^2]\}^2} \xrightarrow{n \rightarrow \infty} 0,$$

όπου $G_n(a, b) = E[H_n(X_1, a)H_n(X_1, b)]$, τότε:

$$U_n \xrightarrow{d} N(0, n^2 E[H_n(X_1, X_2)^2]/2).$$

Για την απόδειξη του παραπάνω Θεωρήματος παραπέμπουμε τον αναγνώστη στο Θεώρημα 1 του άρθρου των Hall and Heyde (1981).

Ένας συνεπής εκτιμητής της διακύμανσης σ^2 μπορεί να προκύψει αν αντικαταστήσουμε την πραγματική σ.π.π. $f(x)$, με τον συνεπή εκτιμητή πυρήνων $\hat{f}(x; h)$. Δηλαδή:

$$E[f(x)] = \int f^2(x) dx,$$

εκτιμάται από την ποσότητα:

$$\begin{aligned} \int \hat{f}^2(x; h) dx &= \int \left\{ \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right\}^2 dx \\ &= \frac{1}{n^2 h^2} \int \left\{ \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right\}^2 dx \\ &= \frac{1}{n^2 h^2} \int \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{x - X_i}{h}\right) \cdot K\left(\frac{x - X_j}{h}\right) dx \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \sum_{j=1}^n \int K\left(\frac{x - X_i}{h}\right) \cdot K\left(\frac{x - X_j}{h}\right) dx. \end{aligned}$$

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

Θέτουμε $u = \frac{x-X_j}{h} \Rightarrow x = X_j + hu$ και $du = \frac{1}{h}dx$.

Οπότε:

$$\begin{aligned} \int \hat{f}^2(x; h) dx &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \int K\left(\frac{X_j - X_i}{h} + u\right) \cdot K(u) du \\ &= (n^2 h)^{-1} \sum_{i=1}^n \sum_{j=1}^n \int K\left(\frac{X_i - X_j}{h} - u\right) \cdot K(u) du \\ &= (n^2 h)^{-1} \sum_{i=1}^n \sum_{j=1}^n K * K\left(\frac{X_i - X_j}{h}\right), \end{aligned}$$

όπου για δύο συνεχείς συναρτήσεις $f, g : \mathbb{R} \rightarrow \mathbb{R}$, η συνέλιξη των f, g συμβολίζεται με $f * g$ και ορίζεται ως εξής

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y) dy, \forall x \in \mathbb{R}.$$

Επομένως το στατιστικό τεστ και η ασυμπτωτική του κατανομή, υπό την $H_0: f(x) = f_0(x; \theta)$, θα είναι:

$$T = nh^{\frac{1}{2}} \frac{(\hat{I}_n - c^*(n))}{\hat{\sigma}} \xrightarrow[H_0]{d} N(0, 1),$$

όπου $\hat{\sigma}^2 = 2\left\{\int \hat{f}^2(x; h) dx\right\} \left[\int \left\{\int K(y)K(y+t) dy\right\}^2 dt\right]$.

Ένας διαφορετικός τρόπος εκτίμησης της διακύμανσης είναι η εκτίμηση της αναμενόμενης τιμής $E[f(x)]$, με τον δειγματικό της μέσο, εκτιμώντας ταυτόχρονα την άγνωστη $f(x)$, με τον εκτιμητή πυρήνων $\hat{f}(x; h)$, δηλαδή όταν

$$E[\hat{f}(x)] = n^{-1} \sum_{i=1}^n \hat{f}(X_i; h).$$

Προκειμένου όμως να διαπιστώσουμε την επιρροή του εύρους h στη σύγκλιση της $\hat{I}_n$, στην κανονική κατανομή, θα παρουσιάσουμε μία εναλλακτική προσέγγιση με βάση το άρθρο της Fan (1994).

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

Έχουμε ότι:

$$\begin{aligned}
 \hat{I}_n &= \int \{\hat{f}(x; h) - f_0(x)\}^2 dx \\
 &= \int \{\hat{f}(x; h) - f(x) + f(x) - f_0(x)\}^2 dx \\
 &= \int \{\hat{f}(x; h) - f(x)\}^2 dx + \int \{f_0(x) - f(x)\}^2 dx \\
 &\quad - 2 \int \{\hat{f}(x; h) - f(x)\} \{f_0(x) - f(x)\} dx \\
 &= I_{1n} + I_{2n} - 2I_{3n}.
 \end{aligned}$$

Το παρακάτω Λήμμα μας παρουσιάζει την ασυμπτωτική κατανομή του $\hat{I}_n$, έχοντας υιοθετήσει τις παρακάτω υποθέσεις:

- i) Ο πυρήνας K είναι φραγμένη, συμμετρική συνάρτηση στο $\mathbb{R}$, που ικανοποιεί $\int |K(u)| du < \infty$, $|u||K(u)| \xrightarrow{|u| \rightarrow \infty} 0$ και είναι τάξης $m \in \mathbb{N}$.
- ii) Η παράμετρος του εύρους ζώνης του πυρήνα $h = h_n$ είναι τέτοια ώστε $h_n \xrightarrow{n \rightarrow \infty} 0$ και $nh \xrightarrow{n \rightarrow \infty} \infty$.
- iii) Η σ.π.π. f και οι m το πλήθος παράγωγοί της, είναι φραγμένες και ομοιόμορφα συνεχείς συναρτήσεις στο $\mathbb{R}$.

Λήμμα (Fan 1994). Έστω $c(n) = (nh)^{-1} \int K^2(y) dy + \int (E\hat{f}(x) - f(x))^2 dx$ και $f^{(m)}(x) = \frac{d^m}{dx^m} f(x)$.

Ορίζουμε:

$$d(n) = \begin{cases} n^{1/2}h^{-m}, & \text{αν } nh^{1+2m} \xrightarrow{n \rightarrow \infty} \infty \\ nh^{1/2}, & \text{αν } nh^{1+2m} \xrightarrow{n \rightarrow \infty} 0 \\ n^{\frac{1+4m}{2(1+2m)}}, & \text{αν } nh^{1+2m} \xrightarrow{n \rightarrow \infty} \lambda, \end{cases}$$

όπου $\lambda > 0$.

Με βάση τις τρεις παραπάνω υποθέσεις, θα ισχύουν:

a)

$$\begin{aligned}
 d(n)(I_{1n} - c(n)) &= 2d(n)(n^2h^2)^{-1} \sum_{1 \leq i < j \leq n}^n H_n(X_i, X_j) \\
 &\quad + 2d(n) \int \{\hat{f}(x) - E\hat{f}(x)\} \{E\hat{f}(x) - f(x)\} dx + o_p(1),
 \end{aligned}$$

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

όπου:

$$H_n(X_i, X_j) = \int \Delta_i(x) \Delta_j(x) dx,$$

$$\Delta_i(x) = K\left(\frac{x - X_i}{h}\right) - E\left[K\left(\frac{x - X_i}{h}\right)\right].$$

β)

$$nh^{1/2} \left\{ (n^2 h^2)^{-1} \sum_{1 \leq i < j \leq n} H_n(X_i, X_j) \right\} \xrightarrow{d} 2^{-1/2} \sigma_2 \cdot Z,$$

$$n^{1/2} h^{-m} \int \{\hat{f}(x) - E\hat{f}(x)\} \{E\hat{f}(x) - f(x)\} dx \xrightarrow{d} k_m \sigma_1 \cdot Z,$$

όπου:

$$\sigma_1^2 = \int \{f^{(m)}(x)\}^2 f(x) dx - \left[\int \{f^{(m)}(x)\} f(x) dx \right]^2 = \text{Var}[f^{(m)}(x)],$$

$$\sigma_2^2 = \left\{ \int f^2(x) dx \right\} \cdot \left[\int \left\{ \int K(y) K(y+t) dy \right\}^2 dt \right],$$

με Z να συμβολίζει την τυπική κανονική κατανομή.

γ)

$$d(n)(I_{1n} - c(n)) \xrightarrow{d} \begin{cases} 2k_m \sigma_1 \cdot Z, & nh^{1+2m} \xrightarrow{n \rightarrow \infty} \infty \\ 2^{1/2} \sigma_2 \cdot Z, & nh^{1+2m} \xrightarrow{n \rightarrow \infty} 0 \\ \left(4k_m^2 \sigma_1^2 \lambda^{\frac{2m}{1+2m}} + 2\sigma_2^2 \lambda^{-\frac{1}{1+2m}} \right)^{1/2} Z, & nh^{1+2m} \xrightarrow{n \rightarrow \infty} \lambda, \end{cases}$$

όπου $\lambda > 0$ και k_m είναι μία ακολουθία συναρτήσεων του $m \in \mathbb{N}$.

Άρα υπό την $H_0: f = f_0$ θα ισχύει $\hat{I}_n = I_{1n}$ και η κατανομή του $\hat{I}_n$ θα είναι η παραπάνω.

Η απόδειξη του παραπάνω Λήμματος βρίσκεται αναλυτικά στο άρθρο της Fan (1994).

Το Λήμμα της Fan μας δείχνει ότι η ασυμπτωτική κατανομή του ολοκληρώματος του τετραγωνικού σφάλματος της $\hat{f}$, I_{1n} , με την ανάλογη κανονικοποίηση, εξαρτάται σε πολύ μεγάλο βαθμό από την ομαλοποίηση που εφαρμόζεται στα δεδομένα. Με άλλα λόγια εξαρτάται από την παράμετρο εκτίμησης h , της πραγματικής f , με τη μέθοδο των πυρήνων. Αυτό μπορούμε να το διαπιστώσουμε

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

αν προσθαφαιρέσουμε την αναμενόμενη τιμή της $\hat{f}(x)$, $E[\hat{f}(x)]$, στο ολοκλήρωμα I_{1n} . Συγκεκριμένα θα έχουμε:

$$\begin{aligned} I_{1n} &= \int \{\hat{f}(x) - f(x)\}^2 dx \\ &= \int \{\hat{f}(x) - E\hat{f}(x) + E\hat{f}(x) - f(x)\}^2 dx \\ &= \int \{\hat{f} - E\hat{f}\}^2 dx + 2 \int \{\hat{f}(x) - E\hat{f}(x)\}\{E\hat{f}(x) - f(x)\} dx \\ &\quad + \int \{E\hat{f}(x) - f(x)\}^2 dx. \end{aligned}$$

Ο τρίτος όρος του παραπάνω αθροίσματος δεν είναι στοχαστικός και άρα η κατανομή του I_{1n} εξαρτάται από τους δύο πρώτους όρους. Ο πρώτος όρος περιλαμβάνει το στοχαστικό κομμάτι της απόκλισης της $\hat{f}$ από την f , ενώ ο δεύτερος όρος περιλαμβάνει τόσο το στοχαστικό, όσο και το συστηματικό κομμάτι της αντίστοιχης απόκλισης. Συγκεκριμένα:

- Αν $nh^{1+2m} \xrightarrow{n \rightarrow \infty} \infty$, τότε τα δεδομένα υπερομαλοποιούνται και το στοχαστικό κομμάτι της απόκλισης γίνεται ασυμπτωτικά αμελητέο σε σύγκριση με το αντίστοιχο συστηματικό. Σε αυτήν την περίπτωση η κατανομή του I_{1n} , όντας κατάλληλα κανονικοποιημένη, εξαρτάται από τον δεύτερο όρο.
- Αν $nh^{1+2m} \xrightarrow{n \rightarrow \infty} 0$, τότε τα δεδομένα υπομαλοποιούνται, έχοντας ως συνέπεια το στοχαστικό κομμάτι της απόκλισης να επικρατεί του συστηματικού μέρους. Συνεπώς, η κατανομή του I_{1n} θα εξαρτάται από τον πρώτο όρο του παραπάνω αθροίσματος.
- Τέλος, αν $nh^{1+2m} \xrightarrow{n \rightarrow \infty} \lambda$ με $\lambda > 0$, τότε τόσο ο πρώτος όσο και ο δεύτερος όρος θα συμβάλουν το ίδιο ασυμπτωτικά. Επομένως, η ασυμπτωτική κατανομή του I_{1n} είναι αυτή του αθροίσματος των δύο πρώτων όρων.

Η ασυμπτωτική κατανομή που παρουσιάζει το Λήμμα, μπορεί να χρησιμοποιηθεί για κατασκευή τεστ ελέγχου υποθέσεων της μορφής:

$$H_0: f(x) = f_0(x), \forall x \text{ έναντι } H_1: f(x) \neq f_0(x), \text{ για κάποιο } x,$$

όπου $f_0(x)$ είναι μία πλήρως γνωστή σ.π.π..

Σε καταστάσεις που η μηδενική υπόθεση είναι πιο σύνθετη, δηλαδή υπάρχει πεπερασμένο πλήθος άγνωστων παραμέτρων στη σ.π.π. f_0 , υπό την H_0 ,

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

χρειάζεται να ληφθούν υπόψιν και οι δύο τελευταίοι όροι του αθροίσματος $\hat{I}_n = I_{1n} + I_{2n} - 2I_{3n}$, προκειμένου να κατασκευαστεί ένα τεστ ελέγχου τέτοιου είδους υποθέσεων. Στο άρθρο της Fan (1994), παρουσιάζονται τρία στατιστικά τεστ για έλεγχο υποθέσεων της μορφής:

$$H_0: f(x) = f_0(x, \beta), \forall x \quad \text{έναντι} \quad H_1: f(x) \neq f_0(x, \beta), \text{ για κάποιο } x,$$

όπου β είναι το διάνυσμα της άγνωστης παραμέτρου. Έπειτα αναλύει την ασυμπτωτική τους κατανομή και δείχνει την επιρροή του εύρους των πυρήνων, $h = h_n$, στη σύγκλιση στην τυπική κανονική κατανομή. Σε αυτήν τη διατριβή θα παρουσιάσουμε έναν έλεγχο καλής προσαρμογής, με χρήση εκτιμήτριας πυρήνων, στον οποίο η σ.π.π. της μηδενικής υπόθεσης ελέγχεται έναντι Pitman εναλλακτικών υποθέσεων.

ΚΕΦΑΛΑΙΟ 1 1.4. Έλεγχος καλής προσαρμογής με τη μέθοδο των πυρήνων

ΚΕΦΑΛΑΙΟ 2

ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΣ ΕΛΕΓΧΟΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΕΝΑΝΤΙ ΕΝΑΛΛΑΚΤΙΚΩΝ ΡΙΤΜΑΝ ΥΠΟΘΕΣΕΩΝ

2.1 Περιγραφή του ελέγχου υποθέσεων

Έστω δείγμα $X_1, \dots, X_n$ ανεξάρτητων και ισόνομων παρατηρήσεων, διάστασης $d = 1$, από μία άγνωστη κατανομή με σ.π.π. f . Ο στόχος μας είναι να ελέγξουμε την υπόθεση (1.6). Το ενδιαφέρον επικεντρώνεται στον έλεγχο υποθέσεων έναντι Pitman εναλλακτικών σύμφωνα με τις οποίες, η H_1 γίνεται:

$$H_n : f(x) \equiv f_n(x) = f(x; \theta_1) + a_n \eta(x), x \in \mathbb{R},$$

όπου $\theta_1 \in \Theta$, $a_n \eta(x)$ ικανοποιεί κάποιες υποθέσεις που θα παρουσιάσουμε μετέπειτα και είναι τέτοιο ώστε, η $f_n(x)$ να μην ανήκει στο παραμετρικό μοντέλο P_Θ . Η συνάρτηση $\eta(x)$ είναι άγνωστη και μπορεί να είναι ανεξάρτητη από το μέγεθος του δείγματος n ή μπορεί να οριστεί ως άθροισμα συναρτήσεων που είναι εξαρτημένες από το n . Συχνά χρησιμοποιούμενες μορφές των εναλλακτικών Pitman είναι οι:

$$H_{1n} : f_{1n}(x) = f(x; \theta_1) + n^{-\beta} \eta(x),$$

$$H_{2n} : f_{2n}(x) = f(x; \theta_1) + n^{-\varepsilon} \sum_{\xi=1}^{\psi} \eta_{\xi} \left(\frac{x - c_{\xi}}{n^{-\gamma}} \right),$$

οι οποίες αντανακλούν είτε ομαλές αποκλίσεις από τη μηδενική υπόθεση (μέσω μίας συνάρτησης $\eta(x)$) είτε τοπικές αποκλίσεις (μέσω συναρτήσεων η_{ξ}) γύρω από σημεία c_{ξ} . Επιπλέον σημειώνουμε ότι:

- $\beta > 0, \gamma > 0$ είναι τέτοια ώστε $hn^\gamma \xrightarrow{n \rightarrow \infty} 0$,
- $\varepsilon > 0$,
- $\eta(x)$ είναι συνεχής και τετραγωνικά ολοκληρώσιμη συνάρτηση,
- $\eta_\xi(x)$ είναι τετραγωνικά ολοκληρώσιμες συναρτήσεις, οι οποίες είναι φραγμένες ασυμπτωτικά και έχουν συνεχείς παραγώγους μέχρι και δεύτερης τάξης,
- c_ξ είναι πραγματικές σταθερές τέτοιες ώστε, $c_\xi \neq c_\zeta$, αν $\xi \neq \zeta$.

2.2 Κατασκευή του ελέγχου καλής προσαρμογής

Έστω τα προσαρμοσμένα σημεία στην ευθεία των πραγματικών αριθμών, $x_j = \Delta \cdot c \cdot j \in \mathbb{R}$, όπου $\Delta = \Delta_n > 0$ μία μονοδιάστατη ποσότητα, που εξαρτάται από το μέγεθος του δείγματος n .

Ορίζουμε το σύνολο:

$$D_n = [-g, g] \cap \mathbb{Z}, \text{ όπου } g \in \mathbb{N} \text{ σταθερά.}$$

Για $j \in D_n$ ορίζουμε:

$$\tilde{f}_j = \# \left\{ i \in \{1, \dots, n\} \mid X_i \in \left(x_j - \frac{\Delta c}{2}, x_j + \frac{\Delta c}{2} \right) \right\}$$

και

$$Y_j = (n\Delta)^{-1} \tilde{f}_j.$$

Σημειώνουμε εδώ ότι η ποσότητα $Y_j = (n\Delta)^{-1} \tilde{f}_j, j \in D_n$ δεν είναι τίποτα άλλο από την εμπειρική εκτίμηση της πραγματικής σ.π.π. f στο σημείο x_j μέσω ιστόγραμματος με εύρος Δ . Έστω επίσης $\hat{\theta}_0$ ένας συνεπής εκτιμητής της παραμέτρου θ_0 , όπως παραδείγματος χάρη ο Ε.Μ.Π.. Συνεπώς η εκτίμηση της $f(x; \theta_0)$, η οποία δεν είναι πλήρως προσδιορισμένη, καθώς η παράμετρος θ_0 είναι άγνωστη, θα είναι $\hat{f}_0(x) = f(x; \hat{\theta}_0)$.

Ορίζουμε την ποσότητα:

$$\tilde{S}_j(h; \hat{\theta}_0) = \sum_{\substack{k \in D_n \\ k \neq j}} K \left(\frac{x_j - x_k}{h} \right) (Y_k - \hat{f}_0(x_k)),$$

όπου K είναι ο πυρήνας, που τυπικά υποθέτουμε ότι είναι μία συνεχής σ.π.π. και $h > 0$ είναι η παράμετρος, που ελέγχει το εύρος του.

Υπό την H_0 , βάση του ισχυρού νόμου των μεγάλων αριθμών και των υποθέσεων A.1 - A.7, που θα αναφέρουμε παρακάτω, θα ισχύει ότι για κάθε $j \in D_n$: $Y_j \xrightarrow{a.s.} f(x_j; \theta_0)$. Επομένως, υπό την H_0 , η ποσότητα $\tilde{S}_j(h; \hat{\theta}_0)$ θα είναι σχετικά μικρή, όταν $n \rightarrow \infty$, ενώ υπό την εναλλακτική υπόθεση, η $\tilde{S}_j(h; \hat{\theta}_0)$ θα τείνει να είναι πιο απομακρυσμένη από το μηδέν. Ο σκοπός είναι να συγκεντρώσουμε αρκετή πληροφορία, που θα μας βοηθήσει στην απόρριψη ή μη της μηδενικής υπόθεσης. Για αυτό ορίζουμε την παρακάτω ποσότητα, η οποία είναι ένα σταθμισμένο άθροισμα των συναρτήσεων $\tilde{S}_j(h; \hat{\theta}_0)$.

$$\begin{aligned} \hat{S}_n(h) &= \sum_{j \in D_n} \tilde{S}_j(h; \hat{\theta}_0) \{Y_j - \hat{f}_0(x_j)\} \\ &= \sum_{j=-g}^g \sum_{\substack{k=-g \\ k \neq j}}^g K\left(\frac{x_j - x_k}{h}\right) \{Y_j - \hat{f}_0(x_j)\} \{Y_k - \hat{f}_0(x_k)\}. \end{aligned}$$

Η στατιστική συνάρτηση, που θα χρησιμοποιηθεί είναι η $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$. Συγκεκριμένα, τα δεδομένα κατανομονται σε διαστήματα, που διαμορφώνονται διαχωρίζοντας το εύρος των δεδομένων σε g το πλήθος διαστήματα ίσου μήκους, $\Delta = \frac{X_{(n)} - X_{(1)}}{g}$, όπου $X_{(n)}$ και $X_{(1)}$ είναι η μέγιστη και η ελάχιστη τιμή των παρατηρήσεων, που διαθέτουμε. Επομένως, θα έχουμε ότι το δεξί άκρο κάθε διαστήματος θα είναι της μορφής $X_{(1)} + k\Delta$, ενώ το κέντρο κάθε διαστήματος θα είναι $x_k = X_{(1)} + (k - \frac{1}{2})\Delta$, $k \in \{1, \dots, g\}$. Η επιλογή της παραμέτρου του εύρους ζώνης του πυρήνα, h , θα αναλυθεί παρακάτω.

2.3 Υποθέσεις

Ορίζουμε για $p \in \mathbb{N}$ και συνάρτηση $F : \mathbb{R} \rightarrow \mathbb{R}$:

$$R_p(F) = \int_{-\infty}^{\infty} F(u)^p du.$$

Επιπλέον για $g \in \mathbb{N}$, ορίζουμε:

$$\sum_{j \neq k} = \sum_{-g \leq j, k \leq g} I_{\{j \neq k\}}.$$

Για την εύρεση της ασυμπτωτικής κατανομής και της αντίστοιχης ισχύος της στατιστικής συνάρτησης $\hat{S}_n(h)$, έχουν χρησιμοποιηθεί οι παρακάτω υποθέσεις:

- A.1) Το δείγμα $X_1, \dots, X_n$, αποτελείται από ανεξάρτητες και ισόνομες παρατηρήσεις, διάστασης $d = 1$, από μία συνεχή σ.π.π. $f(x)$, $x \in \mathbb{R}$.
- A.2) Έστω ότι $\Delta = n^{-\lambda}$, $\lambda \in (\frac{1}{2}, 1)$ και $g = g_n$ είναι τέτοια ώστε: $g\Delta \xrightarrow{n \rightarrow \infty} \infty$ και $\frac{g}{n} \xrightarrow{n \rightarrow \infty} 0$. Επίσης, $h \xrightarrow{n \rightarrow \infty} 0$, $\frac{h}{\Delta} \xrightarrow{n \rightarrow \infty} \infty$, $nh \xrightarrow{n \rightarrow \infty} \infty$ και $nh\Delta^4 \xrightarrow{n \rightarrow \infty} 0$.
- A.3) Οι παραγώγοι πρώτης και δεύτερης τάξης της $f(x)$ είναι φραγμένες και ομοιόμορφα συνεχείς στο $\mathbb{R}$.
- A.4) Ο πυρήνας K είναι συνεχής, συμμετρική και φραγμένη σ.π.π. στο διάστημα $[-1, 1]$.
- A.5) α) Η $f(x; \theta)$ είναι διαφορίσιμη ως προς θ , η πρώτη παράγωγός της είναι συνεχής στο x και στο θ , $0 < E[f'(\cdot; \theta)^2] < \infty$ και $R_2(f_0) < \infty$, όπου $f_0(x) = f(x; \theta_0)$.
β) Η ακολουθία των συνεχών συναρτήσεων $\alpha_n \eta(x)$ είναι τέτοια ώστε $\alpha_n \xrightarrow{n \rightarrow \infty} 0$, $0 < R_2(\eta(x)) < \infty$ και $\int \eta(x) dx = 0$.
- A.6) Υπό την H_0 ισχύει ότι $\|\hat{\theta} - \theta_0\| = O_p(n^{-\frac{1}{2}})$, ενώ υπό την H_1 ισχύει ότι $\|\hat{\theta} - \theta_1\| = O_p(n^{-\frac{1}{2}})$, όπου $\|\cdot\|$ συμβολίζει την Ευκλείδεια νόρμα στον $\mathbb{R}^n$.
- A.7) Όλες οι παραγώγοι πρώτης, δεύτερης και τρίτης τάξης ως προς θ , είναι ομοιόμορφα συνεχείς, φραγμένες και τετραγωνικά ολοκληρώσιμες ως προς τη μετρική Lebesgue στο $\mathbb{R}$.

2.4 Ασυμπτωτική κατανομή

Η ασυμπτωτική κατανομή της στατιστικής συνάρτησης $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$ αποδεικνύεται στο Θεώρημα 1 του άρθρου των Bagkavos et al. (2023), το οποίο είναι το παρακάτω.

Θεώρημα 2.1. Υπό τις συνθήκες A.1-A.7 η ασυμπτωτική κατανομή της στατιστικής συνάρτησης $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$,

i) υπό τη μηδενική υπόθεση $H_0 : f = f_0 \in P_\Theta$ είναι:

$$n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h) \xrightarrow{d} \mathcal{N}(0, \sigma_{0, \theta_0}^2),$$

όπου $\sigma_{0, \theta_0}^2 = 2R_2(K)R_2(f_0)$, $\mu \in f_0(x) = f(x; \theta_0)$.

ii) Υπό την εναλλακτική υπόθεση $H_1 : f = f_1 \notin P_\Theta$, θα ισχύει:

$$n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h) \asymp_p \mu_1,$$

όπου $\mu_1 \asymp nh^{\frac{1}{2}}(1 + n\Delta^4) \xrightarrow{n \rightarrow \infty} \infty$.

Απόδειξη. Τα μέρη (i) και (ii) του Θεωρήματος 2.1 αποδεικνύονται άμεσα από τα μέρη (i) και (ii) του Λήμματος 4.5 αντίστοιχα. $\square$

Ασυμπτωτικά λοιπόν, με επίπεδο σημαντικότητας α , το τεστ απορρίπτει τη μηδενική υπόθεση H_0 , αν $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h) > z_\alpha \sigma_{0,\theta_0}$, όπου z_α είναι το $(1 - \alpha)$ ποσοστημόριο της τυπικής κανονικής κατανομής. Το γεγονός ότι στην κρίσιμη περιοχή δεν υπάρχουν και αρνητικές τιμές είναι εμφανές από τη συμπεριφορά της στατιστικής συνάρτησης, υπό την εναλλακτική υπόθεση. Παρατηρούμε δηλαδή, ότι υπό την H_1 , η στατιστική συνάρτηση παίρνει μεγάλες και θετικές τιμές και για αυτό μιλάμε για κρίσιμη περιοχή της μορφής $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h) > c > 0$.

2.5 Ασυμπτωτική ισχύς του ελέγχου

Η ασυμπτωτική ισχύς της στατιστικής συνάρτησης $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$ αποδεικνύεται στο Πρόσμα 1 του άρθρου των Bagkavos et al. (2023), το οποίο είναι το παρακάτω.

Πρόσμα 2.1. Έστω $\alpha \in (0, 1)$, $\beta > 0$ και $\varepsilon > 0$ είναι σταθερές ανεξάρτητες του μεγέθους του δείγματος n και έστω Φ είναι η α.σ.κ. της τυπικής κανονικής κατανομής. Η ασυμπτωτική ισχύς του στατιστικού $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$ έναντι των εναλλακτικών υποθέσεων H_{1n} και H_{2n} αντίστοιχα, είναι:

i) Υπό την H_{1n} :

$$P\left(n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h) > z_\alpha \sigma_{0,\theta_0} \mid H_{1n}\right) = \begin{cases} \alpha, & \beta > \frac{1}{4}\delta + \frac{1}{2} \\ \Phi(c_0 l - z_\alpha), & \beta = \frac{1}{4}\delta + \frac{1}{2} \\ 1, & \beta < \frac{1}{4}\delta + \frac{1}{2}, \end{cases}$$

όπου $l = \sigma_{0,\theta_0}^{-1} R_2(n)$ και στην περίπτωση $\beta = \frac{1}{4}\delta + \frac{1}{2}$ θα ισχύει $n^{1-2\beta} h^{\frac{1}{2}} \xrightarrow{n \rightarrow \infty} c_0 \neq 0$.

ii) Υπό την H_{2n} , για $\gamma > 0$ τέτοιο ώστε $hn^\gamma \xrightarrow{n \rightarrow \infty} 0$ και $\varepsilon > 0$:

$$P\left(n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h) > z_a \sigma_{0,\theta_0} \mid H_{2n}\right) = \begin{cases} \alpha, & 2\varepsilon + \gamma > \frac{1}{2}\delta + 1 \\ \Phi(d_0 l' - z_a), & 2\varepsilon + \gamma = \frac{1}{2}\delta + 1 \\ 1, & 2\varepsilon + \gamma < \frac{1}{2}\delta + 1, \end{cases}$$

όπου $l' = \sigma_{0,\theta_0}^{-1} \sum_{j=1}^k R_2(n_j)$ και στην περίπτωση $2\varepsilon + \gamma = \frac{1}{2}\delta + 1$ θα ισχύει $n^{1-2\varepsilon-\gamma} h^{\frac{1}{2}} \xrightarrow{n \rightarrow \infty} d_0 \neq 0$.

Η απόδειξη του Πορίσματος 2.1 βρίσκεται στο άρθρο των Bagkavos et al.(2023) και παρουσιάζεται παρακάτω για χάριν πληρότητας της παρουσίασης.

Απόδειξη. i) Σύμφωνα με το Λήμμα 4.5, η ποσότητα $S_{1,n}$ που ορίζεται στη σχέση (4.23) είναι αμελητέα σε σύγκριση με το $S_{2,n}$. Συνεπώς αυτό που χρειάζεται να εξετάσουμε είναι η σχέση των ποσοτήτων $S_{3,n} = S_{4,n}$, $S_{5,n}$, $S_{6,n} = S_{7,n}$ και $S_{8,n}$ με το $S_{2,n}$. Πρώτα θα μελετήσουμε τη στατιστική συνάρτηση $S_{5,n}$. Υπό την $H_{1,n} : f = f_{1,n}$,

$$S_{5,n} \sim n^{-2\beta} \frac{h}{\Delta^2} R_2(\eta),$$

οπότε

$$\frac{S_{5,n}}{\sqrt{\text{Var}[S_{2,n}]}} = O(n^{-2\beta} h \Delta^{-2}) O(h^{-1/2} n \Delta^2) = O(n^{1-2\beta} h^{1/2}).$$

Οι ποσότητες $S_{3,n} = S_{4,n}$, $S_{6,n} = S_{7,n}$ και $S_{8,n}$ είναι όλες μικρότερες, ως προς την τάξη σύγκλισης, από τα $S_{2,n}$ και $S_{5,n}$ για κάθε β . Χρησιμοποιώντας τις σχέσεις (4.52)–(4.54) και το Λήμμα 4.2, υπό την $H_{1,n} : f = f_{1,n}$, προκύπτει $E[W_{3,i}^2] \asymp n^{-2\beta} (h/\Delta)^2$. Κατά συνέπεια, μέσω της σχέσης (4.56) θα έχουμε

$$E[S_{3,n}^2] \asymp \frac{1}{n\Delta^2} O(n^{-2\beta} h^2 \Delta^{-2}) = O(n^{1-2\beta} h^2 \Delta^{-4}),$$

άρα

$$\sqrt{\frac{\text{Var}[S_{3,n}]}{\text{Var}[S_{2,n}]}} = \sqrt{O(n^{1-2\beta} h^2 \Delta^{-4}) O(h^{-1} n^2 \Delta^4)} = O(n^{1/2-\beta} h^{1/2}).$$

Επομένως το $S_{3,n}$ υπερκαλύπτεται από το $S_{5,n}$ όταν $\beta < 1/2$ και από το $S_{2,n}$ όταν $\beta > 1/2 - \varepsilon$, όπου $\varepsilon > 0$ είναι ένας αρκετά μικρός και θετικός πραγματικός αριθμός. Παρόμοιοι υπολογισμοί δείχνουν ότι οι ποσότητες $S_{6,n} = S_{7,n}$ υπερκαλύπτονται από τα $S_{2,n}$ και $S_{5,n}$ και το $S_{8,n}$ είναι πάντα μικρότερης τάξης από το

$S_{2,n}$ λόγω της σχέσης (4.61). Αυτό που μένει να εξετάσουμε τώρα είναι η σχέση του $S_{5,n}$ με το $S_{2,n}$. Η συνθήκη $n^{1-2\beta}h^{1/2} \rightarrow \infty$ είναι ισοδύναμη με τη συνθήκη $\ln n^{1-2\beta}h^{1/2} \rightarrow \infty$ και έχοντας υπόψιν την υπόθεση ότι $(\ln h)/(\ln n) \xrightarrow{n \rightarrow \infty} \delta$, προκύπτουν οι παρακάτω τρεις περιπτώσεις:

$$\beta < \frac{1}{2} + \frac{1}{4}\delta,$$

και αυτό οδηγεί σε ασυμπτωτική σύγκλιση της ισχύος στο 1. Η συνθήκη $n^{1-2\beta}h^{1/2} \rightarrow 0$, αντιστοιχεί στην ανισότητα

$$\beta > \frac{1}{2} + \frac{1}{4}\delta,$$

η οποία έχει ως αποτέλεσμα η απόσταση από τη μηδενική υπόθεση να είναι ασυμπτωτικά αμελητέα. Σε αυτήν την περίπτωση η πιθανότητα απόρριψης της μηδενικής υπόθεσης ισούται ασυμπτωτικά με α . Τέλος προκύπτει η περίπτωση:

$$\beta = \frac{1}{2} + \frac{1}{4}\delta.$$

Σε αυτήν την περίπτωση, θέτοντας $\ell = \sigma_0^{-1}R_2(\eta)$ και υποθέτοντας ότι $n^{1-2\beta}h^{1/2} \rightarrow c_0 \neq 0$, καταλήγουμε στο ότι η πιθανότητα απόρριψης της μηδενικής υπόθεσης ισούται ασυμπτωτικά με $\Phi(c_0\ell - z_\alpha)$.

ii) Όπως και την απόδειξη του μέρους (i) του Πορίσματος 2.1, οι στατιστικές συναρτήσεις $\hat{S}_n(h)$, $S_{2,n}$ και $S_{5,n}$ που ορίζονται από τις σχέσεις (4.23), (4.25) και (4.27) αντίστοιχα, παίζουν κύριο ρόλο στην απόδειξη. Υπό την $H_{2,n}$: $f = f_{2,n}$, όπου

$$f_{2,n}(x) = f_0(x) + n^{-\epsilon} \sum_{\xi=1}^{\psi} \eta_{\xi} \left(\frac{x - c_{\xi}}{n^{-\gamma}} \right) + o(n^{-\epsilon-\gamma}),$$

και λαμβάνοντας υπόψιν την ασυμπτωτική αμεροληψία του ιστογράμματος θα έχουμε υπό την $H_{2,n}$ ότι $E[Y_j] = f_{2,n} + O(n^{-\epsilon})$ ομοιόμορφα πάνω από όλα τα j καθώς $n \rightarrow \infty$. Οπότε το $S_{5,n}$ ασυμπτωτικά γίνεται

$$\begin{aligned}
S_{5,n} &= \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{E[Y_j] - f_0(x_j)\} \{E[Y_k] - f_0(x_k)\} \\
&= (1 + o(1)) \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{f_{2,n}(x_j) - f_0(x_j)\} \{f_{2,n}(x_k) - f_0(x_k)\} \\
&= (1 + o(1)) n^{-2\epsilon} \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \left\{ \sum_{\xi=1}^{\psi} \eta_{\xi} \left(\frac{x_j - c_{\xi}}{n^{-\gamma}} \right) \right\} \\
&\quad \times \left\{ \sum_{\zeta=1}^{\psi} \eta_{\zeta} \left(\frac{x_k - c_{\zeta}}{n^{-\gamma}} \right) \right\}. \quad (2.1)
\end{aligned}$$

Αν αναπτύξουμε τα δύο πεπερασμένα αθροίσματα, ο όρος που θα εμφανιστεί στο τελικό πολλαπλό άθροισμα θα είναι

$$n^{-2\epsilon} K\left(\frac{x_j - x_k}{h}\right) \eta_{\xi} \left(\frac{x_j - c_{\xi}}{n^{-\gamma}} \right) \eta_{\zeta} \left(\frac{x_k - c_{\zeta}}{n^{-\gamma}} \right). \quad (2.2)$$

Υπενθυμίζουμε ότι ο πυρήνας $K(\cdot)$ και οι συναρτήσεις $\eta_{\xi}(\cdot)$ έχουν συμπαγή φορέα και ότι $c_{\xi} \neq c_{\zeta}$ αν $\xi \neq \zeta$. Με βάση λοιπόν ότι έχουμε αναφέρει προκύπτει ότι για μεγάλο n , ο όρος (2.2) είναι διάφορος του μηδενός αν $\xi = \zeta$. Κατά συνέπεια, για να προσεγγίσουμε τη σχέση (2.1), αρκεί να εστιάσουμε το ενδιαφέρον μας στους όρους της μορφής (2.2) όταν $\xi = \zeta$. Επομένως η σχέση (2.1) γίνεται

$$S_{5,n} = (1 + o(1)) n^{-2\epsilon} \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \sum_{\xi=1}^{\psi} \eta_{\xi} \left(\frac{x_j - c_{\xi}}{n^{-\gamma}} \right) \eta_{\xi} \left(\frac{x_k - c_{\xi}}{n^{-\gamma}} \right). \quad (2.3)$$

Για δοθέν $\xi \in \{1, \dots, \psi\}$, θεωρούμε

$$T_{\xi,n} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \eta_{\xi} \left(\frac{x_j - c_{\xi}}{n^{-\gamma}} \right) \eta_{\xi} \left(\frac{x_k - c_{\xi}}{n^{-\gamma}} \right). \quad (2.4)$$

Συνεπώς θα έχουμε ότι $S_{5,n} = (1 + o(1)) n^{-2\epsilon} \sum_{\xi=1}^{\psi} T_{\xi,n}$. Λαμβάνοντας υπόψιν

ότι $x_k = x_j + (k - j)\Delta c$, προκύπτει

$$\begin{aligned} T_{\xi,n} &= \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \eta_{\xi}\left(\frac{x_j - c_{\xi}}{n^{-\gamma}}\right) \eta_{\xi}\left(\frac{x_k - c_{\xi}}{n^{-\gamma}}\right) \\ &= \sum_{j=-g}^g \sum_{k=-g}^g I(j \neq k) K\left(\frac{(k-j)\Delta c}{h}\right) \eta_{\xi}\left(\frac{x_j - c_{\xi}}{n^{-\gamma}}\right) \\ &\quad \times \eta_{\xi}\left(\frac{x_j - c_{\xi}}{n^{-\gamma}} + \frac{(k-j)\Delta c}{n^{-\gamma}}\right). \end{aligned}$$

Τώρα χρησιμοποιούμε τις ακόλουθες υποθέσεις: (i) Ο πυρήνας K έχει συμπαγή φορέα στο διάστημα $[-1, 1]$ και (ii) $hn^{\gamma} \rightarrow 0$, το οποίο συνεπάγεται ότι $h \ll n^{-\gamma}$, έχοντας έτσι ως αποτέλεσμα $\Delta c/h \gg \Delta c/n^{-\gamma}$. Αθροίζοντας λοιπόν από το $k = j - [h/(\Delta c)]$ μέχρι το $k = j + [h/(\Delta c)]$ και χρησιμοποιώντας το Λήμμα 4.2, προκύπτει

$$\begin{aligned} T_{\xi,n} &= (1 + o(1)) \frac{h}{\Delta} \left(\int_{-1}^1 K(u) du \right) \sum_{j=-g}^g \eta_{\xi}\left(\frac{x_j - c_{\xi}}{n^{-\gamma}}\right)^2 \\ &= (1 + o(1)) \frac{hn^{-\gamma}}{\Delta^2} R_2(\eta_{\xi}), \end{aligned} \tag{2.5}$$

καθώς $R_1(K) = 1$. Συνεχίζοντας με το $S_{5,n}$, θα έχουμε ότι

$$\begin{aligned} S_{5,n} &= (1 + o(1)) n^{-2\epsilon} \sum_{\xi=1}^{\psi} T_{\xi,n} \\ &= (1 + o(1)) n^{-2\epsilon} \sum_{\xi=1}^{\psi} \frac{hn^{-\gamma}}{\Delta^2} R_2(\eta_{\xi}) \\ &= (1 + o(1)) n^{-2\epsilon-\gamma} \Delta^{-2} h \sum_{\xi=1}^{\psi} R_2(\eta_{\xi}), \end{aligned} \tag{2.6}$$

και άρα

$$\frac{S_{5,n}}{\sqrt{\text{Var}[S_{2,n}]}} = n^{-2\epsilon-\gamma} \Delta^{-2} h^{-1/2} \Delta^2 n h \sigma_0^{-1} \sum_{\xi=1}^{\psi} R_2(\eta_{\xi})$$

$$= n^{1-2\epsilon-\gamma} h^{1/2} \sigma_0^{-1} \sum_{\xi=1}^{\psi} R_2(\eta_\xi), \quad (2.7)$$

από το οποίο προκύπτει η ασυμπτωτική ισχύ του S_n υπό την $H_{2,n}$: $f = f_{2,n}$, χρησιμοποιώντας τα ίδια επιχειρήματα με αυτά του (i) σκέλους της απόδειξης. Συγκεκριμένα αποδεικνύεται με παρόμοιο τρόπο με το (i) μέρος της απόδειξης ότι (α) οι ποσότητες $S_{1,n}$, $S_{3,n} = S_{4,n}$, $S_{6,n} = S_{7,n}$ και $S_{8,n}$ είναι όλες αμελητέες σε σύγκριση με το $S_{2,n} + S_{5,n}$, χρησιμοποιώντας παρόμοιο σκεπτικό με αυτό του (i) σκέλους της απόδειξης και της απόδειξης του Θεωρήματος 2.1; και (β) στην περίπτωση της ισότητας $2\epsilon + \gamma = \delta/2 + 1$, θα έχουμε ότι $n^{1-2\epsilon-\gamma} h^{1/2} \rightarrow d_0 \neq 0$. $\square$

2.6 Επιλογή εύρους ζώνης h

Προκειμένου να αναζητήσουμε μία μέθοδο επιλογής του εύρους ζώνης h , ώστε να είναι υλοποιήσιμο στην πράξη το στατιστικό τεστ $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$, πρέπει πρώτα να αντιληφθούμε τον τρόπο με τον οποίο λειτουργεί το τεστ. Αρχικά υπενθυμίζουμε τον τύπο της συνάρτησης $\hat{S}_n(h)$.

$$\hat{S}_n(h) = \sum_{j=-g}^g \sum_{\substack{k=-g \\ k \neq j}}^g K\left(\frac{x_j - x_k}{h}\right) \{Y_j - \hat{f}_0(x_j)\} \{Y_k - \hat{f}_0(x_k)\},$$

όπου:

- $Y_j = \frac{1}{n\Delta} \tilde{f}_j = \frac{1}{n\Delta} \sum_{i=1}^n I_{\{X_i \in (x_j - \frac{\Delta}{2}, x_j + \frac{\Delta}{2})\}}$,
- $\hat{f}_0(x) = f(x; \hat{\theta}_0)$.

Παρατηρούμε από τον τύπο της $\hat{S}_n(h)$, ότι δεν είναι τίποτα άλλο από μία διακριτοποιημένη προσέγγιση της L_2 μετρικής, που υπολογίζει τη διαφορά μεταξύ της εμπειρικής σ.π.π. και της σ.π.π., με υπό τη μηδενική υπόθεση, όπου η άγνωστη παράμετρος θ_0 έχει εκτιμηθεί από τον Ε.Μ.Π. $\hat{\theta}_0$. Μία σημαντική διαφορά όμως είναι ότι έχει ενταχθεί μία συνάρτηση βάρους, στη συγκεκριμένη περίπτωση ένας πυρήνας K , που “ζυγίζει” τα κέντρα στα οποία γίνεται η εκτίμηση. Ουσιαστικά σε σημεία που οι διαφορές των δύο συναρτήσεων δεν είναι τόσο σημαντικές, ο πυρήνας παίρνει χαμηλές τιμές, με αποτέλεσμα η διαφορά σε αυτά τα σημεία να είναι αμελητέα και να μην επηρεάζει το συνολικό άθροισμα, που προσεγγίζει

την L_2 μετρική. Αντίστοιχα συμβαίνει και σε σημεία που παρατηρούνται σημαντικές διαφορές μεταξύ των δύο συναρτήσεων, δίνοντας έτσι πληροφορία για τη συνολική απόκλιση της εμπειρικής σ.π.π. με την σ.π.π., υπό τη μηδενική υπόθεση. Σε τέτοιου είδους τεστ λοιπόν, σύμφωνα με τη βιβλιογραφία, η επιλογή του εύρους ζώνης του πυρήνα γίνεται είτε με βελτιστοποίηση της εκτίμησης της σ.π.π. είτε με βελτιστοποίηση της εκτίμησης της L_2 μετρικής. Δυστυχώς αυτές οι δύο ποσότητες δε βελτιστοποιούνται ταυτόχρονα και έτσι να οδηγούμαστε στο εξής πρόβλημα. Αν επιλέξουμε να βελτιστοποιήσουμε την εκτίμηση της σ.π.π., αποδεικνύεται ότι μειώνεται η ισχύς του τεστ. Από την άλλη, αν επιλέξουμε να βελτιστοποιήσουμε την εκτίμηση της L_2 μετρικής, καταλήγουμε σε μείωση του σφάλματος τύπου I, $\alpha = P(\text{Απόρριψη της } H_0 | \text{Ισχύει η } H_0)$, με αποτέλεσμα το πραγματικό επίπεδο σημαντικότητας να αποκλίνει από το θεωρητικό, ενδεχομένως επηρεάζοντας την εγκυρότητα των συμπερασμάτων. Προκειμένου λοιπόν να αντιμετωπίσουμε αυτό το πρόβλημα, διατηρούμε την παράμετρο h σταθερή, για την προσέγγιση της L_2 μετρικής, εκτιμώντας την πραγματική σ.π.π. με το ιστόγραμμα. Επομένως με αυτόν τον τρόπο εισάγουμε δύο παραμέτρους.

- h : η παράμετρος του εύρους ζώνης του πυρήνα, που επηρεάζει μόνο την εκτίμηση της στατιστικής συνάρτησης $\hat{S}_n(h)$ του τεστ, η οποία με τη σειρά της προσεγγίζει την L_2 μετρική.
- Δ : η παράμετρος του μήκους των διαστημάτων, που χρησιμοποιούνται στο ιστόγραμμα.

Συνεπώς έχουμε δύο παραμέτρους, από τις οποίες το h μεταβάλλεται, ενώ το Δ σταθεροποιείται σύμφωνα με την υπόθεση A.2. Το h επιλέγεται έτσι ώστε να επιταγχύνεται κατά το δυνατόν ο ρυθμός σύγκλισης της στατιστικής συνάρτησης $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$ στην κανονική κατανομή. Αυτό υλοποιείται ελαχιστοποιώντας το άνω φράγμα των Berry - Esseen ως προς το εύρος ζώνης του πυρήνα, h .

Λήμμα 2.1. Έστω $X_1, \dots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. $H_n(X_1, X_2)$ όπως έχει οριστεί στο Θεώρημα του Hall, που αναφέραμε στην υποενότητα 1.3 και $(x_j)_{j \in D_n}$ όπως ορίστηκαν στην υποενότητα 2.2.

Τότε θα ισχύει:

$$\sup_{x \in \mathbb{R}} \left| P \left(\frac{S_{2,n}}{\tau} \leq x \right) - \Phi(x) \right| \leq C \left[\left\{ \frac{E [G_n(X_1, X_2)^2]}{(E [H_n(X_1, X_2)^2])^2} \right\}^{\frac{1}{6}} + \left\{ \frac{E [H_n(X_1, X_2)^4]}{n (E [H_n(X_1, X_2)^2])^2} \right\}^{\frac{1}{3}} + \left\{ \frac{E [H_n(X_1, X_2)^4]}{n (E [H_n(X_1, X_2)^2])^2} \right\}^{\frac{1}{5}} \right],$$

όπου $\tau^2 = E[S_{2,n}^2]$ και $S_{2,n}$ ορίζεται στη σχέση (4.25).

Απόδειξη. Στο άρθρο των Hall and Heyde (1981), p.397, παρουσιάζεται το ακόλουθο άνω φράγμα:

$$\sup_{x \in \mathbb{R}} \left| P\left(s_n^{-1} S_n \leq x\right) - \Phi(x) \right| \leq C \left[\left\{ s_n^{-2} \sum_{i=1}^n E[X_i^2 I(|X_i| > s_n)] + E[|s_n^{-2} V_n^2 - 1|] \right\}^{1/3} + \left\{ s_n^{-4} \sum_{i=1}^n E[X_i^4 I(|X_i| \leq s_n)] \right\}^{1/5} \right],$$

όπου C είναι μία σταθερά, S_n είναι μία στοχαστική ποσότητα, $X_i = S_i - S_{i-1}$ είναι μία στοχαστική διαφορά, $s_n^2 = E[S_n^2]$ και $V_n^2 = \sum_{j=1}^n E[X_j^2 | \mathcal{F}_{j-1}]$. Μπορούμε να προσεγγίσουμε το δεξιό μέλος της παραπάνω ανίσωσης ως εξής:

$$\left\{ s_n^{-2} \sum_{i=1}^n E[X_i^2 I(|X_i| > s_n)] + E[|s_n^{-2} V_n^2 - 1|] \right\}^{1/3} \leq \left\{ s_n^{-2} \sum_{i=1}^n E[X_i^2 I(|X_i| > s_n)] \right\}^{1/3} + \{E[|s_n^{-2} V_n^2 - 1|]\}^{1/3},$$

έχοντας χρησιμοποιήσει την ανισότητα $(A+B)^{1/3} \leq A^{1/3} + B^{1/3}$ όταν $A, B \geq 0$. Εφαρμόζοντας την ανισότητα του Chebychev προκύπτει

$$s_n^{-2} \sum_{i=1}^n E[X_i^2 I(|X_i| > s_n)] \leq s_n^{-4} \sum_{i=1}^n E[X_i^4].$$

Χρησιμοποιώντας επίσης την ανισότητα του Liapounov, σύμφωνα με την οποία για κάθε τυχαία μεταβλητή X , η ποσότητα $(E[|X|^p])^{1/p}$ είναι μία αύξουσα συνάρτηση ως προς $p \geq 1$, θα έχουμε ότι

$$\left\{ E[|s_n^{-2} V_n^2 - 1|] \right\}^{1/3} \leq \left\{ E[(s_n^{-2} V_n^2 - 1)^2] \right\}^{1/6},$$

και

$$s_n^{-4} \sum_{i=1}^n E[X_i^4 I(|X_i| \leq s_n)] \leq s_n^{-4} \sum_{i=1}^n E[X_i^4].$$

Αντικαθιστώντας λοιπόν προκύπτει

$$\sup_{x \in \mathbb{R}} \left| P\left(s_n^{-1} S_n \leq x\right) - \Phi(x) \right| \leq C \left[\left\{ s_n^{-4} \sum_{i=1}^n E[X_i^4] \right\}^{1/3} \right]$$

$$+ \left\{ E \left[(s_n^{-2} V_n^2 - 1)^2 \right] \right\}^{1/6} + \left\{ s_n^{-4} \sum_{i=1}^n E [X_i^4] \right\}^{1/5} \right]. \quad (2.8)$$

Ο Hall στο άρθρο του Hall (1984), όπου ασχολείται με το αντίστοιχο Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.) για το ολοκλήρωμα του τετραγωνικού σφάλματος, δείχνει ότι: $s_n^2 = \sum_{i=2}^n E[Y_{n,i}^2] = n(n-1)E[H_n(X_1, X_2)^2]/2$, όπου $Y_{n,i} = \sum_{j=1}^{i-1} H_n(X_i, X_j)$ είναι τώρα οι στοχαστικές διαφορές και X_i, X_j να αποτελούν παρατηρήσεις. Επιπλέον δείχνει ότι

$$\sum_{i=2}^n E[Y_{n,i}^4] \leq C_1 \cdot n^3 E[H_n(X_1, X_2)^2], \quad (2.9)$$

και

$$E[(V_n^2 - s_n^2)^2] \leq C_2 \cdot n^4 E[G_n(X_1, X_2)^2], \quad (2.10)$$

όπου C_1, C_2 να είναι σταθερές. Η απόδειξη του Λήμματος 2.1 προκύπτει αν αντικαταστήσουμε τις σχέσεις (2.9) και (2.10) στην ανισότητα (2.8).

□

Για λόγους απλότητας όμως, επιλέγουμε να ελαχιστοποιήσουμε, ως προς h την ποσότητα:

$$\frac{E[G_n(X_1, X_2)^2] + n^{-1} E[H_n(X_1, X_2)^4]}{\{E[H_n(X_1, X_2)^2]\}^2}, \quad (2.11)$$

που εμφανίζεται στο Θεώρημα του Hall και η σύγκλιση της στο μηδέν αποτελεί κριτήριο, για τη σύγκλιση των U - στατιστικών, στην κανονική κατανομή. Η παραπάνω ποσότητα σχετίζεται αρκετά με το άνω φράγμα των Berry - Esseen, αλλά προφανώς είναι μία πιο απλή έκφραση και για αυτό επιλέγουμε να ελαχιστοποιήσουμε αυτή, ως προς h .

Σημειώνουμε ότι στη σχέση (2.11), θα έχουμε:

$$H_n(X_1, X_2) = \sum_{j \neq k} K \left(\frac{x_j - x_k}{h} \right) A_n(x_j, X_1) A_n(x_k, X_2),$$

και

$$G_n(X_1, X_2) = \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} \left\{ K \left(\frac{x_{j_1} - x_{k_1}}{h} \right) K \left(\frac{x_{j_2} - x_{k_2}}{h} \right) \right.$$

$$\times A_n(x_{k_1}, X_1)A_n(x_{k_2}, X_2)E[A_n(x_{j_1}, X_3)A_n(x_{j_2}, X_3)]\Big\},$$

όπου

$$A_n(x_j, X_i) = I_{\{|x_j - X_i| \leq \Delta c/2\}} - p_j$$

με

$$\begin{aligned} p_j &= E \left[I_{\{|x_j - X_i| \leq \Delta c/2\}} \right] \\ &= E \left[I_{\{x_j - \Delta c/2 \leq X_i \leq x_j + \Delta c/2\}} \right] \\ &= \int_{x_j - \Delta c/2}^{x_j + \Delta c/2} f(x) dx \\ &= \int_{-\Delta c/2}^{\Delta c/2} f(x_j + u) du. \end{aligned}$$

Αποδεικνύεται κάτω από τις υποθέσεις A.1 - A.7, ότι καθώς $n \rightarrow \infty$, θα ισχύουν, με βάση τις σχέσεις (4.42) και (4.36), οι παρακάτω ασυμπτωτικοί τύποι:

$$\frac{E[G_n(X_1, X_2)^2]}{\{E[H_n(X_1, X_2)^2]\}^2} = h \frac{R_4(K)J_1(K)}{R_2(K)^2 R_2(f)^2} (1 + o(1)),$$

όπου

$$J_1(K) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(-y_1)K(y_2 - y_1)K(-y_3)K(y_2 - y_3) dy_1 dy_2 dy_3,$$

και

$$\frac{1}{n} \frac{E[H_n(X_1, X_2)^4]}{\{E[H_n(X_1, X_2)^2]\}^2} = \frac{1}{nh} \frac{R_4(K)}{4R_2(K)^2 R_2(f)} (1 + o(1)).$$

Συνεπώς το πηλίκο (2.11) είναι ασυμπτωτικά ισοδύναμο με:

$$h \frac{R_4(K)J_1(K)}{R_2(K)^2 R_2(f)^2} + \frac{1}{nh} \frac{R_4(K)}{R_2(K)^2 R_2(f)}.$$

Ελαχιστοποιώντας λοιπόν την παραπάνω σχέση, ως προς h , προκύπτει το βέλτιστο εύρος ζώνης:

$$h_{\text{opt}} = n^{-1/2} \sqrt{\frac{R_2(f)R_4(K)}{R_4(f)J_1(K)}}.$$

Για να είναι ο παραπάνω τύπος υλοποιήσιμος στην πράξη, εκτιμούμε τις ποσότητες $R_2(f)$, $R_4(f)$, καθώς δεν γνωρίζουμε την άγνωστη σ.π.π. f .

Συγκεκριμένα θα έχουμε:

$$R_p(f) = \int f(x)^p dx = \int f(x)^{p-1} f(x) dx = E[f(x)^{p-1}] \approx n^{-1} \sum_{i=1}^n f(X_i)^{p-1}.$$

Οπότε ένας εκτιμητής της $R_p(f)$, θα είναι:

$$\hat{\nu}_p = \hat{R}_p(f) = n^{-1} \sum_{i=1}^n f^\dagger(X_i; \hat{h}_a)^{p-1},$$

όπου

$$f^\dagger(x; h_a) = (nh_a)^{-1} \sum_{j=1}^n K\left(\frac{x - X_j}{h_a}\right),$$

είναι ο εκτιμητής πυρήνων της άγνωστης σ.π.π. $f(x)$.

Ο πυρήνας K διαλέγεται να είναι η σ.π.π. της τυπικής κανονικής κατανομής, ενώ το εύρος ζώνης h_a , επιλέγεται να είναι $\hat{h}_a = \left(\frac{4}{3n}\right)^{1/5}$, σύμφωνα με τον τρόπο επιλογής του εύρους ζώνης του πυρήνα στο άρθρο των Zhang et al. (2006).

Επομένως το εκτιμώμενο βέλτιστο εύρος ζώνης h , που οδηγεί σε ικανοποιητικό ρυθμό σύγκλισης της στατιστικής συνάρτησης $n\Delta^2 h^{-\frac{1}{2}} \hat{S}_n(h)$ στην κανονική κατανομή είναι:

$$\hat{h}_{\text{opt}} = n^{-1/2} \sqrt{\frac{\hat{\nu}_2 R_4(K)}{\hat{\nu}_4 J_1(K)}}. \quad (2.12)$$

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΤΗΣ ΙΣΧΥΟΣ

Σε αυτό το κεφάλαιο θα διερευνήσουμε την ισχύ του στατιστικού τεστ, που αναφέραμε στο Κεφάλαιο 2, για δείγματα πεπερασμένου μεγέθους. Έπειτα θα μελετήσουμε τη μεταβολή της ισχύος ως προς την παράμετρο Δ , βασιζόμενοι στο βέλτιστο εύρος ζώνης (2.12). Οι υποθέσεις που ελέγχονται περιλαμβάνουν αρκετά παραδείγματα Pitman εναλλακτικών τύπου 1 και 2 (regular και sharp peak).

3.1 Σχεδιασμός προσομοιώσεων

Για να έχουμε μια καλύτερη εικόνα της απόδοσης της ισχύος του τεστ, θα γίνει η σύγκριση με την αντίστοιχη των ήδη υπάρχων ελέγχων καλής προσαρμογής που χρησιμοποιούνται στην πράξη. Αυτοί οι ελέγχοι, που περιλαμβάνονται στις προσομοιώσεις, βασίζονται στην εμπειρική α.σ.κ. και αποτελούνται από τα τεστ των Kolmogorov-Smirnov (D_n), Anderson-Darling (A_n), Cramér-von Mises (C_n). Επιπρόσθετα θα χρησιμοποιήσουμε το τεστ της Fan (F_n) (βλέπε Fan (1994)) που ορίζεται βάσει της σ.π.π. και της L_2 μετρικής. Η έκδοση του τεστ της Fan που θα εφαρμόσουμε εδώ, ενσωματώνει τη διόρθωση της μεροληψίας (βλέπε Li and Racine (2007)). Τέλος γίνεται η σύγκριση και με το εμπειρικό τεστ του πηλίκου πιθανοφάνειας των Gurevich και Vexler (V_n) (βλέπε Gurevich and Vexler (2010)). Οι στατιστικές συναρτήσεις D_n, A_n και C_n έχουν οριστεί στο Κεφάλαιο 2, ενώ οι υπόλοιπες παρουσιάζονται παρακάτω.

$$F_n = n^{-1} \sum_{i=1}^n f_{-i}(X_i) + \int (K_h * f(x; \theta))^2 dx - 2n^{-1} \sum_{i=1}^n K_h * f_0(X_i; \theta),$$
$$V_n = \min_{1 \leq M \leq n^{1/2}} (2\pi e \hat{S}^2)^2 \prod_{i=1}^n \frac{2M}{n(X_{(i+M)} - X_{(i-M)})},$$

όπου $X_{(i)}$ συμβολίζει την i σε αύξουσα σειρά παρατήρηση, $M < n/2$, θ είναι η παράμετρος της f_0 , $\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ είναι η αμερόπληκτη δειγματική διακύμανση, $\hat{F}$ είναι η εμπειρική α.σ.κ., $K_h(\cdot) = h^{-1}K(\cdot/h)$ και

$$n^{-1} \sum_{i=1}^n f_{-i}(X_i) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K_h(X_i - X_j),$$

$$K_h * f_0(x; \theta) = \int K_h(x - v) f_0(v; \theta) dv.$$

Όλες οι προσομοιώσεις των συναρτήσεων της ισχύος, έναντι Pitman εναλλακτικών, πραγματοποιούνται στη μονοδιάστατη περίπτωση ($d = 1$). Παρ'όλο που τα θεωρητικά αποτελέσματα της διατριβής ισχύουν για την περίπτωση της εκτίμησης της άγνωστης παραμέτρου της σ.π.π. υπό την H_0 με τον Ε.Μ.Π., στις προσομοιώσεις αυτού του κεφαλαίου θα χρησιμοποιήσουμε την αληθινή τιμή της παραμέτρου της κατανομής. Ο λόγος είναι ότι θα διευκολύνει τη σύγκριση της απόδοσης του τεστ με όλα τα στατιστικά τεστ που αναφέραμε παραπάνω. Η απόρριψη ή μη της μηδενικής υπόθεσης βασίζεται στην p -τιμή του κάθε τεστ. Στα τεστ των Kolmogorov Smirnov (D_n) και Cramér– von Mises (C_n) ο υπολογισμός των p -τιμών γίνεται με μεγάλη ακρίβεια, μέσα από αριθμητικές μεθόδους που εφαρμόστηκαν από τους Marsaglia και Csorgo-Faraway, αντίστοιχα (βλέπε Marsaglia et al. (2003), Csörgo and Faraway (1996)). Η p -τιμή για το τεστ της Fan (F_n) υπολογίζεται από τους παρακάτω τύπους, που βασίζονται στο Θεώρημα 4.1 της Fan.

$$p = 1 - \phi \left(\frac{n\sqrt{h}F_n}{\sigma_{F_n}} \right),$$

$$\sigma_{F_n} = \left\{ \frac{2}{hn^2} \sum_{i=1}^n \sum_{j=1}^n K^2 \left(\frac{X_i - X_j}{h} \right) \right\}^{1/2} \left\{ \int \left(\int K(u)K(u+v) du \right)^2 dv \right\}.$$

Τέλος, οι υπολογισμοί για τη συνάρτηση ισχύος της στατιστικής συνάρτησης V_n βασίζεται στο γεγονός ότι η μηδενική υπόθεση απορρίπτεται αν-ν $\ln(V_n) > C_a$, όπου η τιμή C_a υπολογίζεται μέσω μίας μεθόδου επαναδειγματοληψίας Monte Carlo, ως ρίζα της παρακάτω εξίσωσης (βλέπε Gurevich and Vexler (2010)).

$$\begin{aligned} P(\text{Απόρριψη } H_0 \mid \text{Ισχύει } H_0) &= P_{H_0} \{ \ln(V_n) > C_a \} \\ &= P_{Y_1, \dots, Y_n \sim U(0,1)} \{ \ln(V_n) > C_a \} \\ &= a. \end{aligned}$$

Σημειώνουμε ότι $Y_i = F_{H_0}(X_i; \theta_0)$, $i \in \{1, \dots, n\}$, είναι οι μετασχηματισμένες τιμές του δείγματος, οι οποίες χρησιμοποιούνται για τον υπολογισμό του V_n . Το γεγονός ότι η $F_{H_0}(\cdot; \theta_0)$ είναι πλήρως προσδιορισμένη, μας εξασφαλίζει ότι ο έλεγχος της (1.6), είναι ισοδύναμος με τον έλεγχο ότι οι παρατηρήσεις $Y_1, \dots, Y_n$ προέρχονται από την ομοιόμορφη κατανομή $U(0, 1)$, για κάθε σ.π.π. f_0 .

Όλα τα στατιστικά τεστ, καθώς και οι p -τιμές τους, υπολογίζονται στην R. Το στατιστικό D_n υλοποιείται μέσω της εντολής `ks.test`, η οποία είναι ενσωματωμένη στα πακέτα της R. C_n υπολογίζεται από τη συνάρτηση `cvm.test` του πακέτου `gofTest`. Το τεστ F_n εφαρμόζεται μέσω της εντολής `fan.test` του πακέτου `GoFKernel`. Με βάση τους Ghosh and Huang (1991), στην υλοποίηση του F_n χρησιμοποιείται ο ομοιόμορφος πυρήνας και η επιλογή του εύρους ζώνης γίνεται από την εντολή `dpik` και υπογραμμίζει τη σημασία της κατάλληλης επιλογής εύρους ζώνης όταν στόχος είναι η μεγιστοποίηση της ισχύος, και όχι απλώς η καλή εκτίμηση της πυκνότητας. (βλέπε Li and Racine (2007)). Τέλος το στατιστικό V_n υπολογίζεται μέσω της συνάρτησης `dbEmpLikeGOF` του πακέτου της R με την ίδια ονομασία.

Έστω $\hat{h}_{\text{opt}}$ είναι το βέλτιστο εύρος ζώνης (2.12) για διάσταση δείγματος $d = 1$. Για λόγους απλότητας ορίζουμε $S_n = \hat{h}_{\text{opt}}^{-1/2} n \Delta^2 \hat{S}_n(\hat{h}_{\text{opt}}; \cdot)$ να είναι η στατιστική συνάρτηση, που παρουσιάσαμε στο Κεφάλαιο 2, για πλήρως προσδιορισμένες ή για πυκνότητες με παραμετρική εκτίμηση των παραμέτρων τους. Σύμφωνα με την υπόθεση A.2 το εύρος των διαστημάτων του ιστογράμματος ισούται με $\Delta = n^{-3/4}$. Στη συνέχεια, ακολουθώντας την υπόθεση A.4, στη στατιστική συνάρτηση S_n χρησιμοποιείται ο πυρήνας Epanechnikov. Συγκεκριμένα για αυτόν τον πυρήνα θα έχουμε:

$$\begin{aligned} J_1(K) &= \frac{3^4}{4^4} \int_{-1}^1 \int_{-1}^1 \int_{-1}^1 (1 - y_1^2)(1 - (y_2 - y_1)^2)(1 - y_3^2) \\ &\quad \times (1 - (y_3 - y_2)^2) dy_1 dy_2 dy_3 \\ &= \frac{3^4}{4^4} \frac{736}{675} \approx 0.345, \end{aligned}$$

καθώς επίσης και

$$R_4(K) = \left(\frac{3}{4}\right)^4 \int_{-1}^1 (1 - u^2)^4 du = \frac{81}{256} \frac{256}{315} = \frac{9}{35} = 0.2571,$$

και το βέλτιστο εύρος ζώνης γίνεται

$$\begin{aligned}\hat{h}_{\text{opt}} &= n^{-1/2} \sqrt{\frac{\hat{\nu}_2 R_4(K)}{\hat{\nu}_4 J_1(K)}} = n^{-1/2} \sqrt{\frac{\hat{\nu}_2 \cdot 0.2571}{\hat{\nu}_4 \cdot 0.345}} \\ &= 0.8632 \cdot n^{-1/2} \sqrt{\frac{\sum_{i=1}^n f^\dagger(X_i; \hat{h}_a)}{\sum_{i=1}^n f^\dagger(X_i; \hat{h}_a)^3}}.\end{aligned}$$

Για ευκολία συμβολίζουμε $S_n = n\Delta^2 \hat{h}_{\text{opt}}^{-1/2} \hat{S}_n(\hat{h}_{\text{opt}})$. Σε όλα τα παραδείγματα υπολογίζονται δύο καμπύλες ισχύος για τη στατιστική συνάρτηση S_n , η $S_{n,a}$ και η $S_{n,b}$. Αυτές οι καμπύλες αντιστοιχούν στην προσέγγιση της πιθανότητας $P(S_n > z_a \sigma_{0,\theta_0} | H_1)$ για την $S_{n,a}$ και της πιθανότητας $P(S_n > \ell_a^* | H_1)$ για την $S_{n,b}$. Σημειώνουμε ότι $z_a \sigma_{0,\theta_0}$ είναι το ασυμπτωτικό κρίσιμο σημείο ενώ ℓ_a^* είναι το α εμπειρικό ποσοστημόριο με τη μέθοδο bootstrap, το οποίο ικανοποιεί $\ell_\alpha^* = \hat{\sigma}_0 z_{1-\alpha} + o_p(1) = \ell_\alpha + o_p(1)$, με $\ell_\alpha = \sigma_0 z_{1-\alpha} + o_p(1)$, όπου $z_{1-\alpha}$ είναι το $1-\alpha$ ποσοστημόριο της τυπικής κανονικής κατανομής $N(0, 1)$, σ_0^2 είναι η διακύμανση της στατιστικής συνάρτησης S_n , η οποία παρουσιάζεται στην ενότητα 2.4 και $\hat{\sigma}_0^2$ είναι η bootstrap εκδοχή της σ_0^2 (βλέπε Πόρισμα 3 στο άρθρο των Bagkavos et al. (2023)). Το S_n υπολογίζεται μέσω του πακέτου `L2DensityGoFtest` της R, με $h \equiv \hat{h}_{\text{opt}}$.

Παρακάτω παρουσιάζεται η μέθοδος Bootstrap που χρησιμοποιήθηκε για να εκτιμηθεί το κρίσιμο σημείο της κατανομής του στατιστικού ελέγχου S_n , όταν το μέγεθος του δείγματος n είναι πεπερασμένο και η ασυμπτωτική κανονικότητα δεν επαρκεί. Σκοπός δηλαδή είναι η προσέγγιση του κρίσιμου σημείου ℓ_α στο οποίο βασίζεται η κατασκευή της περιοχής απόρριψης του ελέγχου σε επίπεδο σημαντικότητας α το οποίο ορίζεται ως εξής:

$$\begin{aligned}\alpha &= P(\text{Απόρριψη της } H_0 | \text{Ισχύει η } H_0) \\ &= P(S_n > \ell_\alpha | H_0)\end{aligned}$$

Αντί να βασιζόμαστε στον θεωρητικό τύπο της κατανομής, χρησιμοποιούμε την εμπειρική κατανομή του στατιστικού από τα bootstrap δείγματα. Σύμφωνα με τους Gao and Gijbels (2008), είναι εφικτό να προσεγγιστεί το $(1 - \alpha)100$ -ποσοστημόριο της κατανομής της $n\Delta^2 h^{-1/2} \hat{S}_n(h)$ με ℓ_α^* , το οποίο λαμβάνεται από τον ακόλουθο αλγόριθμο bootstrap:

- i) Επιλέγουμε κάθε φορά δείγμα από τις παρατηρήσεις $\mathcal{X} = \{X_1, \dots, X_n\}$ για να λάβουμε M δείγματα bootstrap, που συμβολίζονται με $\mathcal{X}_m^* = \{X_{1m}^*, \dots, X_{nm}^*\}$, όπου για κάθε $m \in \{1, \dots, M\}$, το X_m^* επιλέγεται

τυχαία, με επανατοποθέτηση, από το $\mathcal{X}$. Γράφουμε $\hat{\theta} = \theta(\mathcal{X})$ για τον εκτιμητή του θ που βασίζεται στο αρχικό δείγμα $\mathcal{X}$ και για κάθε m ορίζουμε τον εκτιμητή bootstrap του θ με $\theta_m^* = \theta(\mathcal{X}_m^*)$.

- ii) Για $m \in \{1, \dots, M\}$, χρησιμοποιούμε το $X_m^* = \{X_{1m}^*, \dots, X_{nm}^*\}$ και το $\hat{\theta}_m^*$ από το προηγούμενο βήμα για να υπολογίσουμε το $n\Delta^2 h^{-1/2} \hat{S}_{n,m}^*(h)$, $m \in \{1, \dots, M\}$.
- (iii) Υπολογίζουμε το l_α^* ως το $1 - \alpha$ εμπειρικό ποσοστημόριο των τιμών $n\Delta^2 h^{-1/2} \hat{S}_{n,m}^*(h)$, $m \in \{1, \dots, M\}$. Τότε το l_α^* προσεγγίζει την πιθανότητα $P^*(n\Delta^{2d} h^{-d/2} \hat{S}_{n,m}^*(hp) > l_\alpha^*) = 1 - \alpha$, όπου το P^* υποδηλώνει το μέτρο πιθανότητας bootstrap υπό τη δέσμευση του δείγματος $\mathcal{X}$.
- (iv) Απορρίπτουμε την H_0 σε έναν έλεγχο επιπέδου σημαντικότητας α αν και μόνο αν $n\Delta^{2d} h^{-d/2} \hat{S}_n(hp) > l_\alpha^*$.

Τι είναι ποσοστημόριο;

Έστω ότι T είναι μία τυχαία μεταβλητή. Το ποσοστημόριο τάξης $1 - \alpha$, όπου $\alpha \in (0, 1)$, δηλαδή το $(1 - \alpha)100$ - ποσοστημόριο, ορίζεται ως:

$$\ell_\alpha = \inf \{x \in \mathbb{R} : P(T \leq x) \geq 1 - \alpha\}.$$

Ισοδύναμα:

$$P(T > \ell_\alpha) = \alpha \quad \text{ή} \quad P(T \leq \ell_\alpha) = 1 - \alpha$$

το οποίο ουσιαστικά μας δείχνει ότι το $\alpha 100\%$ των τιμών της τυχαίας μεταβλητής T βρίσκεται πάνω από το σημείο ℓ_α ή ισοδύναμα το $(1 - \alpha)100\%$ των τιμών της T βρίσκεται κάτω από το σημείο ℓ_α .

Το πρόβλημα με την κατανομή του S_n

Στον έλεγχο καλής προσαρμογής του άρθρου, η στατιστική συνάρτηση είναι:

$$S_n = n\Delta^2 h^{-1/2} \hat{S}_n(h),$$

και ασυμπτωτικά ισχύει ότι:

$$S_n \xrightarrow{d} \mathcal{N}(0, \sigma_0^2), \text{ υπό } H_0.$$

Ωστόσο, για πεπερασμένο n , η κατανομή του S_n μπορεί να απέχει σημαντικά από την κανονική. Γι' αυτό, χρησιμοποιείται η bootstrap προσέγγιση.

Bootstrap εκτίμηση του ποσοστημορίου

Για κάθε bootstrap δείγμα $X_m^* = \{X_{1m}^*, \dots, X_{nm}^*\}$:

- Υπολογίζεται ο εκτιμητής παραμέτρων $\hat{\theta}_m^*$.
- Υπολογίζεται το στατιστικό:

$$S_n^{*(m)} = n\Delta^2 h^{-1/2} \hat{S}_n^{*(m)}(h\rho).$$

- Υπολογίζεται το bootstrap ποσοστημόριο ℓ_α^* ως το $(1 - \alpha)100$ εμπειρικό ποσοστημόριο των τιμών $S_n^{*(m)}$, δηλαδή:

$$\frac{1}{M} \sum_{m=1}^M I_{\{S_n^{*(m)} \leq \ell_\alpha^*\}} = 1 - \alpha.$$

Έλεγχος υποθέσεων με bootstrap ποσοστημόριο

Η απόφαση του ελέγχου δίνεται ως εξής:

$$\text{Απόρριψη } H_0 \iff S_n > \ell_\alpha^*.$$

Έτσι, ο έλεγχος έχει ασυμπτωτικά σωστό επίπεδο σημαντικότητας:

$$\Pr(S_n > \ell_\alpha^*) = \alpha + o(1).$$

Θεωρητικό αποτέλεσμα (Πόρισμα 3, Bagkavos et al. (2023))

Το άρθρο δείχνει ότι:

$$\Pr(S_n > \ell_\alpha^*) = \alpha + o(1)$$

δηλαδή ότι η χρήση του bootstrap ποσοστημορίου ℓ_α^* δίνει σωστό επίπεδο σημαντικότητας ασυμπτωτικά.

Επίσης, αναφέρει ότι:

$$\ell_\alpha^* = \hat{\sigma}_0 z_{1-\alpha} + o_p(1) = \ell_\alpha + o_p(1),$$

δηλαδή ότι το bootstrap ποσοστημόριο προσεγγίζει το θεωρητικό κρίσιμο σημείο της κανονικής κατανομής, με πιθανότητα να συγκλίνει στο 1, όπου:

ΚΕΦΑΛΑΙΟ 3 3.2. Αποδοτικότητα των ελέγχων υπό Pitman εναλλακτικές

- $z_{1-\alpha}$: ποσοστημόριο της τυπικής κανονικής κατανομής,
- $\hat{\sigma}_0$: bootstrap εκτίμηση της διακύμανσης σ_0 ,
- $o_p(1)$: τυχαίο μέγεθος που συγκλίνει κατά πιθανότητα στο μηδέν.

Συνεχίζοντας λοιπόν, όλες οι γραφικές παραστάσεις βασίζονται σε δείγμα μεγέθους $n = 40$. Σε κάθε προσομοίωση, η πραγματική πυκνότητα επιλέγεται από τις 15 το πλήθος μίξεις κανονικών κατανομών των Marron και Wand, οι οποίες συμβολίζονται ως NM1, NM2, ..., NM15 (βλέπε Marron and Wand (1992)). Η αναπαραγωγή τυχαίων αριθμών για κάθε μίξη υλοποιείται στο πακέτο `nor1mix` της R. Η διαδικασία αυτή βασίζεται αρχικά στη δημιουργία ενός κατηγορικού δείγματος από την πολυωνυμική κατανομή, όπου το διάνυσμα των πιθανοτήτων των γεγονότων είναι ίσο με τα βάρη του μείγματος. Αυτό αντιστοιχίζει τις τυχαίες τιμές που πρόκειται να παραχθούν με τις συνιστώσες της μίξης. Στη συνέχεια, οι τιμές του δείγματος για κάθε συνιστώσα παράγονται χρησιμοποιώντας μια γεννήτρια τυχαίων αριθμών κανονικής κατανομής, παραμετροποιημένη σύμφωνα με τη μέση τιμή και τη διασπορά της αντίστοιχης συνιστώσας της μίξης. Λεπτομέρειες για την αντίστοιχη εφαρμογή στην εναλλακτική υπόθεση παρουσιάζεται παρακάτω για κάθε τύπο εναλλακτικής. Η συνάρτηση ισχύος για κάθε τεστ προσεγγίζεται από το πλήθος των φορών που απορρίπτεται η μηδενική υπόθεση σε 100 εφαρμογές του τεστ. Στο τέλος εκτυπώνεται ο μέσος όρος 30 τέτοιων αποτελεσμάτων για κάθε τεστ. Σημειώνουμε ότι σε κάθε επανάληψη του πειράματος χρησιμοποιείται το ίδιο δείγμα για τον υπολογισμό όλων των στατιστικών ελέγχων.

3.2 Αποδοτικότητα των ελέγχων υπό Pitman εναλλακτικές

Στις γραφικές παραστάσεις που παρουσιάζονται παρακάτω, συγκρίνονται οι καμπύλες της ισχύος των $S_{n,a}$ και $S_{n,b}$ με τις αντίστοιχες των στατιστικών ελέγχων D_n, C_n, F_n και V_n υπό τις εναλλακτικές υποθέσεις Pitman τύπου 1 και 2.

Υπενθυμίζουμε ότι ο γενικός τύπος των εναλλακτικών υποθέσεων Pitman είναι της μορφής:

$$H_n : f_n(x; \theta_0) = f_0(x; \theta_0) + a_n \eta(x).$$

Αρχικά θα επεξηγήσουμε την επιλογή των κατάλληλων συναρτήσεων για την εναλλακτική υπόθεση Pitman τύπου 1 (regular peak Pitman alternatives). Σύμφωνα με τον ορισμό, που δώσαμε στην υποενότητα 2.1, η εναλλακτική υπόθεση Pitman τύπου 1 είναι της μορφής

$$H_{1n} : f_{1n}(x; \theta_0) = f_0(x; \theta_0) + n^{-\beta} \eta(x).$$

Θέτουμε $\eta(x) = i\{f_{1n}(x; \theta_0) - f_0(x; \theta_0)\}$, όπου η παράμετρος $i \in \{0, 0.2, 0.4, \dots, 2.4\}$ καθορίζει την απόσταση μεταξύ των σ.π.π. υπό τη μηδενική υπόθεση και την αντίστοιχη εναλλακτική Pitman τύπου 1. Από τον τύπο του βέλτιστου εύρους ζώνης (2.12), παρατηρούμε ότι $\hat{h}_{\text{opt}} \sim n^{-1/2}$. Προκειμένου να συγκρίνουμε τη συνάρτηση ισχύος του S_n με τις αντίστοιχες των στατιστικών συναρτήσεων D_n, C_n και F_n , πρέπει ο ρυθμός σύγκλισης, του ελέγχου, στη μηδενική υπόθεση να είναι ίσος με αυτόν των υπόλοιπων ελέγχων, δηλαδή της τάξης $n^{-1/2}$. Για αυτόν το λόγο επιλέγουμε $n^{-\beta} = a_n = n^{-5/8} \hat{h}_{\text{opt}}^{-1/4}$, το οποίο οδηγεί σε $a_n = O_p(n^{-1/2})$, έχοντας ως αποτέλεσμα να ταυτίζεται με το ρυθμό σύγκλισης των στατιστικών τεστ, που αναφέραμε παραπάνω.

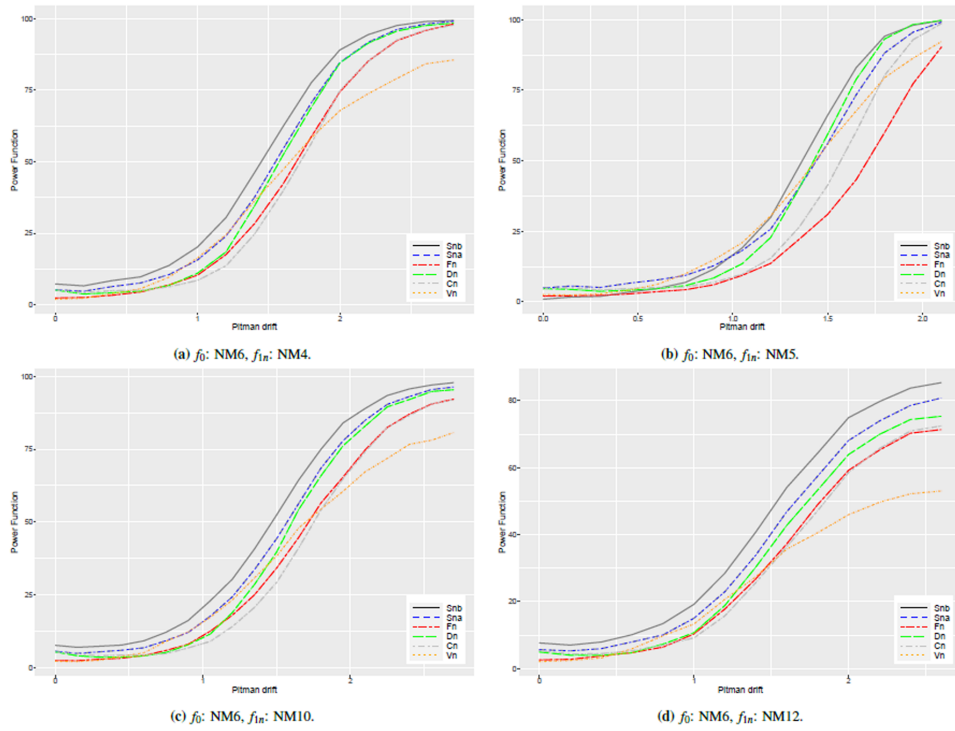
Συνεχίζουμε με την εναλλακτική υπόθεση Pitman τύπου 2 (sharp peak Pitman alternatives), η οποία σύμφωνα με τον ορισμό στην υποενότητα 2.1, για $\psi = 1$ θα είναι της μορφής

$$H_{2n} : f_{2n}(x; \theta_0) = f_0(x; \theta_0) + n^{-\varepsilon} \eta\left(\frac{x - c}{n^{-\gamma}}\right).$$

Έχοντας ως σκοπό ο ρυθμός σύγκλισης του ελέγχου να είναι της τάξης $n^{-1/2}$, θέτουμε $n^{-\varepsilon} = a_n = n^{-1/2} \hat{h}_{\text{opt}}^{-1/4} n^{-\gamma}$ και $\eta(x) = i\{f_{2n}(x; \theta_0) - f_0(x; \theta_0)\}$, όπου η παράμετρος $i \in \{0, 0.2, 0.4, \dots, 2.4\}$ ελέγχει πάλι την απόσταση από τη μηδενική υπόθεση. Σε όλα τα παραδείγματα για εναλλακτικές Pitman τύπου 2, έχουμε επιλέξει $\gamma = 1/4$, το οποίο μας οδηγεί σε ρυθμό σύγκλισης στη μηδενική υπόθεση τάξης $n^{-1/2}$, δηλαδή $a_n = O_p(n^{-1/2})$. Στο πρώτο και τέταρτο αντίστοιχο παράδειγμα παρακάτω έχουμε επιλέξει $c = 1/2$, στο δεύτερο $c = -1/2$ και στο τρίτο $c = 1$.

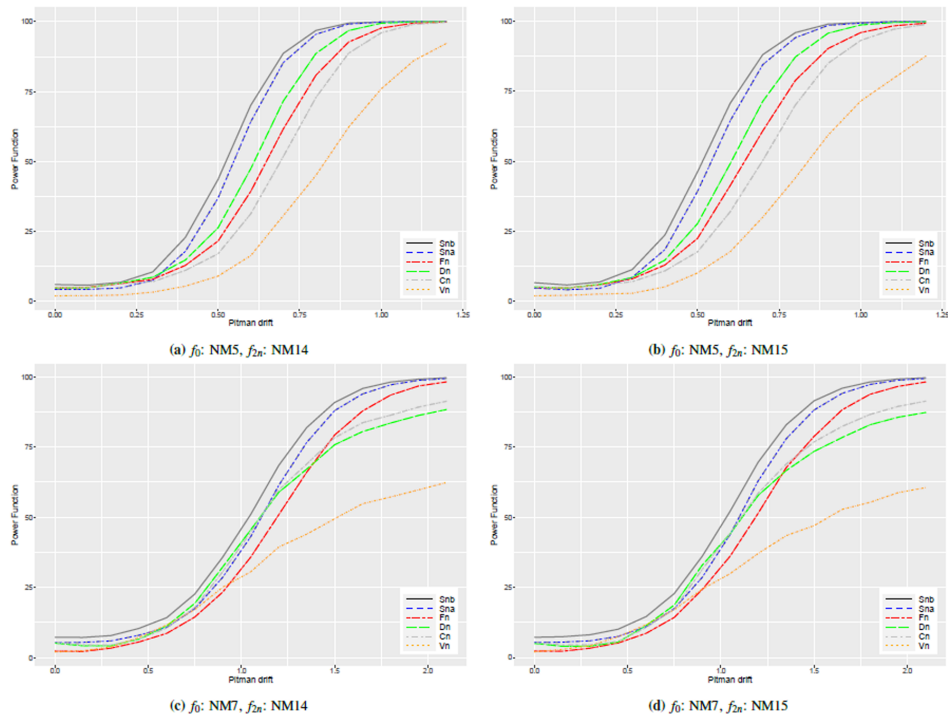
Τα αποτελέσματα των προσομοιώσεων βασίζονται στο άρθρο των Bagkavos et al. (2023) και μεταφέρονται σε αυτό το κεφάλαιο χάριν πληρότητας της παρουσίασης. Παρακάτω λοιπόν παρουσιάζονται οι γραφικές παραστάσεις 3.1 για Pitman εναλλακτικές τύπου 1 και 3.2 για τις αντίστοιχες τύπου 2.

ΚΕΦΑΛΑΙΟ 3 3.2. Αποδοτικότητα των ελέγχων υπό Pitman εναλλακτικές



Σχήμα 3.1: Εμπειρική εκτίμηση της συνάρτησης ισχύος των S_n ($S_{n,a}$ και $S_{n,b}$), D_n , C_n , F_n και V_n υπό την εναλλακτική υπόθεση Pitman τύπου 1. $n = 40$, $a = 5\%$ (βλέπε Bagkavos et al. (2023))

ΚΕΦΑΛΑΙΟ 3 3.2. Αποδοτικότητα των ελέγχων υπό Pitman εναλλακτικές



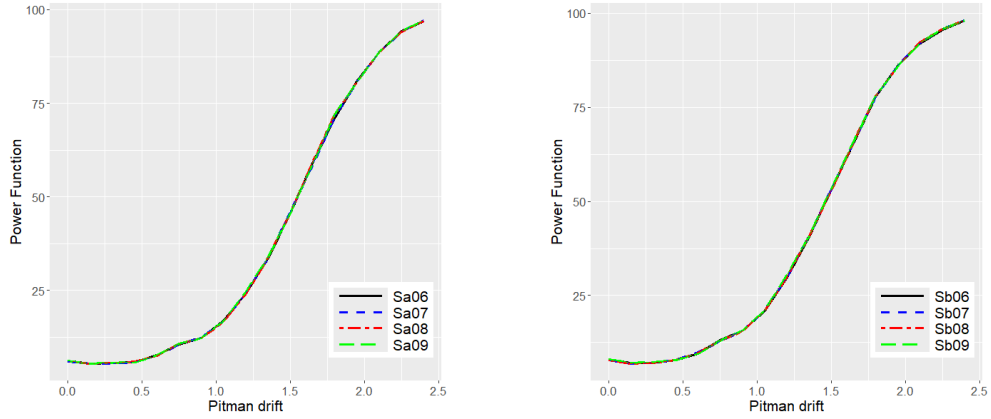
Σχήμα 3.2: Εμπειρική εκτίμηση της συνάρτησης ισχύος των S_n ($S_{n,a}$ και $S_{n,b}$), D_n , C_n , F_n και V_n υπό την εναλλακτική υπόθεση Pitman τύπου 2. $n = 40, a = 5\%$ (βλέπε Bagkavos et al. (2023))

Όλα τα παραδείγματα δείχνουν ότι η στατιστική συνάρτηση S_n είναι σχεδόν πάντα πιο ισχυρή από τους υπόλοιπους ελέγχους. Συγκεκριμένα, η S_n παρουσιάζει ανώτερη απόδοση σε σύγκριση με τις στατιστικές συναρτήσεις D_n και F_n . Πρέπει να σημειωθεί εδώ ότι ο προεπιλεγμένος κανόνας εύρους ζώνης $\mathbf{dpik}$, που χρησιμοποιείται για την υλοποίηση της F_n , είναι ένας κανόνας plug-in που βελτιστοποιεί την εκτίμηση της υποκείμενης πυκνότητας και όχι την ισχύ του ελέγχου. Αυτό είναι ένα γνωστό πρόβλημα των ελέγχων, που χρησιμοποιούν την μέθοδο των πυρήνων και μειώνει την απόδοση τους σε δείγματα πεπερασμένου μεγέθους (βλέπε Ghosh and Huang (1991)). Μια πρώτη προφανής συνέπεια της χρήσης ενός μη βέλτιστου εύρους ζώνης (από την άποψη της ισχύος) είναι ότι το μέγεθος της F_n δεν πλησιάζει το θεωρητικό επίπεδο σε όλα τα παραδείγματα που παρουσιάζονται εδώ. Τώρα, η προσομοίωση δείχνει επίσης σαφώς ανώτερη απόδοση, σε πεπερασμένα δείγματα, της S_n σε σύγκριση με τους ελέγχους C_n και V_n . Το συμπέρασμα που παίρνουμε από τις γραφικές παραστάσεις 3.1 και 3.2 είναι ότι ο κανόνας εύρους ζώνης που επιλέχτηκε και οι εκτιμήσεις της κρίσιμης τιμής που προτείνονται εδώ, εξασφαλίζουν εξαιρετική απόδοση για την S_n όταν χρησιμοποιούνται υπό τις εναλλακτικές υποθέσεις Pitman τύπου 1 και 2, τουλάχιστον για πυκνότητες με ενδιαφέροντα χαρακτηριστικά (π.χ. ακραίες τιμές, υψηλή κυρτότητα, ασυμμετρία, πολυτροπικότητα) όπως οι πυκνότητες που χρησιμοποιήθηκαν εδώ. Ένα ακόμη αξιοσημείωτο συμπέρασμα είναι ότι η καμπύλη ισχύος $S_{n,a}$ έχει σχεδόν πάντα την ίδια απόδοση με την αντίστοιχη $S_{n,b}$. Συνεπώς, η υλοποίηση της S_n με το ασυμπτωτικό κρίσιμο σημείο $z_{\alpha,0,\theta_0}$ αντί του bootstrap κρίσιμου σημείου ℓ_α^* αναμένεται να επιτύχει ουσιαστικά την ίδια απόδοση, ακόμη και για μεγέθη δείγματος τόσο μικρά όσο $n = 40$.

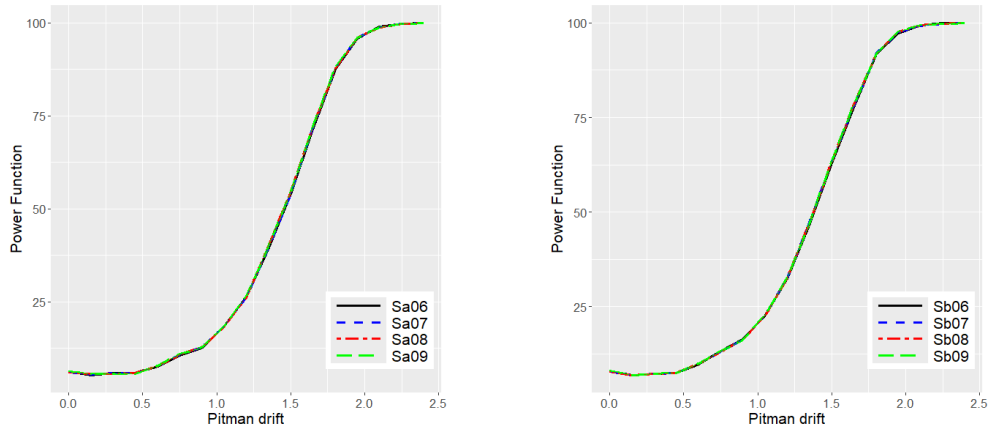
3.3 Μεταβολή της παραμέτρου Δ

Σκοπός της παρούσας προσομοίωσης είναι η διερεύνηση της επίδρασης της παραμέτρου Δ στην ισχύ του στατιστικού ελέγχου, που αναφέραμε στο κεφάλαιο 2. Συγκεκριμένα θα δούμε αν υπάρχει σημαντική διαφορά στη συνάρτηση ισχύος τόσο του $S_{n,a}$ όσο και του $S_{n,b}$, καθώς θα μεταβάλλουμε την παράμετρο Δ . Σύμφωνα με την υπόθεση A.2, η παράμετρος Δ είναι της μορφής $\Delta = n^{-\lambda}$, $\lambda \in (\frac{1}{2}, 1)$. Επομένως ένα πρώτο βήμα είναι να μεταβάλλουμε κάθε φορά την παράμετρο λ και να εξετάζουμε τις διαφορές μεταξύ των καμπυλών της ισχύος. Παρακάτω παρουσιάζονται σε μία γραφική παράσταση οι καμπύλες της ισχύος του S_n , $S_{n,a}$ και $S_{n,b}$ για διάφορες μεταβολές του λ . Σε κάθε προσομοίωση αναγράφεται ποιές μίξεις κανονικών κατανομών ελέγχονται. Συμβολίζουμε με $Sa06$, $Sa07$, $Sa08$ και

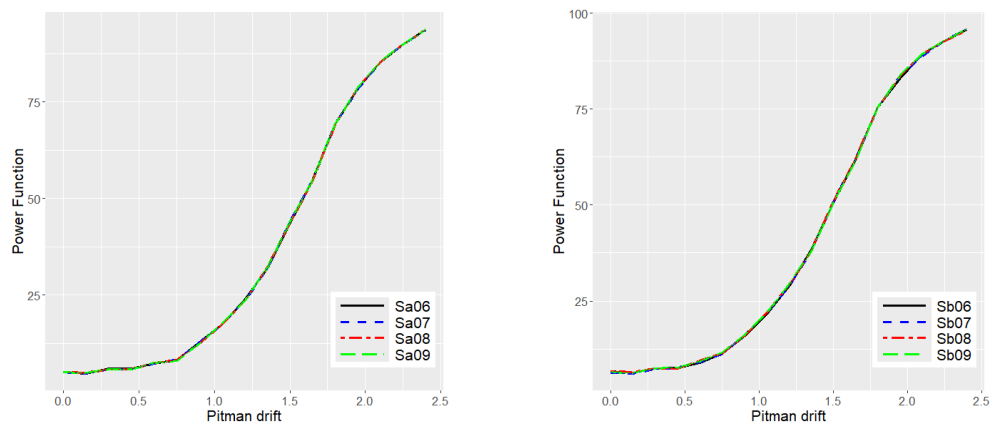
$Sa09$ το στατιστικό $S_{n,a}$, στο οποίο η παράμετρος λ έχει επιλεγεί να είναι 0.6, 0.7, 0.8 και 0.9 αντίστοιχα. Ακολουθήθηκε πανομοιότυπη διαδικασία και για το $S_{n,b}$.



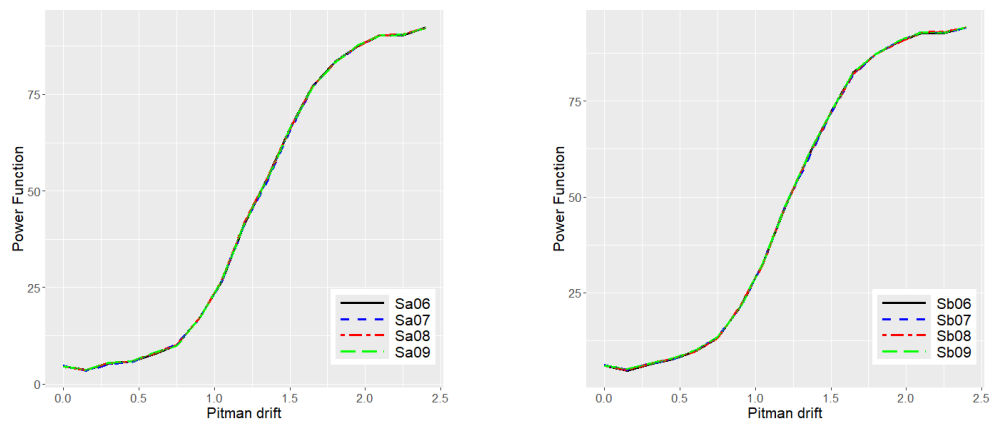
Σχήμα 3.3: $f_0 : NM6, f_{1n} : NM4$



Σχήμα 3.4: $f_0 : NM6, f_{1n} : NM5$



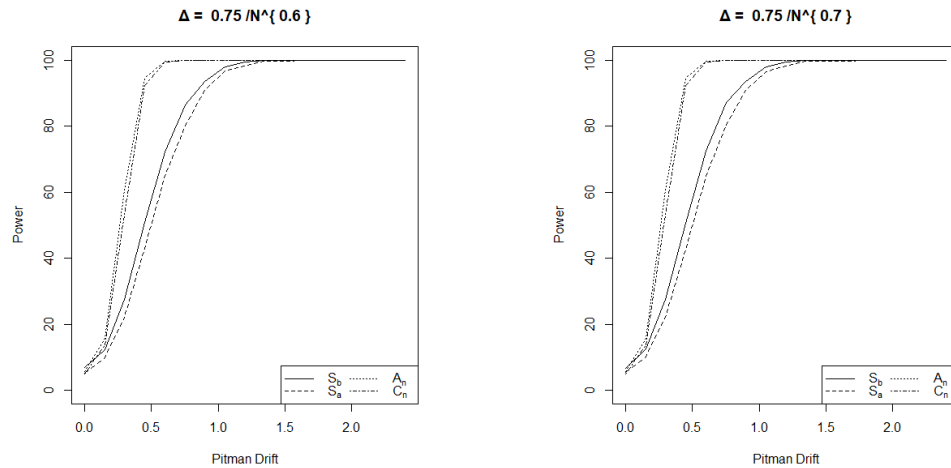
Σχήμα 3.5: $f_0 : NM6, f_{1n} : NM10$

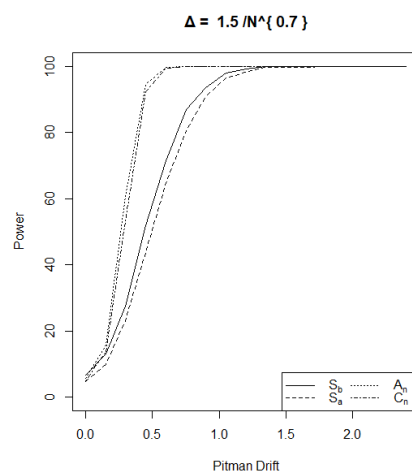
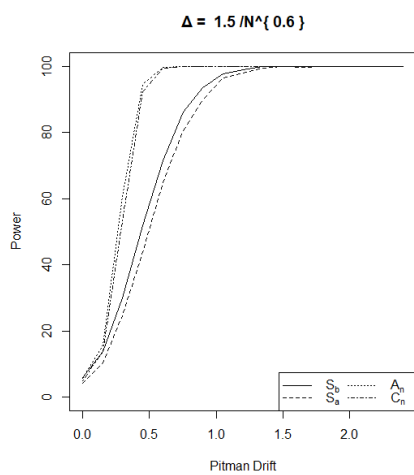
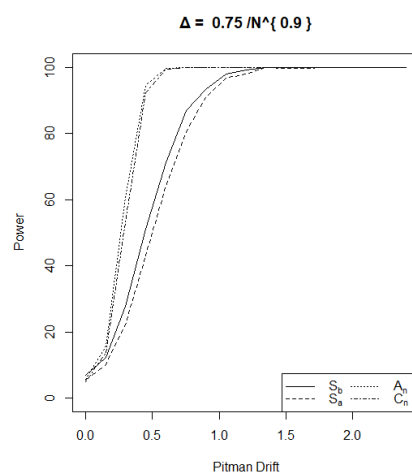
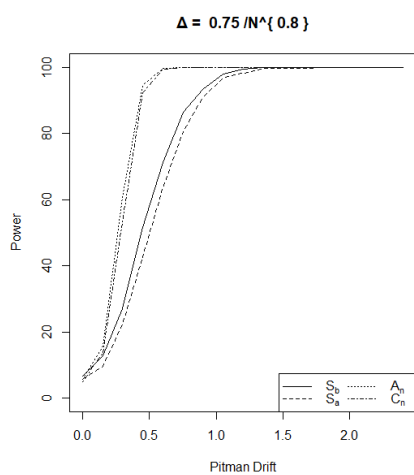


Σχήμα 3.6: $f_0 : NM14, f_{2n} : NM15$

Κάθε μία από τις παραπάνω γραφικές παραστάσεις είναι η εμπειρική εκτίμηση της συνάρτησης ισχύος του S_n , με $\Delta = n^{-\lambda}$, $\lambda \in \{0.6, 0.7, 0.8, 0.9\}$ υπό την εναλλακτική υπόθεση Pitman τύπου 1 και τύπου 2 στο τελευταίο σχήμα, όπου $n = 40$ και $a = 5\%$.

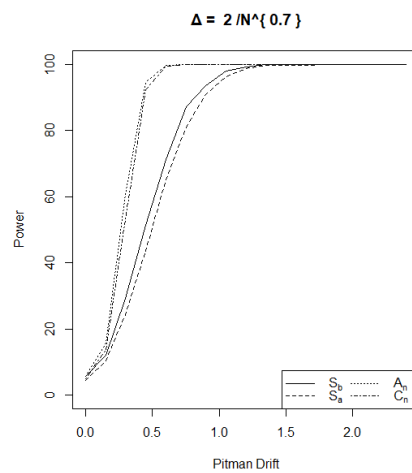
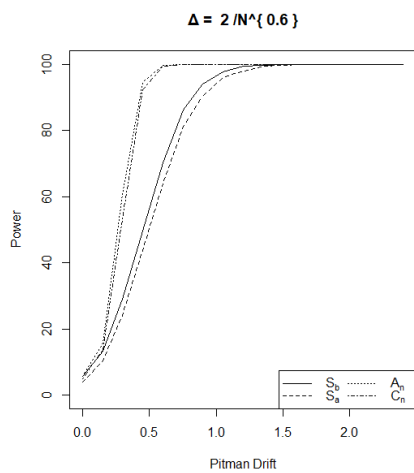
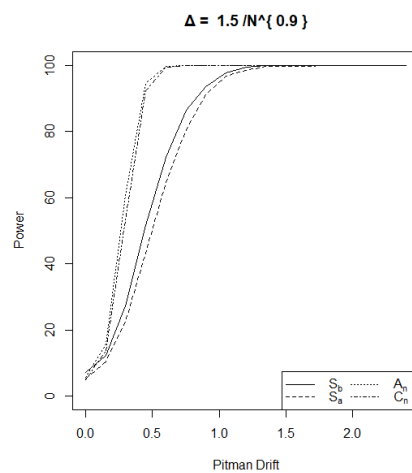
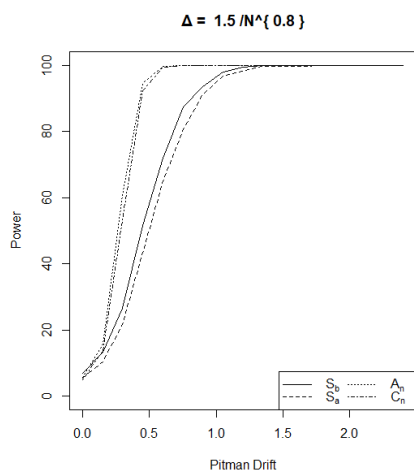
Είναι εμφανές από τα παραπάνω σχήματα ότι η παράμετρος Δ δεν επηρεάζει την ισχύ του S_n , καθώς οι καμπύλες της ισχύος, $S_{n,a}$ και $S_{n,b}$, για κάθε μεταβολή του Δ παραμένουν ουσιαστικά αμετάβλητες, με αποτέλεσμα οι γραφικές παραστάσεις της ισχύος να ταυτίζονται σχεδόν πλήρως. Συνεχίζοντας την εξέταση της επίδρασης του Δ στην ισχύ του S_n , εισάγεται μία εναλλακτική μορφή για την παράμετρο Δ , της μορφής $\Delta = \frac{c}{n^\lambda}$, η οποία διατηρεί τον ρυθμό σύγκλισης του τεστ αμετάβλητο. Συγκεκριμένα θέτουμε $\Delta = \frac{c}{n^\lambda}$, με $c \in \{0.75, 1.5, 2\}$ και $\lambda \in \{0.6, 0.7, 0.8, 0.9\}$. Η εφαρμογή αυτή θα γίνει σε συνδιασμούς μηδενικής και εναλλακτικής υπόθεσης Pitman, στους οποίους η ισχύς του S_n είναι σημαντικά πιο μικρή από τους υπόλοιπους ελέγχους, έτσι ώστε να εξετάσουμε τυχόν βελτίωση στην καμπύλη της ισχύος. Παρακάτω σε κάθε γραφική παράσταση υπάρχουν ενδεικτικά οι συναρτήσεις ισχύος των Anderson–Darling A_n και Cramér–von Mises (C_n), για να τονιστεί η υπεροχή τους σε σχέση με τις καμπύλες της ισχύος, $S_{n,a}$ και $S_{n,b}$. Επιπλέον, σε κάθε προσομοίωση αναγράφεται η μορφή του Δ που εφαρμόζεται. Τέλος, σε όλες τις γραφικές παραστάσεις έχει οριστεί $n = 40$ και $a = 5\%$.

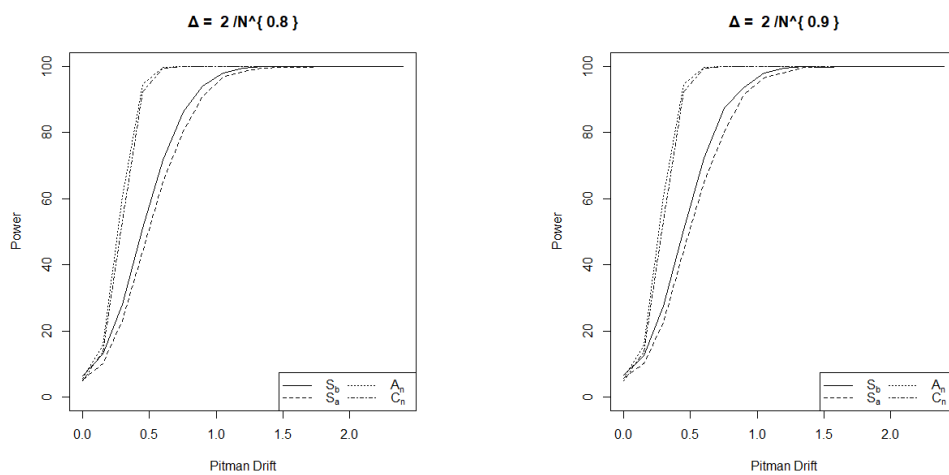




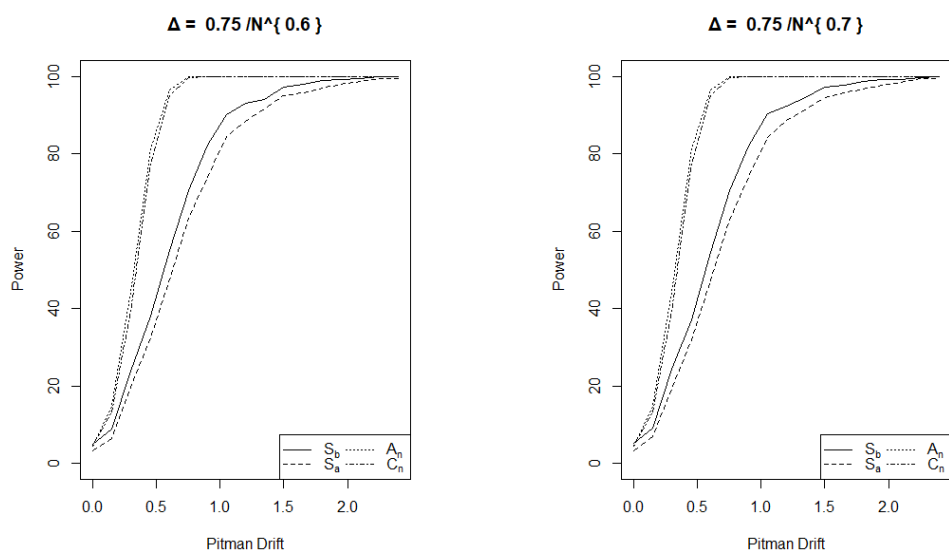
ΚΕΦΑΛΑΙΟ 3

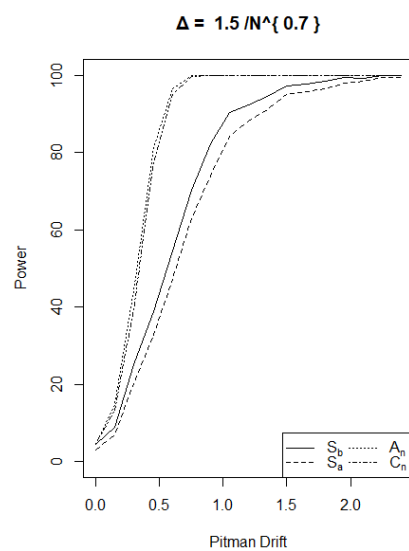
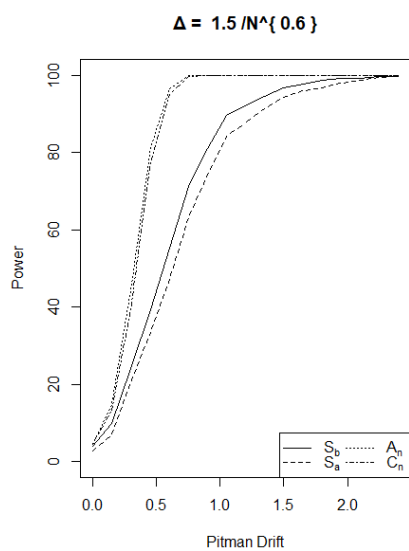
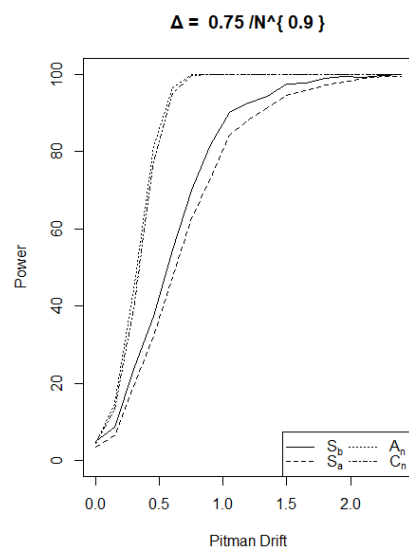
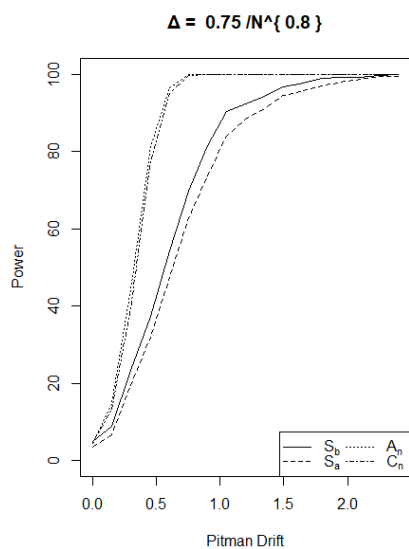
3.3. Μεταβολή της παραμέτρου Δ

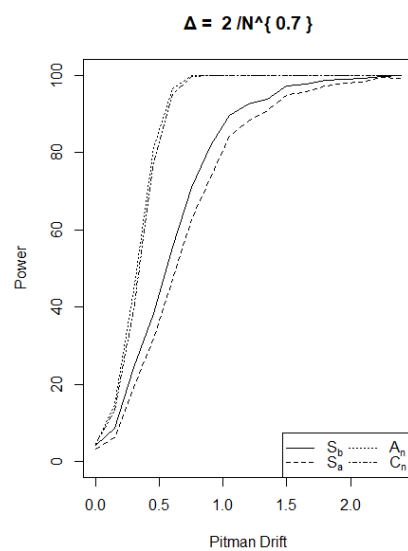
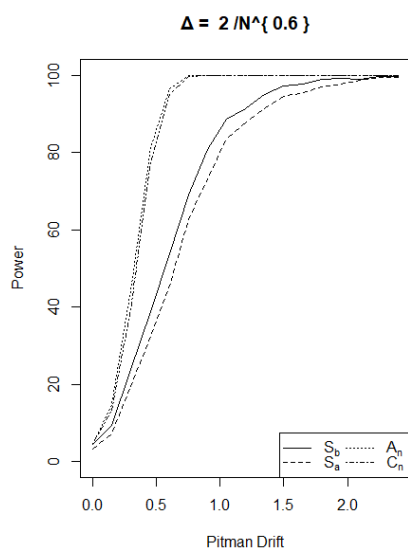
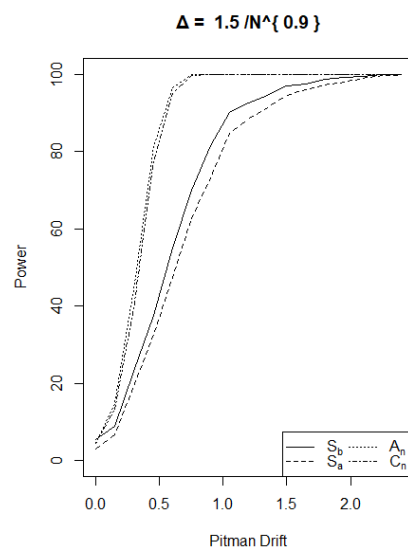
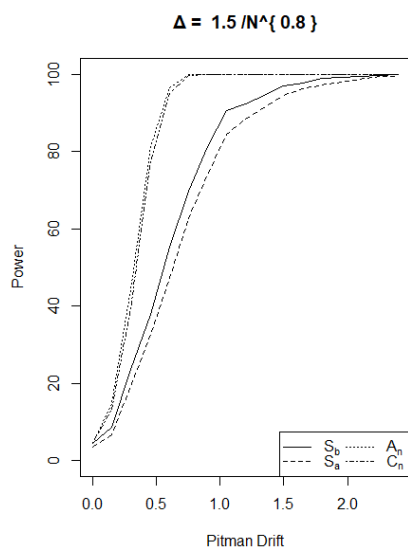


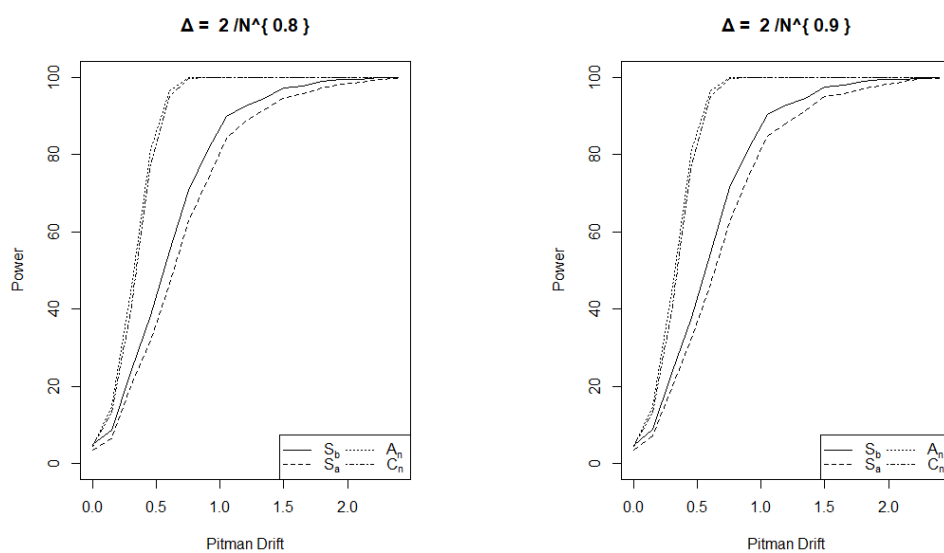


Σχήμα 3.7: $f_0 : NM3, f_{1n} : NM8$









Σχήμα 3.8: $f_0 : NM3, f_{1n} : NM9$

Επίλογος

Συμπερασματικά, από τις παραπάνω γραφικές παραστάσεις προκύπτει ότι η παράμετρος εκτίμησης της άγνωστης σ.π.π., Δ , δεν επηρεάζει την ισχύ του S_n σε δείγματα μικρού μεγέθους, τόσο με την εφαρμογή ασυμπτωτικών κρίσιμων σημείων όσο και με τη χρήση της μεθόδου bootstrap. Αυτό είναι εμφανές διότι σε όλες τις προσομοιώσεις η μεταβολή των καμπυλών της ισχύος $S_{n,a}$ και $S_{n,b}$ είναι αμελητέα, μη προσφέροντας κάποια σημαντική αυξομείωση που να επηρεάζει την αποδοτικότητα του τεστ. Συνεπώς σε πρακτικές εφαρμογές η αρχική μας επιλογή του Δ , είναι ικανοποιητική τόσο για μικρά όσο και για μεγάλα δείγματα, ιδιαίτερα για πυκνότητες με σχήματα σαν αυτά που εξετάσαμε εδώ. Επομένως με αυτόν τον τρόπο δίνουμε μία κατεύθυνση σε μελλοντική έρευνα πάνω σε αυτόν τον έλεγχο καλής προσαρμογής της οποίας η βάση θα είναι τα θεωρητικά αποτέλεσμα του άρθρου των Bagkavos et al. (2023) και θα μπορεί να επεκταθεί μελλοντικά στη μελέτη της περίπτωσης των ελαφρώς εξαρτημένων δεδομένων έναντι Pitman, Kullback-Leibler ή κάποιας άλλης μορφής εναλλακτικών υποθέσεων.

ΚΕΦΑΛΑΙΟ 4

ΠΑΡΑΡΤΗΜΑ

Οι αποδείξεις ακολουθούν πιστά τις αποδείξεις του άρθρου των Bagkavos et al. (2023) και μεταφέρονται σε αυτό το παράρτημα χάριν πληρότητας της παρουσίασης.

Ορίζουμε για κάθε συνάρτηση $F: \mathbb{R} \rightarrow \mathbb{R}$,

$$M(F) = \sup_{x \in \mathbb{R}} |F(x)|.$$

Λήμμα 4.1. Έστω $F: \mathbb{R} \rightarrow \mathbb{R}$ είναι μία φραγμένη, ολοκληρώσιμη συνάρτηση, η οποία έχει συνεχή και φραγμένη παράγωγο $F'(x) = \frac{dF}{dx}(x)$. Έστω επίσης $\Delta > 0$, $c > 0$ είναι θετικές και πραγματικές παράμετροι.

i) $\forall x \in \mathbb{R}$ ισχύει,

$$\left| \int_{-\Delta c/2}^{\Delta c/2} F(x+u)du - c\Delta F(x) \right| \leq \frac{1}{4}c^2 M(F')\Delta^2.$$

ii) $\forall N \in \mathbb{N}$ και $\forall \Delta > 0$ ισχύει,

$$\left| c\Delta \sum_{j=-N}^N F(j\Delta c) - \int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} F(u)du \right| \leq \frac{1}{4}c^2 M(F')(2N+1)\Delta^2.$$

Επίσης, αν $N \rightarrow \infty$ και $\Delta \rightarrow 0$ τότε το παραπάνω άνω φράγμα είναι τάξης $O(N\Delta^2)$.

iii) Έστω ότι $G: \mathbb{R} \rightarrow \mathbb{R}$ είναι μία φραγμένη συνάρτηση με φραγμένη και συνεχή παράγωγο και έστω ότι η συνάρτηση F ικανοποιεί τις παραπάνω συνθήκες και επιπλέον

$$\int_{-\infty}^{\infty} |F(u)||u|du < \infty.$$

ΚΕΦΑΛΑΙΟ 4

Τότε $\forall$ ακολουθία $\{\alpha_N\}_{N \in \mathbb{N}}$ τέτοια ώστε $a_N \xrightarrow{N \rightarrow \infty} 0$, θα ισχύει

$$\begin{aligned} & \left| c\Delta \sum_{j=-N}^N F(j\Delta c)G(a_N j\Delta c) - G(0) \int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} F(u)du \right| \\ & \leq \frac{1}{2}c^2 \{3M(F')M(G) + a_N M(F)M(G')\}(2N+1)\Delta^2 \\ & \quad + a_N M(G') \int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} |F(u)||u|du. \end{aligned} \quad (4.1)$$

Επιπλέον, το παραπάνω άνω φράγμα είναι τάξης $O(N\Delta^2 + a_N)$.

Απόδειξη. i) Η πρώτη παράγωγος της συνάρτησης F είναι συνεχής και άρα μπορούμε να εφαρμόσουμε το Θεώρημα Taylor,

$$\begin{aligned} \left| \int_{-\Delta c/2}^{\Delta c/2} F(x+u)du - c\Delta F(x) \right| &= \left| \int_{-\Delta c/2}^{\Delta c/2} \{F(x+u) - F(x)\}du \right| \\ &= \left| \int_{-\Delta c/2}^{\Delta c/2} uF'(x+u^*)du \right| \\ &\leq \int_{-\Delta c/2}^{\Delta c/2} |uF'(x+u^*)|du \\ &\leq M(F') \int_{-\Delta c/2}^{\Delta c/2} u du \\ &= \frac{1}{4}c^2 M(F')\Delta^2, \end{aligned}$$

όπου u^* είναι συνάρτηση του u , η οποία υπάρχει σύμφωνα με το Θεώρημα του Taylor.

ii) Για την απόδειξη του (ii) μέρους έχουμε ότι

$$\begin{aligned} \int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} F(u)du &= \sum_{j=-N}^N \int_{-\frac{1}{2}\Delta c}^{\frac{1}{2}\Delta c} F(j\Delta c + u)du \\ &= \sum_{j=-N}^N \int_{-\frac{1}{2}\Delta c}^{\frac{1}{2}\Delta c} \left\{ F(j\Delta c) + uF'(j\Delta c + u_j^*) \right\} du \\ &= c\Delta \sum_{j=-N}^N F(j\Delta c) + \sum_{j=-N}^N \int_{-\frac{1}{2}\Delta c}^{\frac{1}{2}\Delta c} uF'(j\Delta c + u_j^*)du, \end{aligned}$$

ΚΕΦΑΛΑΙΟ 4

όπου για κάθε j η ποσότητα u_j^* είναι συνάρτηση του u , η οποία υπάρχει σύμφωνα με το Θεώρημα του Taylor. Επομένως προκύπτει

$$\begin{aligned} \left| c\Delta \sum_{j=-N}^N F(j\Delta c) - \int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} F(u)du \right| &= \left| \sum_{j=-N}^N \int_{-\frac{1}{2}\Delta c}^{\frac{1}{2}\Delta c} uF'(j\Delta c + u_j^*)du \right| \\ &\leq \sum_{j=-N}^N \int_{-\frac{1}{2}\Delta c}^{\frac{1}{2}\Delta c} \left| uF'(j\Delta c + u_j^*) \right| du \\ &\leq (2N+1)M(F') \int_{-\frac{1}{2}\Delta c}^{\frac{1}{2}\Delta c} |u|du \\ &= \frac{1}{4}c^2M(F')(2N+1)\Delta^2, \end{aligned}$$

και το μέρος (ii) αποδείχθηκε.

iii) Τέλος για την απόδειξη του (iii) μέρους παρατηρούμε ότι

$$\begin{aligned} \sum_{j=-N}^N F(j\Delta c)G(a_N j\Delta c) &= \sum_{j=-N}^N F(j\Delta c)G(0) \\ &\quad + \sum_{j=-N}^N F(j\Delta c)\{G(a_N j\Delta c) - G(0)\}. \end{aligned} \quad (4.2)$$

Από το Λήμμα 4.1(ii) έχουμε ότι

$$\begin{aligned} \left| c\Delta \sum_{j=-N}^N F(j\Delta c)G(0) - G(0) \int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} F(u)du \right| \\ \leq \frac{1}{2}c^2|G(0)|M(F')(2N+1)\Delta^2. \end{aligned} \quad (4.3)$$

Ο δεύτερος όρος του αθροίσματος στο δεξί μέλος της ισότητας (4.2) μπορεί να γραφεί και ως εξής

$$\int_{-(N+\frac{1}{2})\Delta c}^{(N+\frac{1}{2})\Delta c} F(u)\{G(a_N u) - G(0)\}du + R, \quad (4.4)$$

έχοντας χρησιμοποιήσει το Λήμμα 4.1(ii), όπου

$$|R| = \frac{1}{4}c^2M(F'^*)(2N+1)\Delta^2, \quad (4.5)$$

ΚΕΦΑΛΑΙΟ 4

και

$$F^*(u) = F(u)\{G(a_N u) - G(0)\}.$$

Από τον κανόνα της παραγώγισης κατά παράγοντες έχουμε ότι

$$\begin{aligned} M(F^*)' &= M[\{G(a_N u) - G(0)\}F'(u) + a_N F(u)G'(a_N u)] \\ &\leq 2M(F')M(G) + a_N M(F)M(G'). \end{aligned} \quad (4.6)$$

Έπειτα εφαρμόζουμε το Θεώρημα Taylor στην απόλυτη τιμή του πρώτου όρου του αθροίσματος της σχέσης (4.4) και προκύπτει το παρακάτω άνω φράγμα

$$\begin{aligned} \int_{-\infty}^{\infty} \left| F(u)\{G(a_N u) - G(0)\} \right| du &\leq \int_{-\infty}^{\infty} a_N \left| F(u)uG'(a_N u^*) \right| du \\ &\leq a_N M(G') \int_{-\infty}^{\infty} |uF(u)| du. \end{aligned} \quad (4.7)$$

Επομένως για να προκύψει το άνω φράγμα της ανισότητας (4.1), αθροίζουμε το δεξί μέλος της ανισότητας (4.3), το δεξί μέλος της ισότητας (4.5) χρησιμοποιώντας την ανισότητα (4.6) και τέλος το δεξί μέλος της ανισότητας (4.7).

□

Λήμμα 4.2. Έστω K είναι μία συνάρτηση πυκνότητας πιθανότητας με συμπαγή φορέα στο κλειστό διάστημα $[-1, 1]$, είναι συμμετρική γύρω από το 0 και έχει συνεχή παράγωγο. Υποθέτουμε επίσης ότι για τις ακολουθίες $h = h_n$, $\Delta = \Delta_n$ και $g = g_n$ ισχύουν: $h \rightarrow 0$, $\Delta \rightarrow 0$, $g \rightarrow \infty$, $gh \rightarrow \infty$, $h/\Delta \rightarrow \infty$ και $g/n \rightarrow 0$, καθώς $n \rightarrow \infty$. Τέλος υποθέτουμε ότι $0 \neq c < \infty$.

i) Καθώς $n \rightarrow \infty$,

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \sim \frac{2}{c} \frac{hg}{\Delta} R_1(K).$$

ii) Επιπλέον έστω $F: \mathbb{R} \rightarrow \mathbb{R}$ είναι ολοκληρώσιμη συνάρτηση με $R_1(F) \neq 0$ και έχει συνεχή παράγωγο. Τότε καθώς $n \rightarrow \infty$,

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) F(x_j) \sim \frac{1}{c^2} \frac{h}{\Delta^2} R_1(K) R_1(F). \quad (4.8)$$

iii) Έστω ότι F_1 και F_2 ικανοποιούν τις υποθέσεις της συνάρτησης F του μέρους

ΚΕΦΑΛΑΙΟ 4

(ii) και επιπλέον το γινόμενο τους $F_1 F_2$ είναι ολοκληρώσιμο με $R_1(F_1 F_2) \neq 0$. Τότε καθώς $n \rightarrow \infty$,

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) F_1(x_j) F_2(x_k) \sim \frac{1}{c^2} \frac{h}{\Delta^2} R_1(K) R_1(F_1 F_2). \quad (4.9)$$

Απόδειξη. i) Παρατηρούμε ότι

$$\begin{aligned} \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) &= \sum_{j=-g}^g \sum_{\{k: |j-k| \leq h/\Delta c \text{ \& } |k| \leq g\}} K\left(\frac{x_j - x_k}{h}\right) \\ &= \sum_{j=-g}^g \sum_{\{k: |j-k| \leq h/(\Delta c)\}} K\left(\frac{x_j - x_k}{h}\right), \end{aligned} \quad (4.10)$$

όπου η δεύτερη ισότητα ισχύει για μεγάλο n καθώς $g\Delta \rightarrow \infty$, με αποτέλεσμα να ισχύει $2h/(\Delta c) \leq g$ για n αρκετά μεγάλο. Στη συνέχεια γράφουμε το σύνολο που αθροίζουμε ως προς k ως εξής:

$$\left\{ k : \max\left(j - \frac{h}{\Delta c}, -g\right) \leq k \leq \min\left(j + \frac{h}{\Delta c}, g\right) \right\}, \quad (4.11)$$

τότε, υπό τη συνθήκη $2h/(\Delta c) \leq g$, το j στο σύνολο (4.11) για να είναι της μορφής $\{m : |m| \leq h/(\Delta c)\}$ πρέπει $|j| \leq g - h/(\Delta c) = g\{1 - o(1)a\}$, όπου a είναι μία παράμετρος ομοιόμορφη στο n και ο υπολοίπομενος όρος είναι της τάξης $o(1)$, λόγω της υπόθεσης $g\Delta \rightarrow \infty$. Το κυρίαρχο μέρος του αθροίσματος στη σχέση (4.10) είναι

$$\begin{aligned} \sum_{j=-g}^g \sum_{\{m: |m| \leq h/(\Delta c)\}} K\left(m \frac{\Delta c}{h}\right) &= \frac{h}{c\Delta} \sum_{j=-g}^g \frac{c\Delta}{h} \sum_{\{m: |m| \leq h/(\Delta c)\}} K\left(m \frac{\Delta c}{h}\right) \\ &\sim (2g+1) \frac{h}{c\Delta} \int_{-1}^1 K(u) du \\ &\sim \frac{2gh}{c\Delta} R_1(K). \end{aligned}$$

Παραπάνω χρησιμοποιήσαμε το γεγονός ότι οι συναρτήσεις K και K' είναι φραγμένες καθώς και το ομοιόμορφο φράγμα του Λήμματος 4.1(ii).

ii) Χρησιμοποιώντας τον ίδιο λογισμό με αυτόν του Λήμματος 4.1(i) και

ΚΕΦΑΛΑΙΟ 4

έπειτα εφαρμόζοντας το Λήμμα 4.1(ii), προκύπτει

$$\begin{aligned} \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) F(x_j) &= \sum_{j=-g}^g F(x_j) \sum_{\{k: |j-k| \leq h/\Delta c \text{ \& } |k| \leq g\}} K\left(\frac{x_j - x_k}{h}\right) \\ &\sim \frac{h}{c\Delta} \frac{1}{c\Delta} c\Delta R_1(K) \sum_{j=-g}^g F(j\Delta c) \\ &\sim \frac{h}{c^2 \Delta^2} R_1(K) R_1(F), \end{aligned}$$

όπου υποθέτουμε ότι $R_1(F) \neq 0$.

iii) Ακολουθώντας το ίδιο σκεπτικό με αυτό των αποδείξεων για τα μέρη (i) και (ii) του Λήμματος 4.2 και χρησιμοποιώντας τα μέρη (ii) και (iii) του Λήμματος 4.1, προκύπτει

$$\begin{aligned} \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) F_1(x_j) F_2(x_k) &= \sum_{j=-g}^g F_1(x_j) \sum_{\{k: |j-k| \leq h/\Delta c \text{ \& } |k| \leq g\}} K\left(\frac{x_j - x_k}{h}\right) \\ &\quad \times F_2\left(x_j - h \frac{k\Delta c}{h}\right) \\ &\sim \frac{1}{c} \frac{h}{\Delta} \frac{1}{c\Delta} c\Delta R_1(K) \sum_{j=-g}^g F_1(j\Delta c) F_2(j\Delta c) \\ &\sim \frac{1}{c^2} \frac{h}{\Delta^2} R_1(K) R_1(F_1 F_2), \end{aligned}$$

όπου υποθέτουμε ότι $R_1(F_1 F_2) \neq 0$.

□

Λήμμα 4.3. Έστω $f: \mathbb{R} \rightarrow \mathbb{R}$ και $K: [-1, 1] \rightarrow \mathbb{R}$ είναι συναρτήσεις πυκνότητας πιθανότητας με φραγμένες και συνεχείς παραγώγους.

i) Καθώς $n \rightarrow \infty$,

$$\begin{aligned} \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) E[A_n(x_{j_1}, X_1) A_n(x_{k_1}, X_1) \\ \times A_n(x_{j_2}, X_1) A_n(x_{k_2}, X_1)] \sim 4^2 h^2 R_3(f). \end{aligned} \quad (4.12)$$

ΚΕΦΑΛΑΙΟ 4

ii) Καθώς $n \rightarrow \infty$,

$$\sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) E[A_n(x_{j_1}, X_1) A_n(x_{j_2}, X_1)] \\ \times E[A_n(x_{k_1}, X_2) A_n(x_{k_2}, X_2)] \sim 4h R_2(K) R_2(f). \quad (4.13)$$

Απόδειξη. i) Αρχικά εστιάζουμε στο αριστερό μέλος της σχέσης (4.13) και παρατηρούμε ότι προκύπτουν τρεις περιπτώσεις:

Περίπτωση 1: Δύο ισότητες. Δηλαδή $j_1 = j_2 \neq k_1 = k_2$ ή $j_1 = k_2 \neq j_2 = k_1$.

Περίπτωση 2: Μία ισότητα και δύο ανισότητες. Δηλαδή $j_1 = j_2$ και $k_1 \neq k_2$ ή $j_1 = k_2$ και $j_1 \neq k_1$ ή $k_1 = j_2$ και $j_1 \neq k_2$ και τέλος $k_1 = k_2$ και $j_1 \neq j_2$.

Περίπτωση 3: Όλοι οι δείκτες j_1, k_1, j_2 και k_2 είναι διαφορετικοί.

Στην Περίπτωση 1 έχουμε ότι

$$E[A_n(x_{j_1}, X_1)^2 A_n(x_{k_1}, X_1)^2] = p_{j_1}^2 p_{k_1} + p_{j_1} p_{k_1}^2 - 3p_{j_1}^2 p_{k_1}^2. \quad (4.14)$$

Στην Περίπτωση 2 όπου $k_1 \neq j_1 = j_2 \neq k_2$ και $k_1 \neq k_2$ προκύπτει ότι

$$E[A_n(x_{j_1}, X_1)^2 A_n(x_{k_1}, X_1) A_n(x_{k_2}, X_1)] = p_{j_1} p_{k_1} p_{k_2} - 3p_{j_1}^2 p_{k_1} p_{k_2}, \quad (4.15)$$

ενώ στην Περίπτωση 3,

$$E[A_n(x_{j_1}, X_1) A_n(x_{k_1}, X_1) A_n(x_{j_2}, X_1) A_n(x_{k_2}, X_1)] = -3p_{j_1} p_{k_1} p_{j_2} p_{k_2}. \quad (4.16)$$

Από τον ορισμό της πιθανότητας p_j και εφαρμόζοντας το Λήμμα 4.1(i) προκύπτει ότι $p_j = c\Delta f(x_j) + O(\Delta^2)$ ομοιόμορφα στο j . Εφαρμόζοντας επαναλαμβανόμενα το Λήμμα 4.2(iii) βλέπουμε ότι η συνεισφορά του πρώτου όρου του αθροίσματος στο δεξί μέλος της σχέσης (4.14) είναι

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) c^3 \Delta^3 f(x_j)^2 f(x_k) = O\left(\frac{h}{\Delta} \Delta^{-1} \Delta^3\right) = O(h\Delta),$$

και με παρόμοιους υπολογισμούς αποδεικνύεται ότι η συνεισφορά του δεύτερου και τρίτου όρου του αθροίσματος στο δεξί μέλος της σχέσης (4.14), είναι αντίστοιχα της τάξης $O(h\Delta)$ και $O(h\Delta^2)$. Στη συνέχεια εστιάζουμε στην Περίπτωση 2, όπου με αναλυτικούς υπολογισμούς παρουσιάζεται η συνεισφορά του πρώτου όρου του αθροίσματος στο δεξί μέλος της σχέσης (4.15).

$$\sum_{j \neq k_1} \sum_{j \neq k_2} K\left(\frac{x_j - x_{k_1}}{h}\right) K\left(\frac{x_j - x_{k_2}}{h}\right) p_j p_{k_1} p_{k_2} = c^3 \Delta^3 \frac{h^2}{c^2 \Delta^2} \sum_{j=-g}^g f(x_j) \frac{c\Delta}{h}$$

ΚΕΦΑΛΑΙΟ 4

$$\begin{aligned}
& \times \sum_{\{k_1: |j-k_1| \leq h/(\Delta c)\}} K\left(\frac{x_j - x_{k_1}}{h}\right) f(x_{k_1}) \frac{c\Delta}{h} \sum_{\{k_2: |j-k_2| \leq h/(\Delta c)\}} K\left(\frac{x_j - x_{k_2}}{h}\right) \\
& \times f(x_{k_2}) \sim c\Delta h^2 \sum_{j=-g}^g f(x_j) \frac{c\Delta}{h} \sum_{\{m_1: |m_1| \leq h/(\Delta c)\}} K\left(m_1 \frac{\Delta c}{h}\right) f\left(x_j - hm_1 \frac{\Delta c}{h}\right) \\
& \times \frac{c\Delta}{h} \sum_{\{m_2: |m_2| \leq h/(\Delta c)\}} K\left(m_2 \frac{\Delta c}{h}\right) f\left(x_j - hm_2 \frac{\Delta c}{h}\right) \sim \\
& h^2 c\Delta R_1(K)^2 \sum_{j=-g}^g f(x_j)^3 \sim h^2 R_1(K)^2 R_3(f). \tag{4.17}
\end{aligned}$$

όπου οι δύο τελευταίοι όροι της σχέσης (4.17) προκύπτουν από το Λήμμα 4.1(iii). Σημειώνουμε ότι ο δεύτερος όρος του αθροίσματος του δεξιού μέλους της σχέσης (4.15) στην Περίπτωση 2 αλλά και ο όρος του δεξιού μέλους της σχέσης (4.16) στην Περίπτωση 3, είναι της τάξης $O(h^2\Delta)$ και επομένως είναι ασυμπτωτικά αμελητέοι ως προς τον πρώτο όρο του αθροίσματος του δεξιού μέλους της σχέσης (4.15), του οποίου η συνεισφορά παρουσιάζεται στη σχέση (4.17). Τέλος για να λάβουμε το αποτέλεσμα στο δεξί μέλος της σχέσης (4.13), πολλαπλασιάζουμε με 4 διότι υπάρχουν, στην Περίπτωση 2, 4 τέτοιες υποπεριπτώσεις όπως εξηγήσαμε στην αρχή της απόδειξης.

ii) Για τους δείκτες του αθροίσματος διακρίνουμε 3 περιπτώσεις:

Περίπτωση 1: δύο ισότητες. Δηλαδή $j_1 = j_2$ και $k_1 = k_2$.

Περίπτωση 2: μία ισότητα και δύο ανισότητες. Δηλαδή $j_1 = j_2$ και $k_1 \neq k_2$ ή $j_1 \neq j_2$ και $k_1 = k_2$.

Περίπτωση 3: όλοι οι δείκτες είναι διαφορετικοί μεταξύ τους. Δηλαδή $j_1 \neq j_2$ και $k_1 \neq k_2$.

Έπειτα από υπολογισμούς προκύπτει ότι

$$E[A_n(x_j, X_1)A_n(x_k, X_1)] = \begin{cases} p_j - p_j^2, & \text{αν } j = k \\ -p_j p_k, & \text{αν } j \neq k. \end{cases}$$

Επομένως χρησιμοποιώντας παρόμοιες τεχνικές με αυτές του μέρους (i) του Λήμματος 4.3 και εφαρμόζοντας ταυτόχρονα το Λήμμα 4.2 (iii), αποδεικνύεται ότι: στην Περίπτωση 1 θα ισχύουν

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right)^2 p_j p_k \sim h R_2(K) R_2(f), \tag{4.18}$$

ΚΕΦΑΛΑΙΟ 4

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right)^2 p_j^2 p_k = O(h\Delta) = o(h), \quad (4.19)$$

$$\sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right)^2 p_j^2 p_k^2 = O(h\Delta^2) = o(h), \quad (4.20)$$

στην Περίπτωση 2 θα προκύπτει ότι

$$\sum_{j \neq k_1} \sum_{j \neq k_2} K\left(\frac{x_j - x_{k_1}}{h}\right) K\left(\frac{x_j - x_{k_2}}{h}\right) p_j p_{k_1} p_{k_2} = O(h^2) = o(h), \quad (4.21)$$

ενώ στην Περίπτωση 3 αποδεικνύεται ότι

$$\sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) p_{j_1} p_{j_2} p_{k_1} p_{k_2} = O(h^2 \Delta^2) = o(h). \quad (4.22)$$

Ο κυρίαρχος όρος εμφανίζεται στη σχέση (4.18) και η απόδειξη του μέρους (ii) έχει ολοκληρωθεί. $\square$

Λήμμα 4.4. Από τον ορισμό των $\hat{S}_n(h)$ και Y_j στην υποενοότητα 2.2, έχουμε ότι

$$\begin{aligned} \hat{S}_n(h) &= \sum_{j=-g}^g \sum_{\substack{k=-g \\ k \neq j}}^g K\left(\frac{x_j - x_k}{h}\right) \{Y_j - \hat{f}(x_j)\} \{Y_k - \hat{f}(x_k)\} \\ &= S_{1,n} + S_{2,n} + S_{3,n} + S_{4,n} + S_{5,n} + S_{6,n} + S_{7,n} + S_{8,n}, \end{aligned} \quad (4.23)$$

όπου

$$S_{1,n} = \frac{1}{(nc\Delta)^2} \sum_{i=1}^n \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) A_n(x_j, X_i) A_n(x_k, X_i), \quad (4.24)$$

$$S_{2,n} = \frac{1}{(nc\Delta)^2} \sum_{1 \leq i \neq r \leq n} \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) A_n(x_j, X_i) A_n(x_k, X_r), \quad (4.25)$$

$$S_{3,n} = \frac{1}{nc\Delta} \sum_{i=1}^n \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) A_n(x_j, X_i) \{E[Y_k] - f_*(x_k)\} = S_{4,n}, \quad (4.26)$$

ΚΕΦΑΛΑΙΟ 4

λόγω της συμμετρίας του πυρήνα K ,

$$S_{5,n} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{E[Y_j] - f_*(x_j)\} \{E[Y_k] - f_*(x_k)\}, \quad (4.27)$$

$$S_{6,n} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{Y_j - f_*(x_j)\} \{f_*(x_k) - \hat{f}(x_k)\} = S_{7,n}, \quad (4.28)$$

επίσης λόγω της συμμετρίας του K , και τέλος

$$S_{8,n} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{f_*(x_j) - \hat{f}(x_j)\} \{f_*(x_k) - \hat{f}(x_k)\}. \quad (4.29)$$

Σημειώνουμε ότι $f_* = f(x, \theta_0)$ υπό την H_0 και $f_* = f(x, \theta_1)$ υπό την H_1 , όπου θ_0 και θ_1 ορίζονται στην υπόθεση A.6.

Απόδειξη. Αρχικά παρατηρείται ότι

$$\begin{aligned} \hat{S}_n(h) &= \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{Y_j - \hat{f}(x_j)\} \{Y_k - \hat{f}(x_k)\} \\ &= \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{Y_j - f_*(x_j) + f_*(x_j) - \hat{f}(x_j)\} \\ &\quad \times \{Y_k - f_*(x_k) + f_*(x_k) - \hat{f}(x_k)\} \\ &= S_n^*(h) + S_{6,n} + S_{7,n} + S_{8,n}, \end{aligned}$$

όπου

$$S_n^*(h) = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{Y_j - f_*(x_j)\} \{Y_k - f_*(x_k)\}.$$

Αν εκφράσουμε το Y_j ως μία συνάρτηση $A_n(x_j, X_i)$ και αναπτύξουμε το άθροισμα, προκύπτει

$$\begin{aligned} S_n^*(h) &= \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \left[\frac{1}{n\Delta} \sum_{i=1}^n A_n(x_j, X_i) + \{E[Y_j] - f_*(x_j)\} \right] \\ &\quad \times \left[\frac{1}{n\Delta} \sum_{r=1}^n A_n(x_k, X_r) + \{E[Y_k] - f_*(x_k)\} \right] \\ &= S_{1,n} + S_{2,n} + S_{3,n} + S_{4,n} + S_{5,n}, \end{aligned}$$

και το Λήμμα αποδείχθηκε. $\square$

Εφαρμογή του Θεωρήματος του Hall

Υπενθυμίζουμε το Θεώρημα του Hall που αναφέραμε στην υποενότητα 1.4.

Θεώρημα Hall (1984). Έστω ένα δείγμα $X_1, \dots, X_n$ ανεξάρτητων και ισόνομων τυχαίων παρατηρήσεων.

Ορίζουμε U - στατιστικό την παρακάτω στατιστική συνάρτηση:

$$U_n = \sum_{1 \leq i < j \leq n} H_n(X_i, X_j) \quad (4.30)$$

όπου:

- H_n είναι συμμετρική συνάρτηση,
- $E[H_n(X_1, X_2)|X_1] \xrightarrow{a.s.} 0$,
- $E[H_n(X_1, X_2)^2] < \infty$, $\forall n \in \mathbb{N}$.

Αν ισχύει:

$$\frac{E[G_n(X_1, X_2)^2] + n^{-1}E[H_n(X_1, X_2)^4]}{\{E[H_n(X_1, X_2)^2]\}^2} \xrightarrow{n \rightarrow \infty} 0, \quad (4.31)$$

όπου $G_n(a, b) = E[H_n(X_1, a)H_n(X_1, b)]$, τότε:

$$U_n \xrightarrow{d} N(0, n^2 E[H_n(X_1, X_2)^2]/2).$$

Στόχος μας είναι να εφαρμόσουμε το Θεώρημα του Hall (1984) στη στατιστική συνάρτηση $S_{2,n}$ που ορίστηκε στη σχέση (4.25). Αν ορίσουμε

$$H_n(X_1, X_2) = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) A_n(x_j, X_1) A_n(x_k, X_2), \quad (4.32)$$

τότε μέσω της (4.25) προκύπτει $S_n = 2(nc\Delta)^{-2}U_n$. Επίσης η συνθήκη (4.31) δεν μεταβάλλεται όταν γίνεται η μετατροπή $H_n \rightarrow cH_n$ με $c \neq 0$ να είναι σταθερά. Επομένως η σταθερά $2/(nc\Delta)^2$ στο $S_{2,n}$ δεν παίζει κάποιο ρόλο στη συνθήκη (4.31), οπότε μπορούμε να δουλέψουμε απευθείας με τα U_n και H_n αντί με το $S_{2,n}$. Από τον ορισμό του $G_n(a, b)$ και τη (4.32) έχουμε ότι

$$\begin{aligned} G_n(a, b) &= E[H_n(X_1, a)H_n(X_1, b)] \\ &= \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) \\ &\quad \times A_n(x_{k_1}, a) A_n(x_{j_2}, b) E[A_n(x_{j_1}, X_3) A_n(x_{j_2}, X_3)], \end{aligned}$$

ΚΕΦΑΛΑΙΟ 4

οπότε

$$G_n(X_1, X_2) = \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) A_n(x_{k_1}, X_1) \\ \times A_n(x_{k_2}, X_2) E[A_n(x_{j_1}, X_3) A_n(x_{j_2}, X_3)]. \quad (4.33)$$

Σημειώνουμε ότι λόγω της ανεξαρτησίας των X_i και του ορισμού του $A_n(x_j, X_i)$, θα ισχύει ότι $E[H_n(X_1, X_2)] = 0 = E[G_n(X_1, X_2)]$.

Υπολογισμός του $E[H_n(X_1, X_2)^2]$. Υψώνοντας τη (4.32) στο τετράγωνο, μετά παίρνοντας την αναμενόμενη τιμή της και χρησιμοποιώντας το Λήμμα 4.3 μέρος (ii) προκύπτει

$$E[H_n(X_1, X_2)^2] = \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) \\ \times E[A_n(x_{j_1}, X_1) A_n(x_{j_2}, X_1)] E[A_n(x_{k_1}, X_2) A_n(x_{k_2}, X_2)]. \\ \sim h R_2(K) R_2(f). \quad (4.34)$$

Υπολογισμός του $E[H_n(X_1, X_2)^4]$. Ξεκινώντας από τη σχέση (4.32), έχουμε ότι

$$E[H_n(X_1, X_2)^4] = \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} \sum_{j_3 \neq k_3} \sum_{j_4 \neq k_4} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) \\ \times K\left(\frac{x_{j_3} - x_{k_3}}{h}\right) K\left(\frac{x_{j_4} - x_{k_4}}{h}\right) E[A_n(x_{j_1}, X_1) A_n(x_{j_2}, X_1) A_n(x_{j_3}, X_1) \\ \times A_n(x_{j_4}, X_1)] E[A_n(x_{k_1}, X_2) A_n(x_{k_2}, X_2) A_n(x_{k_3}, X_2) A_n(x_{k_4}, X_2)]. \quad (4.35)$$

Για να υπολογίσουμε την ποσότητα

$$T_0 = E[A_n(x_{j_1}, X_1) A_n(x_{j_2}, X_1) A_n(x_{j_3}, X_1) A_n(x_{j_4}, X_1)],$$

διακρίνουμε τις εξής περιπτώσεις.

Περίπτωση 1: Όλοι οι δείκτες ίσοι ($j_1 = j_2 = j_3 = j_4$). Υπάρχει μόνο μία τέτοια περίπτωση και τότε θα ισχύει, $T_0 = f(x_j) c \Delta + O(\Delta^2)$.

Περίπτωση 2: Τρεις δείκτες ίσοι, ένας διαφορετικός ($j_1 = j_2 = j_3 \neq j_4$). Υπάρχουν 4 τέτοιες περιπτώσεις. Στην αναφερόμενη περίπτωση θα ισχύει, $T_0 = -f(x_{j_1}) f(x_{j_4}) c^2 \Delta^2 + O(\Delta^2)$.

ΚΕΦΑΛΑΙΟ 4

Περίπτωση 3: Δύο ζευγάρια δεικτών ίσα ($j_1 = j_2 \neq j_3 = j_4$). Υπάρχουν 3 τέτοιες περιπτώσεις. Στην αναφερόμενη περίπτωση θα έχουμε, $T_0 = \{f(x_{j_1})^2 f(x_{j_3}) + f(x_{j_1})f(x_{j_3})^2\}\Delta^2 + O(\Delta^3)$.

Περίπτωση 4: Ένα ζευγάρι δεικτών ίσο και οι υπόλοιποι δείκτες διαφορετικοί ($j_3 \neq j_1 = j_2 \neq j_4 \neq j_3$). Υπάρχουν 6 τέτοιες περιπτώσεις. Σε αυτές τις περιπτώσεις προκύπτει, $T_0 = f(x_{j_1})f(x_{j_3})f(x_{j_4})\Delta^3 + O(\Delta^4)$.

Περίπτωση 5: Όλοι οι δείκτες διαφορετικοί. Υπάρχει μόνο μία τέτοια περίπτωση και τότε έχουμε,

$$T_0 = f(x_{j_1})f(x_{j_2})f(x_{j_3})f(x_{j_4})\Delta^4 + O(\Delta^5).$$

Από την *Περίπτωση 1* προκύπτει ότι αν $j_1 = j_2 = j_3 = j_4$ και $k_1 = k_2 = k_3 = k_4$, τότε η συνεισφορά στην αναμενόμενη τιμή $E[H_n(X_1, X_2)^4]$ είναι ασυμπτωτικά ανάλογη με την ποσότητα $hR_4(K)R_2(f)$. Επίσης χρησιμοποιώντας παρόμοια επιχειρήματα για τον υπολογισμό της $E[H_n(X_1, X_2)^2]$, καταλήγουμε στο συμπέρασμα ότι όλες οι υπόλοιπες συνεισφορές οδηγούν σε τάξη μεγέθους $o(h)$. Κατά συνέπεια, $E[H_n(X_1, X_2)^4] \sim hR_4(K)R_2(f)(1 + o(1))$ και επομένως

$$\begin{aligned} \frac{1}{n} \frac{E[H_n(X_1, X_2)^4]}{\{E[H_n(X_1, X_2)^2]\}^2} &= \frac{1}{n} \frac{hR_4(K)R_2(f)(1 + o(1))}{\{hR_2(K)R_2(f)(1 + o(1))\}^2} \\ &= \frac{1}{nh} \frac{R_4(K)(1 + o(1))}{R_2(K)^2 R_2(f)} \rightarrow 0, \end{aligned} \quad (4.36)$$

διότι $h \rightarrow 0$ και $nh \rightarrow \infty$. Συνεπώς ο δεύτερος όρος του αθροίσματος της συνθήκης (4.31) του Θεωρήματος του Hall συγκλίνει στο 0.

Υπολογισμός του $E[G_n(X_1, X_2)^2]$. Τώρα θα ασχοληθούμε με τον πρώτο όρο του αθροίσματος της συνθήκης (4.31), ο οποίος είναι

$$E[G_n(X_1, X_2)^2] \left\{ E[H_n(X_1, X_2)^2] \right\}^{-2},$$

όπου η συνάρτηση G_n ορίζεται στη σχέση (4.33). Χρησιμοποιώντας τον ορισμό της H_n στη σχέση (4.32), υπενθυμίζουμε ότι θα έχουμε

$$\begin{aligned} G_n(X_1, X_2) &= \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) A_n(x_{k_1}, X_1) \\ &\quad \times A_n(x_{k_2}, X_2) E[A_n(x_{j_1}, X_3) A_n(x_{j_2}, X_3)]. \end{aligned} \quad (4.37)$$

Επομένως μετά από πράξεις προκύπτει

$$E[G_n(X_1, X_2)^2] = \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} \sum_{j_3 \neq k_3} \sum_{j_4 \neq k_4} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right)$$

ΚΕΦΑΛΑΙΟ 4

$$\begin{aligned} & \times K\left(\frac{x_{j_3} - x_{k_3}}{h}\right) K\left(\frac{x_{j_4} - x_{k_4}}{h}\right) E[A_n(x_{j_1}, X_1) A_n(x_{j_3}, X_1)] E[A_n(x_{j_2}, X_2) \\ & \times A_n(x_{j_4}, X_2)] E[A_n(x_{k_1}, X_3) A_n(x_{k_2}, X_3)] E[A_n(x_{k_3}, X_4) A_n(x_{k_4}, X_4)]. \end{aligned} \quad (4.38)$$

Παρατηρούμε ότι ο όρος που συνεισφέρει περισσότερο στο παραπάνω πολλαπλό άθροισμα εμφανίζεται όταν $j_1 = j_3$, $j_2 = j_4$, $k_1 = k_2$ και $k_3 = k_4$. Επομένως σε αυτήν την περίπτωση το δεξί μέλος της σχέσης (4.38) είναι ασυμπτωτικά ισοδύναμο με την ποσότητα

$$\begin{aligned} & \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} \sum_{j_3 \neq k_3} \sum_{j_4 \neq k_4} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_1}}{h}\right) \\ & \times K\left(\frac{x_{j_1} - x_{k_3}}{h}\right) K\left(\frac{x_{j_2} - x_{k_3}}{h}\right) f(x_{j_1}) f(x_{k_1}) f(x_{j_2}) f(x_{k_3}) \Delta^4, \end{aligned} \quad (4.39)$$

όπου έχουμε χρησιμοποιήσει τα εξής: $E[A_n(x_j, X_i)^2] = p_j(1 - p_j)$ με $p_j = O(\Delta) = f(x_j)c\Delta + O(\Delta^2)$, όπου οι τάξεις σύγκλισης $O(\Delta)$ και $O(\Delta^2)$ είναι ομοιόμορφες ως προς n και j .

Τώρα κρατάμε σταθερό το j_1 και αθροίζουμε ως προς τα k_1 , j_2 και k_3 . Θέτουμε $x_{k_1} = x_{j_1} + m_1\Delta c/2$, $x_{j_2} = x_{j_1} + m_2\Delta c/2$, $x_{k_3} = x_{j_1} + m_3\Delta c/2$, όπου $m_1 = k_1 - j_1$, $m_2 = j_2 - j_1$ και $m_3 = k_3 - j_1$. Τότε το άθροισμα της σχέσης (4.39) γίνεται ασυμπτωτικά ισοδύναμο με

$$\begin{aligned} & c^4 \Delta^4 \sum_{j_1=-g}^g \sum_{m_1=-\infty}^{\infty} \sum_{m_2=-\infty}^{\infty} \sum_{m_3=-\infty}^{\infty} K\left(-m_1 \frac{\Delta c}{2h}\right) K\left((m_2 - m_1) \frac{\Delta c}{2h}\right) \\ & \times K\left(-m_3 \frac{\Delta c}{2h}\right) K\left((m_2 - m_3) \frac{\Delta c}{2h}\right) f(x_{j_1}) f\left(x_{j_1} + hm_1 \frac{\Delta c}{2h}\right) \\ & \times f\left(x_{j_1} + hm_2 \frac{\Delta c}{2h}\right) f\left(x_{j_1} + hm_3 \frac{\Delta c}{2h}\right). \end{aligned}$$

Κρατώντας σταθερό τον δείκτη j_1 , το εσωτερικό τριπλό άθροισμα γίνεται ασυμπτωτικά ισοδύναμο με

$$\begin{aligned} & h^3 c \Delta f(x_{j_1})^4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(-y_1) K(y_2 - y_1) K(-y_3) K(y_2 - y_3) dy_1 dy_2 dy_3 \\ & = h^3 c \Delta f(x_{j_1})^4 I_0(K), \end{aligned} \quad (4.40)$$

ΚΕΦΑΛΑΙΟ 4

όπου $I_0(K)$ είναι το παραπάνω τριπλό ολοκλήρωμα που περιλαμβάνει τον πυρήνα K . Επομένως αθροίζοντας ως προς το j_1 , η συνεισφορά της (4.40) στην αναμενόμενη τιμή της (4.38) θα είναι ασυμπτωτικά ισοδύναμη με την ποσότητα

$$h^3 R_4(f) I_0(K). \quad (4.41)$$

Παρατηρούμε ότι η (4.41) είναι τάξης $O(h^3)$, οπότε

$$\frac{E[G_n(X_1, X_2)^2]}{\{E[H_n(X_1, X_2)^2]\}^2} = \frac{h^3 R_4(f) I_0(K)(1 + o(1))}{\{h R_2(K) R_2(f)(1 + o(1))\}^2} = O(h) \rightarrow 0, \quad (4.42)$$

διότι $h \rightarrow 0$. Κατά συνέπεια, ο πρώτος όρος του αθροίσματος της συνθήκης (4.31) του Θεωρήματος του Hall θα συγχλίνει και αυτός στο 0, με την προϋπόθεση ότι όλες οι υπόλοιπες συνεισφορές στην αναμενόμενη τιμή $E[G_n(X_1, X_2)^2]$ είναι μικρότερης τάξης από την αντίστοιχη της ποσότητας (4.41).

Με βάση τους ορισμούς των συναρτήσεων $H_n(X_1, X_2)$, $G_n(X_1, X_2)$ στις σχέσεις (4.32) και (4.33) αντίστοιχα, ικανοποιούνται οι υποθέσεις του Θεωρήματος του Hall. Επιπλέον, οι σχέσεις (4.36) και (4.42) ικανοποιούν τη συνθήκη (4.31). Σύμφωνα λοιπόν με όσα έχουν ειπωθεί, χρησιμοποιώντας τη σχέση (4.34) και παρατηρώντας ότι $E[S_{2,n}] = 0$, προκύπτει άμεσα

$$n\Delta^2 h^{-1/2} c^2 S_{2,n} \xrightarrow{d} N(0, \sigma^2), \quad (4.43)$$

όπου $\sigma^2 = 2R_2(K)R_2(f)$. Επίσης, υπό την H_0 : $f = f(\cdot; \theta_0) = f_0$, η (4.43) ισχύει για $\sigma^2 = \sigma_0^2 = 2R_2(K)R_2(f_0)$, ενώ υπό την H_1 : $f = f_1 \neq f_0$, η (4.43) ισχύει για $\sigma^2 = \sigma_1^2 = 2R_2(K)R_2(f_1)$.

Λήμμα 4.5. *ι) Έστω η μηδενική υπόθεση H_0 : $f(x) = f(x; \theta_0) = f_0(x)$ για κάποιο $\theta_0 \in \Theta$, όπου $\{f(x; \theta) : \theta \in \Theta\}$ ορίζει ένα παραμετρικό μοντέλο. Θέτουμε*

$$\tau_n^2 = \text{Var}(S_{2,n}) \sim 2R_2(K)R_2(f_0)h/(n^2\Delta^4 c^4),$$

και έστω ισχύουν οι υποθέσεις A.1-A.7, που αναφέραμε στην υποενότητα 2.4. Τότε με βάση τους ορισμούς του Λήμματος 2.1, θα ισχύουν

$$S_{1,n}/\tau_n \asymp_p h^{1/2}, S_{3,n}/\tau_n = S_{4,n}/\tau_n \asymp_p n^{1/2}h^{1/2}\Delta^2, S_{5,n}/\tau_n \asymp nh^{1/2}\Delta^4, \quad (4.44)$$

$$S_{6,n}/\tau_n = S_{7,n}/\tau_n \asymp_p h^{1/2}(1 + n^{1/2}\Delta^2) \text{ και } S_{8,n}/\tau_n \asymp_p h^{1/2}. \quad (4.45)$$

ΚΕΦΑΛΑΙΟ 4

Επιπλέον, υπό τη μηδενική υπόθεση H_0 , η στατιστική συνάρτηση $\hat{S}_n(h)$ που ορίστηκε στην υποενότητα 2.2, ικανοποιεί

$$\hat{S}_n(h)/\tau_n = S_{2,n}/\tau_n + R_n \xrightarrow{d} N(0, 1), \quad (4.46)$$

όπου

$$R_n \asymp_p h^{1/2}(1 + n^{1/2}\Delta^2 + n\Delta^4) \asymp_p h^{1/2}(1 + n\Delta^4) \rightarrow 0. \quad (4.47)$$

ii) Υπό την H_1 : $f = f_1 \notin \{f(\cdot; \theta) : \theta \in \Theta\}$ και με την προϋπόθεση ότι ισχύουν οι υποθέσεις A.1-A.7, θα ικανοποιείται

$$n\Delta^2 h^{-1/2} \hat{S}_n(h) \asymp_p \mu_1, \quad (4.48)$$

όπου $\mu_1 \asymp nh^{1/2} \rightarrow \infty$.

Απόδειξη. i) Αρχικά παρατηρούμε ότι $E[S_{2,n}] = 0$ και με βάση τη σχέση (4.34) και το Λήμμα 4.3 (ii), θα ισχύει

$$\sqrt{\text{Var}(S_{2,n})} \sim \sqrt{\frac{2hR_2(K)R_2(f)}{n^2c^4\Delta^4}} \asymp \frac{h^{1/2}}{n\Delta^2}. \quad (4.49)$$

Χρησιμοποιώντας τη σχέση (4.24) και το Λήμμα 4.2 (iii), προκύπτει

$$|E[S_{1,n}]| \asymp \frac{1}{n\Delta^2} \Delta^2 \frac{h}{\Delta^2} = \frac{h}{n\Delta^2},$$

από το οποίο καταλήγουμε στο ότι

$$\frac{|E[S_{1,n}]|}{\sqrt{\text{Var}(S_{2,n})}} = O(h^{1/2}) \rightarrow 0, \quad (4.50)$$

διότι $h \rightarrow 0$. Επίσης χρησιμοποιώντας το Λήμμα 4.3 (i) και μετά από πράξεις, θα έχουμε ότι

$$\sqrt{\frac{\text{Var}(S_{1,n})}{\text{Var}(S_{2,n})}} = \sqrt{\frac{h^2/(n^2\Delta^4)}{h/(n^2\Delta^4)}} = O(h^{1/2}). \quad (4.51)$$

Επομένως το $S_{1,n}$ είναι ασυμπτωτικά αμελητέο σε σχέση με το $S_{2,n}$ υπό την H_0 .

Τώρα θα ασχοληθούμε με τις στατιστικές συναρτήσεις $S_{3,n}$ και $S_{4,n}$ που ορίστηκαν στη σχέση (4.26), υπενθυμίζοντας ότι $S_{3,n} = S_{4,n}$ λόγω της συμμετρίας γύρω από το $0 \in \mathbb{R}$ του πυρήνα K . Μπορούμε λοιπόν το $S_{3,n}$ να το γράψουμε στη μορφή

$$S_{3,n} = \frac{1}{nc\Delta} \sum_{i=1}^n W_{3,i}, \quad (4.52)$$

ΚΕΦΑΛΑΙΟ 4

όπου

$$W_{3,i} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) A_n(x_j, X_i) \{E[Y_k] - f_0(x_k)\}, \quad (4.53)$$

και λόγω της ανεξαρτησίας και ισονομίας των τυχαίων μεταβλητών $X_1, \dots, X_n$, η από κοινού κατανομή των $W_{3,i}$ θα εξαρτάται από το n . Οπότε θα ισχύουν $E[W_{3,i}] = 0$ και

$$E[W_{3,i}^2] = \sum_{j_1 \neq k_1} \sum_{j_2 \neq k_2} K\left(\frac{x_{j_1} - x_{k_1}}{h}\right) K\left(\frac{x_{j_2} - x_{k_2}}{h}\right) E[A_n(x_{j_1}, X_i) A_n(x_{j_2}, X_i)] \times \{E[Y_{k_1}] - f_0(x_{k_1})\} \{E[Y_{k_2}] - f_0(x_{k_2})\}. \quad (4.54)$$

Έχοντας υποθέσει ότι η $f_0(x)$ και όλες οι παράγωγοι της, πρώτης και δεύτερης τάξης, είναι φραγμένες, καταλήγουμε στο ότι οι διαφορές $E[Y_j] - f_0(x_j)$ μπορούν να γραφούν ως εξής:

$$|E[Y_j] - f_0(x_j)| = \frac{1}{c\Delta} p_j - f_0(x_j).$$

Αξιοποιώντας το γεγονός ότι $p_j = P(X \in (x_j - \Delta c/2, x_j + \Delta c/2))$ και εφαρμόζοντας ανάπτυγμα Taylor για τη συνάρτηση f_0 στο σημείο $u - x_j$ και γύρω από το x_j , καταλήγουμε στο ότι η παραπάνω διαφορά ισούται με

$$\left| \frac{1}{c\Delta} \int_{-\Delta c/2}^{\Delta c/2} \left\{ f_0(x_j) + u f_0'(x_j) + \frac{1}{2} u^2 f_0''(x_j + u^*) - f_0(x_j) \right\} du \right|,$$

και συνεχίζοντας προκύπτει

$$|E[Y_j] - f_0(x_j)| = \frac{1}{c\Delta} \int_{-\Delta c/2}^{\Delta c/2} \left| \frac{1}{2} u^2 f_0''(x_j + u^*) \right| du \asymp \Delta^2. \quad (4.55)$$

Σημειώνουμε ότι έχουμε συμβολίσει $f_0'(x) = \frac{df_0}{dx}(x)$ και $f_0''(x) = \frac{d^2 f_0}{dx^2}(x)$. Επιστρέφοντας στη σχέση (4.54) διακρίνουμε δύο περιπτώσεις: Περίπτωση 1 στην οποία $j_1 = j_2 = j$ και Περίπτωση 2 στην οποία $j_1 \neq j_2$. Στην Περίπτωση 1 θα έχουμε ότι $E[A_n(x_j, X_i) A_n(x_j, X_i)] = p_j - p_j^2$ και στην Περίπτωση 2 θα ισχύει $E[A_n(x_{j_1}, X_i) A_n(x_{j_2}, X_i)] = -p_{j_1} p_{j_2}$. Σημειώνουμε ότι υπό την H_0 , θα ισχύει $E[Y_j] = p_j/(c\Delta) = f_0(x_j) + O(\Delta^2)$ για $-g \leq j \leq g$. Χρησιμοποιώντας κάποιες εφαρμογές των Λημμάτων 4.1 και 4.2, προκύπτει $E[W_{3,i}^2] \asymp h^2 \Delta^2$, ομοιόμορφα πάνω από όλα τα i και το n . Συνεπώς

$$E[S_{3,n}^2] \asymp \frac{n}{(n\Delta)^2} h^2 \Delta^2 = n^{-1} h^2. \quad (4.56)$$

ΚΕΦΑΛΑΙΟ 4

Οπότε

$$\frac{Var(S_{3,n})}{Var(S_{2,n})} = O(n^{-1}h^2)O(h^{-1}n^2\Delta^4) = O(nh\Delta^4) \rightarrow 0. \quad (4.57)$$

Επομένως υπό τις υποθέσεις A.1-A.7 και υπό την H_0 , τα $S_{3,n}$ και $S_{4,n}$ είναι αμελητέα σε σύγκριση με το $S_{2,n}$. Με παρόμοιους υπολογισμούς αποδεικνύεται ότι υπό την H_1 , τα $S_{3,n}$ και $S_{4,n}$ συμπεριφέρονται ως εξής:

$$\frac{S_{3,n}}{\sqrt{Var(S_{2,n})}} = \frac{S_{4,n}}{\sqrt{Var(S_{2,n})}} \asymp nh \xrightarrow{n \rightarrow \infty} \infty. \quad (4.58)$$

Παρατηρούμε ότι η αλλαγή στην τάξη σύγκλισης από τη σχέση (4.57) στη σχέση (4.58) συμβαίνει λόγω της αφαίρεσης του όρου Δ^4 . Τώρα θα ασχοληθούμε με το $S_{8,n}$ που ορίζεται από τη σχέση (4.29). Αρχικά εφαρμόζουμε Θεώρημα Taylor στη συνάρτηση $\hat{f}(x_j) = f(x_j; \hat{\theta})$ γύρω από το $\theta = \theta_0$,

$$\hat{f}(x_j) = f_0(x_j) + (\hat{\theta} - \theta_0)^T \nabla_{\theta} f_0(x_j) + \frac{1}{2}(\hat{\theta} - \theta_0)^T \nabla_{\theta} \nabla_{\theta}^T f_0(x_j; \theta^*)(\hat{\theta} - \theta_0), \quad (4.59)$$

όπου το θ^* υπάρχει από το Θεώρημα Taylor, $d^2 f/d\theta^2$ είναι συνεχής και φραγμένη,

$$\sup_{\theta \in \Theta} \sup_{x \in \mathbb{R}} \left\| \nabla_{\theta} \nabla_{\theta}^T f(x; \theta) \right\| < \infty$$

και Θ είναι ο παραμετρικός χώρος. Εφαρμόζοντας τη (4.59) στη (4.29), βλέπουμε ότι ο κυρίαρχικός όρος είναι

$$(\theta - \theta_0)^T W_{8,n}(\theta - \theta_0), \quad (4.60)$$

όπου

$$W_{8,n} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \nabla_{\theta} f_0(x_j) \{\nabla_{\theta} f_0(x_k)\}^T.$$

Με βάση το Λήμμα 4.2, προκύπτει $\|W_{8,n}\| = O(h\Delta^{-2})$ το οποίο έχει ως αποτέλεσμα η σχέση (4.60) να είναι τάξης $O_p(n^{-1}h\Delta^{-2})$. Χρησιμοποιώντας παρόμοια επιχειρήματα με βάση το Λήμμα 4.2 προκύπτει

$$\frac{S_{8,n}}{\sqrt{Var(S_{2,n})}} = O_p(n^{-1}h\Delta^{-2})O(h^{1/2}n^{-1}\Delta^{-2}) = O_p(h^{1/2}), \quad (4.61)$$

το οποίο συγκλίνει στο 0, διότι $h \rightarrow 0$. Τώρα εξετάζουμε το $S_{6,n}$ που ορίζεται στη σχέση (4.28). Αρχικά γράφουμε τη διαφορά $Y_j - f_0(x_j)$ ως εξής:

$$Y_j - f_0(x_j) = Y_j - E[Y_j] + E[Y_j] - f_0(x_j) = \frac{1}{nc\Delta} \sum_{i=1}^n A_n(x_j, X_i) + E[Y_j] - f_0(x_j),$$

ΚΕΦΑΛΑΙΟ 4

και λόγω αυτού, μπορούμε να γράψουμε $S_{6,n} = S'_{6,n} + S''_{6,n}$, όπου

$$S'_{6,n} = \frac{1}{nc\Delta} \sum_{i=1}^n \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) A_n(x_j, X_i) \{f_0(x_k) - \hat{f}(x_k)\},$$

και

$$S''_{6,n} = \sum_{j \neq k} K\left(\frac{x_j - x_k}{h}\right) \{E[Y_j] - f_0(x_j)\} \{f_0(x_k) - \hat{f}(x_k)\}.$$

Χρησιμοποιώντας παρόμοιους υπολογισμούς με αυτούς που κάναμε για να βρούμε την τάξη σύγκλισης των $W_{3,i}$ και $S_{3,n}$, καταλήγουμε υπό την H_0

$$\frac{S'_{6,n}}{\sqrt{Var(S_{2,n})}} = O_p(h^{1/2}), \quad (4.62)$$

και με βάση τη σχέση (4.59) προκύπτει

$$\frac{S''_{6,n}}{\sqrt{Var(S_{2,n})}} = \begin{cases} O_p(n^{1/2}h^{1/2}\Delta^2), & \text{υπό την } H_0 \\ O_p(n^{1/2}h^{1/2}), & \text{υπό την } H_1. \end{cases} \quad (4.63)$$

Τέλος θα ασχοληθούμε με τη στατιστική συνάρτηση $S_{5,n}$ που ορίζεται στη σχέση (4.27). Εφαρμόζοντας Θεώρημα Taylor στην $f(x_j + u)$ γύρω από το σημείο x_j , θα έχουμε υπό την H_0 ότι

$$\begin{aligned} E[Y_j] - f_0(x_j) &= \frac{1}{c\Delta} \int_{-\Delta c/2}^{\Delta c/2} \left\{ f_0(x_j + u) - f_0(x_j) \right\} du \\ &= \frac{1}{c\Delta} \int_{-\Delta c/2}^{\Delta c/2} \left\{ u f'_0(x_j) + \frac{1}{2} u^2 f''_0(x_j) \right\} du + R_0 \\ &= \frac{1}{c\Delta} \int_{-\Delta c/2}^{\Delta c/2} \left\{ u f'_0(x_j) + \frac{1}{2} u^2 f''_0(x_j) \right\} du + R_0 \\ &= \frac{1}{c\Delta} \left(\frac{u^3}{6} \right)_{-\Delta c/2}^{\Delta c/2} f''_0(x_j) + R_0 \\ &= \frac{1}{c\Delta} \frac{\Delta^3 c^3}{24} f''_0(x_j) + R_0 \\ &= \frac{\Delta^2 c^2}{24} f''_0(x_j) + R_0. \end{aligned}$$

Υποθέτοντας ότι η τρίτη παράγωγος, ως προς x , της f_0 είναι συνεχής και ομοιόμορφα φραγμένη, θα έχουμε από το Θεώρημα Taylor ότι $R_0 = O(\Delta^3)$. Επομένως

ΚΕΦΑΛΑΙΟ 4

με βάση το Λήμμα 4.2 (iii) και υπό τη H_0 και τις υποθέσεις A.1-A.7, προκύπτει

$$S_{5,n} \sim \frac{h}{c^2 \Delta^{-2}} R_2 \left(\frac{c^2}{24} f_0'' \right), \quad (4.64)$$

και άρα

$$\frac{S_{5,n}}{\sqrt{\text{Var}(S_{2,n})}} \asymp \frac{\Delta^4 h}{\Delta^2} \Big/ \sqrt{\frac{h}{n^2 \Delta^4}} = O(nh^{1/2} \Delta^4).$$

ii) Υπό την εναλλακτική υπόθεση H_1 , με βάση το Λήμμα 4.2 (iii) θα έχουμε ότι

$$S_{5,n} \sim \frac{h}{c^2 \Delta^2} R_2(f_1 - f_0),$$

και κατά συνέπεια, χρησιμοποιώντας τη σχέση (4.43)

$$\frac{\sqrt{\text{Var}(S_{2,n})}}{S_{5,n}} = O(n^{-1} h^{1/2} \Delta^{-2}) O(h^{-1} \Delta^2) = O(n^{-1} h^{-1/2}) \rightarrow 0. \quad (4.65)$$

□

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Aït-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies*, 9:385–426.
- Bachmann, D. and Dette, H. (2005). A note on the Bickel–Rosenblatt test in autoregressive time series. *Statistics and Probability Letters*, 74:221–234.
- Bagkavos, D., Patil, P., and Wood, A. (2023). Nonparametric goodness-of-fit testing for a continuous multivariate parametric model. *Journal of Multivariate Analysis*.
- Berk, R. and Jones, D. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Probability Theory and Related Fields*, 47:47–59.
- Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, 1:1071–1095.
- Cao, R. and Lugosi, G. (2005). Goodness-of-fit tests based on the kernel density estimator. *Scandinavian Journal of Statistics*, 32:599–616.
- Chebana, F. (2004). On the optimization of the weighted Bickel–Rosenblatt test. *Statistics and Probability Letters*, 68:333–345.
- Chen, F., Jiménez-Gamero, M., Meintanis, S., and Zhu, L. (2022). A general Monte Carlo method for multivariate goodness-of-fit testing applied to elliptical families. *Computational Statistics and Data Analysis*, 175:107548.
- Csörgö, S. and Faraway, J. (1996). The exact and asymptotic distributions of Cramér–von Mises statistics. *Journal of the Royal Statistical Society Series B*, 58:221–234.

- d'Agostino, R. and Stephens, M. (1986). *Goodness-of-fit Techniques*. Marcel Dekker, New York.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling And Its Applications*. Chapman and Hall, London.
- Fan, Y. (1994). Testing the goodness-of-fit of a parametric density function by kernel method. *Econometric Theory*, 10:316–356.
- Fan, Y. (1995). Bootstrapping a consistent nonparametric goodness-of-fit test. *Econometric Reviews*, 14:367–382.
- Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis*, 62:36–63.
- Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed parameters. *Econometric Theory*, 14:604–621.
- Fragiadakis, K. and Meintanis, S. (2011). Goodness-of-fit tests for multivariate Laplace distributions. *Mathematical and Computer Modelling*, 53:769–779.
- Gao, J. and Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association*, 103:1584–1594.
- Gao, J. and King, M. (2004). Adaptive testing in continuous-time diffusion models. *Econometric Theory*, 20:844–882.
- Ghosh, B. and Huang, W. (1991). The power optimal kernel of the Bickel–Rosenblatt test for goodness of fit. *Annals of Statistics*, 19:999–1009.
- González-Manteiga, W. and Crujeiras, R. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, 22:361–411.
- Gouriéroux, C. and Tenreiro, C. (2001). Local power properties of kernel-based goodness of fit tests. *Journal of Multivariate Analysis*, 78:161–190.
- Gurevich, G. and Vexler, A. (2010). Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics and Data Analysis*, 54:531–545.
- Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, 14:1–16.

- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.
- Hall, P. and Heyde, C. (1981). Rates of convergence in the martingale central limit theorem. *Annals of Probability*, 9:395–404.
- Hallin, M., Mordant, G., and Segers, J. (2021). Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15:1328–1371.
- Härdle, W. and Mammen, E. (2003). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21:1926–1947.
- Jager, L. and Wellner, J. (2007). Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, 35:2018–2053.
- Jiménez-Gamero, M., Alba-Fernández, V., Muñoz-García, J., and Chalco-Cano, Y. (2009). Goodness-of-fit tests based on empirical characteristic functions. *Computational Statistics and Data Analysis*, 53:3957–3971.
- Khmaladze, E. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli*, 22:563–588.
- Klar, B., Lindner, F., and Meintanis, S. (2012). Specification tests for the error distribution in GARCH models. *Computational Statistics and Data Analysis*, 56:3587–3598.
- Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer, New York.
- Li, Q. and Racine, J. (2007). *Nonparametric Econometrics*. Princeton University Press, Princeton, New Jersey.
- Liang, H.-Y. and Jing, B.-Y. (2007). The lil for the Bickel–Rosenblatt test statistic. *Journal of Statistical Planning and Inference*, 137:1829–1837.
- Lin, L., Lee, S., and Guo, M. (2014). The Bickel–Rosenblatt test for continuous time stochastic volatility models. *Test*, 23:195–218.
- Marron, J. and Wand, M. (1992). Exact mean integrated squared error. *Annals of Statistics*, 20:712–736.
- Marsaglia, G., Tsang, W., and Wang, J. (2003). Evaluating Kolmogorov’s distribution. *Journal of Statistical Software*, 8. DOI: 10.18637/jss.v008.i18.

- Miecznikowski, J., Vexler, A., and Shepherd, L. (2013). dbemlikegof: an R package for nonparametric likelihood ratio tests for goodness-of-fit and two-sample comparisons based on sample entropy. *Journal of Statistical Software*, 54. DOI: 10.18637/jss.v054.i03.
- Pagan, A. and Ullah, A. (2005). *Nonparametric Econometrics*. Cambridge University Press.
- Powell, J., Stock, J., and Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrics*, 57:1403–1430.
- Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Annals of Statistics*, 3:1–14.
- Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690.
- Song, K.-S. (2002). Goodness-of-fit tests based on Kullback–Leibler discrimination information. *IEEE Transactions on Information Theory*, 48:1103–1117.
- Tenreiro, C. (2007). On the asymptotic behaviour of location-scale invariant Bickel–Rosenblatt tests. *Journal of Statistical Planning and Inference*, 137:103–116.
- Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Chapman and Hall, London.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.
- Zhang, H., Jin, J., and Wu, Z. (2017). Distributions and statistical power of optimal signal-detection methods in finite cases. arXiv:1702.07082v1.
- Zhang, X., King, M., and Hyndman, R. (2006). A bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 11:3009–3031.
- Zheng, J. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75:263–289.
- Κουρούκλης, Σ., Πετρόπουλος, Κ., and Πιπερίγκου, Β. (2015). Θέματα Παραμετρικής Στατιστικής Συμπερασματολογίας: Εκτιμητική και Διαστήματα Εμπιστοσύνης.