
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

# "ΕΦΑΡΜΟΓΗ ΤΟΥ ΕΛΕΓΧΟΥ ΜΕΓΙΣΤΗΣ ΕΝΤΡΟΠΙΑΣ ΣΕ ΚΑΤΑΝΟΜΕΣ ΜΕ ΠΑΧΙΕΣ ΟΥΡΕΣ"

Διπλωματική Εργασία

της

**Ελευθερίας Λαζάρου**

**Επιβλέπουσα Καθηγήτρια :** Φιλία Βόντα

Καθηγήτρια Ε.Μ.Π

**Αθήνα, Μάρτιος 2021**

**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**
**Σ**ΧΟΛΗ **Ε**ΦΑΡΜΟΣΜΕΝΩΝ **Μ**ΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ **Φ**ΥΣΙΚΩΝ **Ε**ΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

# "ΕΦΑΡΜΟΓΗ ΤΟΥ ΕΛΕΓΧΟΥ ΜΕΓΙΣΤΗΣ ΕΝΤΡΟΠΙΑΣ ΣΕ ΚΑΤΑΝΟΜΕΣ ΜΕ ΠΑΧΙΕΣ ΟΥΡΕΣ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

**ΕΛΕΥΘΕΡΙΑΣ ΛΑΖΑΡΟΥ**

**Επιβλέπουσα Καθηγήτρια :** Φιλία Βόντα
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Μαρτίου 2021.

| *(Υπογραφή)* | *(Υπογραφή)* | *(Υπογραφή)* |
|---|---|---|
| ..... | ..... | ..... |
| **Φιλία Βόντα**<br>Καθηγήτρια Ε.Μ.Π. | **Χρυσηίς Καρώνη**<br>Καθηγήτρια Ε.Μ.Π. | **Αλέξανδρος Καραγρηγορίου**<br>Καθηγητής Πανεπιστημίου Αιγαίου |

**Αθήνα, Μάρτιος 2021**

*(Υπογραφή)*

.....

**Ελευθερία Λαζάρου**

Διπλωματούχος Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Ε.Μ.Π.



Η παρούσα Διπλωματική Εργασία, η οποία εκπονήθηκε στα πλαίσια του Μ.Δ.Ε. : ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ, και τα λοιπά αποτελέσματά της, αποτελούν συνιδιοκτησία του Ε.Μ.Π. και της φοιτήτριας, ο καθένας από τους οποίους έχει το δικαίωμα ανεξάρτητης χρήσης και αναπαραγωγής τους (στο σύνολο ή τμηματικά) για διδακτικούς και ερευνητικούς σκοπούς, σε κάθε περίπτωση αναφέροντας τον τίτλο και την συγγραφέα και το Ε.Μ.Π. όπου εκπονήθηκε η Διπλωματική εργασία, καθώς και τον επιβλέποντα και την επιτροπή κρίσης.

**Αθήνα, Μάρτιος 2021**

# Ευχαριστίες

Σε αυτό το σημείο αισθάνομαι την ανάγκη να εκφράσω τις θερμές μου ευχαριστίες σε ορισμένους ανθρώπους, η συμβολή και η συμπαράσταση των οποίων ήταν πολύτιμη και καθοριστική στην εκπόνηση της παρούσας διπλωματικής εργασίας.

Ιδιαίτερες ευχαριστίες οφείλω, καταρχάς, στην Καθηγήτρια και επιβλέπουσα της διπλωματικής μου εργασίας κυρία Βόντα Φιλία για την επιστημονική καθοδήγηση, τις πολύτιμες συμβουλές και παρατηρήσεις επί της οργάνωσης, της δομής και του περιεχομένου της παρούσας εργασίας αλλά και για τον επιδέξιο τρόπο που επεσήμανε λάθη ή παραλήψεις. Κυρίως την ευχαριστώ για την εμπιστοσύνη που μου έδειξε εξ' αρχής αναθέτοντάς μου το συγκεκριμένο θέμα, τη συμπαράστασή της και τη θέλησή της να γίνει επιβλέπουσά μου, παρά τον περιορισμένο χρόνο της, καθώς και τη συνεχή της υποστήριξη και το αμείωτο ενδιαφέρον που έδειξε από την αρχή μέχρι το τέλος.

Επιπλέον, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους Sangyeol Lee, Φιλία Βόντα και Αλέξανδρο Καραγρηγορίου, καθώς χωρίς την επιστημονική τους έρευνα, δεν θα μου δινόταν η ευκαιρία να στοχαστώ, να διερωτηθώ και να εξετάσω σε βάθος το θέμα της διπλωματικής μου, στηριζόμενη και μελετώντας πολλά σημεία της δικιάς τους έρευνας.

Παράλληλα κρίνω απαραίτητο να εκφράσω την εκ των προτέρων εκτίμησή μου προς τα υπόλοιπα μέλη της επιτροπής για τις χρήσιμες υποδείξεις και εύστοχες παρατηρήσεις τους στο σύνολο της διπλωματικής μου εργασίας.

Κλείνοντας, ευχαριστώ ειλικρινά την οικογένειά μου για την ηθική τους στήριξη και τους ευγνωμονώ που στάθηκαν δίπλα μου, δίνοντάς μου την ελπίδα και τη δύναμη να μην τα παρατήσω και να συνεχίσω να προσπαθώ για το καλύτερο.

Αθήνα, Μάρτιος 2021

Ελευθερία Λαζάρου

*Στην μνήμη της μητέρας μου,*

# ΠΕΡΙΛΗΨΗ:

Στη στατιστική βιβλιογραφία, υπάρχει μια ευρεία κλίμακα από στατιστικούς ελέγχους που βασίζονται στην συνάρτηση εμπειρικής κατανομής και κατηγοριοποιούνται ανάλογα με το σκοπό που εξυπηρετούν, όπως είναι ο έλεγχος Pearson, Kolmogorov-Smirnov, Anderson-Darling, Cramér-von Mises, Watson, Kuiper και Lilliefors. Ωστόσο, τα τελευταία χρόνια, τα μέτρα εντροπίας είναι αρκετά διαδεδομένα στους ελέγχους καλής προσαρμογής. Το 2011 οι Lee, Vonta και Karagrigoriou πρότειναν ένα διαφορετικό μέτρο πληροφορίας που χρησιμοποιείται σαν έλεγχος καλής προσαρμογής και βασίζεται στη μεγιστοποίηση της εντροπίας, αποτελώντας ουσιαστικά μία γενίκευση της εντροπίας των Forte και Hughes. Στην εργασία αυτή παρουσιάζεται η ασυμπτωτική κατανομή της στατιστικής ελεγχοσυνάρτησης της μέγιστης εντροπίας κάτω από τη μηδενική υπόθεση και εξετάζεται μέσω προσομοιώσεων η συμπεριφορά του ελέγχου με τυχαία βάρη από την Ομοιόμορφη κατανομή $[0,1]$, και κάνοντας παράλληλα χρήση του θεωρήματος μετασχηματισμού ολοκληρώματος πιθανότητας.

Λαμβάνοντας λοιπόν υπόψη την καλή επίδοση του ελέγχου, ιδιαίτερο ερευνητικό ενδιαφέρον παρουσιάζει η εξέταση της προσαρμογής του, σε δεδομένα που προέρχονται από κατανομές με παχιές ουρές, στις οποίες η προσοχή μας επικεντρώνεται στις ουρές της κατανομής. Παίρνοντας λοιπόν δύο περιπτώσεις βαρών, σύμφωνα με τις οποίες ιδιαίτερη έμφαση δίνεται στις παρατηρήσεις που βρίσκονται στην ουρά της κατανομής, εξετάσαμε την συμπεριφορά-προσαρμογή του στατιστικού ελέγχου με βάση τη μέγιστη εντροπία. Προκειμένου να αντιμετωπιστούν μικρά προς μεσαία μεγέθη δείγματος, προσαρμόσαμε κατάλληλα τον έλεγχο, υπολογίζοντας τις κρίσιμες τιμές από την εμπειρική κατανομή της ελεγχοσυνάρτησης. Επομένως, αναλύοντας τα αποτελέσματα του μεγέθους και της ισχύος των προσομοιωμένων δεδομένων, εξετάζουμε αν η Monte Carlo μελέτη προσομοίωσης για μια πληθώρα κατανομών με παχιές ουρές μαρτυρεί μια ικανοποιητική προσαρμογή του προτεινόμενου ελέγχου, για μικρά και μεσαία μεγέθη δείγματος.

**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ : "** *έλεγχος υποθέσεων*, *μηδενική υπόθεση*, *εναλλακτική υπόθεση*, *έλεγχος καλής προσαρμογής*, *μέγεθος*, *ισχύς*, *σφάλμα τύπου I*, *σφάλμα τύπου II*, *έλεγχος Anderson-Darling*, *έλεγχος Cramér-von Mises*, θεώρημα μετασχηματισμού ολοκληρώματος πιθανότητας, *εμπειρική συνάρτηση κατανομής*, *Brownian bridge*, *έλεγχος μέγιστης εντροπίας, κατανομές με παχιές ουρές.* **"**

# ABSTRACT:

In statistical literature, there is a wide range of statistical tests based on the empirical distribution function which are categorized according to the purpose they serve, such as the Pearson test, the Kolmogorov-Smirnov test, the Anderson-Darling test, the Cramér-von Mises test, the Watson test, the Kuiper test and the Lilliefors test. However, in recent years, the measures of entropy are widely used in goodness of fit tests. In 2011, Lee, Vonta and Karagrigoriou suggested a different measure of information, based on entropy maximization and used in goodness of fit tests. In fact, it is a generalization of Forte and Hughes entropy. In that work, the asymptotic distribution of the test statistic of the maximum entropy test was presented, and the behavior of the test was examined by the use of random weights form the Uniform distribution $[0,1]$ and simultaneously by the use of the probability integral transformation theorem.

Thus, taking into consideration the good performance of the test, a research interest was raised as to how well the proposed test performs on data from distributions with heavy tails, in which, all our attention is focused on the tails of distribution. So, our purpose was to examine the maximum entropy test fit, by taking two choices of weights, according to which the observations located on the distribution tails, are of primary importance. In order to deal with small and medium sample sizes, we simulated data appropriately, calculating the critical values of the test through the empirical distribution of the test statistic. Consequently, we examined the size and the power of the test based on simulated data generated from heavy-tailed distributions and concluded that the test performs satisfactorily for small and medium sample sizes.

**KEY WORDS :** *" hypothesis testing, null hypothesis, alternative hypothesis, goodness of fit tests, size, power, type I error, type II error,* probability integral transformation theorem*, Anderson-Darling test, Cramér-von Mises test, empirical distribution function, Brownian bridge, maximum entropy test, heavy-tailed distributions"*

# ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ:

## ΚΕΦΑΛΑΙΟ 3 :

### *"ΕΛΕΓΧΟΣ ΜΕ ΒΑΣΗ ΤΗΝ ΜΕΓΙΣΤΗ ΕΝΤΡΟΠΙΑ"*



## ΒΙΒΛΙΟΓΡΑΦΙΑ :

# ΕΙΣΑΓΩΓΗ:

# "ΜΙΑ ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ ΕΛΕΓΧΟΥΣ ΥΠΟΘΕΣΕΩΝ"

## 1. ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΚΑΙ ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

Οι τεχνικές της κλασσικής στατιστικής συμπερασματολογίας, βασίζονται σε υποθέσεις που εκφράζονται σχετικά με τη φύση του πληθυσμού που μας ενδιαφέρει και πιο συγκεκριμένα σχετικά με κάποιο χαρακτηριστικό του πληθυσμού. Στην παραμετρική στατιστική υποθέτουμε ότι έχουμε πληροφορίες για τη μορφή της συνάρτησης κατανομής (μοντέλο πιθανότητας) του χαρακτηριστικού του πληθυσμού που μας ενδιαφέρει ενώ στην απαραμετρική στατιστική δεν υποθέτουμε τίποτα εκ των προτέρων για την συνάρτηση κατανομής. Οι τεχνικές που χρησιμοποιούνται σε κάθε ενότητα της στατιστικής συμπερασματολογίας είναι διαφορετικές. Στην παραμετρική στατιστική υποθέτουμε ότι η συναρτησιακή μορφή της κατανομής είναι γνωστή εκτός από κάποια παράμετρο ή παραμέτρους. Σε αυτή την ενότητα οι υποθέσεις που διατυπώνονται είναι σχετικά με την άγνωστη παράμετρο ενώ στην απαραμετρική στατιστική οι υποθέσεις που διατυπώνονται είναι σχετικά με την ίδια την κατανομή η οποία είναι άγνωστη. Για κάθε σετ υποθέσεων υπάρχουν πολλοί έλεγχοι που μπορούν να εφαρμοστούν.

Ένα σημαντικό πρόβλημα στη στατιστική, συνδέεται με την εξαγωγή πληροφοριών σχετικά με την κατανομή του πληθυσμού, από τον οποίο έχουμε πάρει το δείγμα, καθώς πολλές φορές, δεν είμαστε σε θέση να προσδιορίσουμε την κατανομή αυτή. Γενικά είναι σημαντικό στον έλεγχο υποθέσεων να τίθεται μια λογική υπόθεση για ένα χαρακτηριστικό του πληθυσμού ***[11].***Για παράδειγμα, θα θέλαμε να ξέρουμε αν ένα σετ από παρατηρήσεις $x_1, x_2, \dots x_n$ είναι συμβατό με την υπόθεση ότι τα $X_i$ i=1,..n, είναι ένα τυχαίο δείγμα από μια κανονική κατανομή, (*hypothesis of goodness of fit*). ***[16].***

Στο επίκεντρο λοιπόν της έρευνας της στατιστικής συμπερασματολογίας είναι η μορφή της κατανομής του πληθυσμού, η έλλειψη γνώσης της οποίας, δηλώνει αβεβαιότητα και επομένως την ανάγκη για στατιστική ανάλυση. Στη στατιστική συμπερασματολογία, υπάρχει μια πληθώρα κριτηρίων που ουσιαστικά στηρίζονται σε υποθέσεις για τη μορφή της κατανομής που ακολουθεί ένας πληθυσμός, από τον οποίο προκύπτει το δείγμα μας και άρα οι παρατηρήσεις. Ουσιαστικά δηλαδή, κάποιος ακολουθεί μια διαδικασία, στηριζόμενος στο δείγμα που έχει, κάνοντας έναν στατιστικό έλεγχο, για να καταλήξει αν πρέπει να απορρίψει ή όχι, μια

στατιστική υπόθεση που έχει διατυπωθεί πριν αρχίσει η συγκεκριμένη διαδικασία, και η οποία σχετίζεται με την πρόβλεψη-εικασία για το τι κατανομή ακολουθεί το τυχαίο δείγμα από τον πληθυσμό που μας ενδιαφέρει . *[13],[29],[39].*

***Έλεγχος υποθέσεων (hypothesis testing)*** λοιπόν, είναι μία μέθοδος για την αξιολόγηση ενός ισχυρισμού ή μιας υπόθεσης για ένα πληθυσμό, χρησιμοποιώντας δεδομένα που προέρχονται από ένα δείγμα του πληθυσμού αυτού. *[29].* Στην πραγματικότητα, αποτελεί μια στατιστική διαδικασία, σύμφωνα με την οποία οι ερευνητές συλλέγουνε δεδομένα από ένα δείγμα, μελετώντας με αυτό τον τρόπο, τη συμπεριφορά των δειγμάτων, με σκοπό να μάθουν περισσότερα για τα χαρακτηριστικά ενός πληθυσμού, ο οποίος τις περισσότερες φορές είναι αρκετά μεγάλος και συνεπώς δύσκολος να μελετηθεί. *[13],[29].* Ο έλεγχος υποθέσεων αποτελείται από δύο ισχυρισμούς, ο ένας ονομάζεται **μηδενική υπόθεση (null hypothesis)** και ο άλλος **εναλλακτική υπόθεση (alternative hypothesis)** και συμβολίζονται με $H_0$ και $H_1$ αντίστοιχα. *[10],[13],[16],[29],[35].*

Έστω ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ από κάποια τυχαία μεταβλητή $X$ η οποία περιγράφει κάποιο χαρακτηριστικό του πληθυσμού που μας ενδιαφέρει και ακολουθεί την κατανομή $P$ . Το τυχαίο δείγμα ακολουθεί την κατανομή $P$ που ανήκει σε μια γενική οικογένεια κατανομών $\boldsymbol{P}$. Θεωρούμε δύο υπό-οικογένειες της $\boldsymbol{P}$ την $\boldsymbol{P_0}$ και την $\boldsymbol{P_1}$, ώστε $\boldsymbol{P_0} \cap \boldsymbol{P_1} = \emptyset$.
Για την μηδενική υπόθεση έχουμε $H_0: P \in \boldsymbol{P_0}$
ενώ για την εναλλακτική υπόθεση $H_1: P \in \boldsymbol{P_1}$
όπου το $\boldsymbol{P_i} \subset \boldsymbol{P}$ για $i = 0,1$,

και επίσης $\boldsymbol{P_0} \cup \boldsymbol{P_1} = \boldsymbol{P}$ *[16], [39].*

Δηλαδή οι υποθέσεις αυτές είναι συμπληρωματικές (mutually exclusive), άρα ο συνδυασμός τους καλύπτει όλα τα πιθανά ενδεχόμενα, οπότε τα δεδομένα θα μας οδηγούν στο ποια είναι η κατάλληλη για να επιλεγεί κάθε φορά, καθώς είτε η μία είτε η άλλη θα είναι ¨αληθινή¨. Γενικά υπάρχουν οι **απλές υποθέσεις** που περιλαμβάνουν το $=$, οι **μονόπλευρες (one-sided)** που συμβολίζονται με $>$ ή $\geq$, είτε με $<$ ή $\leq$ και οι **αμφίπλευρες (two-sided)** που συμβολίζονται με $\neq$ . *[10],[13],[29],[39].*

Πάντα σαν μηδενική υπόθεση $H_0$ θέτουμε την πεποίθηση εκείνη για την οποία αμφιβάλουμε, αυτόν δηλαδή τον ισχυρισμό που αμφισβητείται και τον οποίο θέλουμε να εξετάσουμε μέσω ενός ελέγχου, για να επιβεβαιώσουμε ή όχι, μια prior πεποίθηση που έχουμε για τα δεδομένα.*[13],[35].*Σκοπός λοιπόν είναι να ελέγξουμε αν αυτή η πεποίθηση είναι σωστή ή όχι, θεωρώντας αρχικά ότι ισχύει μέχρις ότου να έχουμε επαρκή στατιστικά ευρήματα που να συνηγορούν υπέρ της απόρριψης της, έναντι της εναλλακτικής. Επομένως, μέσω της ***εναλλακτικής υπόθεσης $\boldsymbol{H_1}$***, εκφράζουμε την υποσυνείδητη πεποίθηση ότι η αρχική μας υπόθεση, δηλαδή η μηδενική, είναι λανθασμένη. *[29].*

Ο έλεγχος καλής προσαρμογής παραδείγματος χάριν που θα δούμε σε επόμενα κεφάλαια, παίρνει την εξής μορφή :

$$H_0 : \textit{τα δεδομένα προέρχονται από μια συγκεκριμένη κατανομή } F.$$

$$\textit{Vs} \qquad H_1 : \textit{τα δεδομένα δεν προέρχονται από αυτή την συγκεκριμένη κατανομή } F.$$

Σκοπός των ελέγχων καλής προσαρμογής, είναι η σύγκριση των δεδομένων με την κατανομή που έχουμε υποθέσει. Μέσω της καλής προσαρμογής ενός στατιστικού μοντέλου, εξετάζεται πόσο καλά αυτό προσαρμόζεται σε ένα σετ παρατηρήσεων, λαμβάνοντας υπόψη τη διαφορά ανάμεσα στις παρατηρούμενες τιμές και τις τιμές που προσδοκούμε κάτω από αυτό το μοντέλο. ***[23].*** Ένας στατιστικός έλεγχος λοιπόν, είναι στην ουσία μια μαθηματική τεχνική που βασίζεται στην κατανομή πιθανότητας του δείγματος, με σκοπό να φτάσει σε ένα συμπέρασμα, σχετικά με την ορθότητα της μηδενικής ή εναλλακτικής υπόθεσης. Χρησιμοποιώντας λοιπόν τη στατιστική συμπερασματολογία, η οποία περιλαμβάνει ένα σύνολο από στατιστικούς ελέγχους, αξιολογούμε τη στατιστική σημαντικότητα μιας συγκεκριμένης υπόθεσης, αποσκοπώντας στην εξαγωγή συμπερασμάτων για τον πληθυσμό. ***[10],[16],[29],[35].***

Γενικά οι παρατηρήσεις, μας δίνουν πληροφορίες για την κατανομή από την οποία που έχουν προέλθει, ή και κατευθυντήριες γραμμές για το ποια είναι η σωστή απόφαση σε κάθε έλεγχο υποθέσεων. Σε κάθε τέτοιον έλεγχο το ζητούμενο είναι να καθορίσουμε έναν κανόνα, ο οποίος για κάθε σετ τιμών των παρατηρήσεων, να καθορίζει ποια απόφαση είναι η κατάλληλη. ***[16].*** Ο αναλυτής επιθυμεί φυσικά να επιλέξει μια διαδικασία από όλες τις διαθέσιμες τεχνικές της στατιστικής συμπερασματολογίας που υπάρχουν και η πρώτη του επιλογή είναι αν θα χρησιμοποιήσει μια παραμετρική ή μια απαραμετρική διαδικασία.***[10].*** Αν μπορεί να υποτεθεί π.χ. ότι τα δεδομένα προέρχονται από κανονική κατανομή, θα πρέπει να χρησιμοποιηθούν παραμετρικοί έλεγχοι, ενώ αν δεν έχουμε πληροφορίες για την κατανομή, οι μη- παραμετρικοί έλεγχοι είναι κατάλληλοι.

Η στατιστική συνάρτηση που χρησιμοποιείται στη διαδικασία λήψης της απόφασης απόρριψης ή μη της μηδενικής υπόθεσης, δηλαδή που αποτελεί τον κανόνα όπως ειπώθηκε και πιο πάνω και που βασίζεται στο τυχαίο δείγμα, ονομάζεται ***ελεγχοσυνάρτηση (test function).***

## 2. ΚΡΙΣΙΜΟ ΣΗΜΕΙΟ-ΠΕΡΙΟΧΗ ΑΠΟΡΡΙΨΗΣ ΚΑΙ ΑΠΟΔΟΧΗΣ

Κάθε έλεγχος έχει ανάλογα με την μορφή της μηδενικής υπόθεσης άλλη ***περιοχή απόρριψης,*** όπως ονομάζεται, της μηδενικής υπόθεσης. Αν συμβολίσουμε την ελεγχοσυνάρτηση που θα χρησιμοποιηθεί στον έλεγχο με $T(X_1, X_2, \dots, X_n)$ , τότε ο κανόνας θα μπορούσε να λέει ότι απορρίπτουμε την $H_0$ αν $T(X_1, X_2, \dots, X_n) \geq c$ , σε διαφορετική περίπτωση δεν την απορρίπτουμε. Στόχος μας λοιπόν σε ένα έλεγχο είναι να βρούμε τη τιμή της σταθεράς c η οποία ονομάζεται ***κρίσιμο σημείο***. Για τον προσδιορισμό της σταθεράς, χρησιμοποιούμε το επίπεδο σημαντικότητας $\alpha$, το

οποίο καθορίζεται από τον ερευνητή και αποτελεί το ποσοστό που ανεχόμαστε μια λανθασμένη απόρριψη της μηδενικής υπόθεσης όταν αυτή ισχύει. Δηλαδή έχουμε ότι

$$P[\mathrm{T}(\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}) \geq c \text{ κάτω από την } H_0] = a \qquad \textit{[39].}$$

Βρίσκοντας λοιπόν το κρίσιμο σημείο, μπορούμε να βρούμε την ***περιοχή απόρριψης*** της μηδενικής υπόθεσης, όταν $T(X_1, X_2, \dots, X_n) \geq c$ και αντίστοιχα την ***περιοχή αποδοχής*** της μηδενικής υπόθεσης όταν $T(X_1, X_2, \dots, X_n) < c$.

Ανάλογα με τις τιμές που παίρνει το δείγμα μας η τιμή της ελεγχοσυνάρτησης μπορεί να πέσει στην περιοχή απόρριψης οπότε και απορρίπτουμε την μηδενική υπόθεση, ή μπορεί να πέσει στο χωρίο αποδοχής οπότε λέμε ότι δεν έχουμε αρκετά τεκμήρια για να απορρίψουμε την μηδενική υπόθεση υπέρ της εναλλακτικής υπόθεσης.

## 3. P-VALUE

Στη στατιστική συμπερασματολογία είναι καλό, όχι μόνο να προσδιορίσουμε αν δεν απορρίπτουμε ή απορρίπτουμε μια υπόθεση σε ένα δοσμένο επίπεδο σημαντικότητας, αλλά παράλληλα να καθορίσουμε το μικρότερο επίπεδο σημαντικότητας, δηλαδή :

$$\hat{p} = \hat{p}(x) = \inf\{a : x \in S_a\}\,, \text{ όπου } S_\alpha \text{ είναι η κρίσιμη περιοχή}$$

στο οποίο η υπόθεση θα απορριφθεί για συγκεκριμένες παρατηρήσεις του δείγματος. Αυτή η ποσότητα ονομάζεται ***p-value*** και δίνει μια ιδέα στο πως τα δεδομένα μπορούν να καθορίσουν μια απόφαση.*[16].* Δηλαδή η p-value είναι το κατάλληλο μέτρο έκφρασης της πιθανότητας, που δείχνει πόσο ισχυρές είναι οι αποδείξεις που προκύπτουν από το δείγμα εναντίον της $H_{0.}$ , αφού είναι το μικρότερο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η μηδενική υπόθεση. *[10],[29].*

Όσο πιο μικρή είναι η $p - value$ τόσο ισχυρότερες ενδείξεις εναντίον της υπόθεσης $H_0$ έχει ο ερευνητής, ή αλλιώς τόσο πιο σημαντική είναι η τιμή της στατιστικής συνάρτησης ελέγχου που δίνει το δείγμα. Τα δεδομένα σε αυτή την περίπτωση, δεν υποστηρίζουν τη μηδενική υπόθεση και άρα το αποτέλεσμα είναι στατιστικά σημαντικό υπέρ της εναλλακτικής υπόθεσης, συνεπώς η στατιστική απόφαση είναι ότι η μηδενική απόφαση απορρίπτεται. Από την άλλη πλευρά, όταν η $p - value$ έχει μεγάλη τιμή, το αποτέλεσμα από το δείγμα δεν δίνει πειστικές ενδείξεις ότι ο ισχυρισμός για τον πληθυσμό κάτω από την μηδενική υπόθεση είναι λανθασμένος και άρα η στατιστική απόφαση είναι ότι η μηδενική απόφαση δεν μπορεί να απορριφθεί. *[10].* Γενικά λέμε ότι απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας $\alpha\%$ αν η p-value είναι μικρότερη ή ίση από το επίπεδο σημαντικότητας (δηλαδή $p - value \leq \alpha$), και το αποτέλεσμα λέγεται στατιστικά σημαντικό σε επίπεδο σημαντικότητας $\alpha$ *[29].*

## 4. ΕΙΔΗ ΣΦΑΛΜΑΤΟΣ

Ο έλεγχος υποθέσεων είναι μια συμπερασματική διαδικασία, το οποίο σημαίνει ότι χρησιμοποιεί περιορισμένη πληροφορία σαν βάση για να βγάλει ένα γενικό συμπέρασμα. Ο στατιστικός έλεγχος λοιπόν με βάση το δείγμα, το οποίο παρέχει περιορισμένη πληροφορία για τον πληθυσμό, σκοπό έχει να εξάγει συμπεράσματα για αυτόν. *[13].* Επομένως, επειδή παρατηρούμε μόνο ένα δείγμα και όχι όλο τον πληθυσμό, μερικές φορές, στην απόφαση που θα πάρουμε, υπάρχει πάντα η πιθανότητα να προκύψει ένα μη σωστό συμπέρασμα. *[13],[29].* Γενικά υπάρχουν τέσσερα είδη διαφορετικών αποφάσεων που μπορούμε να πάρουμε σε έναν έλεγχο υποθέσεων τα οποία παρουσιάζονται στον παρακάτω πίνακα :

| | | **ΑΠΟΦΑΣΗ** | |
|---|---|---|---|
| | | **Δεν απορρίπτω** την μηδενική υπόθεση($\boldsymbol{H_0}$) | **Απορρίπτω** την μηδενική υπόθεση($\boldsymbol{H_0}$) |
| **ΠΡΑΓΜΑΤΙΚΗ ΚΑΤΑΣΤΑΣΗ** | Η μηδενική απόφαση($\boldsymbol{H_0}$) είναι **σωστή** | **ΣΩΣΤΗ ΑΠΟΦΑΣΗ (true negative)** 1-α | **ΣΦΑΛΜΑ ΤΥΠΟΥ Ι (false positive)** α |
| | Η μηδενική απόφαση($\boldsymbol{H_0}$) είναι **λανθασμένη** | **ΣΦΑΛΜΑ ΤΥΠΟΥ ΙΙ (false negative)** β | **ΣΩΣΤΗ ΑΠΟΦΑΣΗ (true positive)** 1-β |

Όταν αποφασίσουμε να μην απορρίψουμε μια μηδενική υπόθεση, δύο είναι τα ενδεχόμενα: είτε η υπόθεση να είναι σωστή, είτε λανθασμένη. Όμοια όταν απορρίψουμε τη μηδενική υπόθεση, τα ενδεχόμενα είναι, είτε αυτή να είναι όντως λανθασμένη είτε να είναι σωστή. Σωστή απόφαση παίρνουμε όταν δεν απορρίψουμε μια αληθινή μηδενική υπόθεση, και αντίστοιχα μια λάθος απόφαση παίρνουμε όταν δεν απορρίψουμε μια λανθασμένη μηδενική υπόθεση. *[10],[29].* ***Σφάλμα τύπου ΙΙ ( type II error)***, είναι η πιθανότητα να μην απορριφθεί η μηδενική υπόθεση, ενώ στην πραγματικότητα είναι λάθος. Ένα σφάλμα τύπου ΙΙ συμβαίνει όταν τα δειγματικά αποτελέσματα δεν είναι στην κρίσιμη περιοχή. Αντίστοιχα, ***σφάλμα τύπου Ι (type I error)*** είναι η πιθανότητα να απορριφθεί η μηδενική απόφαση, ενώ στην πραγματικότητα είναι σωστή.*[10],[13],[16],[29].* Οπότε το **σφάλμα τύπου Ι**, μπορούμε να το πούμε και λανθασμένα θετικό (**false positive**), ενώ το **σφάλμα τύπου ΙΙ** είναι λανθασμένα αρνητικό (**false negative**).

Σε ένα παραμετρικό έλεγχο υποθέσεων παραδείγματος χάρη σχετικά με την άγνωστη παράμετρο $\theta$ όπου

$$H_0 : \theta \in \theta_0 \; vs \; H_1 : \theta \in \theta_1 \text{ , τότε :}$$

- Η πιθανότητα να εμφανιστεί σφάλμα τύπου ΙΙ είναι :

$$P[\sigma\phi\acute{a}\lambda\mu a \; \tau\acute{u}\pi ou \; II] = P[a\pi o\delta o\chi\acute{\eta} \; \tau\eta\varsigma \; H_0 | \iota\sigma\chi\acute{u}ει \; \eta \; H_1] = P_\Theta[a\pi o\delta o\chi\acute{\eta} \; \tau\eta\varsigma \; H_0] =$$
$$= P_\Theta[\underline{X} \in \pi\varepsilon\rho\iota o\chi\acute{\eta} \; a\pi o\delta o\chi\acute{\eta}\varsigma] \quad \text{, όπου } \theta \in \theta_1 \text{ ,} \quad \textit{[39].}$$

- Η πιθανότητα να εμφανιστεί σφάλμα τύπου I είναι :

$$P[\sigma\phi\acute{\alpha}\lambda\mu\alpha\ \tau\acute{\upsilon}\pi\upsilon\ I] = P[\alpha\pi\acute{o}\rho\rho\iota\psi\eta\ \tau\eta\varsigma\ H_0 | \iota\sigma\chi\acute{\upsilon}\epsilon\iota\ \eta\ H_0] = P_{\Theta}[\alpha\pi\acute{o}\rho\rho\iota\psi\eta\ \tau\eta\varsigma\ H_0]$$
$$= P_{\Theta}\left[\underline{X}\ \in \kappa\rho\acute{\iota}\sigma\iota\mu\eta\ \pi\epsilon\rho\iota o\chi\acute{\eta}\right]$$

όπου θ ∈ $\theta_0$ ***[39].***

***ΠΑΡΑΤΗΡΗΣΗ :*** Η μεγαλύτερη δυνατή τιμή σφάλματος τύπου I είναι ίση με

$$sup_{\theta\in\theta_0} P_{\theta}[\underline{X} \in \kappa\rho\acute{\iota}\sigma\iota\mu\eta\ \pi\epsilon\rho\iota o\chi\acute{\eta}]$$

και λέγεται ***μέγεθος του ελέγχου (size of the test)*** ***[16], [39].***

Η πιθανότητα της απόρριψης μιας λανθασμένης μηδενικής υπόθεσης, που ουσιαστικά είναι και η σωστή απόφαση, η απόφαση δηλαδή που στοχεύουμε όταν ξεκινάμε έναν έλεγχο καθώς η μηδενική υπόθεση είναι υπό αμφισβήτηση, λέγεται ***ισχύς (power)του ελέγχου*** και ισχύει ότι *power= 1-β όπου*

*β=P(σφάλματος τύπου II).* ***[13],[29].***

Ουσιαστικά φανερώνει το ποσοστό σωστών απορρίψεων της μηδενικής υπόθεσης. Δηλαδή: $\pi(\theta) = 1 - P[\sigma\phi\acute{\alpha}\lambda\mu\alpha\ \tau\acute{\upsilon}\pi\upsilon\ II] = 1 - P_{\theta}\left[\underline{X}\ \in \pi\epsilon\rho\iota o\chi\acute{\eta}\ \alpha\pi o\delta o\chi\acute{\eta}\varsigma\right] =$

$= P_{\theta}\left[\underline{X}\ \in \kappa\rho\acute{\iota}\sigma\iota\mu\eta\ \pi\epsilon\rho\iota o\chi\acute{\eta}\right]$ , για κάθε $\theta \in \theta_1$ ***[39].***

Ο τέλειος έλεγχος είναι να μην υπάρχει κανένα είδους λάθους, δηλαδή ούτε false positive ούτε false negative, ωστόσο επειδή υπάρχει πάντα αβεβαιότητα, σε όλους τους στατιστικούς ελέγχους υπάρχει πάντα η πιθανότητα να συμβεί σφάλμα τύπου I ή II. Το επιθυμητό λοιπόν, είναι να διεξαχθεί ένα τεστ με τέτοιο τρόπο που να κρατά τις πιθανότητες και των δύο τύπων σφάλματος στο ελάχιστο.***[10],[16].*** Γενικά όμως για ένα δεδομένο μέγεθος δείγματος, όταν η πιθανότητα του ενός τύπου σφάλματος μειωθεί, η πιθανότητα να συμβεί ο άλλος τύπος αυξάνεται και συνήθως όχι με τον ίδιο ρυθμό και για αυτό το λόγο δεν είναι εφικτός ο ταυτόχρονος έλεγχός τους.***[10].*** Συνήθως λοιπόν, προσπαθούμε να ελέγξουμε την πιθανότητα σφάλματος τύπου I (ή διαφορετικά το μέγεθος του ελέγχου), διαλέγοντας ένα επίπεδο σημαντικότητας $\alpha$ $(0 < \alpha < 1)$ , όπου δηλαδή

$$sup_{\theta\in\theta_0} P_{\theta}\left[\underline{X} \in \kappa\rho\acute{\iota}\sigma\iota\mu\eta\ \pi\epsilon\rho\iota o\chi\acute{\eta}\right] = \alpha \quad \textbf{[16].}$$

και στη συνέχεια προσπαθούμε να βρούμε κανόνες-ελέγχους που σχετίζονται με την ελαχιστοποίηση της πιθανότητας του σφάλματος II ή αντίστοιχα που μεγιστοποιούν την ισχύ. Στόχος μας γενικά είναι η μεγιστοποίηση της ισχύος ενός ελέγχου.

Γενικά το επίπεδο σημαντικότητας ή αλλιώς alpha level για έναν έλεγχο υποθέσεων, εκφράζει τη πιθανότητα ένας έλεγχος να οδηγηθεί σε σφάλμα τύπου I. Συγκεκριμένα η πιθανότητα να γίνει σφάλμα τύπου I είναι ίση με το επίπεδο σημαντικότητας, άρα για την αποφυγή του σφάλματος, το $\alpha$ θα πρέπει να κρατηθεί στο μέγεθος του δυνατού, σε χαμηλά επίπεδα.***[13].*** Κατ' επέκταση για την αποφυγή του σφάλματος τύπου II, θα πρέπει να επιλέξουμε μια μεγάλη τιμή για το $\alpha$.

# ΚΕΦΑΛΑΙΟ 1:

# "ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ ΕΛΕΓΧΟΥΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ"

## 1.1. ΠΑΡΑΜΕΤΡΙΚΟΙ ΚΑΙ ΑΠΑΡΑΜΕΤΡΙΚΟΙ ΕΛΕΓΧΟΙ

Με τη διεξαγωγή ενός ελέγχου υποθέσεων επιθυμούμε να αντλήσουμε πληροφορίες από τα δεδομένα που έχουμε στη διάθεσή μας, με στόχο την εξαγωγή συμπερασμάτων σχετικά με τον πληθυσμό. Αυτό που θα μας απασχολήσει στην παρούσα εργασία, είναι ο προσδιορισμός της μορφής της κατανομής που ακολουθούν τα δεδομένα. Βασική προϋπόθεση για την επίτευξη του σκοπού αυτού λοιπόν, είναι η επιλογή των κατάλληλων ελέγχων. Υπάρχουν περιπτώσεις όπου η μορφή της κατανομής που ακολουθούν τα δεδομένα είναι πλήρως προσδιορισμένη, άρα καθορισμένες παράμετροι προσδιορίζουν αυτή την κατανομή.*[12].* Ωστόσο, υπάρχουν και περιπτώσεις, όπου λόγω μη επαρκούς πλήθους παρατηρήσεων ή σχετικής πληροφορίας, ο προσδιορισμός της κατανομής των δεδομένων δεν είναι εφικτός.*[10],[11],[40].* Στόχος μας λοιπόν είναι να ορίσουμε ελέγχους οι οποίοι μπορούν να υλοποιηθούν σε ένα ευρύ φάσμα κατανομών.*[12].*

Στη στατιστική συμπερασματολογία, ***παραμετρικοί έλεγχοι*** ονομάζονται οι έλεγχοι στους οποίους, ο προσδιορισμός της μορφής της κατανομής των παρατηρήσεων του δείγματος, άρα κατ΄επέκταση και του πληθυσμού, στηρίζεται σε συγκεκριμένες υποθέσεις, με πιο συχνή εκείνη της κανονικότητας. *[10],[11],[40].* Παραδείγματος χάριν αν η κανονική προσέγγιση θεωρηθεί λογική, τότε η ανάλυση θα πραγματοποιηθεί χρησιμοποιώντας ελέγχους που στηρίζονται στην θεωρία της κανονικότητας. Σε αντίθετη περίπτωση, συνήθως μετασχηματίζουμε τα δεδομένα, έτσι ώστε με αυτό τον τρόπο να μπορούν να θεωρηθούν κανονικά κατανεμημένα και συνεπώς να μπορούν να χρησιμοποιηθούν παραμετρικοί έλεγχοι για τα μετασχηματισμένα δεδομένα.

Γενικά, είναι εύκολα αντιληπτό ότι η εγκυρότητα των παραμετρικών ελέγχων εξαρτάται από την εγκυρότητα των υποθέσεων, στις οποίες ο έλεγχος έχει βασιστεί. Στην πράξη ωστόσο, αυστηρές υποθέσεις σπάνια μπορούν να είναι τελείως δικαιολογημένες. Μπορεί επομένως, να θεωρηθούν μη αξιόπιστοι και ασταθείς όταν εφαρμοστούν σε δεδομένα που ακολουθούν κατανομές με έστω και μικρή

σχετική απόκλιση από τη μορφή της κατανομής που έχουμε υποθέσει. *[10],[40].* Μάλιστα, συχνά υπάρχει η τάση πολλές κατανομές να τις αντιμετωπίζουμε σαν κανονικές, χρησιμοποιώντας τα θεωρήματα και τα αποτελέσματα της αντίστοιχης θεωρίας, με αποτέλεσμα την δημιουργία σοβαρών σφαλμάτων.

Κατάλληλοι για την αποφυγή τέτοιων σφαλμάτων είναι οι ***distribution-free έλεγχοι***, καθώς η μορφή της υποκείμενης κατανομής δεν παίζει σημαντικό ρόλο στην "λειτουργία" τους. *[40].* Γενικά η εγκυρότητα των συμπερασμάτων τους, δεν στηρίζεται πάνω σε συγκεκριμένη κατανομή πιθανότητας του πληθυσμού, αλλά σε ασθενείς υποθέσεις για την κατανομή των παρατηρήσεων, η οποία μπορεί να προσδιοριστεί χωρίς τη γνώση του ποια ακριβώς κατανομή ακολουθεί ο βασικός πληθυσμός. Στους μη παραμετρικούς ελέγχους συνήθως η κατανομή θεωρείται συνεχής, ενώ περιστασιακά υποθέτουμε και συμμετρία. Στις περισσότερες ωστόσο περιπτώσεις, δεν χρειάζεται να υποθέσουμε τίποτα. *[10],[11],[12],[40].*

Σκοπός μας γενικά, είναι να εφαρμόσουμε ελέγχους που να μη βασίζονται σε λεπτομερείς υποθέσεις για την κατανομή του πληθυσμού από τον οποίο έχουμε πάρει το δείγμα μας, αλλά σε μια ευρεία κατηγορία κατανομών, όπως για παράδειγμα σε όλες τις συνεχείς. Στους ***απαραμετρικούς ελέγχους***, κάνουμε ελάχιστες υποθέσεις περί του βασικού μας πληθυσμού. Δηλαδή δεν υποθέτουμε ότι η κατανομή του πληθυσμού ανήκει σε παραμετρικές οικογένειες, οι οποίες περιγράφονται με τη χρήση παραμέτρων που υπεισέρχονται στην κατανομή πιθανότητας**.** Σε γενικές γραμμές οι έλεγχοι αυτοί σχετίζονται με τη διαχείριση τυπικών στατιστικών προβλημάτων, όταν η π.χ. η υπόθεση της κανονικότητας έχει αντικατασταθεί από γενικές υποθέσεις σχετικές με τη μορφή της κατανομής.*[10], [11],[12],[15],[40].*

Στην πραγματικότητα λοιπόν, αυτοί οι δύο όροι χρησιμοποιούνται ως ισοδύναμοι, αλλά ουσιαστικά δεν είναι ταυτόσημοι, καθώς οι distribution-free έλεγχοι συνδέονται με την κατανομή της στατιστικής ελεγχοσυνάρτησης, ενώ οι απαραμετρικοί έλεγχοι με τον τύπο της υπόθεσης που θα εξεταστεί. *[10],[11],[12], [40].* Γενικά ένας απαραμετρικός έλεγχος μπορεί να είναι distribution-free ή η ελεγχοσυνάρτηση να ακολουθεί μια συγκεκριμένη κατανομή, αλλά με τις παραμέτρους της κατανομής ακαθόριστες.

Συμπερασματικά ένας απαραμετρικός έλεγχος απαιτεί μόνο πολύ γενικές υποθέσεις, ενώ ο παραμετρικός απαιτεί ισχυρές υποθέσεις για την κατανομή του πληθυσμού. *[10],[11].*

## 1.2. ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΑΠΑΡΑΜΕΤΡΙΚΩΝ ΕΛΕΓΧΩΝ

Η σφαίρα εφαρμογής των απαραμετρικών ελέγχων είναι πιο ευρεία, καθώς είναι συχνά (όχι πάντα) πιο εύκολοι να εφαρμοστούν, ακόμα και σε περιπτώσεις για τις οποίες είναι αδύνατο ή μη πρακτικό, να υποθέσουμε κανονική κατανομή. *[10],[11], [15].* Γενικά πολλοί από αυτούς μπορούν να χρησιμοποιηθούν σε κατηγορικά δεδομένα, σε σχέση με τους παραμετρικούς ελέγχους που χρειάζονται τις ίδιες τις

παρατηρήσεις και άρα το πραγματικό τους μέγεθος. ***[15],[40].*** Επομένως αφού μόνο η τάξη των δεδομένων χρειάζεται να παρατηρηθεί, όλη η διαδικασία της ανάλυσης, μπορεί να είναι λιγότερο χρονοβόρα, και το μέγεθος των δειγμάτων δε χρειάζεται να είναι μεγάλο και το επίπεδο των μετρήσεων που χρειάζεται για την ανάλυση είναι συχνά λιγότερο επεξεργασμένο. ***[10],[11].*** Ένα λογικό λοιπόν συμπέρασμα, είναι ότι οι απαραμετρικοί έλεγχοι είναι σχετικά ανεπηρέαστοι από τις απομακρυσμένες παρατηρήσεις.***[15]***

Γενικά παρόλο που με την πρώτη ματιά, πολλοί από τους απαραμετρικούς ελέγχους φαίνεται να θυσιάζουν πολλή από τη βασική πληροφορία που δίνουν τα δείγματα, έρευνες σχετικά με την αποτελεσματικότητα, αναφέρουν ότι κάτι τέτοιο δεν ισχύει. Στην πραγματικότητα, συνήθως οι απαραμετρικοί έλεγχοι είναι ελάχιστα μόνο λιγότερο αποδοτικοί σε σχέση με τους παραμετρικούς, ενώ σε κάποιες περιπτώσεις μπορεί να θεωρηθούν και επιεικώς πιο αποδοτικοί.***[15].***

## 1.3. ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ ΕΛΕΓΧΟΥΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ

Ένα σημαντικό πρόβλημα που υπάρχει στη στατιστική ανάλυση, σχετίζεται με την εξαγωγή πληροφορίας από ένα τυχαίο δείγμα σχετικά με τη μορφή του γενικού πληθυσμού από τον οποίο έχει προέλθει το δείγμα. ***[10],[11].*** Ένας ερευνητής λοιπόν, συχνά επιθυμεί να εξετάσει από ποια κατανομή πιθανότητας, η οποία ανήκει σε μια συγκεκριμένη οικογένεια κατανομών, προέρχονται τα δεδομένα που έχει στη διάθεση του. Κατάλληλο γι αυτό είναι ο έλεγχος καλής προσαρμογής. ***[14].***

Γενικά η κατανομή πιθανότητας είναι απαραίτητη ως ένα μοντέλο που περιγράφει κάποιο φαινόμενο, το οποίο μπορεί να ερευνηθεί εμπειρικά. Για *παράδειγμα*, εικάζουμε ότι ο αριθμός των εργοστασιακών ατυχημάτων που συμβαίνουν καθημερινά σε ένα συγκεκριμένο εργοστάσιο είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή Poisson. Αυτή η υπόθεση μπορεί να εξεταστεί, παρατηρώντας τον αριθμό των ατυχημάτων για κάποιες συνεχόμενες μέρες και εξετάζοντας έπειτα, αν είναι λογικό να υποθέτουμε ότι η "κρυμμένη" κατανομή είναι η Poisson. ***[10],[33].***Αυτό που μας ενδιαφέρει λοιπόν είναι να καθορίσουμε, αν ένα συγκεκριμένο μοντέλο είναι κατάλληλο ή όχι, για ένα δοσμένο τυχαίο φαινόμενο. Οι στατιστικοί έλεγχοι λοιπόν που καθορίζουν αν ένας δεδομένος μηχανισμός είναι κατάλληλος, ονομάζονται ***έλεγχοι καλής προσαρμογής (goodness-of-fit tests).*** ***[33].*** Ουσιαστικά δηλαδή, είναι έλεγχοι σχεδιασμένοι να συγκρίνουν τη συμβατότητα του δείγματος, με τον τύπο του δείγματος που θα περιμέναμε να έχουμε από την κατανομή που υποθέτουμε, ώστε να μπορούμε να δούμε αν η θεωρητική συνάρτηση κατανομής προσαρμόζεται στα δεδομένα του δείγματος. ***[10],[11],[17].***

Αν στην υπόθεση δηλώνεται μόνο κάποια γενική οικογένεια από τις κατανομές πιθανότητας, οι άγνωστες παράμετροι θα πρέπει να υπολογίζονται από τα δεδομένα, πριν εφαρμοστούν οι έλεγχοι. ***[10].*** Μια δυσκολία στον έλεγχο μιας τέτοιας υπόθεσης, είναι ότι οι περιπτώσεις των εναλλακτικών υποθέσεων τυπικά

είναι πάρα πολλές, και ο ισχυρισμός είναι αναπόφευκτα αρκετά γενικός με αποτέλεσμα η απόρριψη της μηδενικής υπόθεσης να μη δίνει κάποια πληροφορία για το ποιος είναι ο τύπος του πληθυσμού, αλλά μόνο για το τι δεν είναι. *[10],[11],[16].*Ουσιαστικά δηλαδή τα συμπεράσματα για αυτούς τους ελέγχους ασχολούνται με τη μορφή και τον τύπο του βασικού πληθυσμού και κατηγοριοποιούνται ως απαραμετρικοί, εξαιτίας του τύπου της υπόθεσης.*[10],[40].*

Η κλασσική προσέγγιση για να κάνουμε έναν έλεγχο καλής προσαρμογής, για μια μηδενική υπόθεση, σύμφωνα με την οποία, ένα δείγμα έχει μια συγκεκριμένη κατανομή πιθανότητας, είναι με το να χωρίσουμε το πεδίο τιμών των δεδομένων σε ένα πεπερασμένο αριθμό υποδιαστημάτων. Οι συχνότητες των δειγματικών τιμών-δεδομένων που πέφτουν μέσα σε κάθε περιοχή, εξετάζονται και συγκρίνονται με τις προσδοκώμενες θεωρητικές συχνότητες κάτω από συγκεκριμένη κατανομή πιθανότητας, με αποτέλεσμα αν υπάρχει σημαντική διαφορά, η μηδενική υπόθεση να απορρίπτεται. *[17],[33].*

Γενικά υπάρχουν διαθέσιμοι πολλοί έλεγχοι καλής προσαρμογής, όπως και αρκετές προϋποθέσεις που πρέπει να λάβει κανείς υπόψη, προκειμένου να επιλέξει τον κατάλληλο κάθε φορά, καθώς δεν υπάρχει κάποιος γενικός κανόνας που να μας το υποδεικνύει. *[14].* Οι πιο γνωστοί έλεγχοι που χρησιμοποιούνται ευρέως είναι ο $X^2$ του Pearson και εκείνοι που βασίζονται στην συνάρτηση εμπειρικής κατανομής, όπως είναι ο έλεγχος των Kolmogorov-Smirnov, των Anderson-Darling, και των Cramér-von Mises. Ακόμα, αρκετά δημοφιλής έλεγχος καλής προσαρμογής είναι εκείνος που βασίζεται στην εντροπία. Ο $\mathrm{X}^2$ είναι ο πιο παλιός, πολύ εύκολος στη χρήση, αλλά γενικά λιγότερο ισχυρός σε σχέση με τους ελέγχους που στηρίζονται στην εμπειρική κατανομή. *[6].*

## 1.4. Ο $\mathrm{X}^2$ ΕΛΕΓΧΟΣ ΤΟΥ PEARSON

Η ιδέα του Pearson ουσιαστικά ήταν να μειώσει το γενικό πρόβλημα του ελέγχου προσαρμογής σε ένα πολυωνυμικό πρόβλημα που στηρίζεται στην σύγκριση της κατανομής του παρατηρούμενου δείγματος με την κατανομή πιθανότητας που υποθέτουμε κάτω από την μηδενική υπόθεση. Καθορίζει λοιπόν, πόσο καλά η θεωρητική κατανομή προσαρμόζεται στην εμπειρική κατανομή. Αυτή η μείωση γενικά “πετάει” κάποια πληροφορία και γι αυτό το λόγο θεωρείται λιγότερο ισχυρός έλεγχος σε σχέση με τους υπόλοιπους. *[6],[40].*

Η θεωρητική κατανομή σε αυτόν τον έλεγχο μπορεί να είναι πλήρως ορισμένη, είτε μπορεί κάποιες παράμετροι της να είναι άγνωστες, με αποτέλεσμα να πρέπει να εκτιμηθούν από τις παρατηρήσεις μας. *[40].*

### 1.4.1. Η ΠΕΡΙΠΤΩΣΗ ΤΗΣ ΑΠΛΗΣ ΥΠΟΘΕΣΗΣ

Στον $X^2$ έλεγχο για την εξέταση μιας απλής μηδενικής υπόθεσης, είναι επιθυμητό να εξετάζουμε τη μηδενική υπόθεση όπου ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ μεγέθους $n$, προέρχεται από έναν πληθυσμό με πλήρως καθορισμένη συνάρτηση κατανομής $F(x)$, ενάντια στην εναλλακτική υπόθεση ότι δεν προέρχεται από τη συγκεκριμένη κατανομή. Τα δεδομένα του δείγματος $X_j$ χωρίζονται σε $m$ $(m \geq 2)$ διαστήματα, για παράδειγμα τα $\Delta_1, \Delta_2, \dots, \Delta_m$ , και έπειτα οι παρατηρήσεις που βρίσκονται μέσα στα διαστήματα, συγκρίνονται με τις προσδοκώμενες παρατηρήσεις του κάθε διαστήματος κάτω από τη μηδενική υπόθεση. Αν $N_i$ για $i = 1, \dots, m,$ είναι ο αριθμός των παρατηρήσεων του δείγματος $X_j, j = 1, \dots n$ που ανήκουν στο διάστημα $\Delta_i, i = 1, \dots, m,$ δηλαδή $N = (N_1, N_2, \dots, N_m)^T$ είναι ένα τυχαίο διάνυσμα από συχνότητες με $n = \sum_{i=1}^{m} N_i$ τότε τα $N_i$ ακολουθούν τη Διωνυμική κατανομή με παραμέτρους $n$ και

$$p_i = P\left(X_j \text{ "πέφτει" στο } \Delta_i\right) = \int_{\Delta_i} d\,F(x),$$

όταν η μηδενική υπόθεση είναι σωστή. ***[6],[14],[25],[26].***
Άρα $p = (p_1, p_2, \dots, p_m)^T$ είναι ένα διάνυσμα πιθανοτήτων με $\sum_{i=1}^{m} p_i = 1$.

Ωστόσο μια σημαντική παρατήρηση είναι ότι οι διαφορές $N_i - np_i$ ανάμεσα στις παρατηρούμενες και στις προσδοκώμενες συχνότητες, εκφράζουν έλλειψη προσαρμογής των δεδομένων στην $F$. Για αυτό το λόγο κιόλας ήταν αναγκαία και η εύρεση μιας συνάρτησης αυτών των διαφορών, για να χρησιμοποιηθεί σαν ένα μέτρο προσαρμογής. Για την εύρεση λοιπόν αυτής της συνάρτησης γενικά υπάρχουν τρία σημαντικά στάδια, τα οποία είναι τα εξής.

Αρχικά οι ποσότητες $N_i - np_i$ , σε μεγάλα δείγματα, έχουν προσεγγιστικά μια πολυμεταβλητή κανονική κατανομή. Στη συνέχεια αν $Y = (Y_1, Y_2, \dots, Y_p)^T$ ακολουθεί μια nonsingular p-διάστατη κανονική κατανομή $N_p(\mu, \Sigma)$, όπου $\mu$ είναι το p-διάνυσμα των μέσων τιμών και Σ είναι ένας $p \times p$ πίνακας συνδιακύμανσης του $Y$, τότε η τετραγωνική μορφή $(Y - \mu)^T \Sigma^{-1} (Y - \mu)$ που εμφανίζεται στον εκθέτη της συνάρτησης πυκνότητας, ακολουθεί τη $X_p^2$ κατανομή ως συνάρτηση του $Y$. Τέλος ο υπολογισμός δείχνει ότι αν $Y = (N_1 - np_1, N_2 - np_2, \dots, N_{m-1} - np_{m-1})^T$ , η τετραγωνική μορφή είναι ***[6],[14],[31].***

$$X^2 = X_n^2 = X_n^T X_n \equiv \sum_{i=1}^{m} \frac{(N_i - np_i)^2}{np_i} \qquad (1.1)$$

το οποίο γράφεται συχνά και

$$X^2 \equiv \sum_{i=1}^{m} \frac{(observed_i - expected_i)^2}{expected_i} \equiv \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

όπου το $O_i = N_i$ αναφέρεται ως οι παρατηρούμενες συχνότητες και το $E_i = np_i$ ως οι αναμενόμενες συχνότητες. Προσεγγιστικά ακολουθεί την $X_{m-1}^2$ κατανομή κάτω από την μηδενική υπόθεση για μεγάλα δείγματα. Δηλαδή $X^2 \xrightarrow{d} X_{m-1}^2$. Αυτό

προκύπτει από το κεντρικό οριακό θεώρημα και τις παραπάνω σχέσεις καθώς η σχέση (1.1) γίνεται:

$$X^2 \equiv \sum_{i=1}^{m} \frac{(N_i - np_i)^2}{np_i} = \frac{1}{n}\,(N-np)^T \Sigma^{-1}(N-np)$$

Βλέπουμε επίσης ότι παρόλο που η $X^2$ εξαρτάται από τις $m$ τιμές $N_1, N_2, \dots, N_m$, χάνει τον ένα βαθμό ελευθερίας, λόγω της γραμμικής σχέσης $n = \sum_{i=1}^{m} N_i$. Η παραπάνω ασυμπτωτική κατανομή γενικά είναι καλή όταν όλες οι αναμενόμενες συχνότητες είναι $E_i = np_i \geq 1$ για κάθε διάστημα $\Delta_i$, $i = 1, \dots, m$ και το πολύ το 20% από αυτές να είναι μικρότερες του 5. Η παραπάνω ελεγχοσυνάρτηση ονομάζεται ***chi-square*$(X^2)$ *ελεγχοσυνάρτηση του*** Pearson **.**

Για να κάνουμε λοιπόν τον έλεγχο καλής προσαρμογής υπολογίζουμε την ελεγχοσυνάρτηση $T = X^2$ από τη σχέση (1.1) απορρίπτοντας την μηδενική υπόθεση όταν η τιμή της $T$ είναι μεγάλη. Γενικά σε επίπεδο σημαντικότητας $\alpha$, απορρίπτουμε την μηδενική υπόθεση $H_0$ αν

$$K : T = X^2 \geq X^2_{\alpha, m-1}$$

διαφορετικά δε μπορούμε να την απορρίψουμε. ***[14],[15],[25],[31],[33],[37],[40].***

### 1.4.2. Η ΠΕΡΙΠΤΩΣΗ ΤΗΣ ΣΥΝΘΕΤΗΣ ΥΠΟΘΕΣΗΣ

Στον έλεγχο του Pearson υπάρχει περίπτωση, μία ή περισσότερες από τις παραμέτρους να είναι άγνωστες και ουσιαστικά αυτό αποτελεί μια θεμελιώδη διαφορά στην εφαρμογή του, ανάμεσα στην απλή και στην σύνθετη υπόθεση. Θέλοντας λοιπόν να εξετάσουμε μια σύνθετη υπόθεση είναι αναγκαίο να επιλέξουμε μια κατανομή $F(\cdot\,|\theta^*)$ όπου η συνάρτηση κατανομής του δείγματος $X_j$ να είναι μέλος μιας παραμετρικής οικογένειας $\{F(\cdot\,|\theta) : \theta \in \Omega\}$, όπου Ω είναι p-διάστατος παραμετρικός χώρος. Πρώτα επομένως πρέπει να εκτιμήσουμε από τα δεδομένα, τις παραμέτρους $\theta_1, \theta_2, \dots, \theta_r$ όπου $r < m$, αντικαθιστώντας τες από εκτιμητές $\tilde{\boldsymbol{\theta}} := \tilde{\theta}_n$ (συνάρτηση των $X_1, X_2, \dots, X_n$), οι οποίοι βασίζονται στα διανύσματα συχνοτήτων και στη συνέχεια να εξετάσουμε την προσαρμογή της κατανομής $F\left(\cdot\,\middle|\tilde{\theta}_n\right)$. Επομένως οι εκτιμημένες πιθανότητες γίνονται

$$p_i\left(\tilde{\theta}_n\right) = \int dF\left(x\middle|\tilde{\theta}_n\right) \;\; \textit{ορισμένες στα διαστήματα } \Delta_i$$

Συνεπώς ο στατιστικός έλεγχος των Pearson-Fisher γίνεται

$$X^2\left(\tilde{\boldsymbol{\theta}}\right) \equiv \sum_{i=1}^{m} \frac{\left(N_i - np_i\left(\tilde{\boldsymbol{\theta}}\right)\right)^2}{np_i\left(\tilde{\boldsymbol{\theta}}\right)} \qquad (1.2)$$

Ο Pearson πίστευε ότι αν ο $\tilde{\boldsymbol{\theta}}$ ήταν συνεπής εκτιμητής του $\theta$, η ασυμπτωτική κατανομή του παραπάνω στατιστικού ελέγχου $X^2$ θα συνέπιπτε με την κατανομή του στατιστικού ελέγχου $X^2$,που αναφέρθηκε στην ***ενότητα 1.4.1***. Ωστόσο το εκτιμημένο $\theta$ αλλάζει την κατανομή του $X^2$ συνεπώς ο Fisher ισχυρίστηκε ότι η ασυμπτωτική κατανομή του στατιστικού ελέγχου $X^2(\tilde{\theta}_n)$ δεν είναι η $X^2_{m-1}$ αλλά

εξαρτάται από τη μέθοδο εκτίμησης που χρησιμοποιείται για τον υπολογισμό ενός εκτιμητή $\theta^*$ για την άγνωστη πραγματική τιμή του $\theta$. ***[14],[25].***

Αρχικά η μέθοδος που προτάθηκε ως η καλύτερη, είναι η μέθοδος εκτίμησης της μέγιστης πιθανοφάνειας για διακριτά μοντέλα που βασίζεται στις συχνότητες $N_i$, ή ισοδύναμα ο εκτιμητής της μέγιστης πιθανοφάνειας που βασίζεται σε κατηγοριοποιημένα δεδομένα. Ο εκτιμητής μέγιστης πιθανοφάνειας του θ (εκτιμητής $\hat{\theta}_n$) είναι η λύση της εξίσωσης :

$$\sum_{i=1}^{m} \frac{N_i}{p_i(\theta)} \frac{\vartheta p_i(\theta)}{\vartheta \theta_k} = 0 \text{ , όπου } k = 1, \dots, r$$

που προκύπτει από την παράγωγο του λογαρίθμου της πολυωνυμικής συνάρτησης πιθανοφάνειας, δίνοντας έτσι τη δυνατότητα στον Fisher να διακρίνει ότι ο στατιστικός έλεγχος του λόγου των πιθανοφανειών (likelihood ratio statistical test)

$$G^2 \equiv 2 \sum_{i=1}^{m} N_i log \frac{N_i}{np_i}$$

είναι ασυμπτωτικά ίσος με τον $X^2$. Επιπρόσθετα παρατήρησε ότι ένας εκτιμητής ασυμπτωτικά ισοδύναμος με τα κατηγοριοποιημένα δεδομένα μέγιστης πιθανοφάνειας μπορεί να προκύψει επιλέγοντας $\theta$ που να ελαχιστοποιεί το $X^2(\theta)$ για τα παρατηρούμενα $N_i$. Αυτός ο minimum $X^2$ εκτιμητής $\tilde{\theta}_n$, είναι η λύση της εξίσωσης :

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \{\frac{N_i}{p_i(\theta)}\}^{\mathbf{2}} \frac{\vartheta p_i(\theta)}{\vartheta \theta_k} = 0 \text{ , όπου } k = 1, \dots, r$$

Τέλος ο Neyman παρατήρησε ότι άλλος εκτιμητής ασυμπτωτικά ισοδύναμος με τον $\tilde{\theta}_n$ μπορεί να βρεθεί ελαχιστοποιώντας το τροποποιημένο $X^2$έλεγχο

$$X_m^2 = \sum_{i=1}^{m} \frac{(N_i - np_i(\theta))^2}{N_i}$$

Αυτός ο minimum τροποποιημένος $X^2$ εκτιμητής, $\bar{\theta}_n$ είναι η λύση της εξίσωσης :

$$\sum_{i=1}^{m} \frac{p_i(\theta)}{N_i} \frac{\vartheta p_i(\theta)}{\vartheta \theta_k} = 0 \text{ , όπου } k = 1, \dots, r$$

***[6],[14],[16],[32].***

Επομένως ουσιαστικά έχουμε τρία είδη εκτιμητών για κατηγοριοποιημένα δεδομένα, προκειμένου να κατασκευάσουμε μια ακολουθία $\{\theta_n^*\}$ εκτιμητών για το $\theta$. Τον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\theta}_n$, τον minimum $X^2$ εκτιμητή $\tilde{\theta}_n$ και τον minimum τροποποιημένο $X^2$ $\bar{\theta}_n$. Αυτοί οι εκτιμητές κάτω από γενικές συνθήκες, έχουν τις ίδιες ασυμπτωτικές ιδιότητες, συγκεκριμένα το $X^2(\theta_n^*)$ έχει την ίδια οριακή κατανομή για $\theta_n^* = \tilde{\theta}_n, \hat{\theta}_n$ ή $\bar{\theta}_n$ .Επομένως θα λέμε ότι τέτοιες ακολουθίες εκτιμητών είναι ασυμπτωτικά ισοδύναμες και θα χρησιμοποιούμε τον συμβολισμό $\tilde{\boldsymbol{\theta}} := \tilde{\theta}_n$ όταν αναφερόμαστε σε κάποιαν από αυτές. ***[14],[25].***

Όταν το δείγμα $n$ είναι μεγάλο τότε η ελεγχοσυνάρτηση $T = X^2$ προσεγγιστικά θα ακολουθεί την κατανομή $X^2$ με $m - r - 1$ βαθμούς ελευθερίας, κάτω από τη μηδενική υπόθεση, όπου $r$ είναι ο αριθμός των παραμέτρων που εκτιμούνται από το δείγμα. Αφαιρούμε δηλαδή επιπλέον το πλήθος των παραμέτρων που εκτιμήθηκαν. Δηλαδή απορρίπτουμε την $H_0$ όταν

$$K : T = X^2 \geq X^2_{a,m-r-1}$$

σε διαφορετική περίπτωση δεν την απορρίπτουμε. ***[37],[40].***

## 1.5. ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ: ΑΠΛΗ ΜΗΔΕΝΙΚΗ ΥΠΟΘΕΣΗ

Όπως είδαμε και στην ***ενότητα 1.4.1*** το πρόβλημα του ελέγχου καλής προσαρμογής μιας κατανομής, με μηδενική υπόθεση $H_0: F = F_0$, μπορεί να αντιμετωπιστεί χωρίζοντας το εύρος των δεδομένων σε ξένα διαστήματα και εξετάζοντας την απλή υπόθεση

$$H_0 : p = p^0 \qquad (1.3)$$

όπου $p$ το διάνυσμα των παραμέτρων μιας πολυωνυμικής κατανομής.

Αν λοιπόν ορίσουμε $P = \{E_i\}\ για\ i = 1, \dots, m$ τη διαμέριση των πραγματικών αριθμών σε m διαστήματα καθώς και $p = (p_1, p_2, \dots, p_m)^T$, $p^0 = (p_1^0, p_2^0, \dots., p_m^0)^T$ τις πραγματικές και υποθετικές πιθανότητες των διαστημάτων $\mathrm{E_i}$ για $\mathrm{i} = 1, \dots, \mathrm{m}$ αντίστοιχα, έχουμε $p_i = P_F(E_i)$ για $\mathrm{i} = 1, \dots, \mathrm{m}$ και

$p_i^0 = P_{F_0}(E_i) = \int dF_0$ ορισμένο στα διαστήματα $E_i$ για $i = 1, \dots m$.

Υποθέτουμε ότι το $Y_1, Y_2, \dots, Y_n$ είναι τυχαίο δείγμα από την κατανομή $F$, η οποία είναι τελείως γνωστή. Έστω

$$N_i = \sum_{j=1}^{n} I_{E_i}(Y_j), \quad \text{όπου } I_{E_i}(Y_j) = \begin{cases} 1 & \text{αν } Y_j \in E_i \\ 0 & \text{διαφορετικά} \end{cases}$$

και $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_m)^T$ με $\hat{p}_i = \frac{N_i}{n}$ για $i = 1, \dots., m$ να είναι η απόλυτη και η σχετική συχνότητα στα διαστήματα αντίστοιχα. ***[25],[32].***

Για <u>*παράδειγμα*</u>, υποθέτουμε ότι έχουμε ένα τυχαίο δείγμα με n ανθρώπους (για παράδειγμα 100 γυναίκες που δουλεύουν για μια μεγάλη επιχείρηση). Για κάθε εργαζόμενο μετράμε ένα συγκεκριμένο χαρακτηριστικό, ο οποίος θα πρέπει να έχει μόνο ένα από τα m πιθανά αποτελέσματα (για παράδειγμα ταξινόμηση στη δουλειά: διοικητικό στέλεχος, μάνατζερ, τεχνικός υπάλληλος, υπάλληλος γραφείου, ταμίας κτλ). Δεδομένου ενός μοντέλου συμπεριφοράς του πληθυσμού (για παράδειγμα η κατανομή των γυναικών σε πέντε κατηγορίες εργασίας είναι παρόμοιες με αυτή που έχει ήδη υπολογιστεί για τους άνδρες εργαζόμενους ή για τις εργαζόμενες γυναίκες, τα τελευταία πέντε προηγούμενα χρόνια), μπορούμε να υπολογίσουμε πόσα άτομα, εμείς προσδοκούμε για κάθε κατηγορία. Η εφαρμογή του μοντέλου μπορεί να υπολογιστεί, συγκρίνοντας τις *προσδοκώμενες συχνότητες*

*(expected frequencies)* για κάθε κατηγορία, με τις *παρατηρούμενες συχνότητες (observed frequencies)* από το δείγμα μας. ***[32].***

Για την σύγκριση λοιπόν αυτών των συχνοτήτων, χρησιμοποιούμε την οικογένεια των στατιστικών ελέγχων απόκλισης-ισχύος ***(power-divergence test statistics)***, που δίνεται από τον παρακάτω τύπο των Cressie και Read

$$T_n^{\lambda}(\hat{p}, p^0) = \frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{m} \hat{p}_i \left( \left( \frac{\hat{p}_i}{p_i^0} \right)^{\lambda} - 1 \right) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{m} N_i \left( \left( \frac{N_i}{np_i^0} \right)^{\lambda} - 1 \right), \quad (1.4)$$

και άρα παίρνουμε την σχέση :

$$\frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{m} observed_i \left[ \left( \frac{observed_i}{expected_i} \right)^{\lambda} - 1 \right]$$

όπου λ είναι μια παράμετρος με πραγματικές τιμές, οι οποίες επιλέγονται από τον αναλυτή. Οι στατιστικοί έλεγχοι $T_n^0(\hat{p}, p^0)$ και $T_n^{-1}(\hat{p}, p^0)$ ορίζονται ως τα όρια του $T_n^{\lambda}(\hat{p}, p^0)$ όταν $\lambda \to 0$ και $\lambda \to -1$, αντίστοιχα. Γενικά η παραπάνω σχέση της οικογένειας της ελεγχοσυνάρτησης της απόκλισης-ισχύος υπολογίζεται για όλες τις επιλογές του $\lambda$ στο διάστημα $-\infty < \lambda < \infty$, ενώ με την επιλογή συγκεκριμένων τιμών του $\lambda$ οδηγούμαστε σε γνωστούς στατιστικούς ελέγχους. Αν ωστόσο η προσδοκώμενη και η παρατηρούμενη συχνότητα ταιριάζουν απόλυτα για κάθε πιθανό αποτέλεσμα, η παραπάνω σχέση είναι μηδέν (για κάθε επιλογή του λ). Σε κάθε άλλη περίπτωση είναι θετική και γίνεται μεγαλύτερη καθώς οι προσδοκώμενες και οι παρατηρούμενες συχνότητες αποκλίνουν.***[4],[6],[25],[27], [28],[32].***

Ο στατιστικός έλεγχος του Pearson είναι μια ειδική περίπτωση του στατιστικού ελέγχου του power-divergence (αν βάλουμε $\lambda = 1$) :
Εξετάζοντας λοιπόν την υπόθεση $(1.3)$ έχουμε :

$$X^2 \equiv \sum_{i=1}^{m} \frac{(N_i - np_i^0)^2}{np_i^0} \qquad (1.5)$$

Καθώς και το στατιστικό τεστ του λόγου των πιθανοφανειών, $G^2$ είναι μια ειδική περίπτωσή του (αν έχουμε το όριο $\lambda \to 0$).
Επομένως για τον έλεγχο της υπόθεσης $(1.3)$ έχουμε

$$G^2 \equiv 2 \sum_{i=1}^{m} observed_i log \frac{observed_i}{expected_i} \equiv 2 \sum_{i=1}^{m} N_i log \frac{N_i}{np_i^0}$$

Άλλοι στατιστικοί έλεγχοι που προκύπτουν, για διάφορες τιμές του $\lambda$ είναι:

- Για $\lambda = -1$ ο τροποποιημένος στατιστικός έλεγχος του λόγου των πιθανοφανειών

$$T_n^{-1}(\hat{p}, p^0) = 2n \sum_{i=1}^{m} p_i^0 log \frac{p_i^0}{\hat{p}_i} = 2 \sum_{i=1}^{m} N_i log \frac{np_i^0}{N_i}$$

- Για $\lambda = -2$ ο τροποποιημένος $X^2$ στατιστικός έλεγχος

$$T_n^{-2}(\hat{p}, p^0) = n \sum_{i=1}^{m} \frac{(p_i^0 - \hat{p}_i)^2}{\hat{p}_i} = \sum_{i=1}^{m} \frac{(np_i^0 - N_i)^2}{N_i}$$

- Για $\lambda = -1/2$ ο στατιστικός έλεγχος Freeman-Tukey

$$T_n^{-1/2}(\hat{p}, p^0) = 8n \left(1 - \sum_{i=1}^{m} \sqrt{p_i^0 \hat{p}_i}\right) = 8n \left(1 - \sum_{i=1}^{m} \sqrt{\frac{p_i^0 N_i}{n}}\right)$$

- Για $\lambda = 2/3$ ο στατιστικός έλεγχος Cressie-Read

$$T_n^{2/3}(\hat{p}, p^0) = \frac{9}{5} n \left(\sum_{i=1}^{m} \hat{p}_i \left(\frac{\hat{p}_i}{p_i^0}\right)^{2/3} - 1\right)$$

Ωστόσο είναι εφικτό να θεωρήσουμε μια πιο γενική οικογένεια στατιστικών ελέγχων, για να εξετάσουμε την υπόθεση **(1.3)**, που περιέχει την προαναφερθείσα οικογένεια, άρα και την ελεγχοσυνάρτηση **(1.4)** ως μια ειδική περίπτωση, η οποία είναι το στατιστικό τεστ ϕ -divergence, που ορίζεται από την παρακάτω σχέση :

$$T_n^{\phi}(\hat{p}, p^0) = \frac{2n}{\phi''(1)} D_{\phi}(\hat{p}, p^0) = \frac{2n}{\phi''(1)} \sum_{i=1}^{m} p_i^0 \, \phi\left(\frac{\hat{p}_i}{p_i^0}\right), \quad \phi \in \Phi^*$$

όπου $\Phi^*$ είναι ο χώρος που περιλαμβάνει όλες τις κυρτές συναρτήσεις $\phi(x)$, $x \geq 0$, για $x = 1, \phi(1) = 0$, ενώ για $x = 0$, $0\phi\left(\frac{0}{0}\right) = 0$ και $0\phi\left(\frac{u}{0}\right) = \lim_{u \to \infty} \frac{\phi(u)}{u}, u > 0$. Η $\phi(x)$ είναι δύο φορές συνεχώς παραγωγίσιμη για $x > 0$ με $\phi''(1) \neq 0$. ***[4],[25],[27],[28].***

### 1.5.1. ΑΣΥΜΠΤΩΤΙΚΗ ΚΑΤΑΝΟΜΗ ΕΛΕΓΧΟΣΥΝΑΡΤΗΣΗΣ ΚΑΤΩ ΑΠΟ ΤΗΝ ΑΠΛΗ ΜΗΔΕΝΙΚΗ ΥΠΟΘΕΣΗ

Συνοπτικά λοιπόν όπως είδαμε, ο Pearson (1900) λοιπόν απέδειξε ότι $X^2 \xrightarrow{L} X^2_{m-1}$ καθώς $n \to \infty$ με το $X^2$ να δίνεται από τη σχέση **(1.5)**. Στη συνέχεια, οι Cressie και Read (1984), βρήκαν την ασυμπτωτική κατανομή του στατιστικού ελέγχου της απόκλισης-ισχύος, $T_n^{\lambda}(\hat{p}, p^0)$, {το οποίο για $\lambda = 1$ συμπίπτει με τον στατιστικό έλεγχο $X^2$}, κάτω από την υπόθεση $H_0: p = p^0$ για κάθε $\lambda \in R$ και απέδειξαν συγκεκριμένα ότι $T_n^{\lambda}(\hat{p}, p^0) \xrightarrow{L} X^2_{m-1}$ καθώς $n \to \infty$, ενώ έπειτα οι Zografos et al. (1990) επέκτειναν το αποτέλεσμα για την οικογένεια $T_n^{\phi}(\hat{p}, p^0)$ κάτω από την υπόθεση $H_0: p = p^0$ και απέδειξαν ότι $T_n^{\phi}(\hat{p}, p^0) \xrightarrow{L} X^2_{m-1}$ *καθώς* $n \to \infty$ για κάθε $\phi \in \Phi^*$. ***[25].***

Σε αυτή την ενότητα λοιπόν θα δούμε τα συμπεράσματα που προκύπτουν από αυτή την επέκταση. Πιο συγκεκριμένα, θα δούμε την ασυμπτωτική κατανομή

του $T_n^{\Phi}(\hat{p}, p^0)$ κάτω από την $H_0: p = p^0$, κάτω από την εναλλακτική υπόθεση $H_1: p = p^* \neq p^0$ και τέλος κάτω από συναφείς(contiguous) εναλλακτικές υποθέσεις.

Γενικά οι ***αποδείξεις*** όλων των θεωρημάτων που ακολουθούν στις ***ενότητες 1.5.1*** και ***1.6.1*** είναι αρκετά μεγάλες και ξεφεύγουν από τα πλαίσια της παρούσας εργασίας, για αυτό το λόγο παραλείπονται. Κάποιος μπορεί να τις βρει ανατρέχοντας σε σχετική βιβλιογραφία (παραδείγματος χάριν, *Pardo L., (2006)).*

***ΘΕΩΡΗΜΑ 1.5.1 :***

*Κάτω από τη μηδενική υπόθεση $H_0: p = p^0 = (p_1^0, p_2^0, \dots, p_m^0)^T$, η ασυμπωματική κατανομή του στατιστικού ελέγχου ϕ-divergence $T_n^{\Phi}(\hat{p}, p^0)$, είναι η $X^2$ με $m-1$ βαθμούς ελευθερίας.*

***ΠΑΡΑΤΗΡΗΣΕΙΣ:***

- Αν θεωρήσουμε τη συνάρτηση $\varphi(x) = x\Phi(x^{-1})$ με $\Phi \in \Phi^*$ τότε $\varphi \in \Phi^*$, οπότε από το *θεώρημα 1.5.1* έχουμε ότι $T_n^{\varphi}(\hat{p}, p^0) \xrightarrow{L} X_{m-1}^2$, καθώς $n \to \infty$. Αν επιπρόσθετα λάβουμε υπόψη ότι $\varphi''(1) = \Phi''(1)$ έχουμε ότι

$$T_n^{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} D_{\varphi}(\hat{p}, p^0) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{m} p_i^0 \, \varphi\left(\frac{\hat{p}_i}{p_i^0}\right) =$$

$$= \frac{2n}{\Phi''(1)} \sum_{i=1}^{m} p_i^0 \frac{\hat{p}_i}{p_i^0} \Phi\left(\frac{p_i^0}{\hat{p}_i}\right) = T_n^{\Phi}(p^0, \hat{p})$$

Επομένως συμπεραίνουμε ότι η $\mathrm{X}^2$ κατανομή με $m-1$ βαθμούς ελευθερίας, είναι η ασυμπτωτική κατανομή για τον στατιστικό έλεγχο ϕ-divergence $T_n^{\Phi}(p^0, \hat{p})$, κάτω από τη μηδενική υπόθεση $H_0: \mathrm{p} = \mathrm{p}^0$. ***[25].***

- Αν πάρουμε την περίπτωση της απόκλισης των Kullback-Leibler, έχουμε ότι

$$T_n^0(\hat{p}, p^0) = 2nD_{Kull}(\hat{p}, p^0) \xrightarrow{L} X_{m-1}^2 \; καθώς \; n \to \infty$$

το οποίο είναι ο ***έλεγχος του λόγου των πιθανοφανειών***. Και ακόμα έχουμε

$$T_n^0(p^0, \hat{p}) = 2nD_{Kull}(p^0, \hat{p}) \xrightarrow{L} X_{m-1}^2 \; καθώς \; n \to \infty$$

το οποίο είναι ο ***τροποποιημένος έλεγχος του λόγου των πιθανοφανειών*** ***[25].***

Γενικά αν το δείγμα μας είναι αρκετά μεγάλο, στηριζόμενοι στο *θεώρημα 1.5.1* μπορούμε να χρησιμοποιήσουμε το ποσοστό $100(1-\alpha)$, της $\mathrm{X}^2$ κατανομής με $\mathrm{m}-1$ βαθμούς ελευθερίας, ορισμένο από την εξίσωση $Pr\left(X_{m-1}^2 \geq X_{m-1,a}^2\right) = a$ για να ορίσουμε τον παρακάτω ***κανόνα απόφασης***

***" Θα λέμε ότι απορρίπτουμε την μηδενική υπόθεση $H_0$, σε επίπεδο σημαντικότητας α, αν***

$$\boldsymbol{T_n^{\Phi}(\hat{p}, p^0) > X_{m-1,\alpha}^2} \; \left(\text{ή } \boldsymbol{T_n^{\Phi}(p^0, \hat{p}) > X_{m-1,\alpha}^2}\right). \text{"} \qquad (1.6)$$

το οποίο αποτελεί τον έλεγχο καλής προσαρμογής που βασίζεται στη ϕ-divergence.

Στο επόμενο θεώρημα παρουσιάζεται μια προσέγγιση της συνάρτησης ισχύος για τον έλεγχο της διαδικασίας της σχέσης $(1.6)$.

***ΘΕΩΡΗΜΑ 1.5.2 :***

*Αν* $p^* = (p_1^*, p_2^*, \dots, p_m^*)^T$ *μια κατανομή πιθανότητας με* $p^* \neq p^0$. *Η ισχύς του ελέγχου με τον κανόνα απόφασης που δίνεται από τη σχέση* $(1.6)$ *στο* $p^* = (p_1^*, p_2^*, \dots, p_m^*)^T$ *είναι η εξής :*

$$\beta_{n,\phi}(p_1^*, p_2^*, \dots, p_m^*) = 1 - \boldsymbol{\Phi}_n\left(\frac{1}{\sigma_1(p^*)}\left(\frac{\phi''(1)}{2\sqrt{n}}X^2_{m-1,\alpha} - \sqrt{n}D_\phi(p^*, p^0)\right)\right)$$

*όπου* $\boldsymbol{\Phi}_n$ *τείνει ομοιόμορφα στην συνάρτηση της τυπικής κανονικής κατανομής* $\boldsymbol{\Phi}(x)$ *και*

$$\sigma_1{}^2(p^*) = \sum_{i=1}^{m} p_i^*\left(\phi'\left(\frac{p_i^*}{p_i^0}\right)\right)^2 - \left(\sum_{i=1}^{m} p_i^*\,\phi'\left(\frac{p_i^*}{p_i^0}\right)\right)^2 \qquad (1.7) \quad \blacksquare$$

Γενικά από την αποδοτικότητα του Pitman (***Pitman asymptotic relative efficiency***) έχουμε ότι για ένα διάνυσμα πιθανότητας $p^* \neq p^0$, ισχύει ότι:

$$\lim_{n\to\infty} \beta_{n,\phi}(p^*) = \Pr\left(\, T_n^{\phi}(\hat{p}, p^0) > X^2_{m-1,\alpha} \;/H_1 : p = p^*\right) = 1$$

Με σκοπό λοιπόν να έχουμε μια πιο μεγάλη ασυμπτωτική ισχύ, με τιμή ορίου δηλαδή, μικρότερη από 1, ο Cochran πρότεινε τη χρήση ενός σετ από τοπικές συναφείς εναλλακτικές υποθέσεις οι οποίες καθώς το $n$ αυξάνεται, ορίζονται ως: $p_n \equiv p^0 + \frac{d}{\sqrt{n}}$, όπου $d \equiv (d_1, d_2, \dots, d_m)^T$ είναι ένα καθορισμένο $m \times 1$ διάνυσμα τέτοιο ώστε $\sum_{i=1}^{m} d_j = 0$.

Καθώς λοιπόν $n \to \infty$, η ακολουθία των διανυσμάτων πιθανότητας $\{p_n\}$ $με$ $n \in N$ συγκλίνει στο διάνυσμα πιθανότητας $p^0$ της μηδενικής υπόθεσης με ρυθμό $O\left(n^{-1/2}\right)$. Επομένως λέμε ότι :

$$H_{1,n} : p = p_n = p^0 + \frac{d}{\sqrt{n}} \qquad (1.8)$$

είναι μια αλληλουχία από συναφείς εναλλακτικές υποθέσεις, συναφείς με τη μηδενική υπόθεση $p^0$. Το ενδιαφέρον μας λοιπόν επικεντρώνεται στο να μελετήσουμε την ασυμπτωτική συμπεριφορά του ελέγχου ισχύος, κάτω από συναφείς εναλλακτικές υποθέσεις. Δηλαδή: ***[4],[5],[25],[32].***

$$\beta_{n,\phi}(p_n) = \Pr\left(\, T_n^{\phi}(\hat{p}, p^0) > X^2_{m-1,\alpha} \;/\, H_{1,n} : p = p_n\right) (1.9)$$

***ΠΑΡΑΤΗΡΗΣΗ:***

- Ενδιαφέρον παρουσιάζει το γεγονός ότι αν θεωρήσουμε ένα σημείο $p^* \neq p^0$, μπορούμε να γράψουμε $p^* = p^0 + \frac{1}{\sqrt{n}}(\sqrt{n}(p^* - p^0))$ και αν ορίσουμε

$p_n = p^0 + \frac{d}{\sqrt{n}}$ με $d = \sqrt{n}(p^* - p^0)$, μπορούμε να πάρουμε μια προσέγγιση της συνάρτησης ισχύος στο σημείο $p^*$, χρησιμοποιώντας την σχέση $(1.9)$. ***[5],[25].***

***ΘΕΩΡΗΜΑ 1.5.3 :***

*Η ασυμπτωτική κατανομή του στατιστικού ελέγχου* ϕ*-divergence* $T_n^{\Phi}(\hat{p}, p^0)$ *κάτω από τις συναφείς εναλλακτικές υποθέσεις* $(1.8)$ *είναι η noncentral* $X^2$ *κατανομή με* $m-1$ *βαθμούς ελευθερίας και με noncentrality παράμετρο* $\delta$ *που δίνεται από τον τύπο*

$$\delta = d^T diag((p^0)^{-1})d \quad \blacksquare$$

Σε αυτή την περίπτωση οπότε, θα έχουμε ότι

$$\lim_{n\to\infty} \beta_{n,\phi}(p_n) = 1 - G_{X^2_{m-1}(\delta)}(X^2_{m-1,\alpha})$$

όπου η $G_{X^2_{m-1}(\delta)}$ είναι η συνάρτηση κατανομής μιας noncentral $X^2$ τυχαίας μεταβλητής με $m-1$ βαθμούς ελευθερίας, η οποία δίνει την κατάλληλη προσέγγιση της συνάρτησης ισχύος του στατιστικού ελέγχου $X^2$ καθώς το $n \to \infty$. ***[25],[32].***

Μια διαφορετική ασυμπτωτική προσέγγιση έχει την επόμενη έκφραση.

$$\beta_{n,\phi}(p^*) \approx 1 - \mathbf{\Phi}\left(\frac{1}{\sigma_1(p^*)}\left(\frac{\phi''(1)}{2\sqrt{n}}X^2_{m-1,\alpha} - \sqrt{n}D_\phi(p^*, p^0)\right)\right)$$

όπου $\sigma_1(p^*)$ δίνεται από τη σχέση $(1.7)$ και $\mathbf{\Phi}$ η συνάρτηση τυπικής κανονικής κατανομής. ***[5],[25].***

## 1.6. ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ: ΣΥΝΘΕΤΗ ΜΗΔΕΝΙΚΗ ΥΠΟΘΕΣΗ

Αν έχουμε να εξετάσουμε μια σύνθετη υπόθεση

$$H_0 : F = F_\theta \qquad (1.10)$$

όπου η συνάρτηση κατανομής F είναι μέλος μιας παραμετρικής οικογένειας $\{F_\theta\}_{\theta\in\Theta}$, όπου το $\Theta$ είναι ένα ανοιχτό υποσύνολο του $R^{m_0}$. Όπως είδαμε και στην ***ενότητα 1.5***, μια προσέγγιση σε αυτό το πρόβλημα, είναι να θεωρήσουμε ένα διακριτό στατιστικό μοντέλο που να σχετίζεται με το γνήσιο μας μοντέλο. Για να το κάνουμε αυτό, θεωρούμε τη διαμέριση $P = \{E_i\}$ *για* $i = 1, \dots, m$ του αρχικού μας δειγματικού χώρου. Τώρα οι πιθανότητες των διαστημάτων της διαμέρισης, $E_i$ για i $= 1, \dots,$ m, εξαρτούνται από την άγνωστη παράμετρο $\theta$, δηλαδή

$$p_i(\theta) = P_\theta(E_i) = \int dF_\theta \quad \text{στα διαστήματα } \mathrm{E_i} \text{ για i} = 1, \dots, \mathrm{m}$$

Επομένως τώρα η υπόθεση $(1.10)$ μπορεί να εξεταστεί από την υπόθεση

$$H_0 : p = p(\theta_0) \in T \text{ για κάποιο άγνωστο } \theta_0 \in \Theta \qquad (1.11)$$
$$VS \qquad H_1 : p \in \Delta_m - T$$

όπου $T = \{p(\theta) = (p_1(\theta), p_2(\theta), \dots, p_m(\theta))^T \in \Delta_m : \theta \in \Theta\}$,

με ανοιχτό $\Theta \subset R^{m_0}$ και $m_0 < m-1$. ***[25],[27].***

Χρησιμοποιώντας λοιπόν έναν εκτιμητή $\tilde{\theta}$ για την εκτίμηση του $\theta$, εξετάζουμε την υπόθεση $(1.11)$ , βασιζόμενοι στον στατιστικό έλεγχο $X^2$ της σχέσης $(1.2)$.

Αντίστοιχα με τον έλεγχο καλής προσαρμογής για απλές υποθέσεις, θα χρησιμοποιούμε την $\phi$-οικογένεια στατιστικών ελέγχων, για τον έλεγχο της υπόθεσης $(1.11)$, εξετάζοντας την ασυμπτωτική τους συμπεριφορά κάτω από διάφορες καταστάσεις. Θα στηριχτούμε επομένως, στον παρακάτω τύπο :

$$T_n^{\phi_1}(\hat{\theta}_{\phi_2}) = \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\hat{p}, p(\hat{\theta}_{\phi_2})\right)$$

όπου $\phi_1$ και $\phi_2 \in \Phi^*$ είναι δύο φορές συνεχώς παραγωγίσιμες για $x>0$ με δεύτερες παραγώγους να είναι $\phi_1''(1) \neq 0$ και $\phi_2''(1) \neq 0$. ***[4],[5],[25],[26],[27].***

**ΠΑΡΑΤΗΡΗΣΗ:**

Αν έχουμε $\phi_1(x) = \frac{1}{2}(x-1)^2$ και $\phi_2(x) = x\log x - x + 1$, θα πάρουμε τον έλεγχο του Pearson, $X^2$ , ενώ για $\phi_1(x) = \phi_2(x) = x\log x - x + 1$ , προκύπτει ο στατιστικός έλεγχος του λόγου των πιθανοφανειών, $G^2$. Τέλος ο minimum στατιστικός έλεγχος $X^2$προκύπτει για $\phi_1(x) = \phi_2(x) = \frac{1}{2}(x-1)^2$, ενώ ο minimum $G^2$ έλεγχος, για $\phi_1(x) = x\log x - x + 1$ και $\phi_2(x) = \frac{1}{2}(x-1)^2$. ***[5],[27],[28].***

### 1.6.1. ΑΣΥΜΠΤΩΤΙΚΗ ΚΑΤΑΝΟΜΗ ΕΛΕΓΧΟΣΥΝΑΡΤΗΣΗΣ ΚΑΤΩ ΑΠΟ ΤΗΝ ΣΥΝΘΕΤΗ ΜΗΔΕΝΙΚΗ ΥΠΟΘΕΣΗ

Με την ίδια λογική με την περίπτωση της απλής υπόθεσης, το πρώτο αποτέλεσμα-συμπέρασμα που παρουσιάζεται αφορά την ασυμπτωτική κατανομή κάτω από τη μηδενική υπόθεση που δίνεται στη σχέση $(1.11)$, ενώ το δεύτερο κάτω από εναλλακτική υπόθεση, διαφορετική από αυτή της σχέσης$(1.11)$ και τέλος το τρίτο θεώρημα αφορά την ασυμπτωτική κατανομή κάτω από συναφείς εναλλακτικές υποθέσεις. Αρχικά ωστόσο θα δώσουμε ένα χρήσιμο θεώρημα

***ΘΕΩΡΗΜΑ 1.6.1 :***

*Αν $\phi \in \Phi^*$, είναι δύο φορές συνεχώς διαφορίσιμη συνάρτηση στο $x>0$, με $\phi''(1) > 0$ και $\pi = p(\theta_0)$. Κάτω από τις συνθήκες κανονικότητας του Birch και θεωρώντας ότι η συνάρτηση $p : \Theta \to \Delta_m$ έχει συνεχείς δεύτερες παραγώγους κοντά στο $\theta_0$, ισχύει ότι*

$$\hat{\theta}_\phi = \theta_0 + I_F(\theta_0)^{-1} A(\theta_0)^T diag\left(\pi^{-1/2}\right)(\hat{p}-\pi) + o(\|\hat{p}-\pi\|)$$

*όπου ο εκτιμητής $\hat{\theta}_\phi$ είναι ο μοναδικός κοντά στο $\theta_0$ .Επίσης έχουμε τον πίνακα $I_F(\theta_0)^{-1} = A(\theta_0)^T A(\theta_0)^{-1}$. Ο $I_F(\theta_0)$ είναι ο πίνακας πληροφορίας Fisher για το πολυωνυμικό μοντέλο.* ***[4],[25].***

***ΘΕΩΡΗΜΑ 1.6.2 :***

*Κάτω από την υπόθεση* $(1.11)$ *και λαμβάνοντας υπόψη τις συνθήκες του θεωρήματος 1.6.1, έχουμε ότι :*

$$T_n^{\phi_1}(\hat{\theta}_{\phi_2}) = \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\hat{p}, p(\hat{\theta}_{\phi_2})\right) \xrightarrow{L} X^2_{m-r-1}, \text{καθώς το } n \to \infty$$

*όπου* $\phi_1$ *και* $\phi_2 \in \Phi^*$. ■

Στηριζόμενοι λοιπόν στο *θεώρημα 1.6.2*, μπορούμε να διατυπώσουμε τον εξής ***κανόνα απόφασης*** σύμφωνα με τον οποίο

"***Απορρίπτουμε τη μηδενική υπόθεση που δίνεται στη σχέση*** $(1.10)$ ***σε επίπεδο σημαντικότητας*** $\alpha$ ***αν***

$$T_n^{\phi_1}(\hat{\theta}_{\phi_2}) > X^2_{m-r-1,\alpha} \quad " \quad (1.12)$$

Ορίζουμε ως $q = (q_1, q_2, \dots, q_m)^T$ ένα σημείο στην εναλλακτική υπόθεση. Αν λοιπόν η εναλλακτική υπόθεση $q$ είναι σωστή, έχουμε ότι $\hat{p}$ τείνει κατά πιθανότητα στο $q$. Συμβολίζουμε με $\theta_\alpha$ το σημείο στο $\Theta$, που επαληθεύει το εξής :

$$\theta_\alpha = arg \min_{\theta \in \Theta} D_{\phi_2}(q, p(\theta))$$

Και έχουμε ότι $p(\hat{\theta}_{\phi_2})$ τείνει στο $p(\theta_\alpha)$ κατά πιθανότητα.

***ΣΗΜΕΙΩΣΗ :*** Κάνουμε την εξής χρήσιμη υπόθεση

$$\sqrt{n}\left(\hat{p} - p(\hat{\theta}_{\phi_2}) - (q, p(\theta_\alpha))\right) \xrightarrow{L} N(0, \Sigma) \quad \text{καθώς το } n \to \infty \quad (1.13)$$

Κάτω από την εναλλακτική υπόθεση $q$ όπου

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \text{με} \quad \Sigma_{11} = diag(q) - qq^T \quad \text{και} \quad \Sigma_{12} = \Sigma_{21}$$

***[25],[26],[27],[28].***

***ΘΕΩΡΗΜΑ 1.6.3 :***

*Η ασυμπτωτική ισχύς για τον έλεγχο της σχέσης* $(1.12)$ *για την εναλλακτική υπόθεση q, λαμβάνοντας υπόψη την συνθήκη της σχέσης* $(1.13)$ *δίνεται από τον παρακάτω τύπο :*

$$\beta_{n,\phi_1}(q) = 1 - \mathbf{\Phi}_n\left(\frac{1}{\sigma_{\phi_1}(q)}\left(\frac{\phi_1''(1)}{2\sqrt{n}} X^2_{m-r-1,\alpha} - \sqrt{n} D_{\phi_1}(q, p(\theta_a))\right)\right)$$

*όπου* $\sigma^2_{\phi_1}(q) = Z^T \Sigma_{11} Z + 2Z^T \Sigma_{12} S + S^T \Sigma_{22} S$ *και*

$$Z^T = \left(\frac{\partial D_{\phi_1}(u, p(\theta_a)}{\partial u}\right) \quad \text{όπου } u = q$$

$$S^T = \left(\frac{\partial D_{\phi_1}(q, w)}{\partial w}\right) \quad \text{όπου } w = p(\theta_a)$$

*όπου* $\boldsymbol{\Phi}_n(x)$ *είναι μια ακολουθία από συναρτήσεις κατανομών που τείνουν ομοιόμορφα στην συνάρτηση της τυποποιημένης κανονικής κατανομής* $\boldsymbol{\Phi}(x)$. ***[25],[27].***

Ωστόσο επειδή το παραπάνω θεώρημα δεν είναι εύκολο στην εφαρμογή του, θα θεωρήσουμε μια αλληλουχία από συνεχόμενες εναλλακτικές υποθέσεις που προσεγγίζουν την μηδενική υπόθεση σε ένα βαθμό τάξης $O\left(n^{-1/2}\right)$. Θεωρούμε λοιπόν τις συνεχόμενες ως προς $n$ εναλλακτικές υποθέσεις

$$H_{1,n}: p_n = p(\theta_0) + \frac{1}{\sqrt{n}} d \qquad (1.14)$$

όπου $d \equiv (d_1, d_2, \dots d_m)^T$ είναι ένα καθορισμένο διάνυσμα τέτοιο ώστε $\sum_{i=1}^{m} d_j = 0$ με $p_n \neq p(\theta_0)$ όπου $\theta_0$ άγνωστο και $\theta_0 \in \Theta$. Λαμβάνοντας υπόψη όλα αυτά μπορούμε να διατυπώσουμε το επόμενο θεώρημα.

***ΘΕΩΡΗΜΑ 1.6.4 :***

*Η ασυμπτωτική κατανομή του* $\Phi_1$*-divergence στατιστικού ελέγχου* $T_n^{\Phi_1}\left(\hat{\theta}_{\Phi_2}\right)$*, κάτω από τις συναφείς εναλλακτικές υποθέσεις, που δίνονται στη σχέση* $(1.14)$ *είναι η noncentral* $X^2$ *με* $m - r - 1$ *βαθμούς ελευθερίας και με* $\delta$ *noncentrality παράμετρο που δίνεται από την παρακάτω σχέση*

$$\delta = d^T diag\left(p(\theta_0)^{-1/2}\right)\left(I - A(\theta_0)\left(A(\theta_0)^T A(\theta_0)\right)^{-1} A(\theta_0)^T\right) \times diag\left(p(\theta_0)^{-1/2}\right) d$$

***[5],[24],[25],[28].***

# ΚΕΦΑΛΑΙΟ 2:

# "ΔΙΑΦΟΡΟΙ ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ"

Πέρα από τον έλεγχο του Pearson, υπάρχει μια πληθώρα από ελέγχους, οι οποίοι κατηγοριοποιούνται ανάλογα με το σκοπό που εξυπηρετούν. Σε αυτό το κεφάλαιο παρουσιάζεται μια ευρεία κλίμακα από στατιστικούς ελέγχους που βασίζονται στην συνάρτηση εμπειρικής κατανομής. Ανάμεσα σε αυτούς, θα δούμε τον πιο παλιό, των Kolmogorov-Smirnov, αλλά και άλλους ελέγχους που πρόσφατα προστέθηκαν σε αυτή την κλίμακα, όπως ο Anderson-Darling, ο Cramér-von Mises, ο Watson, ο Kuiper και ο Lilliefors. Όλοι είναι απαραμετρικοί και distribution-free.

Στη συνέχεια θα αναφερθούμε σε μέτρα εντροπίας, τα οποία είναι αρκετά διαδεδομένα τα τελευταία χρόνια στους ελέγχους καλής προσαρμογής, καθώς και σε έναν έλεγχο που βασίζεται πάνω σε αυτά.

## 2.1. ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΜΠΕΙΡΙΚΗ ΣΥΝΑΡΤΗΣΗ ΚΑΤΑΝΟΜΗΣ

Η ***εμπειρική συνάρτηση κατανομής [empirical distribution function (edf)]*** παίζει ένα πολύ κεντρικό ρόλο σε πολλές τεχνικές των ελέγχων καλής προσαρμογής. Στην πραγματικότητα είναι μία εκτιμήτρια της συνάρτησης κατανομής $F$ μιας τυχαίας μεταβλητής $X$ , και ουσιαστικά βασίζεται στην ερμηνεία της συνάρτησης κατανομής $F$. Πιο συγκεκριμένα, για κάθε $x$ έχουμε $F(x) = \Pr\{X \leq x\}$. Επομένως η $F(x)$ είναι μια πιθανότητα, για κάθε $x$, και επειδή οι πιθανότητες είναι εύκολο να εκτιμηθούν, η $F(x)$ έχει επίσης έναν απλό εκτιμητή.

Αν λοιπόν, $S_n = \{X_1, X_2, \dots, X_n\}$ ορίζει ένα τυχαίο δείγμα με $n$ παρατηρήσεις, τότε η $F(x)$ εκτιμάται, για κάθε $x \in R$ ως :

$$\widehat{F_n}(x) = \frac{1}{n}\#\{X_i \in S_n : X_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(X_i \leq x) \quad (2.1)$$

όπου για κάθε $x$, η συνάρτηση $I(X_i \leq x)$ είναι μια δείκτρια συνάρτηση που ορίζεται ως $I(X_i \leq x) = \begin{cases} 1 \ \alpha\nu \ X_i \leq x \\ 0 \ \alpha\nu \ X_i > x \end{cases}$ . Η $\widehat{F_n}$ είναι η εμπειρική συνάρτηση κατανομής και ισούται με το πλήθος των παρατηρήσεων του δείγματος, που είναι μικρότερο ή ίσο από μια συγκεκριμένη τιμή του x, διαιρούμενο από το μέγεθος του δείγματος $n$. Από αυτόν τον ορισμό είναι ξεκάθαρο ότι λειτουργεί ως μια συνάρτηση βηματική, που με βήμα $1/n$ , μεταπηδά σε καθεμιά από τις δειγματικές παρατηρήσεις $x = X_i$. ***[2],[10],[34],[37].***

Η εμπειρική συνάρτηση κατανομής μπορεί επίσης να κατασκευαστεί ως εξής. Αν όλες οι παρατηρήσεις είναι διαφορετικές (κάτι τέτοιο συμβαίνει με πιθανότητα 1, όταν η $F$ είναι συνεχής) , άρα οι $n$ παρατηρήσεις μπορούν να ταξινομηθούν, ως $X_1 < X_2 < \cdots < X_n$, μπορεί να οριστεί ως:

$$\begin{cases} \widehat{F_n}(x) = 0 & \alpha\nu\ x < X_1 \\ \widehat{F_n}(x) = {}^{i}/_{n} & \alpha\nu\ X_i \le x < X_{i+1}\ \gamma\iota\alpha\ i = 1, \dots n-1 \\ \widehat{F_n}(x) = 1 & \alpha\nu\ X_n \le x \end{cases}$$

Γενικά ένα χαρακτηριστικό είναι ότι είναι στενά συνδεδεμένη με τη Διωνυμική κατανομή, κάτι το οποίο είναι εμφανές από τη σχέση(2.1), αφού για κάθε $x$, η δείκτρια $I(X_i \le x)$ είναι μια Bernoulli τυχαία μεταβλητή, με παράμετρο $F(x)$, επομένως η $n\widehat{F_n}(x)$ είναι μια Διωνυμική τυχαία μεταβλητή με παραμέτρους $n$ και $F(x)$. Συνεπώς για κάθε $x$, η ακριβής κατανομή της $\widehat{F_n}(x)$ είναι γνωστή. Έχει μέση τιμή $nF(x)$ και διασπορά $nF(x)\big(1-F(x)\big)$, κάτι το οποίο υποδηλώνει ότι η $\widehat{F_n}(x)$ είναι αμερόληπτη εκτιμήτρια της $F(x)$. Κάποιες από τις ασυμπτωτικές ιδιότητες των συναρτήσεων εμπειρικής κατανομής είναι οι εξής :

**I.** Από τον ισχυρό νόμο των μεγάλων αριθμών προκύπτει ότι για κάθε $x$, έχουμε $\widehat{F_n}(x) \xrightarrow{a.s.} F(x)$ καθώς $n \to \infty$ , το οποίο μπορεί να επεκταθεί στο

$$\sup_{x\in R} |\widehat{F_n}(x) - F(x)| \xrightarrow{a.s.} 0$$

Η $\widehat{F_n}$ τείνει ομοιόμορφα στην $F$ με πιθανότητα 1 (θεώρημα Glivenko-Cantelli).

**II.** Από το κεντρικό οριακό θεώρημα για κάθε $x \in R$, έχουμε ότι η $\widehat{F_n}(x)$ έχει ασυμπτωτικά κανονική κατανομή με ρυθμό σύγκλισης $\sqrt{n}$, καθώς $n \to \infty$,

$$\sqrt{n}(\widehat{F_n}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x))$$

***[2],[7],[34],[37].***

Ουσιαστικά αυτό που μας δείχνουν οι παραπάνω ιδιότητες είναι ότι η $\widehat{F_n}$ είναι κοντά στην $F$, για μεγάλο μέγεθος δείγματος. Συνεπώς όταν το ενδιαφέρον μας επικεντρώνεται στον έλεγχο καλής προσαρμογής της μηδενικής υπόθεσης $H_0 : F(x) = G(x)$, είναι λογικό να εκτιμήσουμε κατά μια έννοια, πόσο διαφέρει η εμπειρική συνάρτηση κατανομής, από την κατανομή $G$ που έχουμε υποθέσει. Και ουσιαστικά αυτό ακριβώς κάνει ένας στατιστικός έλεγχος που στηρίζεται στην εμπειρική κατανομή, μετρώντας την απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής ενός δείγματος από την συνάρτηση κατανομής $G$, που έχουμε υποθέσει ότι ακολουθεί το τυχαίο δείγμα. ***[37].*** Δηλαδή :

$$T_n = c(n)d\big(\widehat{F_n}, G\big) \qquad (2.2)$$

όπου $c(n)$ είναι ένας παράγοντας κλίμακας που εξαρτάται από το μέγεθος του δείγματος $n$ και σκοπό έχει να δημιουργηθεί μια μη εκφυλισμένη ασυμπτωτική κατανομή για την ελεγχοσυνάρτηση $T_n$, κάτω από τη μηδενική υπόθεση. Ενώ $d(..,..)$ είναι κάποια μετρική στο χώρο των συναρτήσεων κατανομής που ορίζει την

απόσταση μεταξύ δύο κατανομών ή την συνάρτηση απόκλισης μεταξύ δύο κατανομών. Γενικά ισχύει ότι:

$$d(F,G) = 0 \; \alpha\nu \; \kappa\alpha\iota \; \mu\acute{o}\nu o \; \alpha\nu \; \eta \; H_0 \; \epsilon\acute{\iota}\nu\alpha\iota \; \alpha\lambda\eta\theta\iota\nu\acute{\eta}$$

*και* $d(\widehat{F_n}, G)$ *είναι συνεπής εκτιμήτρια του* $d(F,G)$

Μια κλασσική επιλογή είναι $d = d_k$, η μετρική του Kolmogorov-Smirnov. Ωστόσο υπάρχουν διάφορες επιλογές για την συνάρτηση απόστασης $d$. ***[16],[37].***

### 2.1.1. ΕΜΠΕΙΡΙΚΗ ΣΤΟΧΑΣΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Παρόλο που εστιάζουμε τη μελέτη μας στην εμπειρική κατανομή $\widehat{F_n}(x)$, στην πράξη αποδεικνύεται ότι είναι πιο βολικό να δουλεύουμε με την ***εμπειρική στοχαστική διαδικασία***, η οποία ορίζεται ως εξής :

$$\boldsymbol{B}_n(x) = \sqrt{n}(\widehat{F_n}(x) - F(x))$$

Και η εμπειρική συνάρτηση κατανομής $\widehat{F_n}(x)$ και η εμπειρική στοχαστική διαδικασία $\boldsymbol{B}_n(x)$ είναι στοχαστικές διαδικασίες. Όταν η $F$ είναι η συνάρτηση κατανομής μιας ομοιόμορφης κατανομής πιθανότητας, η $\boldsymbol{B}_n$ συχνά αναφέρεται σαν μια ομοιόμορφη εμπειρική διαδικασία. Επειδή η $\boldsymbol{B}_n$ εξαρτάται από το τυχαίο δείγμα, είναι μια τυχαία συνάρτηση, που εμπίπτει στον γενικότερο ορισμό της συνάρτησης $(2.2)$. ***[34],[37].***

Ένα πολύ σημαντικό θεώρημα, που έχει σχέση με την εμπειρική διαδικασία, είναι αυτό του Donsker, σύμφωνα με το οποίο

$$\boldsymbol{B}_n(x) \xrightarrow{d} U(F)$$

όπου $U$ είναι η standard Brownian bridge διαδικασία στο [0,1].
Αυτό το θεώρημα είναι γενικά σημαντικό στη μελέτη ασθενούς σύγκλισης των στοχαστικών διαδικασιών.***[34].***

***ΠΑΡΑΔΕΙΓΜΑ:*** Θεωρούμε ένα στατιστικό πρόβλημα ελέγχου καλής προσαρμογής, όπου οι τυχαίες μεταβλητές $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}$, είναι ανεξάρτητες και ακολουθούν την ίδια κατανομή $F$. Θέλουμε να εξετάσουμε αν $H_0 : F = F_0$ VS $H_1 : F \neq F_0$ όπου η $F_0$ είναι μια συγκεκριμένη συνεχής κατανομή. Ο Kolmogorov πρότεινε, για την εξέταση τέτοιου ελέγχου να δουλέψει κάποιος με την ποσότητα

$$\boldsymbol{D}_n(x) = \sqrt{n}\sup_{x\in R}\left(\widehat{F_n}(x) - F_0(x)\right) \qquad (2.3)$$

απορρίπτοντας την μηδενική υπόθεση όταν η τιμή της $\boldsymbol{D}_n$ είναι μεγάλη. Ωστόσο για να υπολογιστεί η p-value του ελέγχου, θα πρέπει η κατανομή της ελεγχοσυνάρτησης κάτω από την μηδενική κατανομή, η κατανομή δηλαδή της $\boldsymbol{D}_n$ κάτω από τη μηδενική υπόθεση, να οριστεί.

Γενικά μια πολύ ενδιαφέρουσα και σημαντική ιδιότητα για την κατανομή της ελεγχοσυνάρτησης $\boldsymbol{D}_n$ κάτω από τη μηδενική υπόθεση, είναι ότι παραμένει η ίδια, όταν η $F_0$ είναι συνεχής. Συνεπώς μπορούμε να υπολογίσουμε αυτή την κατανομή, υποθέτοντας ότι η $F_0$ είναι η συνάρτηση κατανομής πιθανότητας μιας τυχαίας

μεταβλητής που είναι ομοιόμορφα κατανεμημένη στο [0,1]. Δηλαδή η κατανομή της ελεγχοσυνάρτησης $\boldsymbol{D}_n$ κάτω από τη μηδενική υπόθεση είναι ίδια με αυτής της $\sup|U_n(t)|\ για\ κάθε\ t \in [0,1]$ ,

$$\text{όπου } U_n(t) \coloneqq \sqrt{n}\,(\widehat{F}_n(t) - t) \ \text{ με } \ \widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(\,\xi_i \le t)\,, \qquad με \ \ t \in [0,1]$$

και $\xi_1, \xi_2, \dots, \xi_n$ τυχαίες μεταβλητές, ανεξάρτητες που ακολουθούν την ομοιόμορφη κατανομή στο (0,1). Η διαδικασία λέγεται ομοιόμορφη εμπειρική διαδικασία και γενικά υπάρχουν πολλοί έλεγχοι, πέραν του Kolmogorov που βασίζονται πάνω σε αυτή την ιδέα. ***[16],[34],[37].***

Αυτό που έκανε ο Donsker ουσιαστικά είναι να αποδείξει ότι μια ομοιόμορφη εμπειρική διαδικασία $\{U_n(t): t \in [0,1]\}$ συγκλίνει στην Brownian bridge διαδικασία $\{U(t): t \in [0,1]\}$, δηλαδή ότι

$$U_n \xrightarrow{d} U \text{ και συνεπώς } \sup_{x\in[0,1]} |U_n(x)| \xrightarrow{d} \sup_{x\in[0,1]} |U(x)|$$

Πολύ περιληπτικά αυτό προκύπτει από το κεντρικό οριακό θεώρημα για πολλές μεταβλητές, καθώς για κάθε $k \ge 1$,και $0 < t_1, \dots . , t_k < 1$ έχουμε ότι το διάνυσμα $(U_n(t_1), U_n(t_2), \dots, U_n(t_k)$ συγκλίνει σε κατανομή $N_k(0, \Sigma)$, όπου στο κελί $(i, j)$ του πίνακα διασποράς-συνδιασποράς έχουμε $\Sigma_{ij} = t_i \wedge t_j - t_i t_j$. Αυτή η κατανομή είναι η ίδια με την κατανομή του $U(t_1), U(t_2), \dots, U(t_k)$ όπου το $U$ είναι η Brownian bridge. ***[16],[34].***

## 2.2. ΕΛΕΓΧΟΙ ΠΟΥ ΣΤΗΡΙΖΟΝΤΑΙ ΣΤΗΝ ΕΜΠΕΙΡΙΚΗ ΣΥΝΑΡΤΗΣΗ ΚΑΤΑΝΟΜΗΣ

Σε αυτή την υποενότητα, παρουσιάζονται αναλυτικά, κάποιοι από τους πιο σημαντικούς ελέγχους που στηρίζονται στην εμπειρική συνάρτηση κατανομής, χωρισμένοι ουσιαστικά σε δύο μεγάλες ομάδες που θα εξηγηθούν πιο κάτω. Η πρώτη ομάδα στατιστικών ελέγχων, στηρίζεται στον υπολογισμό της μέγιστης απόκλισης που υπολογίζεται ανάμεσα στην υποτιθέμενη και την εμπειρική συνάρτηση κατανομής. Σε αυτή την ομάδα ανήκει η οικογένεια των ελέγχων Kolmogorov-Smirnov. Η δεύτερη ομάδα στηρίζεται στον υπολογισμό όλων των αποστάσεων μεταξύ της υποτιθέμενης και της εμπειρικής συνάρτησης κατανομής. Η οικογένεια των ελέγχων Cramér-von Mises ανήκει σε αυτή την ομάδα. Επιπλέον δίνεται η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση για καθέναν ξεχωριστά. Για τους πίνακες με τα κρίσιμα σημεία του κάθε ελέγχου, κάποιος μπορεί να ανατρέξει σε αντίστοιχη βιβλιογραφία.

Σε γενικές γραμμές, για τον έλεγχο της απλής μηδενικής υπόθεσης $F = F_0$, ένα λογικό-φυσικό σημείο εκκίνησης είναι το να στηρίξουμε έναν στατιστικό έλεγχο στην μέτρηση της διαφοράς μεταξύ της εμπειρικής συνάρτησης κατανομής $\widehat{F}_n$ και της $F_0$. Συγκεκριμένα αν $d$ είναι κάποια μετρική στο χώρο των συναρτήσεων

κατανομής, τότε η $d(\hat{F}_n, F_0)$ θα μπορούσε να χρησιμοποιείται σε κάποιον έλεγχο. ***[16].***

### 2.2.1. ΕΛΕΓΧΟΣ KOLMOGOROV-SMIRNOV

Υποθέτουμε ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ από έναν πληθυσμό, που ακολουθεί μια άγνωστη κατανομή $F$, την οποία συγκρίνουμε με μια συγκεκριμένη κατανομή $F_0$, θέλοντας έτσι να εξετάσουμε αν είναι λογικό να πούμε ότι η $F_0$ είναι και η πραγματική συνάρτηση κατανομής του τυχαίου δείγματος. Δηλαδή έχουμε τον έλεγχο: ***[2].***

$$H_0 : F = F_0 \quad vs \quad H_1 : F \neq F_0 \qquad (2.4)$$

Ξέρουμε ήδη πως μπορούμε να μελετήσουμε έναν τέτοιο έλεγχο καθώς στο **κεφάλαιο 1**, μιλήσαμε για τον $X^2$ έλεγχο καλής προσαρμογής και τη γενική ιδέα της εφαρμογής του. Ωστόσο, αν η κατανομή $F_0$ είναι συνεχής, δεν θεωρείται ο καλύτερος, καθώς προϋποθέτει ομαδοποίηση των δεδομένων (χωρίζουμε το πεδίο τιμών των παρατηρήσεων σε $m$ διαστήματα $E_i$ για $i = 1, \dots, m$), τις περισσότερες κιόλας φορές αυτό γίνεται αυθαίρετα, με συνέπεια την απώλεια πληροφορίας, η οποία περιέχεται στις τιμές του δείγματος. Σε αυτή την περίπτωση λοιπόν, συνήθως προτιμάται ο έλεγχος Kolmogorov-Smirnov, αφού βασίζεται στην εμπειρική συνάρτηση κατανομής του δείγματος, άρα δεν προϋποθέτει κάποια ομαδοποίηση των δεδομένων, καθώς κάθε μέτρηση χωριστά συμμετέχει στην σύγκριση. ***[40].***

Γενικά μέσω αυτού του ελέγχου, εξετάζεται η καλή προσαρμογή ενός τυχαίου δείγματος σε μία δεδομένη κατανομή και στηρίζεται στο ότι η εμπειρική συνάρτηση κατανομής που υπολογίζεται από το δείγμα μας, θα προσεγγίζει την πραγματική (αναμενόμενη) συνάρτηση κατανομής που ακολουθεί ένας πληθυσμός, από τον οποίο έχουμε πάρει το δείγμα, κάτω από την μηδενική υπόθεση.

Επομένως η ***στατιστική ελεγχοσυνάρτηση των Kolmogorov-Smirnov*** είναι η εξής:

$$\boldsymbol{G}_n \equiv \sqrt{n} sup_{x \in R} \left| \hat{F}_n(x) - F_0(x) \right| = sup_{x \in R} |\boldsymbol{B}_n(x)|$$

όπου ουσιαστικά η $\boldsymbol{G}_n$ έχει τη μορφή της σχέσης $(2.2)$ με $d$ να είναι μια συνάρτηση supremum. Δηλαδή $\boldsymbol{G}_n \equiv \sqrt{n} d_k\left(\hat{F}_n, F_0\right)$ όπου $d_k$ είναι η απόσταση των Kolmogorov-Smirnov, η οποία στην πραγματικότητα μετράει τη "συμφωνία", κατά απόλυτη τιμή της μεγαλύτερης κατακόρυφης διαφοράς ανάμεσα στο γράφημα της εμπειρικής και της αναμενόμενης κάτω από την μηδενική συνάρτησης κατανομής. Με άλλα λόγια, ο στατιστικός έλεγχος των Kolmogorov-Smirnov είναι η μεγαλύτερη απόκλιση-απόσταση μεταξύ των δύο κατανομών. Αν η μεγαλύτερη απόκλιση είναι "μικρή", τότε συνεπάγεται ότι όλες οι αποκλίσεις είναι μικρές. ***[10],[37]***

Θεωρούμε λοιπόν την ελεγχοσυνάρτηση

$$D_n = d_k\left(\hat{F}_n, F_0\right) = sup_{x \in R} \left| \hat{F}_n(x) - F_0(x) \right|$$

και θα λέμε ότι τα δεδομένα δεν υποστηρίζουν την μηδενική υπόθεση, όταν αυτή η στατιστική συνάρτηση παίρνει "ασυνήθιστα" μεγάλες τιμές, συμπεραίνοντας ότι η πραγματική, αλλά άγνωστη συνάρτηση κατανομής της $F(x)$ δεν δίνεται από την

$F_0(x)$ που έχουμε υποθέσει κάτω από τη μηδενική μέσω της σχέσης (2.4). Αυτό συμβαίνει καθώς αν η $F_0(x)$ είναι η πραγματική συνάρτηση κατανομής πιθανότητας, θα πρέπει να υπάρχει μια λογική διαφορά ανάμεσα στην $\hat{F}_n(x)$ και την $F_0(x)$ για όλες τις τιμές του $x$, καθώς η $\hat{F}_n(x)$ αντιπροσωπεύει την εικόνα της πραγματικής κατανομής του δείγματος. Ισοδύναμα η απόκλιση ανάμεσα στην $\hat{F}_n(x)$ και την $F_0(x)$ πρέπει να είναι μικρή για όλα τα $x$. ***[2],[10],[22].***

Εκτός από την εναλλακτική υπόθεση της σχέσης (2.4) για την εξέταση του ελέγχου προσαρμοστικότητας, υπάρχουν και άλλα δύο είδη εναλλακτικών υποθέσεων που μπορεί να υποθέσει κανείς. Η εναλλακτική υπόθεση

$$H'_1: F(x) > F_0(x) \quad (2.5)$$

καθώς και η

$$H''_1: F(x) < F_0(x) \quad (2.6) \quad .$$

Πριν αρχίσουμε την αντιμετώπιση των ελέγχων με τον έλεγχο των Kolmogorov-Smirnov, ιδιαίτερο ενδιαφέρον παρουσιάζουν τα επόμενα χρήσιμα θεωρήματα:

***ΘΕΩΡΗΜΑ 2.2.1:***

*Αν η $F_0(x)$ είναι μια συνεχής κατανομή, τότε η κατανομή του*

$$D_n = sup_{x\in R}\left|\hat{F}_n(x) - F_0(x)\right|$$

*δεν εξαρτάται από την $F_0$.*

***ΑΠΟΔΕΙΞΗ :***

Ορίζουμε την αντίστροφή συνάρτηση κατανομής της $F$ ως εξής :

$$F_0{}^{-1}(y) = in\, f\{x: F_0(x) \geq y\}, \quad 0 < y < 1$$

Ισχύει ότι $P\left(sup_{x\in R}\left|\hat{F}_n(x) - F_0(x)\right| \leq z\right) = P\left(sup_{0<y<1}\left|\hat{F}_n(F_0^{-1}(y)) - y\right| \leq z\right)$ αν θέσουμε όπου $y = F_0(x)$ ή $x = F_0{}^{-1}(y)$. Από τον ορισμό της εμπειρικής συνάρτησης κατανομής προκύπτει ότι :

$$\hat{F}_n(F_0^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, F_0^{-1}(y)]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(F_0(X_i))$$

Επομένως

$$P\left(sup_{x\in R}\left|\hat{F}_n(x) - F_0(x)\right| \leq z\right) = P(sup_{0<y<1}|\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(F_0(X_i)) - y| \leq z)$$

Λαμβάνοντας τώρα υπόψη ότι η κατανομή της $F_0(X_i)$ είναι ομοιόμορφη στο διάστημα [0,1] καθώς η συνάρτηση πυκνότητας πιθανότητας της $F_0(X_1)$ είναι :

$$P(F_0(X_1) \leq z) = P\left(X_1 \leq F_0^{-1}(z)\right) = F_0\left(F_0^{-1}(z)\right) = z, \;\; για\; z \in (0,1)$$

Άρα οι μεταβλητές $U_i = F_0(X_i)$ για $i = 1, \dots, n$ είναι ανεξάρτητες και ακολουθούν την ομοιόμορφη κατανομή στο διάστημα [0,1].

$$P\left(sup_{x\in R}\left|\hat{F}_n(x) - F_0(x)\right| \leq z\right) = P\left(sup_{0<y<1}\left|\frac{1}{n}\sum_{i=1}^{n} I(U_i \leq y) - y\right| \leq z\right)$$

που δεν εξαρτάται από την $F_0$, αλλά από το μέγεθος του δείγματος $n$. ***[40],[41].***

***ΘΕΩΡΗΜΑ 2.2.2:*** *(μετασχηματισμός ολοκληρώματος πιθανότητας)*

*Έστω η συνεχής τυχαία μεταβλητή $X$ με συνάρτηση κατανομής $F(x)$. Τότε η τυχαία μεταβλητή $Y = F(X)$ ακολουθεί την ομοιόμορφη κατανομή στο (0,1).*

***ΑΠΟΔΕΙΞΗ :***

Ορίζουμε, όπως στην απόδειξη του προηγούμενου θεωρήματος, την αντίστροφη συνάρτηση κατανομής $F^{-1}(y) = in\, f\{x: F(x) \geq y\}$, $για\ 0 < y < 1$

Λόγω συνέχειας της $F$ ισχύει ότι $F\left(F^{-1}(y)\right) = y$, για κάθε $y \in (0,1)$, ενώ λόγω μονοτονίας έχουμε το ενδεχόμενο :

$$\{X \leq F^{-1}(y)\} \Rightarrow \left\{F(X) \leq F(F^{-1}(y)\right) = \{F(X) \leq y\}$$

Αλλά

$$\{F(X) \leq y\} = \{X \leq F^{-1}(y)\}\, U\, \{X > F^{-1}(y)\ και\ F(X) = y\}$$

Συνεπώς

$$P[F(X) \leq y] = P[X \leq F^{-1}(y)] + P[X > F^{-1}(y)] * P[F(X) = y]$$

Ωστόσο επειδή η $F$ είναι συνεχής, $P[F(X) = y] = 0$, προκύπτει ότι :

$$P[F(X) \leq y] = P[X \leq F^{-1}(y)]$$

Αν η $F_Y(y)$ είναι συνάρτηση κατανομής της $Y$, τότε

$$F_Y(y) = P[Y \leq y] = P[F(X) \leq y] = P[X \leq F^{-1}(y)] = F\left(F^{-1}(y)\right) = y$$

Άρα $F_Y(y) = y$ για κάθε $0 < y < 1$, που είναι η συνάρτηση κατανομής της $U(0,1)$.

***ΠΑΡΑΤΗΡΗΣΗ :***

Εάν $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ είναι το διατεταγμένο δείγμα $n$ παρατηρήσεων από συνεχή κατανομή $F(x)$, τότε από το μετασχηματισμό του ολοκληρώματος πιθανότητας προκύπτει ότι τα

$$Y_i = F\left(X_{(i)}\right)\ για\ i = 1, \dots, n$$

είναι διατεταγμένες τυχαίες μεταβλητές από την ομοιόμορφη κατανομή στο (0,1). Η τυχαία μεταβλητή $Y_i$ ακολουθεί την κατανομή Βήτα με παραμέτρους $i$ και $n - i + 1$. ***[15],[40],[41].*** ■

Έστω $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ ( $x_{(1)} < x_{(2)} < \dots < x_{(n)}$) οι διατεταγμένες τιμές του τυχαίου δείγματος $X_1, X_2, \dots, X_n$, όπου $x_{(1)} = -\infty$ και $x_{(n)} = +\infty$. Η εμπειρική κατανομή $\hat{F}_n$ είναι σταθερή στα διαστήματα $[x_{(i-1)}, x_{(i)})$, ενώ παρουσιάζει άλματα ύψους $1/n$ στα σημεία $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ (step function). Εφόσον λοιπόν η $F_0$ είναι αύξουσα συνάρτηση , η μέγιστη τιμή της συνάρτησης $\hat{F}_n(x) - F_0(x)$, θα λαμβάνεται πάνω σε κάποιο από τα σημεία $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. Δηλαδή :

$$D_n^+ = sup_{-\infty<x<\infty}\{\hat{F}_n(x) - F_0(x)\} = \max_{1\leq i\leq n}\left(\hat{F}_n\left(X_{(i)}\right) - F_0\left(X_{(i)}\right)\right)$$

$$= \max_{1\leq i\leq n}\left(\frac{i}{n} - F_0\left(X_{(i)}\right)\right) \geq 0 \qquad (2.7)$$

Χαρακτηριστικό της $D_n^+$ είναι ότι είναι distribution-free, αφού δεν εξαρτάται από την $F_0$, καθώς από την πιο πάνω ***Παρατήρηση*** είδαμε ότι $Y_i = F(X_{(i)})$, όπου η κατανομή των $Y_i$ είναι ανεξάρτητη από την υποτιθέμενη συνεχή κατανομή $F_0$. Τέλος αποτελεί κριτήριο για την μονόπλευρη εναλλακτική $H'_1$ της σχέσης **(2.5)**.

Όμοια για την μέγιστη τιμή της συνάρτησης $F_0(x) - \hat{F}_n(x)$ ισχύει ότι :

$$D_n^- = sup_{-\infty<x<\infty}\{F_0(x) - \hat{F}_n(x)\} = \max_{1\le i\le n}\left(F_0(X_{(i)}) - \hat{F}_n(X^-{}_{(i)})\right)$$
$$= \max_{1\le i\le n}\left(F_0(X_{(i)}) - \frac{i-1}{n}\right) \ge 0 \qquad (2.8)$$

αποτελώντας κριτήριο για την μονόπλευρη εναλλακτική $H''_1$ της σχέσης **(2.6)**.

Για να βρούμε την μέγιστη τιμή για την $D_n^+$, κοιτάμε τις διαφορές μόνο για εκείνα τα $x$ για τα οποία η $\hat{F}_n(x)$ είναι μεγαλύτερη από την $F_0(x)$, ενώ για την $D_n^-$, εκείνα για τα οποία η $\hat{F}_n(x)$ είναι μικρότερη από την $F_0(x)$. Μεγάλη τιμή της $D_n^+$, ή όμοια της $D_n^-$, σηματοδοτεί την απόρριψη της $H_0$. Η $D_n^-$ είναι free-distribution όπως και η $D_n^+$, ενώ τέλος, εξαιτίας της συμμετρίας, οι μονόπλευροι στατιστικοί έλεγχοι των Kolmogorov-Smirnov είναι πανομοιότυπα κατανεμημένοι, συνεπώς ο ίδιος πίνακας μπορεί να χρησιμοποιηθεί για όλους τους ελέγχους. ***[10],[15],[37],[40].***

Επομένως, για τον έλεγχο της μηδενικής υπόθεσης $H_0$ έναντι της αμφίπλευρης εναλλακτικής υπόθεσης της σχέσης **(2.4)**, μπορούμε να εκφράσουμε τον έλεγχο των Kolmogorov-Smirnov και ως :

$$D_n = sup_{-\infty<x<\infty}\{|\hat{F}_n(x) - F_0(x)|\} = \max_{1\le i\le n}\{D_n^+, D_n^-\}$$
$$= \max_{1\le i\le n}\left(\frac{i}{n} - F_0(X_{(i)}), F_0(X_{(i)}) - \frac{i-1}{n}\right)$$

Όταν λοιπόν η $F$ είναι συνεχής, τότε η $D_n$ έχει συνεχή κατανομή. Αν το τυχαίο δείγμα $X_1, X_2, \dots, X_n$ είναι ανεξάρτητο και πανομοιότυπα κατανεμημένο, σύμφωνα με μια συνεχή κατανομή $F_0$, από το *θεώρημα 2.2.2* παρατηρούμε ότι οι τυχαίες μεταβλητές $Y_i = F_0(X_i)$, $i = 1, \dots, n$ είναι ανεξάρτητες και ακολουθούν την ομοιόμορφη κατανομή στο (0,1). Επομένως οι τυχαίες μεταβλητές $Y_{(i)} = F_0(X_{(i)})$ μπορεί να θεωρηθούν ότι ακολουθούν ένα διατεταγμένο δείγμα από την ομοιόμορφη κατανομή στο (0,1), δηλαδή $F_0(x) = \begin{cases} 0 & \text{για } x < 0 \\ x & \text{για } 0 \le x < 1 \\ 1 & \text{για } x \ge 1 \end{cases}$.

Συνεπώς υπό τη μηδενική υπόθεση $H_0$ , οποιαδήποτε και αν είναι η $F_0$, η $D_n$ έχει την ίδια κατανομή με την τυχαία μεταβλητή:

$$\max\left\{\frac{i}{n} - Y_{(i)}, Y_{(i)} - \frac{i-1}{n} \;\; \text{όπου} \;\; i = 1, \dots, n\right\}$$

όπου $Y_1, Y_2, \dots, Y_n$ είναι ανεξάρτητες τυχαίες μεταβλητές από την ομοιόμορφη κατανομή $U(0,1)$, η οποία δεν εξαρτάται από την $F_0$. ***[15],[16].***

Για επίπεδο σημαντικότητας $\alpha$, η κρίσιμη περιοχή είναι $K$: $D_n \geq D_n(a) = c$ , άρα όπου η τιμή του κριτηρίου είναι αρκετά μεγαλύτερη από τη σταθερά, υπάρχουν σοβαρές ενδείξεις, περί απόρριψης της μηδενικής υπόθεσης.
Η σταθερά $D_n(a)$ είναι το άνω α-σημείο της κατανομής της τυχαίας μεταβλητής $D_n$, για την οποία ισχύει $P[D_n \geq D_n(a)|H_0] = a$ και δίνεται από αντίστοιχους πίνακες με τα άνω α-σημεία της τυχαίας μεταβλητής $D_n$, για διάφορα $n$ και επίπεδα σημαντικότητας $\alpha$. ***[40].*** ■

Οι Birnbaum και Tingey (1951), βρήκαν την κατανομή της $D_n^+$, όπου εύκολα προκύπτει ότι είναι ίδια με της $D_n^-$ και η οποία είναι η εξής:

$$\Pr(D_n^+ \leq t) = 1 - \sum_{k=0}^{[n-nt]} \binom{n}{k} t\left(t + \frac{k}{n}\right)^{k-1}\left(1 - \left(\frac{k}{n} + t\right)\right)^{n-k}, \qquad 0 \leq t \leq 1$$

Παρ' όλα αυτά η ακριβής κατανομή της $D_n$ δεν έχει αποδειχθεί. Γενικά προτιμάται η χρήση της οριακής κατανομής, χρησιμοποιώντας τη θεωρία της εμπειρικής στοχαστικής διαδικασίας. Από την ασθενή σύγκλιση $\boldsymbol{B}_n \xrightarrow{w} B$ και σχετικά θεωρήματα, προκύπτει ότι, για $n \to \infty$, και υπό τη μηδενική υπόθεση:

$$\boldsymbol{D}_n \xrightarrow{d} D = sup_{x \in R}|B(x)| \quad (*)$$

όπου $B$ είναι μια Brownian bridge, που ουσιαστικά αντιπροσωπεύει μια τυχαία συνεχή διαδικασία στο διάστημα [0,1].
Σε γενικές γραμμές, κρίσιμες τιμές μπορούν να βρεθούν προσομοιώνοντας το δεξί μέλος της σχέσης $(*)$, ωστόσο η συνάρτηση κατανομής της $D$ υπάρχει και μπορεί να εκφραστεί ως εξής :

$$F_D(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}$$

Επομένως για μεγάλο $n(n \to \infty)$,η κατανομή της τυχαίας κατανομής $\boldsymbol{G}_n = \sqrt{n}D_n$ έχει ασυμπτωτικά (υπό την $H_0$ και για συνεχή συνάρτηση κατανομής $F_0$) τη συνάρτηση κατανομής:

$$P(\boldsymbol{G}_n \leq t) = P\left(\sqrt{n}D_n \leq t\right) \underset{n\to\infty}{\longrightarrow} 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2} \; \gamma\iota\alpha \; \kappa\acute{\alpha}\theta\varepsilon \; t \geq 0.$$

Κατά συνέπεια προκύπτει το εξής θεώρημα :

***ΘΕΩΡΗΜΑ 2.2.3:***

*Αν η $F_0(x)$ είναι συνεχής, τότε για κάθε $t > 0$*

$$P\left(\sqrt{n}sup_{x\in R}\left|\hat{F}_n(x) - F_0(x)\right| \leq t\right) \to H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}$$

*όπου $H(t)$ είναι η ασυμπτωτική συνάρτηση κατανομής πιθανότητας της ελεγχοσυνάρτησης Kolmogorov-Smirnov.* ***[2],[15],[16],[37],[40].***

*Μέχρι τώρα όμως είδαμε, ότι ο έλεγχος του αν ένα σετ από παρατηρήσεις προέρχεται από μια πλήρως προσδιορισμένη συνεχή κατανομή $F_0(x)$ ή όχι ,*

*εκτελείται από τον έλεγχο $X^2$ ή/και τον έλεγχο Kolmogorov-Smirnov. Όταν το μέγεθος του δείγματος είναι μικρό, ο έλεγχος των Kolmogorov-Smirnov και του Kuiper, τον οποίο θα δούμε στη συνέχεια, είναι προτιμότεροι.* ***[6].***

### 2.2.2. ΕΛΕΓΧΟΣ KUIPER

Ουσιαστικά ο έλεγχος του Kuiper αποτελεί μια διαφορετική εκδοχή των Kolmogorov-Smirnov, ενώ είναι ιδιαίτερα χρήσιμος αν οι παρατηρήσεις είναι μετρήσεις, οι οποίες καταγράφουν σημεία που βρίσκονται σε κύκλο, αποδεικνύοντας ότι η τιμή του στατιστικού ελέγχου, δεν εξαρτάται από την επιλογή της αρχής. Όπως λοιπόν και στον Kolmogorov-Smirnov, η κατανομή πιθανότητας της στατιστικής ελεγχοσυνάρτησης του Kuiper είναι ανεξάρτητη από την πλήρως ορισμένη συνάρτηση κατανομής $F_0(x)$ που υποδηλώνεται στη μηδενική υπόθεση. Βασικό χαρακτηριστικό του ελέγχου αυτού, είναι ότι χρησιμοποιεί στην ελεγχοσυνάρτησή του, τη μέγιστη κατακόρυφη απόκλιση της εμπειρικής κατανομής πάνω από τη συνάρτηση πυκνότητας πιθανότητας, δηλαδή την ποσότητα $D_n^+$ και τη μέγιστη κατακόρυφη απόκλιση της εμπειρικής κατανομής κάτω από τη συνάρτηση πυκνότητας πιθανότητας, δηλαδή την ποσότητα $D_n^-$ που ορίζονται από τη σχέση $(2.7)$ και $(2.8)$ αντίστοιχα, δηλαδή :

$$V_n = \sup_{x\in R}\left[\hat{F}_n(x) - F_0(x)\right] - \inf_{x\in R}\left[\hat{F}_n(x) - F_0(x)\right] = D_n^+ + D_n^-$$

Αυτή η μικρή διαφοροποίηση από τον έλεγχο των Kolmogorov-Smirnov, κάνει το κριτήριο τόσο “ευαίσθητο” στις ουρές, όσο και στο κέντρο. ***[36].***

Γενικά υπάρχουν διαθέσιμοι πίνακες με τα κρίσιμα σημεία του στατιστικού ελέγχου. Μεγάλες τιμές της ελεγχοσυνάρτησης υποδεικνύουν μεγάλη απόκλιση από τον ισχυρισμό της μηδενικής υπόθεσης, οδηγώντας τελικά στην απόρριψή της.

Για μεγάλες τιμές $n$, η ασυμπτωτική κατανομή του $V_n$ είναι

$$\lim_{n\to\infty} P[\sqrt{n}V_n \geq z] =$$

$$= \sum_{m=1}^{\infty} 2(4m^2z^2 - 1)e^{-2m^2z^2} - \frac{8z}{3\sqrt{n}}\sum_{m=1}^{\infty} m^2(4m^2z^2 - 3)e^{-2m^2z^2} + O\left(\frac{1}{n}\right)$$

Η ακριβής κατανομή του $V_n$ και στην πάνω και την κάτω ουρά, μαζί με τα αποτελέσματα που προκύπτουν από την ασυμπτωτική κατανομή της, κάνουν εφικτό τον υπολογισμό των σημείων σημαντικότητας, δίνοντας την δυνατότητα στον έλεγχο, να είναι διαθέσιμος για ολόκληρο το πλήθος τιμών $n$.

*Ωστόσο στη σύνθετη υπόθεση, όπου βασικές παράμετροι της κατανομής είναι άγνωστες, οπότε θα πρέπει να εκτιμηθούν από το δείγμα, οι έλεγχοι των Kolmogorov-Smirnov και του Kuiper, δε μπορούν κατά μία έννοια να εφαρμοστούν, αφού η κατανομή του στατιστικού ελέγχου δεν είναι πλέον ανεξάρτητη από τη μορφή της $F(x)$. Επομένως και οι γνωστοί πίνακες με τα κρίσιμα σημεία δε μπορούν να χρησιμοποιηθούν, αποτελώντας έτσι και το βασικό τους μειονέκτημα έναντι στον*

*έλεγχο του Pearson. Λύση έρχεται να δώσει ο έλεγχος Lilliefors, ο οποίος εφαρμόζεται για ορισμένες κατανομές όπως είναι η Εκθετική και η Κανονική, όπου υπάρχουν άγνωστες παράμετροι.**[2],[19],[20],[40].***

### 2.2.3. ΕΛΕΓΧΟΣ LILLIEFORS

Στην πραγματικότητα και ο έλεγχος Lilliefors είναι μια τροποποίηση του ελέγχου των Kolmogorov-Smirnov. Παρότι ο τελευταίος χρησιμοποιείται ως μέσο, ικανό να εξετάσει, αν ένα σετ από παρατηρήσεις προέρχεται από μια πλήρως προσδιορισμένη, συνεχή κατανομή, πολλές φορές είναι δύσκολο να προσδιορίσει κανείς, πλήρως ή μερικώς, τις παραμέτρους μιας "άγνωστης" κατανομής. Επομένως τα αποτελέσματα της στατιστικής ελεγχοσυνάρτησης των Kolmogorov-Smirnov μπορεί να είναι αρκετά "συντηρητικά". Σε αυτή λοιπόν την περίπτωση, οι παράμετροι για τον έλεγχο Lilliefors χρειάζονται να υπολογιστούν, βασιζόμενοι στα δεδομένα του δείγματος, αποτελώντας έτσι και ένα λόγο προτίμησης του ελέγχου σε σχέση με αυτό των Kolmogorov-Smirnov. Δεδομένου λοιπόν ενός δείγματος με $n$ παρατηρήσεις, ορίζουμε τη ***στατιστική ελεγχοσυνάρτηση του Lilliefors***:

$$D_n{}^* = sup_{x\in R}\{|\hat{F}_n(x) - \hat{F}_0(x)|\} \quad (2.9)$$

όπου $\hat{F}_0(x)$ είναι ο εκτιμητής της $F_0$, ο οποίος υπολογίζεται χρησιμοποιώντας τη μέθοδο της μέγιστης πιθανοφάνειας, με σκοπό να εκτιμηθούν οι άγνωστες παράμετροι της παραμετρικής οικογένειας που έχουμε υποθέσει. Αν η $F$ είναι συνεχής, τότε η $D_n{}^*$ έχει συνεχή κατανομή, ωστόσο η κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική κατανομή και η ασυμπτωτική κατανομή της $D_n$ κάτω από την απλή μηδενική υπόθεση δεν ισχύει για την $D_n{}^*$ στην σύνθετη υπόθεση. Γι' αυτό το λόγο δημιουργήθηκαν πίνακες της κατανομής της $D_n{}^*$ κάτω από τη μηδενική υπόθεση, προσομοιωμένοι με τις μεθόδους Monte Carlo, για κάθε παραμετρική οικογένεια. Χρησιμοποιώντας λοιπόν τους πίνακες με τις κρίσιμες τιμές του στατιστικού ελέγχου, αν η τιμή αυτής της διαφοράς είναι αρκετά μεγάλη, ώστε να θεωρηθεί στατιστικά σημαντική, τότε απορρίπτουμε τη μηδενική υπόθεση. ***[11],[15],[30].***

#### 2.2.3.1. ΕΛΕΓΧΟΣ LILLIEFORS ΓΙΑ ΤΗΝ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Με τον έλεγχο Lilliefors συνήθως επιθυμούμε να εξετάσουμε αν η κατανομή ενός τυχαίου δείγματος $X_1, X_2, \dots, X_n$ , μπορεί λογικά να προσεγγιστεί από μια Κανονική κατανομή. Οι εκτιμητές μέγιστης πιθανοφάνειας, της μέσης τιμής και της τυπικής απόκλισης είναι:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \text{η μέση τιμή του δείγματος} \quad (2.10) \quad \text{και}$$

$$\hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}}, \quad \text{η τυπική απόκλιση του δείγματος} \quad (2.11)$$

Επομένως ο εκτιμητής $\hat{F}_0(x)$ της σχέσης **(2.9)** είναι η κανονική κατανομή με μέση τιμή και τυπική απόκλιση, δοσμένες από τις σχέσεις **(2.10)** και **(2.11)** αντίστοιχα. Δηλαδή για κάθε $x$ έχουμε :

$$\hat{F}_0(x) = \mathbf{\Phi}\left(\frac{x - \bar{X}}{S}\right)$$

Για να κατασκευάσουμε τον στατιστικό έλεγχο Lilliefors, χρειάζεται να έχουμε τις $n$ τιμές της $\hat{F}_n(X_{(i)})$ καθώς επίσης και τις $n$ τιμές της :

$$\hat{F}_0(Z_{(i)}) = \mathbf{\Phi}(Z_i) = \mathbf{\Phi}\left(\frac{X_{(i)} - \bar{X}}{S}\right), i = 1, \dots, n$$

Τυποποιούμε δηλαδή τις τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$, χρησιμοποιώντας:

$$Z_{(i)} = \frac{X_{(i)} - \bar{X}}{S}, i = 1, \dots, n$$

Επομένως η μηδενική υπόθεση είναι ισοδύναμη με την υπόθεση ότι το τυχαίο δείγμα $Z_{(1)}, Z_{(2)}, \dots, Z_{(n)}$ ,με $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ προέρχεται από την τυποποιημένη κανονική κατανομή. Αν δηλαδή $\hat{S}_n(z)$ είναι η εμπειρική συνάρτηση κατανομής του τυποποιημένου δείγματος και $\hat{F}_0(z)$ είναι η συνάρτηση της τυποποιημένης κανονικής κατανομής, η ελεγχοσυνάρτηση του Lilliefors ορίζεται ως εξής:

$$D_n{}^* = sup_{z \in R}\{|\hat{S}_n(z) - \hat{F}_0(z)|\}$$

η ακριβής κατανομής της οποίας είναι δύσκολο να προσδιοριστεί. Η υποτεθείσα συνάρτηση πυκνότητας πιθανότητας έχει "μετακινηθεί" πιο κοντά στα δεδομένα λόγω της εκτίμησης των παραμέτρων που στηρίχτηκε σε αυτά, με αποτέλεσμα η μέγιστη διαφορά να έχει γίνει μικρότερη από ότι θα ήταν προ της εκτίμησης των παραμέτρων. Συνεπώς η κατανομή της ελεγχοσυνάρτησης του στατιστικού ελέγχου, θεωρώντας ότι η $H_0$ είναι αληθινή, είναι στοχαστικά μικρότερη από την κατανομή των Kolmogorov-Smirnov. Χρησιμοποιώντας λοιπόν τους πίνακες για τις κρίσιμες τιμές του στατιστικού ελέγχου του Lilliefors όταν η κατανομή της $\hat{F}_0(x)$ είναι η κανονική, θα απορρίπτουμε την μηδενική υπόθεση, για μεγάλες τιμές της στατιστικής συνάρτησης, καθώς θα υπάρχει απόκλιση από την συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής και από την συνάρτηση κατανομής του τυποποιημένου δείγματος.***[2],[11],[15],[19],[30],[41].***

### 2.2.3.2. ΕΛΕΓΧΟΣ LILLIEFORS ΓΙΑ ΤΗΝ ΕΚΘΕΤΙΚΗ ΚΑΤΑΝΟΜΗ

Παρόμοια με την προηγούμενη περίπτωση, αν θέλουμε να εξετάσουμε την περίπτωση που ένα σετ από παρατηρήσεις προέρχονται από την εκθετική κατανομή με άγνωστη τη μέση τιμή του πληθυσμού, υπολογίζουμε τη μέση τιμή του δείγματος για να την χρησιμοποιήσουμε ως εκτιμητή της άγνωστης παραμέτρου. Για κάθε τυχαία μεταβλητή $X_i$, ο μετασχηματισμός είναι ο εξής :

$$Z_i = \frac{X_i}{\bar{X}}, \quad για \ \ i = 1,2, \dots, n$$

Επομένως η ελεγχοσυνάρτηση του Lilliefors στην περίπτωση της εκθετικής κατανομής είναι η:

$$D_n{}^* = sup_{z\in R}\{|\hat{S}_n(z) - \hat{F}_0(z)|\}$$

όπου $\hat{S}_n(z)$ είναι η εμπειρική συνάρτηση κατανομής του δείγματος, που βασίζεται στα $Z_i$ και $\hat{F}_0(z) = 1 - e^{-z}$ για $z \geq 0$, η συνάρτηση κατανομής της εκθετικής κατανομής με παράμετρο 1. Αν η τιμή της $D_n{}^*$ υπερβαίνει την αντίστοιχη κρίσιμη τιμή του πίνακα, τότε η μηδενική υπόθεση, σύμφωνα με την οποία οι παρατηρήσεις προέρχονται από την Εκθετική κατανομή, απορρίπτεται.***[2],[11],[20].***

*Ακολουθούν τρεις ακόμα έλεγχοι που είναι της ίδιας λογικής με τα παραπάνω και ανήκουν στη δεύτερη ομάδα ελέγχων, ωστόσο για τον υπολογισμό τους δεν χρησιμοποιείται μόνο η μέγιστη απόσταση όπως στο Kolmogorov-Smirnov, αλλά όλες οι αποστάσεις-αποκλίσεις της υποτιθέμενης συνάρτησης κατανομής και της εμπειρικής συνάρτησης κατανομής σε όλες τις παρατηρήσεις* $x_1, x_2, \dots, x_n$. ***[40].***

### 2.2.4. ΕΛΕΓΧΟΣ CRAMER-VON MISES

Η δεύτερη ομάδα στατιστικών ελέγχων, που βασίζεται στην εμπειρική συνάρτηση κατανομής, ανήκει στην οικογένεια των Cramér-von Mises και έχει την εξής μορφή:

$$T_n = \int \psi(F_0(x))B_n^2(x)dF_0(x) \quad \textit{ορισμένο στον χώρο } S,$$

Οπότε με την εμπειρική στοχαστική διαδικασία που ορίσαμε στην ***ενότητα 1.1.1*** έχουμε:

$$T_n^2 = n\int_{-\infty}^{\infty}\{\hat{F}_n(x) - F_0(x)\}^2\psi(F_0(x))d\,F_0(x) \quad (2.12)$$

όπου $\psi(\cdot)$ είναι μια μη αρνητική συνάρτηση βάρους, διαλεγμένη με σκοπό να δώσει έμφαση στις τιμές της τετραγωνικής διαφοράς $\hat{F}_n(x) - F_0(x)$, όπου ο έλεγχος είναι επιθυμητό να έχει ευαισθησία. Η υπόθεση πρόκειται να απορριφθεί, αν η τιμή της $T_n^2$ είναι ικανοποιητικά μεγάλη. ***[1],[6],[37].***

Από τη σχέση $(2.12)$ , αν πάρουμε για συνάρτηση βάρους, $\psi(x) = 1$, και έχοντας ένα τυχαίο δείγμα, για το οποίο θέλουμε να εξετάσουμε τη σχέση $(2.4)$, ο έλεγχος των Cramér και Von Mises δίνεται από τον εξής τύπο:

$$W_n^2 = n\int_{-\infty}^{\infty}\{\hat{F}_n(x) - F_0(x)\}^2 d\,F_0(x) \quad (2.13)$$

όπου ως γνωστόν, $\hat{F}_n(x)$ είναι η εμπειρική συνάρτηση κατανομής και $F_0(x)$ η συνάρτηση κατανομής που έχουμε υποθέσει κάτω από τη μηδενική υπόθεση. Η κατανομή του $W_n^2$, κάτω από την $H_0$, είναι ίδια για όλες τις $F_0$, οι οποίες είναι συνεχείς. Συνεπώς τώρα υποθέτουμε ότι $F_0(x) = x$. Η γενική ιδέα είναι να θεωρήσουμε την κατασκευή των βασικών συνιστωσών των Kac-Siegert της εμπειρικής στοχαστικής διαδικασίας.

Εύκολα προκύπτει ότι :

$l_j(x) = \sqrt{2}\sin(\pi jx)$ , για $j = 1,2, \dots$ και $0 \le x \le 1$ είναι οι ιδιοσυναρτήσεις

και $\lambda_j = \frac{1}{\pi^2 j^2}$ οι αντίστοιχες ιδιοτιμές

Ενώ για την εμπειρική διαδικασία $P_n(x) = \sqrt{\psi(x)}B_n(x) = B_n(x) = \sqrt{n}\left(\hat{F}_n(x) - x\right)$, χρησιμοποιώντας τα αποτελέσματα των Kac και Siegert, παρατηρούμε ότι μπορεί να πάρει τη μορφή:

$$P_n(x) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, l_j(x) Z_{n,j}$$

Οι τυχαίες μεταβλητές $Z_{n,j}$ είναι και οι βασικές συνιστώσες του $\sqrt{n}\left(\hat{F}_n(x) - x\right)$

$$Z_{n,j} = \frac{1}{\sqrt{\lambda_j}} \int_0^1 P_n(x) l_j(x)\, dx \qquad (2.14)$$

μπορούν να απλοποιηθούν ως εξής:

$$Z_{n,j} = \frac{1}{\sqrt{\lambda_j}} \int_0^1 P_n(x) l_j(x)\, dx = \sqrt{2} j\pi \int_0^1 \sqrt{n}\left(\hat{F}_n(x) - x\right) \sin(\pi jx)\, dx =$$

$$= \sqrt{2n} j\pi \left[\int_0^1 \hat{F}_n(x) \sin(\pi jx)\, dx - \int_0^1 x \sin(\pi jx)\, dx\right] =$$

$$= \sqrt{2n} j\pi \int_0^1 \hat{F}_n(x) \sin(\pi jx)\, dx = \sqrt{2n} \int_0^1 \cos(\pi jx) d\hat{F}_n(x) =$$

$$= \sqrt{2n} \frac{1}{n} \sum_{i=1}^{n} \cos(j\pi X_i) = \sqrt{\frac{2}{n}} \sum_{i=1}^{n} \cos(j\pi X_i)$$

και οι οποίες είναι εύκολο να επαληθεύσει κανείς, ότι κάτω από τη μηδενική υπόθεση, είναι ασυμπτωτικά τυπικά κανονικά κατανεμημένες και ανεξάρτητες.

Στην πραγματικότητα, αν έχουμε μόνο τη τιμή του $W_n^2$, λίγα μπορούμε να πούμε για το αν η φύση της διαφοράς είναι σημαντική ή όχι. Αυτό που επιθυμούμε είναι να δούμε συγκεκριμένα ποια πλευρά των δεδομένων, δίνει έναυσμα για ένα "σημαντικό" αποτέλεσμα, κάτι το οποίο μπορούμε να πετύχουμε αν το $W_n^2$, μπορεί να χωριστεί σε ένα κατάλληλο σετ από ξεχωριστές συνιστώσες, η καθεμία από τις οποίες μετρά κάποια ξεχωριστή πλευρά των δεδομένων. Γενικά, η λεπτομερής εξέταση της ερμηνείας τους, μας δίνει περισσότερη πληροφορία για την συμπεριφορά του στατιστικού ελέγχου, κάτω από τις εναλλακτικές. Το ερώτημα δηλαδή, του κάτω από ποιες εναλλακτικές, η τιμή του στατιστικού ελέγχου γίνεται μεγάλη, τώρα μεταφράζεται στο ερώτημα κάτω από ποιες εναλλακτικές, οι συνιστώσες $Z_{n,j}$ προσδοκάτε να είναι διάφορες του μηδενός. Τέλος επειδή τα βάρη των συνιστωσών μειώνονται γρήγορα ($\frac{1}{\pi^2 j^2} \to 0$), καθώς το $j \to \infty$, είναι ιδιαίτερα σημαντικό να κατανοήσουμε τις χαμηλότερης-ταξινόμησης συνιστώσες.

Οι βασικές συνιστώσες του στατιστικού ελέγχου των Cramér-von Mises, μπορούν να βρεθούν από την εφαρμογή του θεωρήματος του Parseval, με αποτέλεσμα το

τεστ να μπορεί να αναπαρασταθεί ως ένα άπειρο άθροισμα από ασυμπτωτικά ανεξάρτητες τετραγωνικές συνιστώσες με βάρη. Δηλαδή :

$$T_n = W_n^2 = \int_0^1 B_n^2(x)\,dx = \int_0^1 \left[\sum_{j=1}^{\infty} \sqrt{\lambda_j} l_j(x)\, Z_{n,j}\right]^2 dx =$$

$$= \sum_{j=1}^{\infty}\sum_{m=1}^{\infty} \sqrt{\lambda_j \lambda_m}\, Z_{n,j} Z_{n,m} \int_0^1 l_j(x) l_m(x)\,dx =$$

$$= \sum_{j=1}^{\infty} \lambda_j\, {Z_{n,j}}^2 = \sum_{j=1}^{\infty} \frac{1}{\pi^2 j^2} {Z_{n,j}}^2$$

Επομένως από τη θεωρία συνεπάγεται ότι κάτω από τη μηδενική υπόθεση έχουμε

$$W_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{\pi^2 j^2} Z_j^2$$

όπου τα $Z_1, Z_2, \dots$ είναι μια ακολουθία από ανεξάρτητες τυποποιημένες κανονικές τυχαίες μεταβλητές. Ενώ ταυτόχρονα συνεπάγεται ότι ο έλεγχος είναι κατά σημείο συνεπής στην ισχύ, έναντι σε κάθε εναλλακτική $F \neq F_0$, για την οποία :

$$E_F\left[T_j(X_1)\right] = \int_0^1 \sqrt{2}\cos(\pi j x) dF(x) \neq \int_0^1 \sqrt{2}\cos(\pi j x) dF_0(x) = 0 \quad \textit{για κάποια } j$$

Αλλά το ότι

$$\int_0^1 \cos(\pi j x) dF(x) = 0, \quad \textit{για όλα τα } j = 1,2,\dots$$

υπονοεί ότι $F = F_0$. ***[1],[7],[16],[37].***

Ωστόσο αν $x_1, x_2, \dots, x_n$ οι παρατηρήσεις του δείγματος, μεγέθους $n$, ο στατιστικός έλεγχος της σχέσης **(2. 13)** μπορεί να λάβει την υπολογιστική μορφή:

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(t_i - \frac{2i-1}{2n}\right)^2 \qquad \textbf{(2. 15)}$$

όπου τα $t_i$ δίνονται από την σχέση $t_i = F_0(x_i),\ i = 1,2,\dots,n$

Η ασυμπτωτική κατανομή του κριτηρίου αυτού καθώς και τα κρίσιμα σημεία, έχουν δοθεί από τους Anderson-Darling.

Λόγω της δυσκολίας γραφής της ασυμπτωτικής κατανομής του ελέγχου, υπάρχει τροποποιημένος στατιστικός έλεγχος για την Κανονική κατανομή με άγνωστη μέση τιμή και διασπορά, που είναι ο εξής:

$$W_n^{2^*} = W_n^2 \left(1 + \frac{0{,}5}{n}\right)$$

***[6],[7],[16],[38],[40].***

### 2.2.5. ΕΛΕΓΧΟΣ WATSON

Μια διαφοροποίηση του ελέγχου των Cramér-von Mises αποτελεί ο έλεγχος του Watson, ο οποίος είχε αρχικά προταθεί για τον έλεγχο καλής προσαρμογής κατανομών πάνω σε κύκλο και ορίζεται ως εξής:

$$U_n^2 = n\int_{-\infty}^{\infty}\left\{\hat{F}_n(x) - F_0(x) - \int_{-\infty}^{\infty}\left[\hat{F}_n(y) - F_0(y)\right]dF_0(y)\right\}^2 d\,F_0(x) \quad (2.16)$$

Ο τύπος της σχέσης (2.16) συνήθως χρησιμοποιείται στη μελέτη θεωρητικών ιδιοτήτων του ελέγχου. Ωστόσο επειδή όταν μετράμε σε κύκλο, δεν υπάρχει μια φυσική αρχή, ένας έλεγχος καλής προσαρμογής για τέτοια δεδομένα, θα πρέπει φυσικά να είναι αμετάβλητος-ανεξάρτητος από την επιλογή αυτής της αρχής, κάτι το οποίο φαίνεται στον παρακάτω τύπο, ο οποίος συνεπάγεται από τη σχέση (2.16).

$$U_n^2 = n\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{(\hat{F}_n(x) - \hat{F}_n(y)) - (F_0(x) - F_0(y)\right\}^2 d\,F_0(x)dF_0(y)$$

Γενικά, κάποιος μπορεί να ερμηνεύσει αυτόν τον τύπο, σαν το μέσο όρο των διαφορών της εμπειρικής πιθανότητας μιας παρατήρησης που είναι στο διάστημα $[x, y]$ και της αντίστοιχης πιθανολογούμενης πιθανότητας.

Παρόλο τον αρχικό σκοπό παραγωγής του ελέγχου, είναι πλέον αποδεκτό ότι μπορεί να χρησιμοποιηθεί και για τον έλεγχο καλής προσαρμογής στη γραμμή των πραγματικών αριθμών. Ενώ τέλος, για τις εφαρμογές είναι χρήσιμη η παρακάτω γραφή του ελέγχου:

$$U_n^2 = W_n^2 - n\left(\bar{t} - \frac{1}{2}\right)^2$$

όπου $W_n^2$ είναι ο έλεγχος των Cramér-von Mises που δίνεται από τη σχέση (2.15) και $\bar{t}$ είναι ο μέσος όρος των $t_i$ για $i = 1,2,\dots,n$, όπου $t_i = F_0(x_i)$.

Όπως και στον έλεγχο των Cramér-von Mises, η εμπειρική στοχαστική διαδικασία του κριτηρίου του Watson, μπορεί να γραφτεί ως εξής :

$$P_n(x) = B_n(x) - \int_0^1 B_n(y)\,dy$$

Συνεπώς $U_n^2 = \int_0^1 P_n(x)^2\,dx$.

Έχει ως ιδιοτιμές τις $\lambda_{2,j-1} = \lambda_{2,j} = \frac{1}{4\pi^2 j^2}$, ενώ οι ιδιοσυναρτήσεις είναι $l_{2,j-1}(x) = \sqrt{2}sin(2\pi jx)$ και $l_{2,j}(x) = \sqrt{2}cos(2\pi jx)$, $j = 1,2,\dots$. Ο στατιστικός έλεγχος του Watson μπορεί να αναπτυχθεί με τις βασικές του συνιστώσες ως εξής :

$$U_n^2 = \sum_{j=1}^{\infty}\frac{1}{4\pi^2 j^2}\left({Y_{n,j}}^2 + {Z_{n,j}}^2\right) \quad (2.17)$$

όπου $Y_{n,j} = \sqrt{\frac{2}{n}}\sum_{i=1}^{n}\cos(2\pi jX_i)$ και $Z_{n,j} = \sqrt{\frac{2}{n}}\sum_{i=1}^{n}sin(2\pi jX_i)$ (2.18)

Η ασυμπτωτική κατανομή της ελεγχοσυνάρτησης του ελέγχου Watson, κάτω από τη μηδενική υπόθεση, βασίζεται στις βασικές του συνιστώσες. Δηλαδή επειδή οι συνιστώσες της σχέσης (2.18) συγκλίνουν σε ανεξάρτητες τυπικές κανονικές μεταβλητές, η εξίσωση της σχέσης (2.17) υπονοεί ότι το $U_n^2$ συγκλίνει στο $U^2$, το οποίο ορίζεται ως:

$$U^2 = \sum_{j=1}^{\infty} \frac{1}{4\pi^2 j^2} ({Z_{2,j-1}}^2 + {Z_{2,j}}^2)$$

όπου τα $Z_j$ $(j = 1,2, \dots)$ είναι ανεξάρτητα και ακολουθούν την τυπική κανονική κατανομή. Συνεπώς μπορούμε να γράψουμε

$$U^2 = \sum_{j=1}^{\infty} \frac{1}{4\pi^2 j^2} {X_j}^2$$

όπου τα ${X_j}^2$ είναι ανεξάρτητα, ακολουθώντας την κατανομή $X_2^2$.

Τέλος η ασυμπτωτική κατανομή του $U_n^2$ είναι

$$\lim_{n\to\infty} P[\, U_n^2 \le z] = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\pi^2 z}$$

Για το $U_n^2$ υπάρχουν πίνακες με κρίσιμα σημεία, που έχουν δοθεί από τους Watson και Stephens. ***[6],[37],[40].***

### 2.2.6. ΕΛΕΓΧΟΣ ANDERSON-DARLING

Τώρα αν στη σχέση **(2.12)** πάρουμε την $\psi(x) = \frac{1}{F_0(x)(1-F_0(x))}$, ως συνάρτηση βάρους, κάτι το οποίο φανερώνει ότι τοποθετείται περισσότερο βάρος στις παρατηρήσεις που βρίσκονται στις ουρές της κατανομής, έχουμε τον έλεγχο των Anderson-Darling. Δηλαδή

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[\hat{F}_n(x) - F_0(x)]^2}{F_0(x)\big(1 - F_0(x)\big)} dF_0(x) \qquad \mathbf{(2.19)}$$

Οι Anderson-Darling ανακάλυψαν ότι για ένα σετ από παρατηρήσεις, ο έλεγχος τους παράγει το ίδιο αποτέλεσμα με αυτό των Kolmogorov-Smirnov, ωστόσο είναι καλύτερος στο να εντοπίζει πολύ μικρές αλλαγές, ακόμα και σε μεγάλα μεγέθη δείγματος, ενώ τέλος είναι πιο ευαίσθητος στις διαφορές που εντοπίζονται στις ουρές της κατανομής.

Όπως στην περίπτωση του Cramér-von Mises, εξετάζουμε $F(x) = F_0(x) = x$, θεωρώντας τον έλεγχο των Anderson-Darling, ο οποίος γίνεται ως εξής:

$$A_n^2 = n \int_0^1 \frac{[\hat{F}_n(x) - x]^2}{x(1-x)} dx$$

Είναι εύκολο να αποδειχθεί ότι ο $A_n^2$ μπορεί να γραφτεί με τη μορφή τετραγωνικού στατιστικού ελέγχου με βάρη. Οι συνιστώσες του έχουν την ίδια μορφή με τον έλεγχο των Cramér-von Mises, στη σχέση **(2.14)**, αλλά τώρα με ιδιοσυναρτήσεις και αντίστοιχες ιδιοτιμές που δίνονται από τους τύπους:

$$l_j(x) = 2\sqrt{\frac{1}{j(j+1)}}\sqrt{x(1-x)}\,\frac{d}{dx} L_j(x)$$

όπου τα $L_j$ δηλώνουν τα ορθοκανονικά πολυώνυμα Legendre.

$$\text{και} \qquad \lambda_j = \frac{1}{j(j+1)} \quad \text{αντίστοιχα}$$

Η εμπειρική διαδικασία του ελέγχου, γράφεται με τη μορφή :

$$P_n(x) = \sqrt{\psi(x)} B_n(x) = \frac{B_n(x)}{\sqrt{x(1-x)}}$$

Επομένως οι συνιστώσες παίρνουν την μορφή:

$$Z_{n,j} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_j(x_i)$$

Μετά από παρόμοιους υπολογισμούς με τον Cramér-von Mises έλεγχο, έχουμε ότι

$$T_n = A_n^2 = \sum_{j=1}^{\infty} \frac{1}{j(j+1)} {Z_{n,j}}^2$$

ενώ από τη θεωρία συνεπάγεται ότι

$$A_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} {Z_j}^2$$

Ισοδύναμα μπορούμε να γράψουμε τη σχέση **(2.19)** ως εξής :

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)\{lnt_i + \ln(1 - t_{n-i+1})\}$$

όπου τα $t_i = F_0(x_i)$ , $i = 1,2, \dots, n$ , και οι λογάριθμοι είναι οι φυσικοί λογάριθμοι. Αν λοιπόν η τιμή του $A_n^2$ είναι αρκετά μεγάλη, τότε η μηδενική υπόθεση, θα απορριφτεί. Η διαδικασία αυτή, ίσως φανεί χρήσιμη σε κάποιον που επιθυμεί να απορρίψει τη μηδενική υπόθεση, οποτεδήποτε η πραγματική κατανομή διαφέρει σημαντικά από την υποτιθέμενη και ειδικά όταν αυτή διαφέρει στις ουρές.***[1],[16], [30],[37],[40].***

Αν πάρουμε $n = 1$, η κατανομή είναι η εξής :

$$P[A_1^2 < z] = P\left[1 - \ln\left(t_1(1-t_1)\right) < z\right] = \sqrt{1 - 4e^{-1-z}} \ ,$$
$$\gamma\iota\alpha \ z > \ln(4) - 1 = 0{,}38629.$$

Ωστόσο για $n = 2$, η κατανομή είναι αρκετά δύσκολο να υπολογιστεί, αν ενσωματώσουμε νούμερα. Στην πραγματικότητα, σημεία σημαντικότητας για το $A_n^2$ δεν είναι διαθέσιμα για δείγματα μικρού μεγέθους, παρά μόνο ασυμπτωτικά. Δηλαδή για $n > 2$ και $n \leq 8$, όλες αυτές οι τιμές είναι διαθέσιμες σε προσομοιωμένο πίνακα από τον Lewis. Ο καθορισμός της κατανομής του $A_n^2$ για $n > 8$ με τις μεθόδους Monte Carlo είναι απαγορευτικός, καθώς είναι πολύ χρονοβόρος. Ευτυχώς η σύγκλιση της κατανομής του $A_n^2$ στην ασυμπτωτική του κατανομή είναι πάρα πολύ γρήγορη, ενώ μάλιστα έχει παρατηρηθεί ότι η προσέγγιση είναι ικανοποιητική ακόμα και για $n = 5$.

Από τον Lewis δόθηκε επίσης η ασυμπτωτική κατανομή του ελέγχου, η οποία είναι:

$$\lim_{n\to\infty} P(A_n^2 < z) = \frac{1}{\mathrm{z}}\sum_{\mathrm{j}=0}^{\infty}\binom{-\frac{1}{2}}{\mathrm{j}}(4\mathrm{j}+1)f(z,j)$$

όπου $f(z,j) = \sqrt{2\pi}e^{-t_j}\int_0^{\infty} e^{\frac{z}{8(1+\omega^2)}-\omega^2 t_j}d\omega$, και $t_j = (4j+1)^2\pi^2/(8z)$

Δηλαδή το δύσκολο είναι ο υπολογισμός του $f(z,j)$. ***[18],[21].***

Αυτό που προκύπτει ωστόσο από τα παραπάνω, είναι ότι για την απλή μηδενική υπόθεση, δεν υπάρχει απλή αναλυτική έκφραση της συνάρτησης της ασυμπτωτικής κατανομής και γι αυτό ουσιαστικά είναι διαθέσιμες προσεγγίσεις. Για παράδειγμα για την Κανονική κατανομή με μέση τιμή και διασπορά άγνωστη, έχει προταθεί ένας τροποποιημένος Anderson-Darling στατιστικός έλεγχος, ο οποίος είναι:

$$A_n^{2^*} = A_n^2\left(1 + \frac{0{,}75}{n} + \frac{2.25}{n^2}\right)$$

***[6],[30],[37],[38].***

Στους ελέγχους των Cramér-von Mises και Anderson-Darling, όταν η μηδενική σύνθετη υπόθεση πρέπει να εξεταστεί, πορευόμαστε όπως στον έλεγχο των Kolmogorov-Smirnov, υπολογίζοντας την άγνωστη παράμετρο από τα δεδομένα. Οπότε η στατιστική ελεγχοσυνάρτηση αυτών των ελέγχων καθώς και η εμπειρική διαδικασία, υπολογίζεται με τον εκτιμητή πλέον, έχοντας βέβαια μια επίπτωση στην ασυμπτωτική κατανομή της ελεγχοσυνάρτησης, κάτω από τη μηδενική υπόθεση. ***[37].*** Τέλος η θεωρία για τις κυκλικές κατανομές είναι περισσότερο πολύπλοκη και γι' αυτό το λόγο παραλείπουμε να μιλήσουμε για τη σύνθετη μηδενική υπόθεση.

## 2.3. ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΝΝΟΙΑ ΤΗΣ ΕΝΤΡΟΠΙΑΣ ΚΑΙ ΒΑΣΙΚΕΣ ΙΔΙΟΤΗΤΕΣ

Το 1936, όταν ο Mahalanobis εισήγαγε την έννοια της απόστασης μεταξύ δύο κατανομών πιθανότητας, πολλά μέτρα απόστασης μεταξύ δύο κατανομών έκαναν την εμφάνισή τους στη στατιστική βιβλιογραφία. Μέσω αυτών ουσιαστικά εκφράζεται το γεγονός ότι κάποιες κατανομές πιθανότητας είναι "περισσότερο κοντά" σε σχέση με κάποιες άλλες και συνεπώς μας διευκολύνουν να διακρίνουμε, αν ένα ζεύγος κατανομών είναι "μακριά" ή "κοντά" το ένα από το άλλο. Τέτοια μέτρα απόστασης είναι τα divergence measures (με πιο γνωστό το Kullback-Leibler), ενώ πολύ ενδιαφέρον και σημαντικό εργαλείο στη Θεωρία της Στατιστικής Πληροφορίας είναι τα μέτρα εντροπίας, ιδιαίτερα χρήσιμα στους ορισμούς των μέτρων αποκλίσεων που μελετήθηκαν από τους Burbea και Rao. ***[25].***

Το 1948, πρώτος ο Shannon C.E., βρήκε ένα *μέτρο πληροφορίας (measure of information)* της αβεβαιότητας, επεκτείνοντας με αυτό τον τρόπο προϋπάρχουσες έννοιες πληροφορίας (όπως είναι του Hartley), για ενδεχόμενα που δεν έχουν απαραίτητα την ίδια πιθανότητα εμφάνισης. Το μέτρο πληροφορίας όπως ορίστηκε από τον Shannon, δίνεται από τη σχέση **(2.20)**:

$$H(X) = \sum_i p_i log\frac{1}{p_i} = -\sum_i p_i log\, p_i \qquad \mathbf{(2.20)}$$

Αυτή η ποσότητα που παρουσιάστηκε, προκάλεσε μεγάλη σύγχυση στην Θεωρία Πληροφορίας και στη Θερμοδυναμική, καθώς είχε την ίδια μαθηματική φόρμα με το μέτρο της εντροπίας του Boltzmann. Το μέτρο του Shannon ονομάστηκε ***εντροπία του Shannon (Shannon's entropy).***

Αργότερα το 1984, ο Forte παρουσίασε την εντροπία των Boltzmann-Shannon, η οποία δίνεται από τον τύπο $(2.21)$. Αν $X$ είναι μια τυχαία μεταβλητή, η οποία παίρνει τιμές στο χώρο $\boldsymbol{X} \subset R^n$, υποθέτουμε ότι η μορφή της συνάρτησης κατανομής $F$ της $X$ είναι γνωστή, εκτός ίσως από ένα πεπερασμένο αριθμό παραμέτρων (έστω $\theta$ το διάνυσμα των άγνωστων παραμέτρων). Ορίζουμε τον στατιστικό χώρο $(\boldsymbol{X}, \beta_{\boldsymbol{X}}, P_\theta)_{\theta\in\Theta}$, όπου$\beta_{\boldsymbol{X}}$ είναι ο $\sigma$-χώρος των Borel υποσυνόλων $A \subset \boldsymbol{X}$ και $\{P_\theta\}_{\theta\in\Theta}$ μια οικογένεια συναρτήσεων πιθανότητας ορισμένη στο μετρήσιμο χώρο $(\boldsymbol{X}, \beta_{\boldsymbol{X}})$, όπου $\Theta$ ανοιχτό υποσύνολο του $R^{M_0}$, με $M_0 \geq 1$. Αν υποθέσουμε ότι $P_\theta$ είναι συνεχείς συναρτήσεις πιθανότητας, με ένα $\sigma$-πεπερασμένο μέτρο $\mu$ στον $(\boldsymbol{X}, \beta_{\boldsymbol{X}})$ και $S_{\boldsymbol{x}}$, το στήριγμα της $P_\theta$ , αν $x \in S_{\boldsymbol{x}}$ τότε :

$$H(X) \equiv H(P_\theta) \equiv H(\theta) = -\int_{-\infty}^{\infty} f_\theta(x)\log\big(f_\theta(x)\big)\,d\mu(x) = E_\theta\left[-log f_\theta(X)\right] \quad (2.21)$$

όπου

$$f_\theta(x) = \frac{dP_\theta}{d\mu}(x) = \begin{cases} f_\theta(x) \ , & \text{αν } \mu \text{ είναι μέτρο Lebesgue} \\ P_\theta(X = x) = p_\theta(x), & \text{αν } \mu \text{ είναι αριθμήσιμο μέτρο} \end{cases}$$

Ουσιαστικά αυτή η εντροπία μετρά το βαθμό της αβεβαιότητας που έχει κάποιος σχετικά με την τιμή $x$ μιας τυχαίας μεταβλητής $X$, η οποία παίρνει πραγματικές τιμές, δεδομένης της συνάρτησης πυκνότητας πιθανότητας $f_\theta(x)$.

Μια πολύ σημαντική παρατήρηση είναι ότι για συνεχείς κατανομές, ο ορισμός της απλής εντροπίας του Shannon (σχέση$(2.20)$) , παύει να είναι χρήσιμος. Επομένως ο Jaynes (1963, 1968, 2003) έδωσε την ακόλουθη μορφή:

$$H_c = -\int_{\boldsymbol{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx \ = -E\left(log\frac{p(x)}{q(x)}\right)$$

υποδηλώνοντας στην πραγματικότητα δηλαδή, ότι η $H_c$ είναι ίση με την αρνητική *σχετική εντροπία (relative entropy)*, γνωστή ως *απόκλιση Kullback-Leibler* μεταξύ των πιθανοτήτων πυκνότητας $q(x)$ και $p(x)$. Η ποσότητα αυτή, εκφράζει το ποσό της πληροφορίας που χάνεται όταν μια συνάρτηση πιθανότητας χρησιμοποιείται προκειμένου να εκτιμηθεί μια άλλη. Γενικά, μας ενδιαφέρει να βρούμε μια κατανομή με συνάρτηση πυκνότητας πιθανότητας την $p(x)$, που είναι όσο το δυνατόν πιο "κοντά" στην γνωστή κατανομή με συνάρτηση πυκνότητας πιθανότητας την $q(x)$, ή ισοδύναμα την ελαχιστοποίηση της "απόστασης" Kullback-Leibler $H_c$ ως προς $p$ (η διαδικασία ονομάζεται Principle of Minimum Discrimination Information). Τέλος μπορούμε να αντιληφθούμε εύκολα ότι σχετίζεται άμεσα με την εντροπία Shannon, καθώς είναι γνωστό ότι η σχέση $(2.20)$ για μια διακριτή τυχαία μεταβλητή, δεν θεωρείται "ανάλογη" της εντροπίας Boltzmann-Shannon για μια συνεχή μεταβλητή. Το 1988, λοιπόν οι Forte και Hughes πρότειναν τη συνάρτηση:

$$\bar{H} = -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{x_i - x_{i-1}}\right) \qquad (2.22)$$

ως ένα αρκετά ικανοποιητικό ορισμό για την διακριτή περίπτωση ανάλογη της (2.21). Έστω μια μεταβλητή ορισμένη στο $[a, b]$, αν η "πληροφορία" μας είναι δοσμένη από το εξής σύστημα

$$p_i = P[x_{i-1} < X \leq x_i] = \int_{x_{i-1}}^{x_i} f(x)dx, \quad i = 1,2, \dots, n-1$$

$$\text{και} \quad a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b \qquad (2.23)$$

Τότε

$$\lim_{\max_i |x_i - x_{i-1}| \to 0} -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{x_i - x_{i-1}}\right) = -\int f \log f = H(f)$$

Οπότε, μπορούμε να επιβεβαιώσουμε ότι η σχέση (2.22) είναι πράγματι το διακριτό ανάλογο της συνεχούς Boltzmann-Shannon εντροπίας. Βρήκαμε λοιπόν ότι $\widehat{H}$, η μέγιστη εντροπία κάθε τυχαίας μεταβλητής, με πυκνότητα πιθανότητας $f(x)$ και η οποία ικανοποιεί την σχέση (2.23) είναι η

$$\widehat{H} = -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{x_i - x_{i-1}}\right) = H(s) = \bar{H}$$

όπου $s$ είναι η εξής απλή συνάρτηση:

$$s(x) := \begin{cases} 0, & \alpha\nu \ x \leq x_0 \\ \dfrac{p_i}{x_i - x_{i-1}}, & \alpha\nu \ x_i < x \leq x_{i+1} \\ 0, & \alpha\nu \quad x_n \leq x \end{cases}$$

***[3],[8],[17].***

*Γενικά υπάρχουν κάποιες* ***βασικές ιδιότητες*** *του μέτρου της εντροπίας του Shannon , οι οποίες εμφανίζονται στο παρακάτω αποτέλεσμα:*

***ΠΡΟΤΑΣΗ 2.3.1:***

*Έστω ότι έχουμε $X = (X_1, X_2, \dots, X_n)$ και $Y = (Y_1, Y_2, \dots, Y_n)$ δύο συνεχή τυχαία διανύσματα με από κοινού συναρτήσεις πυκνότητας πιθανότητας $f_1(x), x \in R^n$και $f_2(y), y \in R^m$, αντίστοιχα. Τότε υποθέτουμε ότι $(X, Y)$ είναι επίσης ένα συνεχές τυχαίο διάνυσμα με συνάρτηση πυκνότητας πιθανότητας $f(x, y)$, με $(x, y) \in R^{n+m}$. Η δεσμευμένη πυκνότητα πιθανότητας της $X$ όταν $Y = y$ δίνεται από $f(x,y) \Big/ f_2(y)$.*

*Συνεπώς η δεσμευμένη εντροπία (conditional entropy) του Shannon του $X$, δεδομένου του $Y = y$, ορίζεται ως εξής :*

$$H(X|Y = y) = -\int_{R^n} \frac{f(x,y)}{f_2(y)} \log \frac{f(x,y)}{f_2(y)} dx$$

*Ενώ η δεσμευμένη εντροπία του Shannon του $X$, δεδομένου του $Y$ από*

$$H(X|Y) = -\int_{R^{n+m}} f(x,y) \log \frac{f(x,y)}{f_2(y)} dxdy$$

$$= \int_{R^m} f_2(y)\ H(X|Y=y)dy$$

*υποθέτοντας την ύπαρξη της προηγούμενης εντροπίας. Χρησιμοποιώντας λοιπόν την εντροπία του Shannon, επαληθεύονται οι παρακάτω ιδιότητες.*

- Η εντροπία του Shannon μπορεί να είναι αρνητική.

Πράγματι αν $X$ είναι μια τυχαία μεταβλητή με εκθετική κατανομή, παραμέτρου $\theta > 0$. Τότε έχουμε ότι

$$H(\theta) = -\int_0^\infty \theta e^{-\theta x} \log\left(\theta\, e^{-\theta x}\right) dx = 1 - log\theta$$

Συνεπώς αν $\theta \in (0, e)$, η εντροπία είναι θετική, ενώ αν $\theta \in (e, \infty)$, η εντροπία είναι αρνητική.

- Αν $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_n)$ μια απλή αμφιμονοσήμαντη αντιστοιχία στον $\mathrm{R}^n$ και υποθέσουμε ότι $Y = \varphi(x)$, τότε

$$H(Y) = H(x) - \int_{\mathrm{R}^n} f(x) \log\ |J\big(\varphi(x)\big)| dx$$

όπου $J(y) = \det\left(\frac{\partial \psi_i}{\partial y_j}(y)\right)_{i,j=1,2,..,n}$ είναι η ορίζουσα του πίνακα Jacobi, που αντιστοιχεί με τον αντίστροφο μετασχηματισμό $\psi \equiv (\psi_1, \psi_2, \dots, \psi_n)$ της $\varphi$.

Αυτό αποδεικνύεται ως εξής : Αν $B \subset \mathrm{R}^n$,

$$P(Y \in B) = P\big(X \in \psi(B)\big) = \int_{\psi(B)} f(x)dx = \int_B (f(\,\psi(y))|J(y)|dy$$

όπου $\psi(B) = \{\psi(y) : y \in B\}$. Συνεπώς η συνάρτηση πυκνότητας πιθανότητας του τυχαίου διανύσματος Y δίνεται από $f(\psi(y))|J(y))$ και

$$H(Y) = -\int_{\mathrm{R}^n} f\big(\psi(y)\big)|J(y)|\log\left(f\big(\psi(y)\big)|J(y)|\right)dy =$$
$$= -\int_{\mathrm{R}^n} f(x)\log\big(f(x)\big|J\big(\varphi(x)\big)\big|\big)\,dx =$$
$$= H(x) - \int_{\mathrm{R}^n} f(x)\log\ |J\big(\varphi(x)\big)|dx$$

- Αν $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_n)$ ένας γραμμικός μετασχηματισμός, με $\varphi_i(x) = \sum_{j=1}^n a_{ij}x_j$ για $i = 1,2,\dots,n$, συνεπώς $H(Y) = H(X) + \log|\det(A)|$, όπου $A = \big(a_{ij}\big)_{i,j=1,2,..,n}$.

Αποδεικνύεται ότι αν η $\varphi$ είναι ένας γραμμικός μετασχηματισμός τότε έχουμε ότι $J(y) = detA^{-1}$, οπότε αν χρησιμοποιήσουμε αυτό το συμπέρασμα, στη δεύτερη ιδιότητα, θα πάρουμε το επιθυμητό αποτέλεσμα.

- Ισχύει η ανισότητα του ***λήμματος του Gibbs*** για συνεχή τυχαία διανύσματα

$$H(x) = -\int_{\mathrm{R}^n} f_1(x) log f_1(x) dx \le -\int_{\mathrm{R}^n} f_1(x) log f_2(x) dx$$

Η ισότητα ισχύει αν και μόνο αν $f_1(x) = f_2(x)$ σχεδόν βεβαίως.

Πράγματι αν $A = \{x: f_1(x) > 0\}$ και δεδομένου ότι $x \in A$, έχουμε ότι :

$$log\frac{f_2(x)}{f_1(x)} \leq \frac{f_2(x)}{f_1(x)} - 1$$

όπου η ισότητα ισχύει αν και μόνο αν $f_1(x) = f_2(x)$. Συνεπώς

$$f_1(x) log\frac{f_2(x)}{f_1(x)} \leq f_2(x) - f_1(x)$$

Αν ορίσουμε ότι $l = -\int_{R^n} f_1(x) log f_1(x) dx + \int_{R^n} f_1(x) log f_2(x) dx$

έχουμε ότι :

$$\begin{aligned} l &= -\int_A f_1(x) log f_1(x) dx + \int_A f_1(x) log f_2(x) dx \\ &\leq \int_A f_2(x) dx - \int_A f_1(x) dx \quad (1) \\ &\leq \int_{R^n} f_2(x) dx - \int_{R^n} f_1(x) dx \quad (2) \quad = 0 \end{aligned}$$

Η ισότητα(1) ισχύει αν και μόνο αν $f_1(x) = f_2(x)$, σχεδόν βεβαίως, στον χώρο A, ενώ η ισότητα (2)ισχύει αν και μόνο αν $f_2(x) = 0 = f_1(x)$ σχεδόν βεβαίως, στον χώρο $A^c$. Συνεπώς η ισότητα ισχύει αν και μόνο αν $f_1(x) = f_2(x)$ με πιθανότητα 1.

- Ισχύει ότι : $H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$

καθώς έχουμε

$$\begin{aligned} H(X,Y) &= -\int_{R^n}\int_{R^m} f(x,y)\, log\, f(x,y) dxdy = -\int_{R^n}\int_{R^m} f(x,y) \log\left(f_2(y) f(x|y)\right) dxdy = \\ &= -\int_{R^n}\int_{R^m} f(x,y) log\, (f_2(y)) dxdy - \int_{R^n}\int_{R^m} f(x,y) log\, (f(x|y)) dxdy \\ &= -\int_{R^n}\int_{R^m} f(x,y) log\, (f_2(y)) dxdy - \int_{R^n}\int_{R^m} f(x,y) log \frac{(f(x,y))}{f_2(y)}\, dxdy \\ &= H(Y) + H(X|Y) \end{aligned}$$

Με παρόμοιο τρόπο προκύπτει και η ισότητα $H(X,Y) = H(Y|X) + H(X)$

- Ισχύει ότι : $H(X|(Y_1,Y_2)) \leq H(X|Y_1) \leq H(X)$, δηλαδή η εκ των προτέρων γνώση μειώνει την αβεβαιότητα. Η πρώτη ανισότητα μπορεί να γίνει ισότητα αν και μόνο αν το τυχαίο διάνυσμα $Y_2$ είναι ανεξάρτητο από το $X$, δεδομένου του $Y_1$. Ενώ τέλος η δεύτερη ανισότητα μπορεί να γίνει ισότητα αν και μόνο αν το $Y_1$ είναι ανεξάρτητο από το $X$.

Θεωρούμε τις συναρτήσεις πυκνότητας πιθανότητας $g(y_1), g(y_1,y_2), s\,(x,y_1)$ και $r(x,y_1,y_2)$ των τυχαίων διανυσμάτων $Y_1, (Y_1,Y_2), (X,Y_1)$και $(X,Y_1,Y_2)$ αντίστοιχα. Συνεπώς έχουμε ότι :

$$\begin{aligned} H\big(X|(Y_1,Y_2)\big) &= -\iint g(y_1,y_2)\left(\int \frac{r(x,y_1,y_2)}{g(y_1,y_2)} log \frac{r(x,y_1,y_2)}{g(y_1,y_2)} dx\right) \\ &\leq -\iint g(y_1,y_2)\left(\int \frac{r(x,y_1,y_2)}{g(y_1,y_2)} log \frac{s\,(x,y_1)}{g(y_1)} dx \quad (3)\right. \end{aligned}$$

$$= -\iint s\,(x,y_1)log\frac{s\,(x,y_1)}{g(y_1)}dxdy_1 = H(X|Y_1)$$

όπου η ανισότητα $(3)$ προκύπτει αν εφαρμόσουμε μέρος από την τέταρτη ιδιότητα στις συναρτήσεις

$$\frac{r(x|y_1,y_2)}{g(y_1,y_2)} \quad και \quad \frac{s\,(x|y_1)}{g(y_1)}$$

ενώ η ισότητα ισχύει αν και μόνο αν

$$\frac{r(x,y_1,y_2)}{g(y_1,y_2)} = \frac{s\,(x,y_1)}{g(y_1)}$$

το οποίο είναι ισοδύναμο με

$$\frac{r(x,y_1,y_2)}{g(y_1)} = \frac{s\,(x,y_1)}{g(y_1)}\frac{g(y_1,y_2)}{g(y_1)}$$

Ο πρώτος λόγος στο δεξί μέλος της ισότητας είναι η συνάρτηση πυκνότητας πιθανότητας της $X$ δεδομένου $Y_1 = y_1$, ενώ ο δεύτερος είναι η συνάρτηση πυκνότητας πιθανότητας της $Y_2$ δεδομένου $Y_1 = y_1$. Ο λόγος στο αριστερό μέρος είναι η συνάρτηση πυκνότητας πιθανότητας των $(X, Y_2)$ δεδομένου $Y_1 = y_1$. Συνεπώς αποδείξαμε το πρώτο σκέλος της ιδιότητας, ενώ το δεύτερο αποδεικνύεται με παρόμοιο τρόπο.

- Ο κανόνας της αλυσίδας :

$$H(X_1,X_2,\dots,X_n) = H(X_1) + \sum_{k=2}^{n} H(X_k|X_1,\dots,X_{k-1}) \le \sum_{k=1}^{n} H(X_k) \qquad (2.24)$$

Η ισότητα ισχύει αν και μόνο αν τα $X_1, X_2, \dots, X_n$ είναι αμοιβαία ανεξάρτητα.

Πράγματι χρησιμοποιώντας προηγούμενη ιδιότητα έχουμε ότι

$$H(X_1,X_2,\dots,X_n) = H(X_1) + H(X_2,\dots,X_n|X_1)$$

Όπως κάναμε και στην απόδειξη της προηγούμενης ιδιότητας λοιπόν, είναι εφικτό να αποδείξει κάποιος ότι

$$H(X_2,\dots,X_n|X_1) = H(X_2|X_1) + H(X_3,\dots,X_n|X_1,X_2)$$

Επομένως επαναλαμβάνοντας τον ίδιο ισχυρισμό προκύπτει ότι

$$H(X_1,X_2,\dots,X_n) = H(X_1) + \sum_{k=2}^{n} H(X_k|X_1,\dots,X_{k-1})$$

- Τέλος έχουμε

$$H(X_1,X_2,\dots,X_n|Y) = H(X_1|Y) + \sum_{k=2}^{n} H(X_k|Y,X_1,\dots,X_{k-1}) \le \sum_{k=1}^{n} H(X_k|Y)$$

Πράγματι χρησιμοποιώντας το αποτέλεσμα της ιδιότητας $(2.24)$, προκύπτει η σχέση που θέλουμε να δείξουμε. ***[25].***

# ΚΕΦΑΛΑΙΟ 3:

# "ΕΛΕΓΧΟΣ ΜΕ ΒΑΣΗ ΤΗΝ ΜΕΓΙΣΤΗ ΕΝΤΡΟΠΙΑ"

Το 2011, οι Lee, Vonta και Karagrigoriou πρότειναν ένα διαφορετικό μέτρο πληροφορίας, το οποίο ουσιαστικά γενικεύει τον τύπο της σχέσης (2.22) του **κεφαλαίου 2** και χρησιμοποιείται σαν έλεγχος καλής προσαρμογής. Στόχος μας λοιπόν, σε αυτό το κεφάλαιο είναι να εξετάσουμε, μέσω προσομοιωμένων δεδομένων, αν αυτός ο έλεγχος που βασίζεται στη μέγιστη εντροπία (maximum entropy test), προσαρμόζεται καλά σε δεδομένα που προέρχονται από κατανομές με παχιές ουρές.

Αρχικά επομένως, θα δούμε τη στατιστική ελεγχοσυνάρτηση του ελέγχου, καθώς και την ασυμπτωτική της κατανομή κάτω από τη μηδενική κατανομή. Στη συνέχεια θα επικεντρωθούμε σε προσομοιωμένα δεδομένα, για μικρά προς μεσαία μεγέθη δείγματος, των οποίων η πραγματική κατανομή ανήκει στην κατηγορία των κατανομών με παχιές ουρές, με σκοπό να δούμε τη συμπεριφορά-δυνατότητες του στατιστικού ελέγχου στην συγκεκριμένη περίπτωση κατανομών. Σε γενικές γραμμές, αν ο έλεγχος της εντροπίας συμπεριφέρεται καλά σε προσομοιωμένα δεδομένα, ελπίζουμε ότι θα δίνει σωστά αποτελέσματα και για πραγματικά δεδομένα.

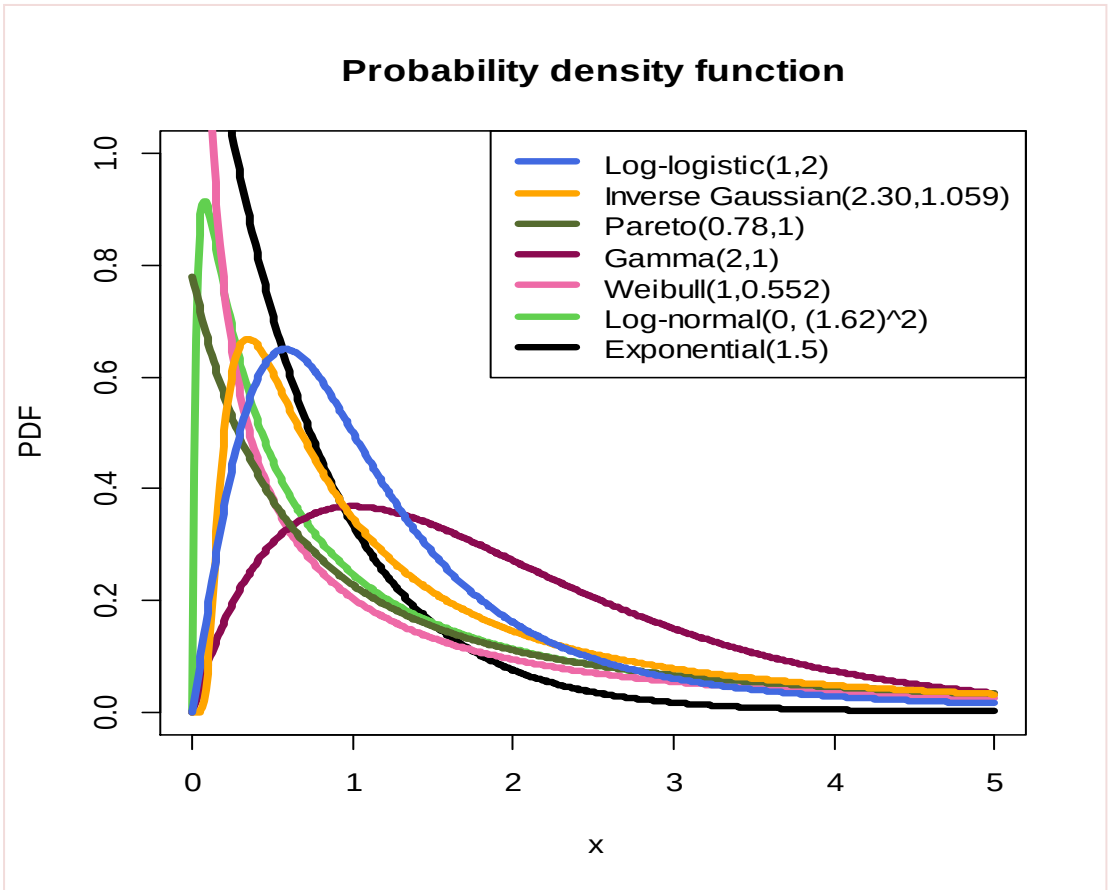
Ωστόσο, πρώτα από όλα, είναι αναγκαίο να αναφερθούμε σε κάποια βασικά χαρακτηριστικά των κατανομών αυτών.

## 3.1. ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΚΑΤΑΝΟΜΕΣ ΜΕ ΠΑΧΙΕΣ ΟΥΡΕΣ

Ένα χαρακτηριστικό γνώρισμα πολλών κατανομών, είναι ότι οι ουρές τους φθίνουν γρήγορα, σε αντίθεση με τις κατανομές όπου για μεγάλες τιμές του $x$ φθίνουν πολύ πιο αργά. Αυτές οι κατανομές περιγράφουν περιπτώσεις γεγονότων που χαρακτηρίζονται από μεγάλες ή απομακρυσμένες τιμές. Με άλλα λόγια, περιγράφουν παραδείγματος χάριν το ενδεχόμενο να συμβεί ένα ακραίο καιρικό φαινόμενο ή μια πολύ μεγάλη οικονομική καταστροφή. Αυτές οι κατανομές ονομάζονται ***κατανομές με παχιές ουρές (heavy-tailed distributions)*** και τείνουν να έχουν πολλά ακραία σημεία *(outliers),* με πολύ υψηλές/απομακρυσμένες τιμές. Με τον όρο ακραίες τιμές, εννοούμε εκείνες που είναι πιο απομακρυσμένες από τις υπόλοιπες παρατηρήσεις. Καθώς λοιπόν, υπάρχει μεγάλη διαφορά μεταξύ μέγιστης και ελάχιστης τιμής, στις κατανομές παχιάς ουράς η διακύμανση είναι πολύ πιο μεγάλη από ότι η μέση τιμή.

Ουσιαστικά οι ουρές αυτών των κατανομών δεν είναι εκθετικά οριοθετημένες, δηλαδή έχουν πιο “παχιές” ουρές σε σχέση με την εκθετική κατανομή. Γενικά όσο πιο “παχιά” είναι η ουρά, τόσο πιο μεγάλη είναι η πιθανότητα να πάρουμε μια ή περισσότερες δυσανάλογες τιμές στο δείγμα. Μπορεί μια κατανομή να έχει “παχιά”, κάποια από τις ουρές της, ενώ υπάρχουν περιπτώσεις όπου και οι δύο ουρές είναι “παχιές”. Ωστόσο σε πολλές εφαρμογές, περισσότερο ενδιαφέρον παρουσιάζει η περίπτωση η κατανομή να έχει παχιά δεξιά ουρά.

Οι περισσότερες heavy-tailed κατανομές που χρησιμοποιούνται στην πράξη, ανήκουν σε μια από τις οικογένειες κατανομών Pareto (και άλλες κατανομές νόμου δύναμης *(power-law distributions)*), Lognormal ή Weibull (με παράμετρο σχήματος (shape parameter) μεγαλύτερη του μηδενός και μικρότερη της μονάδας). Στο **Σχήμα 3.1** παρουσιάζονται ενδεικτικά μερικές κατανομές με παχιές ουρές και πώς αυτές απεικονίζονται σε σχέση με την Εκθετική κατανομή.




***Σχήμα 3.1:*** *“Κατανομές με παχιές ουρές, συγκρινόμενες με την Εκθετική κατανομή”*

Σε γενικές γραμμές οι κατανομές με παχιές ουρές παίζουν σημαντικό ρόλο στην ανάλυση πολλών στοχαστικών συστημάτων. Εμφανίζουν υψηλές μέσες τιμές, μεγάλες διασπορές, ενώ οι τιμές τους δεν μπορεί να είναι αρνητικές, όπως είναι για παράδειγμα στις περιπτώσεις του χρόνου επιβίωσης των ασθενών με καρκίνο, των βροχοπτώσεων στη μετεωρολογία και του μεγέθους των εισοδημάτων στην οικονομία. Υπάρχουν δηλαδή, πολλές στατιστικές ενδείξεις για την καταλληλότητά

τους στον τομέα της φυσικής και των οικονομικών, αποτελούν αναπόσπαστο κομμάτι στην περιγραφή πολλών διαδικασιών κινδύνου, ενώ τέλος βρίσκουν εφαρμογές σε μοντέλα επιδημιολογικής εξάπλωσης. Ενδεικτικά θα πούμε μερικά πράγματα, για κάποιες κατανομές.

Πιο συγκεκριμένα, κατανομές όπως η Weibull, συναντώνται συχνά στα μοντέλα επιβίωσης, όπου η ύπαρξη λογοκριμένων δεδομένων (censored data), κάνουν την επιλογή της κατάλληλης κατανομής ένα αρκετά ενδιαφέρον "πρόβλημα". Επίσης, εμπειρικά έχει αποδειχθεί ότι η Weibull αποτελεί ένα ικανοποιητικό μοντέλο για πολλά φαινόμενα στην ανάλυση αξιοπιστίας, όπως επίσης και οι κατανομές: Εκθετική, Γάμμα και Log-Normal, με πιο απλό αυτό της Εκθετικής και πιο ευέλικτο της Γάμμα. Γενικά είναι αρκετά διαδεδομένες και σε άλλες περιοχές της στατιστικής, όπως στην ανάλυση δεδομένων διάρκειας ζωής, όπου δηλαδή, μελετάμε τον χρόνο μέχρις ότου προκύψει ένα γεγονός, που συνήθως σχετίζεται με ένα ανεπιθύμητο ενδεχόμενο, καθώς και στη βιολογία και στις ιατρικές επιστήμες. Η κατανομή Log-Logistic, η οποία είναι παρόμοια στο σχήμα με τη Log-Normal κατανομή, αλλά έχει πιο παχιές ουρές, χρησιμοποιείται στην ανάλυση επιβίωσης, σαν παραμετρικό μοντέλο για γεγονότα των οποίων ο ρυθμός αυξάνεται αρχικά και μειώνεται στη συνέχεια, όπως για παράδειγμα ο ρυθμός θνησιμότητας από καρκίνο, ενώ τέλος βρίσκει εφαρμογή στη γεωφυσική, στην οικονομία και στην επιστήμη των υπολογιστών. Η κατανομή Inverse Gaussian(IG) (γνωστή και ως Wald) καθώς και η Log-Normal, είναι από τα βασικά μοντέλα στην περιγραφή δεδομένων που εμφανίζουν παχιές ουρές, τα οποία κυριαρχούν ιδιαίτερα σε κλάδους όπως η καρδιολογία, η κοινωνιολογία, και οι υπηρεσίες εύρεσης εργασίας. Ενώ τέλος η κατανομή Pareto χρησιμοποιείται στην περιγραφή κοινωνικού και ποιοτικού ελέγχου, στη γεωφυσική και τον αναλογισμό.

Αν $X$ μια τυχαία μεταβλητή που αντιπροσωπεύει το χρόνο όπου χρειάζεται μια διαδικασία για να ολοκληρωθεί και $F(x)$ η συνάρτηση κατανομής της, ορίζουμε ως $\bar{F}$ τη ***συνάρτηση ουράς*** της $F$, που δίνεται από τη σχέση $\bar{F}(x) = 1 - F(x) = F(x, \infty) = P(X > x)$ για όλα τα $x$. Η πιθανότητα αυτή, δηλαδή του ενδεχομένου η τυχαία μεταβλητή $X$ να πάρει τιμές μεγαλύτερες από μια σταθερά $x$, ονομάζεται *συνάρτηση επιβίωσης (**survival function**)* ή *συνάρτηση αξιοπιστίας **(reliability function).*** Μια άλλη συνάρτηση που συναντάμε συχνά, λέγεται *συνάρτηση κινδύνου (**hazard function**)* ή *αποτυχίας (failure)* και δίνεται από τη σχέση $h(x) = {f(x)}/{\bar{F}(x)}$,

όπου $f(x)$ είναι η συνάρτηση πυκνότητας πιθανότητας της $X$, ενώ $\boldsymbol{H}(x) = -ln\bar{F}(x)$ είναι η ***αθροιστική συνάρτηση κινδύνου***.
Τέλος θα περιγράφουμε ως *ιδιότητα ουράς **(tail property)*** της $F$ κάθε ιδιότητα που εξαρτάται μόνο από το $\{\bar{F}(x) : x \geq x_0\}$ για κάθε (πεπερασμένο) $x_0$.

***ΟΡΙΣΜΟΣ 3.1:*** *Μια κατανομή $F$ ορισμένη στο $\mathcal{R}$, θα λέμε ότι έχει **right-unbounded support**, όπως αναφέρεται στη διεθνή βιβλιογραφία, αν ισχύει :*

$$\bar{F}(x) > 0 \quad για\ όλα\ τα\ x$$

Μια σημαντική *σημείωση* είναι ότι ο ισχυρισμός μιας κατανομής $F$ ως heavy-tailed είναι μια ιδιότητα της ουράς της $F$ και όπως είναι φυσικό κάθε κατανομή με παχιά ουρά, έχει right-unbounded support.

Η προσοχή μας ουσιαστικά εστιάζεται στη συμπεριφορά των ουρών μιας τυχαίας μεταβλητής $X$ που παίρνει πραγματικές τιμές, δηλαδή στις πιθανότητες που έχουν τη μορφή $P(X > x)$ ή $P(X < -x)$ για $x$ μεγάλο ή αρνητικό, που αντιστοιχούν στις δεξιές και αριστερές ουρές αντίστοιχα της κατανομής της $X$. Γενικά η ουρά μιας κατανομής καθορίζει το μέγεθος και τη συχνότητα των ακραίων γεγονότων. Οι κατανομές, με βάση τη συμπεριφορά της ουράς τους, μπορούν να ταξινομηθούν σε δύο βασικές οικογένειες: τις *heavy-tailed* κατανομές και τις *light-tailed* κατανομές, με τις δεύτερες να έχουν πιο ήπια και με μικρότερη συχνότητα ακραία γεγονότα.

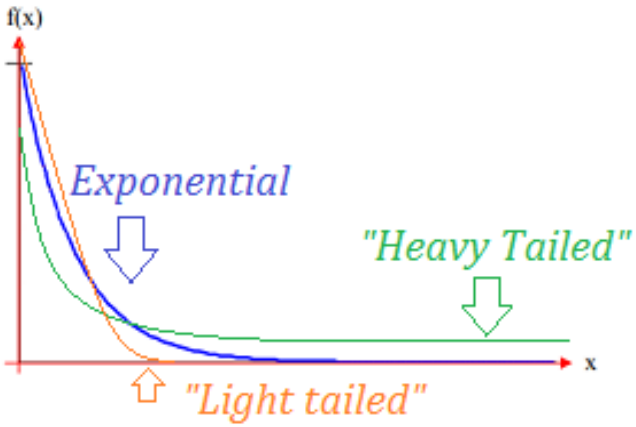
***ΟΡΙΣΜΟΣ 3.2:*** *Ορίζουμε μια κατανομή $F$ να είναι **(right-)heavy-tailed** αν και μόνο αν*

$$\int_{-\infty}^{\infty} e^{\lambda x} F(dx) = \infty \quad για\ όλα\ τα\ \lambda > 0 \quad (3.1)$$

*Δηλαδή αν και μόνο αν η $F$ αποτύχει να έχει κάποια θετικά εκθετική ροπή.*
*Διαφορετικά μια κατανομή $F$ ονομάζεται **light-tailed** αν και μόνο αν*

$$\int_{-\infty}^{\infty} e^{\lambda x} F(dx) < \infty \quad για\ κάποιο\ \lambda > 0 \quad (3.2)$$

Άρα οι light-tailed κατανομές, έχουν πιο "λεπτές" ουρές από την εκθετική κατανομή, κατά συνέπεια πάνε στο μηδέν πιο γρήγορα, επομένως έχουν λιγότερη πληροφορία στις ουρές. Ένα παράδειγμα τέτοιων κατανομών είναι η κατανομή Gumbel (γνωστή και ως log-Weibull), η οποία μπορεί να χρησιμοποιηθεί ως μοντέλο διάρκειας ζωής, παρότι ορίζεται στο διάστημα $(-\infty, \infty)$. Στο **Σχήμα 3.2** μπορούμε να δούμε πως απεικονίζονται γραφικά οι κατανομές.




***Σχήμα 3.2: "Γραφική απεικόνιση Heavy-tailed, Light-tailed και Εκθετικής κατανομής"***

**ΘΕΩΡΗΜΑ 3.1.1:**

*Για κάθε κατανομή $F$ τα παρακάτω είναι ισοδύναμα*

*i.* *Η $F$ είναι heavy-tailed κατανομή*

*ii.* *Η συνάρτηση $\bar{F}$ είναι heavy-tailed*

*iii.* *Η αντίστοιχη σωρευτική συνάρτηση κινδύνου $\boldsymbol{H}$ ικανοποιεί τη σχέση* $lim_{x\to\infty}\, inf\, \frac{\boldsymbol{H}(x)}{x} = 0$

**ΑΠΟΔΕΙΞΗ :**

***(ii)⇒ (iii)*** Υποθέτουμε ότι το $liminf$ στο ***(iii)*** είναι αυστηρά θετικό. Συνεπώς υπάρχει $x_0 > 0$ και $\varepsilon > 0$, τέτοιο ώστε $h(x) \geq \varepsilon x$ για όλα τα $x \geq x_0$ και ουσιαστικά συνεπάγεται ότι $\bar{F}(x) \leq e^{-\varepsilon x}$ για όλα τα $x$, κάτι το οποίο έρχεται σε αντίθεση με το ***(ii)*** που έχουμε ως δεδομένο.

***(iii)⇒ (i)*** Υποθέτουμε ότι η $F$ είναι *light-tailed,* οπότε από τη σχέση $(3.2)$ προκύπτει ότι για κάποιο $\lambda > 0$ και $c > 0$ έχουμε ότι $\bar{F}(x) \leq ce^{-\lambda x}$ για όλα τα $x$, κάτι το οποίο δείχνει ότι $\lim_{x\to\infty}\, inf\, \frac{\boldsymbol{H}(x)}{x} \geq \lambda$, το οποίο έρχεται σε σύγκρουση με το ***(iii)***.

**ΛΗΜΜΑ :**

Αν $F$ είναι μια συνεχής κατανομή με συνάρτηση πυκνότητας πιθανότητας $f$ και είναι heavy-tailed, τότε η συνάρτηση $f(x)$ είναι επίσης heavy-tailed.

*Ουσιαστικά θεωρούμε ότι μια συνάρτηση $f \geq 0$ είναι heavy-tailed αν και μόνο αν*

$$\lim_{x\to\infty} sup f(x)\, e^{\lambda x} = \infty \quad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ \lambda > 0$$

**ΑΠΟΔΕΙΞΗ :**

Έστω ότι η συνάρτηση $f(x)$ δεν είναι heavy-tailed, επομένως υπάρχει $\lambda' > 0$ και $x_0$ τέτοιο ώστε : $$c := sup_{x>x_0} f(x) e^{\lambda' x} < \infty$$

Και συνεπώς για όλα τα $\lambda \in (0, \lambda')$ έχουμε

$$\int_{\mathcal{R}} e^{\lambda x} F(dx) \leq e^{\lambda x_0} + c \int_{x_0}^{\infty} e^{\lambda x} e^{-\lambda' x} dx < \infty$$

Από το οποίο συνεπάγεται ότι ο ορισμός του ολοκληρώματος της σχέσης $(3.1)$ είναι πεπερασμένος για όλα τα $\lambda$ τέτοια ώστε $0 < \lambda < \lambda'$, το οποίο έρχεται σε σύγκρουση με το ότι η κατανομή F είναι heavy-tailed. ■

Γενικά υπάρχουν υποκατηγορίες για τις κατανομές με παχιές ουρές, όπως είναι η *long-tailed κατανομές* και οι *subexponential κατανομές*, όπου στην πράξη, οι πιο συνηθισμένες heavy-tailed κατανομές ανήκουν στη δεύτερη κατηγορία. Γενικά η ταξινόμηση τους στις κατηγορίες συνδέεται με τη συμπεριφορά τους στις μεγάλες τιμές, ωστόσο και στις δύο, η *συνάρτηση αξιοπιστίας* τείνει στο μηδέν, με ρυθμό πιο αργό σε σχέση με αυτόν της εκθετικής κατανομής.

### 3.1.1. LONG-TAILED ΚΑΤΑΝΟΜΕΣ

Μια κατανομή $F$, ορισμένη στο $\mathcal{R}$ θα λέμε ότι έχει ***μακριά ουρά (long-tailed distribution)*** αν $\bar{F}(x) > 0$ και για κάθε σταθερό $y > 0$ ισχύει :

$$\frac{\bar{F}(x+y)}{\bar{F}(x)} \to 1 \quad \text{όταν το } x \to \infty$$

Ιδιότητα ουράς μιας κατανομής είναι και το να είναι *long-tailed.* Σε γενικές γραμμές στον τομέα της στατιστικής και των επιχειρήσεων, μια μακριά ουρά είναι το μέρος της κατανομής που περιλαμβάνει πολλά τυχαία γεγονότα με διάφορες πιθανότητες, μακριά από την κορυφή ή το κύριο μέρος της κατανομής. Τα γεγονότα που βρίσκονται ωστόσο προς το τέλος της ουράς έχουν πολύ χαμηλή πιθανότητα να συμβούν. Το σημείο που θεωρείται ικανοποιητικό και ουσιαστικά ορίζει το πιο μέρος της κατανομής είναι η "μακριά ουρά", είναι αυθαίρετο, ωστόσο σε μερικές περιπτώσεις μπορεί να οριοθετηθεί αντικειμενικά. Συμπερασματικά μια long-tailed κατανομή είναι και heavy-tailed κατανομή, χωρίς να ισχύει πάντα και το αντίθετο, καθώς η παραπάνω σχέση απαιτεί και ένα βαθμό ομαλότητας της $\bar{F}(x)$ που δεν εμφανίζεται σε κάθε heavy-tailed κατανομή.

Η παραπάνω σχέση δίνει ότι : $\lim_{x\to\infty} \frac{\bar{F}(x+y)}{\bar{F}(x)} = 1$,ενώ αν η κατανομή έχει μακριά ουρά, είναι λογικό να ισχύει ταυτόχρονα ότι : $\lim_{x\to\infty} \frac{\bar{F}(x-y)}{\bar{F}(x)} = 1.$

Στην ουσία τα παραπάνω δύο όρια δείχνουν πως η συμπεριφορά της κατανομής $F$ δεν αλλάζει μετά από κάποια μεγάλη τιμή που μπορεί να πάρει το $x$. Πρακτικά δηλαδή η πιθανότητα η τυχαία μεταβλητή $X$ να πάρει μεγαλύτερη τιμή ή μικρότερη τιμή από το $x$, είναι σταθερή.

**ΟΡΙΣΜΟΣ 3.3:** Αν $h$ μια αυστηρά θετική, μη φθίνουσα συνάρτηση, τέτοια ώστε $h(x) \to \infty$, όταν $x \to \infty$, από το θεώρημα ομοιόμορφης σύγκλισης(*Uniform Convergence Theorem*), θα λέμε ότι είναι *η* $\bar{F}$ είναι ***h-insensitive***(ή *h-flat*) αν:

$$sup_{|y|\le h(x)} |\bar{F}(x+y) - \bar{F}(x)| = o\big(\bar{F}(x)\big),$$

$\kappa\alpha\theta\acute{\omega}\varsigma\ \tau o\ \ x \to \infty\ o\mu o\iota\acute{o}\mu o\rho\varphi\alpha\ \sigma\tau o\ |y| \le h(x).$

από την οποία σχέση καταλαβαίνουμε ότι η $\bar{F}(x)$ είναι long-tailed και επομένως κάθε long-tailed κατανομή είναι $h$-insensitive για κάθε συνεχή συνάρτηση $h$. Γενικά αν η συνάρτηση ουράς, $\bar{F}$, μιας κατανομής $F$, είναι *h-insensitive* θα λέμε ότι η $F$ είναι *h-insensitive*(ή *h-flat*) στο $\mathcal{R}$ .

### 3.1.2. SUBEXPONENTIAL ΚΑΤΑΝΟΜΕΣ

Οι ***υπό-εκθετικές κατανομές (subexponential distributions)*** είναι μια ειδική κατηγορία right heavy-tailed κατανομών. Ουσιαστικά το να είναι subexponential μια κατανομή αποτελεί tail property της κατανομής $F$. Το όνομα τους προκύπτει από το

γεγονός ότι η ουρά μιας τέτοιας κατανομής συγκλίνει στο μηδέν πιο αργά από οποιαδήποτε εκθετική κατανομή. Είναι λοιπόν λογικό να ισχύει:

$$\lim_{x\to\infty} sup e^{\lambda x}\,\bar{F}(x) = \infty \quad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ \lambda > 0$$

Πρακτικά αν πάρουμε ένα δείγμα τιμών από μια υπό-εκθετική κατανομή, έχουμε πολύ μεγάλες τιμές στο δείγμα με μη αμελητέα πιθανότητα. Αυτός είναι και ο λόγος που οι κατανομές αυτές χρησιμοποιούνται στην αναλογιστική επιστήμη, ως μοντέλα για την κατανομή των αποζημιώσεων σε ειδικές κατηγορίες ασφαλίσεων, όπως οι ασφαλίσεις πυρός ή οι ασφαλίσεις από φυσικές καταστροφές, όπου οι αποζημιώσεις παίρνουν πολύ μεγάλες τιμές με όχι πολύ μικρή πιθανότητα.

**ΟΡΙΣΜΟΣ 3.4 :** Μια κατανομή $F$ ορισμένη στον θετικό ημιάξονα $\mathcal{R}^+$ με την ιδιότητα $\bar{F}(x) > 0$ για όλα τα $x$, θα λέμε ότι ανήκει στην οικογένεια των υπό-εκθετικών κατανομών αν**:**

$$\lim_{x\to\infty} \frac{\overline{F^{*2}}(x)}{\bar{F}(x)} = 2, \qquad \delta\eta\lambda\alpha\delta\acute{\eta} \qquad \overline{F * F}(x) \sim 2\bar{F}(x) \quad \acute{o}\tau\alpha\nu\ \tau o\ x \to \infty \quad (3.3)$$

Η παραπάνω συνθήκη, μας βεβαιώνει ότι η $F$ είναι heavy-tailed όπως και long-tailed. Με επαγωγικά επιχειρήματα προκύπτει ότι, αν μια κατανομή $F$ στο $\mathcal{R}^+$ είναι subexponential, τότε μπορούμε να γενικεύσουμε τη σχέση ως εξής :

$$\lim_{x\to\infty} \frac{\overline{F^{*n}}(x)}{\bar{F}(x)} = n, \qquad \gamma\iota\alpha\ \acute{o}\lambda o\upsilon\varsigma\ \tau o\upsilon\varsigma\ \alpha\kappa\acute{\varepsilon}\rho\alpha\iota o\upsilon\varsigma\ \alpha\rho\iota\theta\mu o\acute{\upsilon}\varsigma\ \ n \geq 2$$

Επειδή όπως είπαμε παραπάνω, tail property μιας κατανομής είναι και το να ανήκει στην οικογένεια των subexponential, είναι σημαντικό για πολλές εφαρμογές να επεκτείνουμε την έννοια της υπό-εκθετικότητας, για μια κατανομή σε ολόκληρο το $\mathcal{R}$. Κάτι τέτοιο μπορεί να επιτευχθεί είτε απαιτώντας η $F$ να έχει την ίδια ουρά με αυτή μιας subexponential κατανομής στο $\mathcal{R}^+$(θεωρούμε την κατανομή $F^+$, όπου $F^+(x) = F(x)\ \gamma\iota\alpha\ x \geq 0$ και $F^+(x) = 0\ \gamma\iota\alpha\ x < 0$) ή ισοδύναμα απαιτώντας η $F$ να είναι long-tailed και να ικανοποιεί τη σχέση $(3.3)$, καθώς από μόνη της η σχέση και η ιδιότητα ουράς (long-tailed), δεν υποδεικνύουν πλέον υπό-εκθετικότητα της $F$, σε ολόκληρο το $\mathcal{R}$.

**ΠΑΡΑΤΗΡΗΣΗ :** Γενικά όσο η ανάλυση μας κινείται στον θετικό ημιάξονα, αν η κατανομή $F$ ανήκει στην οικογένεια των subexponential κατανομών, τότε θα ανήκει και στην οικογένεια των long-tailed κατανομών και στην οικογένεια των heavy-tailed κατανομών. Δηλαδή αν συμβολίσουμε με $\mathcal{H}$ την κλάση κατανομών με παχιές ουρές, με $\mathcal{L}$ την οικογένεια κατανομών με μακριά ουρά και $S$ την οικογένεια με τις subexponential κατανομές, τότε έχουμε :

$$S \subseteq \mathcal{L} \subseteq \mathcal{H}$$

Το ίδιο βέβαια δεν ισχύει και όταν η μελέτη μας αναφέρεται σε ολόκληρο το $\mathcal{R}$.

***[9].***

## 3.2. ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΟΥ ΕΛΕΓΧΟΥ ΤΗΣ ΜΕΓΙΣΤΗΣ ΕΝΤΡΟΠΙΑΣ

Το 1957, πρώτος ο Jaynes E.T. εισήγαγε την *αρχή της μέγιστης εντροπίας (maximum entropy principle)* ως ένα κριτήριο επιλογής a priori πιθανοτήτων. Σύμφωνα λοιπόν με αυτή την αρχή, για δεδομένη πληροφορία, η κατανομή πιθανότητας που περιγράφει καλύτερα την εκ των προτέρων γνώση μας, είναι αυτή που μεγιστοποιεί την εντροπία του Shannon, υπό τον όρο της δοσμένης πληροφορίας σαν περιορισμό.***[8].*** Ωστόσο, το εναλλακτικό μέτρο πληροφορίας που προτάθηκε, αποτελώντας ουσιαστικά τη γενίκευση του τύπου $(2.22)$, είναι αρκετά χρήσιμο για τον έλεγχο προσαρμογής κατανομών με παχιές ουρές, όπου σε αυτές τις περιπτώσεις, η προσοχή μας επικεντρώνεται στις ουρές της κατανομής και συνεπώς περισσότερο βάρος, θα πρέπει να δοθεί στις παρατηρήσεις που βρίσκονται σε αυτές. ***[3].***

Έστω $(Y_1, Y_2, \dots, Y_n)$ τυχαίο δείγμα από μια κατανομή με άγνωστη συνάρτηση κατανομής $F$ και έστω ότι θεωρούμε τον ακόλουθο έλεγχο προσαρμογής :

$$H_0 : F = F_0 \quad vs. \quad H_1 : F \neq F_0$$

Για συνεχείς κατανομές έχει προταθεί (Lee, Vonta and Karagrigoriou (2011)) η ακόλουθη γενικευμένη μορφή εντροπίας των Forte και Hughes (1988) [σχέση $(2.22)$]:

$$S^w(F) = -\sum_{i=1}^{m} w_i\big(F(s_i) - F(s_{i-1})\big) \log\left(\frac{F(s_i) - F(s_{i-1})}{s_i - s_{i-1}}\right) \qquad (3.4)$$

όπου τα $w$ είναι κατάλληλες συναρτήσεις βάρους για τις οποίες ισχύει: $0 \leq w_i \leq 1$ με $\sum_{i=1}^{m} w_i = 1$, όπου $m$ ο αριθμός των μη επικαλυπτόμενων διαστημάτων από τον διαχωρισμό των δεδομένων και $-\infty < a \leq s_1 \leq s_2 \leq \cdots \leq s_m \leq b < \infty$ είναι προκαθορισμένα σημεία διαμέρισης.***[3],[17].*** Γενικά ο ρόλος των βαρών είναι μείζονος σημασίας στον έλεγχο της μέγιστης εντροπίας. Μπορούμε όμως να αποφύγουμε τη δυσκολία επιλογής βέλτιστων βαρών, ανεξαρτήτως της ύπαρξής τους ή όχι, παίρνοντας τον ομοιόμορφο μετασχηματισμό $s_i = \frac{i}{m}$, μειώνοντας έτσι κατά κάποιον τρόπο την σημασία επιλογής τους, και επιπλέον παίρνοντας το supremum των βαρών, με σκοπό να αντιμετωπιστεί με αυτό τον τρόπο κάθε δυνατή εναλλακτική υπόθεση. ***[17].***

Για μια κατάλληλα επιλεγμένη σταθερά $c$, έχουμε ότι η μηδενική υπόθεση θα απορριφθεί αν $\left|S^w\left(\hat{F}_n\right) - S^w(F_0)\right| \geq c$ ή πιο αυστηρά όταν $\sup_w |S^w\left(\hat{F}_n\right) - S^w(F_0)| \geq c$, όπου $\hat{F}_n$ είναι η εμπειρική συνάρτηση κατανομής που βασίζεται στο δείγμα, ορισμένη από τη σχέση $(2.1)$. ***[3],[17].***

Είναι εύκολα αντιληπτό ότι ο *στατιστικός έλεγχος της μέγιστης εντροπίας (maximum entropy test)* είναι στενά συνδεδεμένος με το μέτρο εντροπίας που δίνεται στη σχέση $(2.21)$. Πράγματι παρατηρούμε ότι η σχέση $(3.4)$ χωρίς βάρη, όταν $m \to \infty$ και $\max_{1 \leq i \leq m} |s_i - s_{i-1}| \to 0$, παίρνει την μορφή :

$$S_{max}(F) = -\sum_{i=1}^{m}\left(\frac{F(s_i)-F(s_{i-1})}{s_i - s_{i-1}}\right)(s_i - s_{i-1})\log\left(\frac{F(s_i)-F(s_{i-1})}{s_i - s_{i-1}}\right)$$

$$\approx -\sum_{i=1}^{m} f(s_i)(s_i - s_{i-1})\log f(s_i)$$

$$\xrightarrow{m\to\infty} -\int_{-\infty}^{\infty} f(x)\log f(x)dx = -E_F\big(\log(f(x))\big) \equiv H(f)$$

όπου $f$ είναι η συνάρτηση πυκνότητας πιθανότητας.

Παρατηρούμε ότι αν η $F_0$ είναι η ομοιόμορφη κατανομή στο $[0,1]$, τότε $S^w(F_0) = 0$. Έτσι χωρίς βλάβη της γενικότητας μπορούμε να εστιάσουμε στην ομοιόμορφη κατανομή στο $[0,1]$, και έτσι ο έλεγχος υποθέσεων να γίνει :

$$H_0 : F = F_0 \equiv U[0,1] \quad vs. \quad H_1 : F \neq F_0 \equiv U[0,1] \qquad (3.5)$$

Πράγματι με τη χρήση του μετασχηματισμού ολοκληρώματος πιθανότητας *(probability integral transformation),* που παρουσιάστηκε στο *θεώρημα 2.2.2* της ***ενότητας 2.2.1*** μπορούμε να κατασκευάσουμε ένα ισοδύναμο σετ τιμών

$$F_0(Y_i) = U_i,$$

ελέγχοντας έπειτα, αν η ομοιόμορφη κατανομή είναι κατάλληλη για τα $U_i$. Με αυτό τον τρόπο ουσιαστικά εξετάζουμε αν τα $Y_i$ ακολουθούν κάποια συνεχή κατανομή $F_0$.

Γενικά παρόλο που η απλή μηδενική υπόθεση εμφανίζεται στην πράξη πιο συχνά, είναι αρκετά σύνηθες και η εξέταση της σύνθετης μηδενικής υπόθεσης, όπου η άγνωστη κατανομή ανήκει σε μια παραμετρική οικογένεια $\{F_\theta\}_{\theta\in\Theta}$, όπου $\Theta$ είναι ένα ανοιχτό υποσύνολο του $\mathcal{R}^k$. Επομένως θεωρούμε μια διαμέριση του αρχικού δειγματικού χώρου σε $m$ ξένα διαστήματα. Σε αυτή την περίπτωση ωστόσο μια σημαντική παρατήρηση, είναι ότι ο μετασχηματισμός ολοκληρώματος πιθανότητας εξαρτάται από την άγνωστη $k$-διάστατη παράμετρο $\theta$. Επομένως απαιτείται ένας συνεπής εκτιμητής $\hat{\theta}$, με σκοπό ο μετασχηματισμός ολοκληρώματος πιθανότητας να εφαρμοστεί για τη δημιουργία τιμών $F_{\hat{\theta}}(Y_i) = U_i$, αν η μηδενική υπόθεση είναι $H_0 : F = F_\theta$. Σε αυτή την περίπτωση, η οριακή κατανομή, που δίνεται από το *θεώρημα 3.2.2* πιο κάτω, μπορεί να επηρεαστεί από την εκτίμηση του $\theta$. Η επιρροή ωστόσο μπορεί να ελαττωθεί όταν το $m$ είναι μεγάλο και $max_i(s_i - s_{i-1})$ είναι μικρό, όπως στην περίπτωση του $X^2$ ελέγχου.

Σχετικά λοιπόν με την μέθοδο εκτίμησης, προκειμένου να έχουμε τον εκτιμητή $\hat{\theta}$, θα χρησιμοποιηθεί η μέθοδος μεγίστης πιθανοφάνειας κάτω από τη μηδενική υπόθεση. Εναλλακτικά ωστόσο, κάποιος μπορεί να θεωρήσει την ευρεία τάξη των εκτιμητών φ-divergence. Συγκεκριμένα αν έχουμε τη διαμέριση του δειγματικού χώρου $\{E_i\}_{i=1,\dots,m}$, τότε ο ελάχιστος (minimum) εκτιμητής φ-divergence του $\theta$ είναι κάθε $\hat{\theta}_\Phi \in \Theta$ που ικανοποιεί την εξής σχέση :

$$d_a(\hat{\theta}_\phi) = \min_{\theta\in\Theta} d_a(\theta) = \min_{\theta\in\Theta} \sum_{i=1}^{m} p_{i0}(\theta)\phi_a\left(\frac{\hat{p}_i}{p_{i0}(\theta)}\right), \qquad \text{με } \phi \in \Phi^*, a > 0$$

όπου $p_{i0}(\theta) = \int_{E_i} dF_\theta$ για $i = 1,2,\dots,m$ , $\hat{p}_i$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της πιθανότητας του $i$-διαστήματος της διαμέρισης και τέλος $\Phi^*$ είναι ο χώρος όλων των κυρτών συναρτήσεων $\phi$ στο $[0,\infty)$, τέτοιων ώστε $\phi(1) = \phi'(1) = 0$ και $\phi''(1) \neq 0$. Ενώ θεωρούμε επίσης τις υποθέσεις : $0\phi\left(\frac{0}{0}\right) = 0$ και $0\phi\left(\frac{u}{0}\right) = \lim_{u\to\infty}\frac{\phi(u)}{u}, u > 0$. Προφανώς, η εκτιμήτρια που προκύπτει, εξαρτάται από τη συνάρτηση $\phi$ που έχει επιλεγεί. Παρατηρούμε λοιπόν ότι αν η $\phi$ έχει την ειδική μορφή :

$$\phi_a(u) = u^{1+a} - \left(1+\frac{1}{a}\right)u^a + \frac{1}{a}\ , a > 0 \quad \text{ή} \quad \phi_a^1(u) = \phi_a(u)\Big/(1+a)$$

για $a \to 0$, η εκτιμήτρια που προκύπτει, είναι η συνήθης εκτιμήτρια μέγιστης πιθανοφάνειας για ομαδοποιημένα δεδομένα.

Θα πρέπει να επισημανθεί ότι αν όλα τα δεδομένα είναι διαθέσιμα (και όχι μόνο τα ομαδοποιημένα), είναι προτιμότερο να στηριχτούμε στη διαδικασία εκτίμησης της μέγιστης πιθανοφάνειας, γιατί παίρνουμε αποδοτικούς εκτιμητές που βρίσκονται πιο εύκολα και πιο απλά. ***[3],[17].***

Στο *θεώρημα 3.2.2* που ακολουθεί, παρουσιάζεται η ασυμπτωτική κατανομή της στατιστικής ελεγχοσυνάρτησης της μέγιστης εντροπίας κάτω από τη μηδενική υπόθεση, στον ορισμό της οποίας τα αυθαίρετα βάρη, προτάθηκαν από τους Lee et al. (2011).

**ΘΕΩΡΗΜΑ 3.2.2 :**

*Έστω $U_1, U_2, \dots, U_n$ ένα τυχαίο δείγμα από μια συνεχή κατανομή με συνάρτηση κατανομής $F$. Κάτω από τη μηδενική υπόθεση που δίνεται στη σχέση (3.5), καθώς $n \to \infty$, έχουμε ότι :*

$$\sqrt{n}\ \sup_{\{w\in W\}}\left|S^w\left(\widehat{F}_n\right)\right| \xrightarrow{d} \sup_{\{w\in W\}}\left|\sum_{i=1}^{m} w_i\left(B(s_i) - B(s_{i-1})\right)\right|, \quad (3.6)$$

*όπου $B(s)$ είναι η Brownian bridge στο $[0,1]$, ενώ $W$ είναι ο χώρος των βαρών για τα οποία ισχύει $0 \le w_i \le 1$, με $\sum_{i=1}^{m} w_i = 1$ και $0 = s_0 \le s_1 \le \cdots \le s_m = 1$.*

**ΑΠΟΔΕΙΞΗ :**

Η απόδειξη του θεωρήματος παραλείπεται. Ωστόσο, κάποιος μπορεί να την βρει ανατρέχοντας στο paper *"A maximum entropy type test of fit"* των Lee, Vonta και Karagrigoriou (2011).

Στην πρακτική εφαρμογή του ελέγχου, θεωρούμε $w_i^{(l)}, l = 1,2,\dots,L$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από την U[0,1], ανεξάρτητες από τις τυχαίες μεταβλητές $U_i$ οι οποίες επίσης ακολουθούν ομοιόμορφη κατανομή στο [0,1]. Έστω $L$ ένας καθορισμένος θετικός ακέραιος αριθμός. Όταν $L \to \infty$, αν πάρουμε

$w_{li} = \frac{w_i^{(l)}}{w_1^{(l)}+w_2^{(l)}+\cdots+w_m^{(l)}}$ όπως οι Lee, Vonta και Karagrigoriou στον προτεινόμενο έλεγχό τους, έχουμε :

$$\max_{1\le l\le L}\left|\sum_{i=1}^{m} w_{li}\big(B(s_i)-B(s_{i-1})\big)\right| \xrightarrow{d} \sup_{\{w\in W\}}\left|\sum_{i=1}^{m} w_i\big(B(s_i)-B(s_{i-1})\big)\right| \qquad (3.7)$$

Κατόπιν αν πάρουμε για ευκολία τα βάρη $s_i = \frac{i}{m}$, $i = 1,2,\dots,m$, μπορούμε να χρησιμοποιήσουμε ως στατιστική ελεγχοσυνάρτηση της μέγιστης εντροπίας, την ποσότητα :

$$S^{w}{}_{max} = \max_{1\le l\le L}\left|\sum_{i=1}^{m} w_{li}\left(E_n\left(\frac{i}{m}\right)-E_n\left(\frac{i-1}{m}\right)\right)\right|$$
$$\approx \sup_{\{w\in W\}}\left|\sum_{i=1}^{m} w_i\left(B\left(\frac{i}{m}\right)-B\left(\frac{i-1}{m}\right)\right)\right| \qquad (3.8)$$

***ΣΗΜΕΙΩΣΗ :*** Η σχέση (3.7) είναι αληθινή, καθώς :

$$P(\overline{\{(w_{l1},w_{l2},\dots,w_{lm}) : l\ge 1\}} = W) = 1$$

Διαφορετικά υπάρχει μια ανοιχτή μπάλα $V$ στον χώρο $W$, ώστε $p := P\big((w_{l1},w_{l2},\dots,w_{lm})\in V^c \text{ για όλα τα } l\ge 1\big) > 0$, το οποίο είναι ανέφικτο στην πραγματικότητα, καθώς $p \le \rho^L$ για κάθε $L$, με $\rho = P((w_{l1},w_{l2},\dots,w_{lm})\in V^c) < 1$ (ή ισοδύναμα $P((w_{l1},w_{l2},\dots,w_{lm})\in V) > 0$), κάτι το οποίο έχει ως επακόλουθο ότι

$$sup_{l\ge 1}\left|\sum_{i=1}^{m} w_{li}\big(B(s_i)-B(s_{i-1})\big)\right| = sup_{\{w\in W\}}\left|\sum_{i=1}^{m} w_i\big(B(s_i)-B(s_{i-1})\big)\right| \; a.s$$

από το οποίο άμεσα συνεπάγεται η σχέση (3.7).

***ΣΧΟΛΙΟ :*** Θα πρέπει να σημειωθεί ότι για πρακτικούς σκοπούς, έχει προταθεί από τους Lee, Vonta, και Karagrigoriou (2011), η χρήση της τιμής $L = 1000$, ως αρκετά ικανοποιητικός αριθμός για τη μελέτη του μεγέθους και της ισχύος του ελέγχου, όταν εφαρμόζεται σε προσομοιώσεις για διάφορες κατανομές και μια πληθώρα εναλλακτικών υποθέσεων.

***[3],[17].*** ■

Όπως αποδείχτηκε λοιπόν το 2011, από τους Lee, Vonta και Karagrigoriou, αν ορίσουμε τυχαία βάρη από την ομοιόμορφη κατανομή $[0,1]$ στον προτεινόμενο έλεγχο, τότε θα πάρουμε έναν αρκετά ικανοποιητικό έλεγχο καλής προσαρμογής για τις συνήθεις κατανομές, κάνοντας παράλληλα χρήση του θεωρήματος μετασχηματισμού ολοκληρώματος πιθανότητας. Ωστόσο αν θέλουμε να αντιμετωπίσουμε κατανομές με κάποιο συγκεκριμένο χαρακτηριστικό, όπως είναι οι παχιές ουρές, με τις οποίες θα ασχοληθούμε στην παρούσα εργασία, καλό είναι να προσαρμόσουμε τον έλεγχό μας. Πρόθεσή μας λοιπόν είναι να

προσαρμόσουμε τα βάρη κατάλληλα, με σκοπό να εφαρμόζουν ικανοποιητικά σε κατανομές με παχιές ουρές, καθώς είναι αναγκαίο να εξετάσουμε αν μια κατανομή προσαρμόζεται καλά, κυρίως στα δεδομένα που βρίσκονται στις ουρές.

Για ***παράδειγμα***, υπάρχουν κλάδοι, όπως αυτός της ασφάλισης και αντασφάλισης, όπου η άνω ουρά είναι ζωτικής σημασίας για την απόκτηση της σωστής εκτίμησης της πιθανότητας που προκαλεί τη ζημιά. Επομένως σε μια τέτοια περίπτωση, θα πρέπει να οριστεί μια κατάλληλη συνάρτηση βάρους $w_i$, με σκοπό στις παρατηρήσεις που βρίσκονται στην πάνω ουρά να δοθεί περισσότερο βάρος σε σχέση με αυτές που βρίσκονται στην κάτω ουρά.

Πιο αναλυτικά κάθε ασφαλιστικός φορέας προσπαθεί να καλυφθεί από την εμφάνιση ενός μη πιθανού ενδεχομένου, το οποίο θα έχει αρνητικές συνέπειες στην βιωσιμότητά του. Για αυτό το λόγο, θα πρέπει η εταιρεία να αντασφαλιστεί, μοιράζοντας μέρος του κινδύνου που αναλαμβάνει, σε άλλους φορείς. Ωστόσο στον τομέα της αντασφάλισης υπάρχει πάντοτε κάποια πιθανότητα να συμβεί μια πολύ μεγάλη ζημία, κάτι το οποίο υποδηλώνεται στον έλεγχο καλής προσαρμογής στις ουρές, καθώς εκεί βρίσκονται οι τιμές του κινδύνου. Αυτός λοιπόν είναι και ο λόγος για τον οποίο στην ανάλυση, πρέπει να χρησιμοποιηθούν κατανομές με παχιές ουρές. Μας ενδιαφέρει δηλαδή η πιθανότητα ουράς $P(X \geq x)$, καθώς $X$ είναι η τυχαία μεταβλητή που εκφράζει το ύψος της ζημιάς. Πολύ μεγάλο ενδιαφέρον λοιπόν έχει το να ξέρουμε και να μπορούμε να υπολογίσουμε ποια είναι η πιθανότητα να προκληθεί μια μεγάλη καταστροφή-ζημιά. ***[3].***

## 3.3. ΠΡΟΣΟΜΟΙΩΣΗ ΓΙΑ ΜΙΚΡΑ ΠΡΟΣ ΜΕΣΑΙΑ ΔΕΙΓΜΑΤΑ

Έχει αποδειχτεί ότι η ασυμπτωτική κατανομή του προτεινόμενου στατιστικού ελέγχου μέγιστης εντροπίας, μπορεί να χρησιμοποιηθεί για μεσαία προς μεγάλα μεγέθη δείγματος, υπολογίζοντας την ισχύ και το μέγεθός του αρκετά ικανοποιητικά. Ωστόσο, δεν ισχύει το ίδιο και για τα μικρά προς μεσαία μεγέθη δείγματος. Επομένως, προκειμένου να αντιμετωπιστούν αυτά τα δείγματα παρουσιάζουμε πιο κάτω, τον έλεγχο του οποίου οι κρίσιμες τιμές υπολογίζονται από την εμπειρική κατανομή της ελεγχοσυνάρτησης.

Ο εμπειρικός έλεγχος στηρίζεται στην ίδια ελεγχοσυνάρτηση με τον αρχικό, ωστόσο, η χρήση πινάκων με κρίσιμες τιμές, οι οποίες εξαρτώνται από τις παραμέτρους της κατανομής που είναι υπό εξέταση κάτω από τη μηδενική υπόθεση και/ή το μέγεθος του δείγματος, είναι αναπόφευκτη. Η εύρεση των εμπειρικών κρίσιμων τιμών στηρίζεται στην εφαρμογή Monte-Carlo προσομοιώσεων.

Στον έλεγχο της μέγιστης εντροπίας, ιδιαίτερη προσοχή οφείλει να δοθεί στον αριθμό των διαστημάτων $m$ που χρησιμοποιούνται, καθώς επίσης και στη μορφή τους, σε συνδυασμό με το μέγεθος του δείγματος. Κίνδυνος ελλοχεύει για παράδειγμα, στη χρήση υπερβολικά πολλών διαστημάτων, στις περιπτώσεις όπου οι παρατηρήσεις βρίσκονται αραιά στο εύρος των δεδομένων. Επίσης, όταν το

μέγεθος του δείγματος είναι πολύ μικρό και ο αριθμός των κλάσεων είναι πολύ μεγάλος, τα αποτελέσματα των προσομοιώσεων χρήζουν μεγάλης προσοχής, λόγω του ενδεχομένου μη εμφάνισης ή μικρού αριθμού εμφάνισης παρατηρήσεων, τουλάχιστον σε κάποια από αυτά. Επομένως, για μικρό μέγεθος δείγματος αποφεύγουμε τη χρήση μεγάλου αριθμού διαστημάτων, ενώ για μεγάλα μεγέθη δείγματος, ο αριθμός των διαστημάτων είναι μικρότερης σημασίας.

Σε γενικές γραμμές, συνιστάται η χρήση ισοπίθανων διαμερίσεων, καθώς ο βέλτιστος αριθμός διαστημάτων εξαρτάται από πολλούς παράγοντες. Ο τύπος της εναλλακτικής υπόθεσης, η "απόσταση" μεταξύ των δεδομένων και της εναλλακτικής υπόθεσης, η μικρότερη δυνατή ισχύ που μπορεί να επιτευχθεί, το επίπεδο σημαντικότητας του ελέγχου και το μέγεθος του δείγματος, είναι κάποιοι από αυτούς. Ο αριθμός που προτείνεται κυμαίνεται από 3 έως 24. Ουσιαστικά η ισχύς του ελέγχου μαζί με το μέγεθος του δείγματος είναι οι παράγοντες κλειδί που καθορίζουν αν η αύξηση του αριθμού τους είναι χρήσιμη. Πρέπει να σημειωθεί, ότι ακόμα και στις περιπτώσεις όπου ο αριθμός των διαστημάτων επιτρέπεται να αυξηθεί, καθώς το μέγεθος του δείγματος αυξάνει, ειδικές υποθέσεις θα πρέπει να διατυπωθούν, προκειμένου να διασφαλίσουμε ικανοποιητικά ασυμπτωτικά αποτελέσματα. Μια τέτοια υπόθεση είναι ότι το όριο του $^{n}/_{M}$ καθώς $n \to \infty$, όπου το $M$ εξαρτάται από το $n$, είναι πεπερασμένο. Στην μελέτη που έγινε από τους Lee et al.(2011), παρατηρήθηκε ότι για δείγματα που είχαν λιγότερες από 100 παρατηρήσεις, επιλέχτηκαν από 3 έως 20 κλάσεις και όταν χρησιμοποιήθηκαν 10 κλάσεις, τα αποτελέσματα που προέκυψαν ήταν αρκετά ικανοποιητικά. ***[3],[17].***

Συνεπώς στη Monte Carlo μελέτη μας επικεντρωνόμαστε στο μέγεθος του δείγματος σε συνδυασμό με την τιμή του $m$, με την προτεινόμενη τιμή $L = 1000$. Γενικά η γλώσσα προγραμματισμού που χρησιμοποιήθηκε σε όλες τις αναλύσεις που έγιναν για να εξάγουμε τα αποτελέσματά μας, είναι η R. Ως μεγέθη δείγματος πήραμε τα $n = 10,20,30,50$ και $100$, ενώ οι τιμές των διαστημάτων που πήραμε είναι $m = 3,4,5$ και $10$. Για κάθε συνδυασμό $n$ και $m$, εφαρμόσαμε τα εξής βήματα :

- Αρχικά ένα δείγμα μεγέθους $n$ επιλέχτηκε από την ομοιόμορφη κατανομή $[0,1]$.
- Η ελεγχοσυνάρτηση του ελέγχου της μέγιστης εντροπίας που δίνεται από τον συνδυασμό των τύπων $(3.4)$ και $(3.6)$ υπολογίστηκε για $L = 1000$.
- Τα προηγούμενα δύο βήματα υπολογίζονται για $M = 1000$ φορές.
- Από τις 1000 διατεταγμένες τιμές των στατιστικών ελεγχοσυναρτήσεων που θα προκύψουν, μας ενδιαφέρει να δούμε το $90^{\circ}$, το $95^{\circ}$ και $99^{\circ}$ ποσοστημόριο του διατεταγμένου δείγματος, τα οποία ουσιαστικά προσεγγίζουν τις κρίσιμες τιμές των αντίστοιχων επιπέδων σημαντικότητας $0.10, 0.05$ και $0.01$ αντίστοιχα.

Προκειμένου λοιπόν να βρεθούν οι στενές σχέσεις που υπάρχουν ανάμεσα στις κρίσιμες τιμές και στις συναρτήσεις των $n$ και $m$, καταλήγοντας στο βέλτιστο μοντέλο, χρησιμοποιούμε τη διαδικασία *stepwise linear regression*. Σύμφωνα με την

οποία, μετά από την πρόσθεση μιας επιπλέον μεταβλητής στο μοντέλο, εξετάζεται η εξασθένηση της σημαντικότητας κάποιας άλλης μεταβλητής που είχε εισαχθεί νωρίτερα, επανεξετάζοντας έτσι την πιθανή αφαίρεση της. Αυτό συμβαίνει γιατί μπορεί η μεταβλητή $x_i$ από μόνη της να είναι η καλύτερη, ωστόσο ο κατάλληλος συνδυασμός δύο άλλων μεταβλητών $x_j$ και $x_k$ να είναι ακόμα καλύτερος.

Στη μελέτη μας θα εξετάσουμε δύο περιπτώσεις βαρών. Σύμφωνα με την **πρώτη περίπτωση**, λαμβάνοντας υπόψη τις απαραίτητες προϋποθέσεις που πρέπει να ισχύουν [δηλαδή: $0 \leq w_i \leq 1$, με $\sum_{i=1}^{m} w_i = 1$], κάνουμε τον εξής συλλογισμό: παίρνουμε $m$ διαστήματα και θέλουμε το βάρος που θα αντιστοιχήσουμε στο πρώτο διάστημα να είναι μικρότερο από το βάρος που θα αντιστοιχήσουμε στο δεύτερο, το βάρος που θα αντιστοιχήσουμε στο δεύτερο να είναι μικρότερο από το βάρος που θα αντιστοιχήσουμε στο τρίτο κ.λπ. με το βάρος που θα αντιστοιχηθεί στο $m$ διάστημα να είναι το μεγαλύτερο από όλα τα βάρη. Έτσι δίνουμε μία έμφαση στις παρατηρήσεις στην ουρά της κατανομής. Οπότε προτείνουμε για βάρη τα $\frac{1}{\frac{m(m+1)}{2}}, \dots, \frac{m}{\frac{m(m+1)}{2}}$ τα οποία είναι μικρότερα της μονάδας και το άθροισμά τους κάνει 1. Επειδή πρέπει να επαναλάβουμε $L = 1000$ φορές την εύρεση τυχαίων βαρών σύμφωνα με τον προτεινόμενο έλεγχο, πήραμε κάθε φορά τυχαίους αριθμούς από τις ομοιόμορφες κατανομές $U\left(i - \frac{1}{2}, i + \frac{1}{2}\right)$ για $i = 1, \dots, m$ δηλαδή μια τιμή από την κατανομή $U(0.5,1.5)$, μια τιμή από την κατανομή $U(1.5,2.5)$ κλπ., και διαιρέσαμε στο τέλος με το άθροισμα των $m$ τυχαίων τιμών.

Με παρόμοιο τρόπο για τη **δεύτερη περίπτωση,** θέλοντας πάλι να δώσουμε έμφαση στην ουρά, παίρνουμε $m$ διαστήματα, θέλοντας το βάρος που θα αντιστοιχήσουμε στο πρώτο διάστημα να είναι μικρότερο από όλα τα βάρη, ενώ αντίστοιχα το βάρος που θα αντιστοιχηθεί στο $m$ διάστημα να είναι το μεγαλύτερο. Προτείνουμε λοιπόν για βάρη, τα $\frac{1}{\frac{m(m+1)(2m+1)}{6}}, \dots, \frac{m^2}{\frac{m(m+1)(2m+1)}{6}}$, τα οποία είναι μικρότερα της μονάδας και το άθροισμά τους κάνει 1, άρα πληρούν τις απαραίτητες προϋποθέσεις. Επαναλαμβάνοντας λοιπόν $L = 1000$ φορές, σύμφωνα με τον προτεινόμενο έλεγχο, πήραμε κάθε φορά τυχαίους αριθμούς από τις ομοιόμορφες κατανομές $U\left(i^2 - \frac{1}{2}, i^2 + \frac{1}{2}\right)$, για $i = 1, \dots, m$, δηλαδή μια τιμή από την κατανομή $U(0.5,1.5)$, μια τιμή από την $U(3.5,4.5)$ κλπ., και στη συνέχεια διαιρέσαμε με το άθροισμα των $m$ τυχαίων τιμών.

Επομένως για να εφαρμόσουμε τον έλεγχο μέγιστης εντροπίας στην πράξη, θεωρούμε βάρη $w_i^{(l)}, l = 1, .., L$, με $i = 1, \dots, m$, τα οποία υπολογίστηκαν όπως περιγράψαμε πιο πάνω.

Θέλοντας να βρούμε κρίσιμες τιμές οι οποίες να μπορούν να υπολογιστούν από κάποιο "τύπο" όταν κάποιος δώσει φυσικά τις τιμές των $n$ και $m$ για τις οποίες ενδιαφέρεται, κάναμε κάποιες πρώτες προσομοιώσεις και εξετάσαμε τις εμπειρικές

κατανομές της ελεγχοσυνάρτησης για διάφορες τιμές των $n$ και $m$ κάτω από τη μηδενική υπόθεση. Υπολογίσαμε τα $90^{\alpha}$, $95^{\alpha}$ και $99^{\alpha}$ ποσοστημόρια των κατανομών αυτών και τα χρησιμοποιήσαμε σαν εξηρτημένη μεταβλητή και μέσω γραμμικών μοντέλων παλινδρόμησης, όπου εξετάστηκαν διάφορες ανεξάρτητες μεταβλητές $\left(n, m, \sqrt{n}, \sqrt{m}, \log(n), \log(m)\right)$, καταλήξαμε στα καλύτερα μοντέλα, βάσει των οποίων θα κάνουμε εκτίμηση των ποσοστημορίων για δεδομένα $n$ και $m$.

Από την εφαρμογή λοιπόν των βημάτων που περιγράφτηκαν παραπάνω, για τις δύο περιπτώσεις βαρών που πήραμε, στους πίνακες πιο κάτω, υπάρχουν οι ανεξάρτητες μεταβλητές που προέκυψαν και θεωρούνται σημαντικές για το βέλτιστο μοντέλο, καθώς επίσης και ο *προσαρμοσμένος συντελεστής προσδιορισμού* $R^2_{adj}$ *(adjusted coefficient of multiple determination)* για κάθε επίπεδο σημαντικότητας. Για πρακτικούς λόγους, οι εκτιμημένες μέσω των μοντέλων, κρίσιμες τιμές των πινάκων χρησιμοποιούνται στις εφαρμογές όταν το $n \leq 100$, ενώ οι κρίσιμες τιμές που προκύπτουν από την ασυμπτωτική κατανομή του *θεωρήματος 3.1*, συνιστούνται στις εφαρμογές όπου το $n > 100$.

| *Percentile* | *Intercept* | $\sqrt{n}$ | $\sqrt{m}$ | log n | log m | $R^2(\%)$ |
|---|---|---|---|---|---|---|
| **90th** | 0.89020660 | 0.04352529 | 0.37723052 | −0.20089538 | −0.62862996 | 97.51 |
| **95th** | 1.00477263 | 0.03176861 | 0.37117699 | −0.17410713 | −0.66413320 | 98.81 |
| **99th** | 1.43430077 | 0.04611534 | 0.66637432 | −0.24877876 | −1.15300073 | 97.80 |

**Πίνακας 3.1** ***Κρίσιμες τιμές για τον έλεγχο της μέγιστης εντροπίας για την πρώτη περίπτωση βαρών***

| *Percentile* | *Intercept* | $\sqrt{n}$ | $\sqrt{m}$ | log n | log m | $R^2(\%)$ |
|---|---|---|---|---|---|---|
| **90th** | 0.83901294 | 0.03275677 | 0.24742079 | −0.14428589 | −0.47539681 | 94.99 |
| **95th** | 1.10954872 | 0.03527129 | 0.22200145 | −0.17597696 | −0.49438062 | 96.25 |
| **99th** | 1.65067126 | 0.02941095 | 0.49670483 | −0.21132799 | −0.99370584 | 96.07 |

**Πίνακας 3.2** ***Κρίσιμες τιμές για τον έλεγχο της μέγιστης εντροπίας για την δεύτερη περίπτωση βαρών***

Γενικά το καλύτερο μοντέλο, είναι αυτό που δίνει τη μικρότερη εκτιμημένη διασπορά, άρα το παίρνουμε όταν το $R^2_{adj}$ πάρει τη μέγιστη τιμή του. Παρατηρούμε ότι στην πρώτη περίπτωση βαρών, οι τιμές του προσαρμοσμένου συντελεστή προσδιορισμού είναι αρκετά υψηλές, δηλαδή κοντά στο 100, άρα το βέλτιστο μοντέλο είναι αρκετά ικανοποιητικό. Στη δεύτερη περίπτωση, οι τιμές του είναι πιο χαμηλές, ενώ παράλληλα παρατηρούμε ότι στο 99° ποσοστημόριο, η μεταβλητή $\sqrt{n}$ έχει αρκετά χαμηλή τιμή συντελεστή ($\approx 0.02$), οπότε δε παίζει σημαντικό ρόλο στο

μοντέλο, δηλαδή δεν είναι ιδιαίτερα σημαντική. Παρ' όλα αυτά την κρατάμε, για λόγους ομοιογένειας στις μεταβλητές, καθώς επίσης και επειδή οι διαφορές των $R^2_{adj}$ του αρχικού και του "βέλτιστου" μοντέλου, παρουσιάζουν αρκετά μικρή μεταβολή.

## 3.4. ΜΕΛΕΤΗ ΠΡΟΣΟΜΟΙΩΣΗΣ

Για την μελέτη μας, και για τις δύο περιπτώσεις βαρών, θεωρούμε διάφορες συνεχείς κατανομές με παχιές ουρές, τόσο κάτω από τη μηδενική υπόθεση, όσο και ως πιθανές εναλλακτικές. Παράλληλα εφαρμόζουμε το θεώρημα μετασχηματισμού ολοκληρώματος πιθανότητας (probability integral transformation), όπου σε αυτή την περίπτωση έχουμε :

$$\boldsymbol{H_0}: \boldsymbol{Y} \sim \boldsymbol{F} = \boldsymbol{F_0} \Leftrightarrow \boldsymbol{H_0}: \boldsymbol{U} = \boldsymbol{F_0}(\boldsymbol{Y}) \sim \boldsymbol{U}[\mathbf{0}, \mathbf{1}]$$

Γενικά επιλέγουμε να επικεντρωθούμε σε κατανομές όπως είναι η Εκθετική, η Lognormal, η Pareto, η Inverse Gaussian(IG), η Weibull, η Log-Logistic και η Γάμμα, οι οποίες εμφανίζονται συχνά στη βίο-ιατρική, στη μηχανική και στην αξιοπιστία, ενώ, κάτω από τη μηδενική υπόθεση, η προσοχή μας επικεντρώνεται στις κατανομές Weibull, Inverse Gaussian και Log-Normal. Κάτω από τη μηδενική υπόθεση, η παράμετρος σχήματος (shape parameter) που θεωρήσαμε στην κατανομή Weibull είναι 0.552, στην Inverse Gaussian $1.059$ και στη Log-normal $0.62$, ενώ κατανομές με τις ίδιες παραμέτρους έχουν χρησιμοποιηθεί ως πιθανές εναλλακτικές στην κάθε περίπτωση. Παράλληλα, ως εναλλακτικές έχουν χρησιμοποιηθεί και οι κατανομές Pareto και Γάμμα με παραμέτρους σχήματος $0.78$ και $2$ αντίστοιχα, ενώ έχει γίνει χρήση και της παραμέτρου $3.57$ για την κατανομή Inverse Gaussian. Σε όλες τις περιπτώσεις, η παράμετρος κλίμακας (scale parameter) έχει θεωρηθεί ίση με 1, καθώς όλες αυτές κατανομές δεν επηρεάζονται από κάποια μεταβολή κλίμακας. Τέλος, παρατηρούμε ότι αν θεωρήσουμε παράμετρο σχήματος ίση με 1, οι κατανομές Γάμμα και Weibull, συμπίπτουν με την Εκθετική κατανομή.

Οι Monte-Carlo προσομοιώσεις στηρίζονται σε 10000 επαναλήψεις, χρησιμοποιώντας τις κρίσιμες τιμές που υπολογίζονται από τα μοντέλα που παρουσιάζονται στους **Πίνακες 3.1** και **3.2**., ενώ τέλος για λόγους πληρότητας, παρουσιάζονται για $n = 100$, εναλλακτικοί έλεγχοι που υπάρχουν διαθέσιμοι στη βιβλιογραφία, οι οποίοι είναι: αυτός των Anderson-Darling (AD) και των Cramér-von Mises (CVM). Γενικά στη μελέτη όλων των ελέγχων, έγινε χρήση των επιπέδων σημαντικότητας $0.10, 0.05$ και $0.01$ και τα αποτελέσματα που προέκυψαν σε κάποιες περιπτώσεις κατανομών, είναι αρκετά ικανοποιητικά. Ωστόσο, για λόγους έκτασης, δεν μπορούν να παρουσιαστούν όλα για την κάθε περίπτωση, και έτσι προβήκαμε στην ενδεικτική επιλογή ενός κάθε φορά.

### 3.4.1. ΠΡΩΤΗ ΠΕΡΙΠΤΩΣΗ ΒΑΡΩΝ

Στην πρώτη περίπτωση που εξετάζουμε, παίρνουμε κάτω από τη μηδενική υπόθεση, την κατανομή ***Weibull*** με παράμετρο $0.552$, ενώ τα αποτελέσματα που παρουσιάζονται στον **Πίνακα 3.3** και στο **Σχήμα 3.3,** αναφέρονται στο επίπεδο σημαντικότητας $\alpha = 0.05$. Παρόμοια αποτελέσματα έχουμε και στα υπόλοιπα επίπεδα.

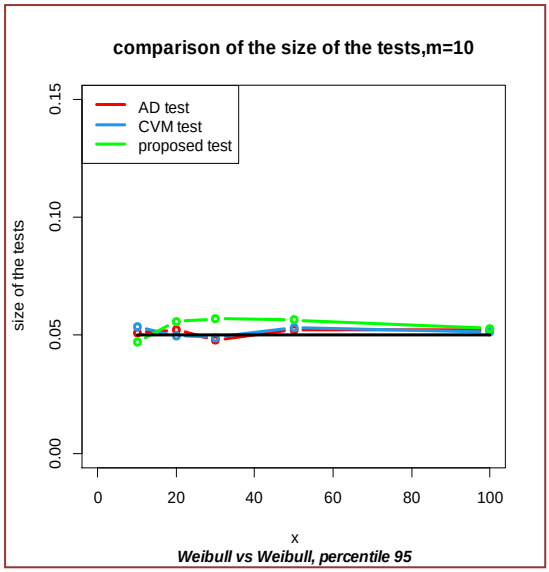
Σύμφωνα λοιπόν με αυτά, το ***μέγεθος*** του προτεινόμενου ελέγχου, σημειώνει αρκετά ικανοποιητικές τιμές, κοντά στο $0.05$, ενώ είναι αρκετά κοντά και στις αντίστοιχες τιμές που δίνουν οι έλεγχοι των Anderson-Darling και Cramér-von Mises. Ενδεικτικά για το διάστημα $m = 10$, κάποιος μπορεί να διαπιστώσει και γραφικά τη συμπεριφορά του ελέγχου της μέγιστης εντροπίας, και πως αυτός προσαρμόζεται σε σχέση κιόλας και με τους άλλους ελέγχους.

Έχοντας πάρει διάφορες κατανομές με παχιές ουρές ως εναλλακτικές υποθέσεις, από την μελέτη των τιμών της ***ισχύος*** του ελέγχου και τις γραφικές παραστάσεις, συμπεραίνουμε τη δυνατότητα του ελέγχου μέγιστης εντροπίας, να σημειώνει μια σημαντικά καλή προσαρμογή. Καθώς το μέγεθος του δείγματος αυξάνει, η ισχύς αυξάνεται, σε όλες τις περιπτώσεις των εναλλακτικών υποθέσεων, επομένως συμπεραίνουμε ότι οι κατανομές διαφέρουν πολύ, καθώς η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική. Τουλάχιστον για μικρά μεγέθη δείγματος είναι μικρή, ωστόσο αυτή η διαφορά μειώνεται γρήγορα καθώς το μέγεθος του δείγματος αυξάνεται, μέχρις ότου $n = 100$, όπου όλες οι παρατηρούμενες τιμές ισχύος είναι ίσες με τη μονάδα. Τέλος είναι σημαντικό να επισημανθεί, ότι ο έλεγχος της μέγιστης εντροπίας, βοηθάει στη διαφοροποίηση των κατανομών Inverse Gaussian και Log-Normal, καθώς συχνά οι τιμές τους δεν είναι πολύ διαφοροποιημένες μεταξύ τους, όταν βρίσκονται υπό διαδικασία ελέγχου.

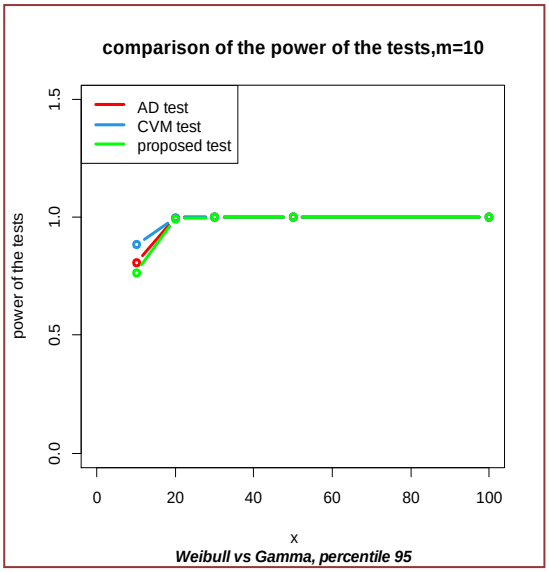
Πιο αναλυτικά, στην περίπτωση της *Inverse Gaussian*, ο προτεινόμενος έλεγχος προσαρμόζεται καλά, ενώ όμοια ή και καλύτερα από τους εναλλακτικούς ελέγχους προσαρμόζεται στα διαστήματα $m = 4$ και $5$, ειδικά στο επίπεδο σημαντικότητας $\alpha = 0.10$. Όταν η κατανομή *Pareto* είναι εναλλακτική, ο έλεγχος συμπεριφέρεται όπως ο Cramér-von Mises σε όλα τα επίπεδα, ωστόσο ο έλεγχος των Anderson-Darling είναι καλύτερος. Ενώ τέλος, στην περίπτωση της κατανομής *Log-Normal*, σχεδόν σε όλα τα διαστήματα και επίπεδα σημαντικότητας είναι καλύτερος από τους εναλλακτικούς ελέγχους, με καλύτερη προσαρμογή ίσως στο επίπεδο $\alpha = 0.10$, αφού η ισχύς σε αυτή την περίπτωση είναι υψηλή, ακόμα και για μικρά μεγέθη δείγματος. Στην περίπτωση της *Γάμμα*, συμπεριφέρεται όμοια με τους εναλλακτικούς ελέγχους, σε όλα τα επίπεδα σημαντικότητας.

| | | Maximum entropy test | | | | | Alternative test ($n = 100$) | |
|---|---|---|---|---|---|---|---|---|
| **m** | **shape** | **10** | **20** | **30** | **50** | **100** | **AD** | **CVM** |
| **3** | **Weibull 0.552** | 0.0499 | 0.0473 | 0.0605 | 0.0533 | 0.0567 | 0.0535 | 0.0540 |
| **4** | | 0.0468 | 0.0496 | 0.0511 | 0.0541 | 0.0573 | 0.0524 | 0.0522 |
| **5** | | 0.0571 | 0.0483 | 0.0475 | 0.0552 | 0.0486 | 0.0450 | 0.0468 |
| **10** | | 0.0470 | 0.0559 | 0.0570 | 0.0565 | 0.0529 | 0.0525 | 0.0513 |
| **3** | **Log-Normal 0.62** | 0.5269 | 0.9987 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.8628 | 0.9983 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.8617 | 0.9992 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.7416 | 0.9979 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Inverse Gaussian 1.059** | 0.7046 | 0.9867 | 0.9997 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.9701 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.9197 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.3946 | 0.9993 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Gamma 2** | 0.7961 | 0.9955 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.8486 | 0.9973 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.8342 | 0.9952 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.7651 | 0.9938 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Pareto 0.78** | 0.4233 | 0.6737 | 0.8783 | 0.9798 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.4093 | 0.7218 | 0.8816 | 0.9849 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.4350 | 0.7147 | 0.8933 | 0.9874 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.4055 | 0.7182 | 0.8905 | 0.9897 | 1.0000 | 1.0000 | 1.0000 |

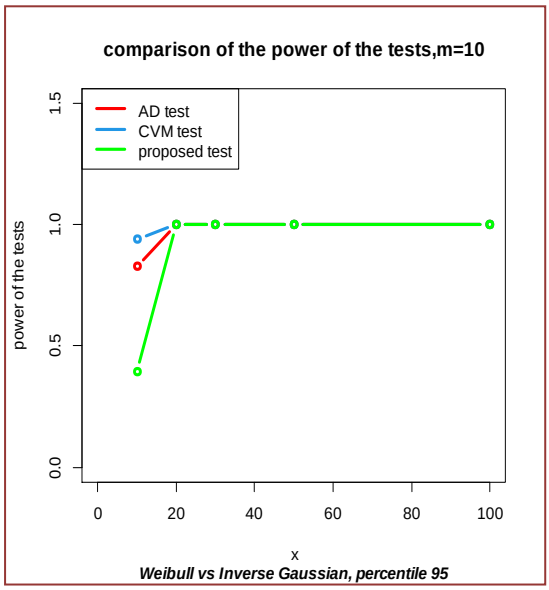
**Πίνακας 3.3 :** ***Weibull null model:* μέγεθος και ισχύ του ελέγχου της μέγιστης εντροπίας για $n = 10, 20, 30, 50, 100$ και οι εναλλακτικοί έλεγχοι για $n = 100$**



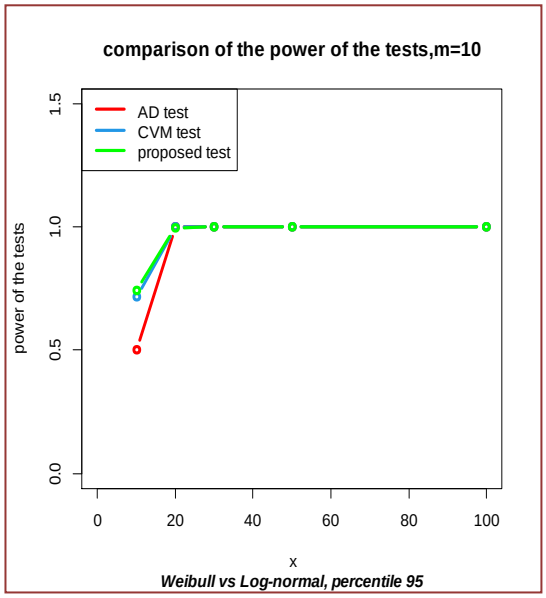

*Weibull vs Weibull, percentile 95*



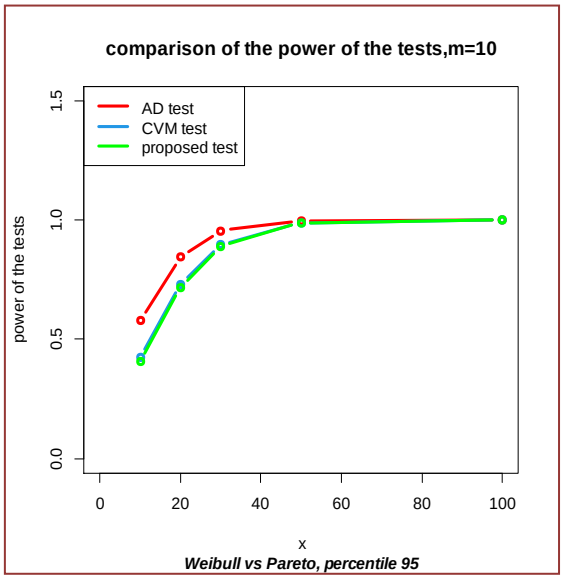

*Weibull vs Gamma, percentile 95*




*Weibull vs Inverse Gaussian, percentile 95*




*Weibull vs Log-normal, percentile 95*




*Weibull vs Pareto, percentile 95*

***Σχήμα 3.3:*** ***Γραφική παράσταση της κατανομής Weibull με εναλλακτικές υποθέσεις την Weibull, Gamma, Log-Normal,Pareto και Inverse Gaussian, για*** $m = 10$ ***σε επίπεδο σημαντικότητας*** $\alpha = 0.05$

Στον **Πίνακα 3.4** απεικονίζονται τα αποτελέσματα μεγέθους και ισχύος, της δεύτερης περίπτωσης που εξετάζουμε, όπου κάτω από τη μηδενική υπόθεση έχουμε πάρει την κατανομή ***Log-Normal*** με παράμετρο $0.62$ ,σε επίπεδο σημαντικότητας $\alpha = 0.05$.

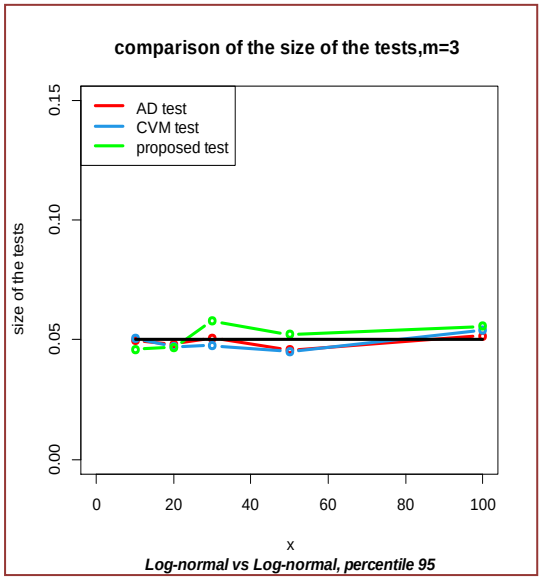
Γενικά τα αποτελέσματα για το ***μέγεθος*** του ελέγχου της μέγιστη εντροπίας είναι πολύ ικανοποιητικά, αρκετά κοντά στο αντίστοιχο επίπεδο σημαντικότητας που παίρνουμε κάθε φορά. Ενδεικτικά μπορεί να το παρατηρήσει κάποιος για $m = 3$, στο **Σχήμα 3.4**. Παρόμοια αποτελέσματα έχουμε και στα υπόλοιπα επίπεδα.

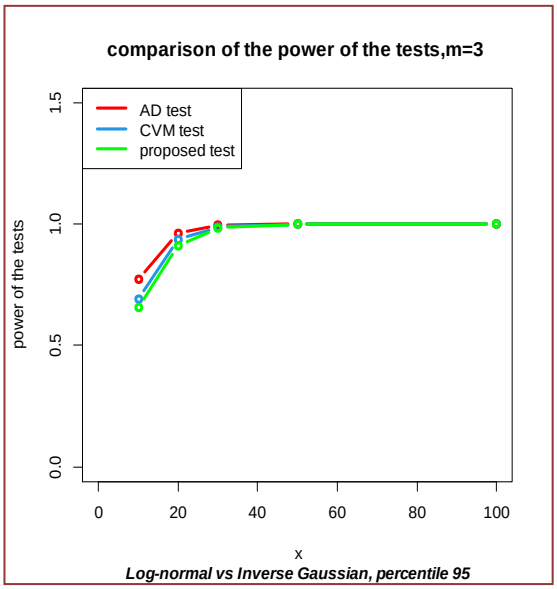
Τώρα όσον αφορά την ***ισχύ*** του ελέγχου, για όλες τις περιπτώσεις των εναλλακτικών κατανομών που έχουμε πάρει, παρατηρούμε σταδιακή αύξησή της, φτάνοντας τη μέγιστη τιμή της, καθώς αυξάνει το μέγεθος του δείγματος. Αυτό έχει ως αποτέλεσμα λοιπόν, να δημιουργείται διαφοροποίηση μεταξύ της πραγματικής και της υποθετικής κατανομής. Συγκεκριμένα, για την κατανομή *Gamma*, η προσαρμογή του προτεινόμενου ελέγχου είναι πολύ καλή για όλα τα επίπεδα σημαντικότητας. Σε όλα τα διαστήματα συμπεριφέρεται όμοια ή και καλύτερα από τον έλεγχο των Cramér-von Mises, ωστόσο, ο έλεγχος Anderson-Darling είναι ισχυρότερος όλων. Όμοια είναι τα αποτελέσματα και για την κατανομή *Inverse Gaussian*. Στην περίπτωση της *Pareto* ως εναλλακτικής κατανομής, παρατηρούμε ότι σε όλα επίπεδα σημαντικότητας, οι εναλλακτικοί έλεγχοι είναι καλύτεροι, ειδικά ο έλεγχος των Anderson-Darling. Για το διάστημα $m = 10$ στα επίπεδα $\alpha = 0.01$ και $0.05$, ο προτεινόμενος έλεγχος είναι καλύτερος από τον έλεγχο Cramér-von Mises, ενώ στο $\alpha = 0.10$ συμπεριφέρεται όμοια με αυτόν.

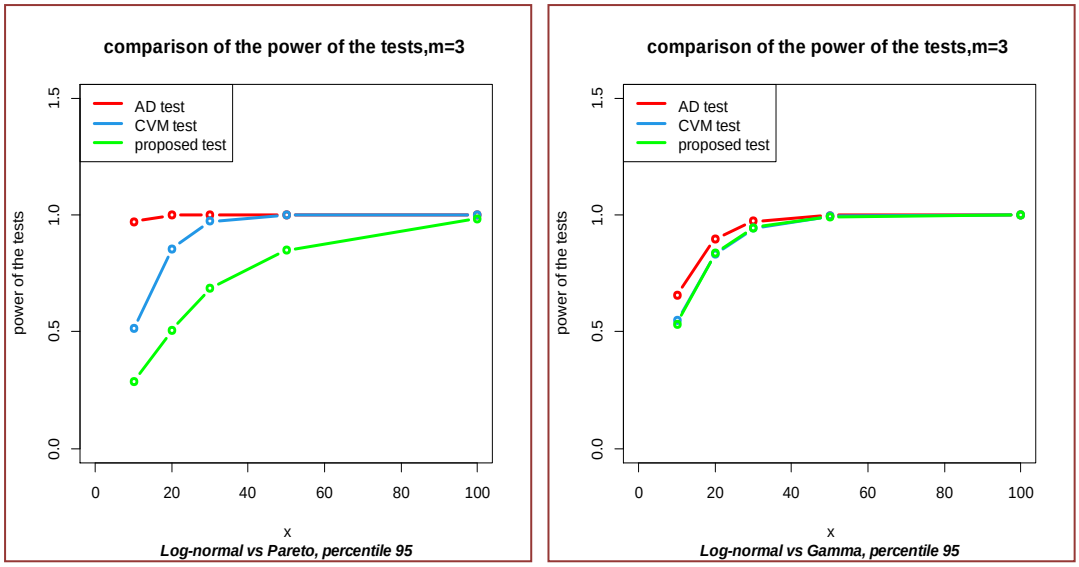
| | | **Maximum entropy test** | | | | | **Alternative test** $(n = 100)$ | |
|---|---|---|---|---|---|---|---|---|
| **m** | **shape** | **10** | **20** | **30** | **50** | **100** | **AD** | **CVM** |
| **3** | **Log-Normal 0.62** | 0.0461 | 0.0469 | 0.0578 | 0.0523 | 0.0557 | 0.0517 | 0.0539 |
| **4** | | 0.0492 | 0.0435 | 0.0505 | 0.0513 | 0.0534 | 0.0492 | 0.0473 |
| **5** | | 0.0509 | 0.0501 | 0.0517 | 0.0508 | 0.0512 | 0.0508 | 0.0495 |
| **10** | | 0.0470 | 0.0542 | 0.0573 | 0.0585 | 0.0522 | 0.0513 | 0.0497 |
| **3** | **Inverse** | 0.6573 | 0.9116 | 0.9847 | 0.9997 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.6852 | 0.9384 | 0.9891 | 0.9997 | 1.0000 | 1.0000 | 1.0000 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **5** | **Gaussian 3.57** | 0.7271 | 0.9408 | 0.9921 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.6806 | 0.9422 | 0.9921 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Pareto 0.78** | 0.2879 | 0.5058 | 0.6886 | 0.8487 | 0.9858 | 1.0000 | 1.0000 |
| **4** | | 0.3987 | 0.6701 | 0.8262 | 0.9631 | 0.9992 | 1.0000 | 1.0000 |
| **5** | | 0.5203 | 0.7407 | 0.8989 | 0.9825 | 0.9999 | 1.0000 | 1.0000 |
| **10** | | 0.6262 | 0.8609 | 0.9575 | 0.9962 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Gamma 2** | 0.5335 | 0.8360 | 0.9469 | 0.9941 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.5846 | 0.8620 | 0.9584 | 0.9972 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.6395 | 0.8703 | 0.9645 | 0.9981 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.5905 | 0.8778 | 0.9684 | 0.9983 | 1.0000 | 1.0000 | 1.0000 |

**Πίνακας 3.4 :** ***Log-Normal null model:* μέγεθος και ισχύ του ελέγχου της μέγιστης εντροπίας για $n = 10, 20, 30, 50, 100$ και οι εναλλακτικοί έλεγχοι για $n = 100$**






***Σχήμα 3.4 : Γραφική παράσταση της κατανομής Log-Normal με εναλλακτικές υποθέσεις την Log-normal, Inverse Gaussian, Gamma και Pareto, για $m = 3$ σε επίπεδο σημαντικότητας $\alpha = 0.05$***

Σε αυτή την περίπτωση εξετάζουμε κάτω από τη μηδενική υπόθεση, την κατανομή ***Inverse Gaussian*** με παράμετρο $1.059$ .Τα αποτελέσματα που παρουσιάζονται για αυτή την περίπτωση (**Πίνακας 3.5, Σχήμα 3.5**) αφορούν το $\alpha = 0.05$, ωστόσο σε όλα τα επίπεδα, παρατηρήθηκαν παραπλήσια αποτελέσματα.

Όσον αφορά το ***μέγεθος*** του ελέγχου της μέγιστης εντροπίας είναι αρκετά ικανοποιητικό και σημαντικά κοντά στο αντίστοιχο επίπεδο σημαντικότητας που έχουμε πάρει κάθε φορά. Επομένως, αφού το μέγεθος κρατιέται όσο πιο κοντά γίνεται στο επίπεδο σημαντικότητας, το σφάλμα τύπου I είναι ελεγχόμενο. Επίσης, όπως ήταν επόμενο, παρατηρούμε ότι όταν το μέγεθος του δείγματος είναι πολύ μικρό και ο αριθμός των κλάσεων είναι πολύ μεγάλος, οι τιμές είναι μικρότερες, αφού τα διαστήματα όπως έχουν κατασκευαστεί, μπορεί να περιέχουν ένα μικρό αριθμό παρατηρήσεων.

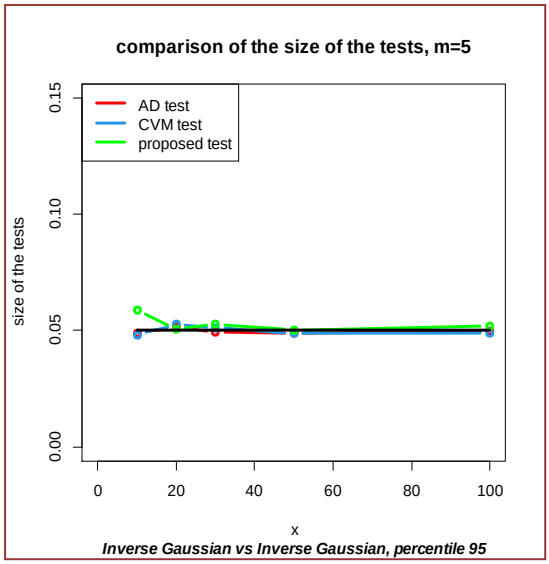
Τώρα παρατηρώντας την ***ισχύ*** του ελέγχου, οδηγούμαστε στο συμπέρασμα ότι η επίδοση του προτεινόμενου ελέγχου είναι καλή σε όλα τα επίπεδα σημαντικότητας. Η προσαρμογή του προτεινόμενου ελέγχου στις περιπτώσεις των κατανομών *Log-Normal* και *Gamma* είναι υπερβολικά ικανοποιητική, ενώ η συμπεριφορά του ελέγχου είναι όμοια με τους εναλλακτικούς ελέγχους. Η ισχύς πιάνει τη μέγιστη τιμή της από πολύ μικρά μεγέθη δείγματος, συμβάλλοντας έτσι στη διαφοροποίηση της υποθετικής και της πραγματικής κατανομής. Παράλληλα, εκδηλώνεται η δυνατότητα του ελέγχου να διακρίνει και να διαφοροποιεί τις κατανομές Inverse Gaussian και Log-Normal, κάτι το οποίο είναι αρκετά δύσκολο, καθώς οι συγκεκριμένες κατανομές παρουσιάζουν αρκετές ομοιότητες όταν βρίσκονται υπό εξέταση. Παρόμοια αποτελέσματα έχουμε και όταν η *Weibull* είναι η εναλλακτική

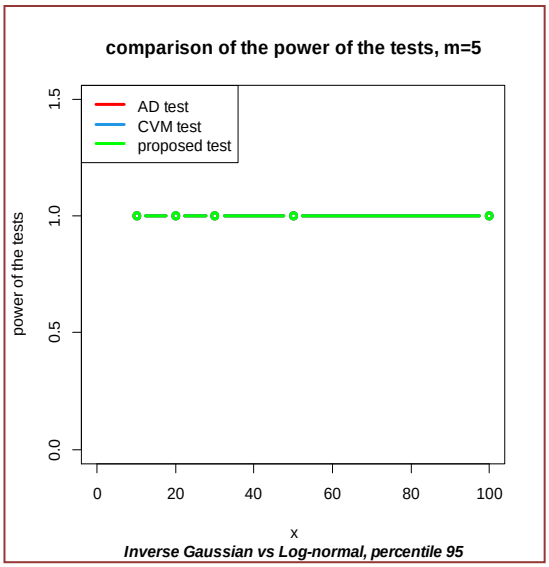
κατανομή. Η ισχύς του ελέγχου είναι μικρή για μικρό μέγεθος δείγματος, και καθώς το μέγεθος του δείγματος αυξάνεται, αυξάνει και αυτή, πλησιάζοντας και φτάνοντας τη μονάδα. Σε όλα τα επίπεδα, παρατηρείται ότι η προσαρμογή του ελέγχου είναι ικανοποιητική, ωστόσο οι εναλλακτικοί έλεγχοι είναι καλύτεροι. Μόνο ίσως για $m = 10$, ο προτεινόμενος έλεγχος συμπεριφέρεται όπως ο έλεγχος Cramér-von Mises, ή και καλύτερα. Τέλος, όταν ο έλεγχος εφαρμοστεί με εναλλακτική υπόθεση την κατανομή *Pareto*, συμπεριφέρεται όπως οι εναλλακτικοί έλεγχοι. Η προσαρμογή του ελέγχου είναι πολύ καλή, ειδικά στα επίπεδα $\alpha = 0.05$ και $\alpha = 0.10$, αφού μεγιστοποιείται η ισχύς του, έχοντας αρκετά υψηλές τιμές ακόμα και για μικρά μεγέθη δείγματος, συμβάλλοντας με αυτό τον τρόπο στη διαφοροποίηση της υποθετικής κατανομής και της κατανομής Pareto.

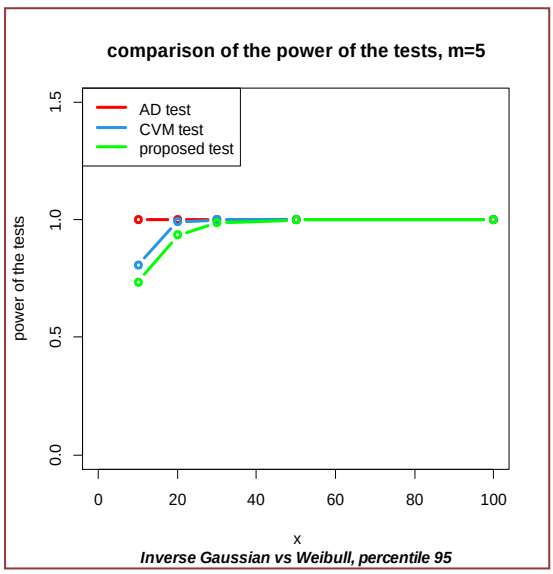
| | | **Maximum entropy test** | | | | | **Alternative test** $(n = 100)$ | |
|---|---|---|---|---|---|---|---|---|
| **m** | **shape** | **10** | **20** | **30** | **50** | **100** | **AD** | **CVM** |
| **3** | **Inverse Gaussian 1.059** | 0.0477 | 0.0504 | 0.0606 | 0.0551 | 0.0528 | 0.0505 | 0.0502 |
| **4** | | 0.0489 | 0.0474 | 0.0511 | 0.0500 | 0.0534 | 0.0512 | 0.0497 |
| **5** | | 0.0590 | 0.0506 | 0.0527 | 0.0503 | 0.0518 | 0.0497 | 0.0488 |
| **10** | | 0.0449 | 0.0499 | 0.0552 | 0.0577 | 0.0489 | 0.0465 | 0.0485 |
| **3** | **Log-Normal 0.62** | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Weibull 0.552** | 0.4598 | 0.7306 | 0.8856 | 0.9749 | 0.9995 | 1.0000 | 1.0000 |
| **4** | | 0.6033 | 0.8834 | 0.9695 | 0.9983 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.7358 | 0.9355 | 0.9879 | 0.9998 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.8340 | 0.9786 | 0.9975 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Gamma** | 0.9996 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

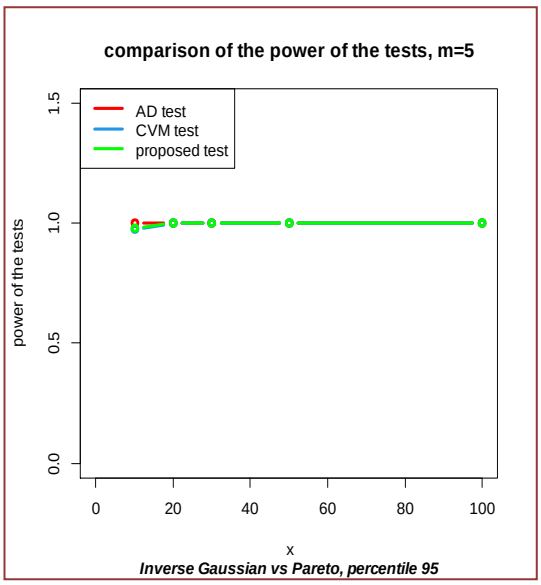
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **5** | **2** | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Pareto 0.78** | 0.9099 | 0.9974 | 0.9997 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.9598 | 0.9995 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.9818 | 0.9997 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.9935 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

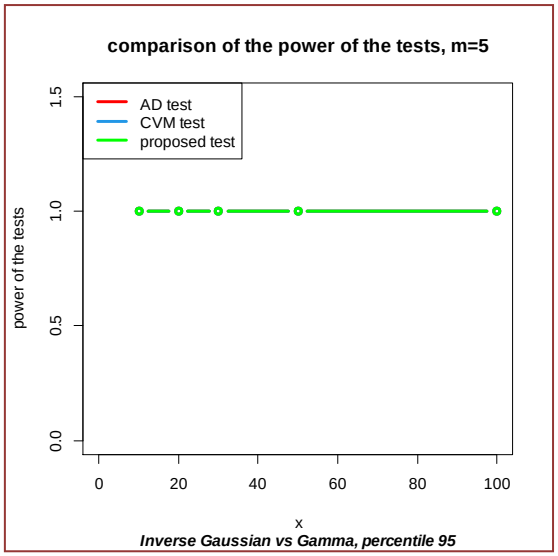
**Πίνακας 3.5 :** ***Inverse Gaussian null model*: μέγεθος και ισχύ του ελέγχου της μέγιστης εντροπίας για $n = 10, 20, 30, 50, 100$ και οι εναλλακτικοί έλεγχοι για $n = 100$**






***Σχήμα 3.5:*** ***Γραφική παράσταση της κατανομής Inverse Gaussian με εναλλακτικές υποθέσεις την Inverse Gaussian, Weibull, Gamma, Pareto και Log-Normal, για*** $m = 5$ ***σε επίπεδο σημαντικότητας*** $\alpha = 0.05$

### 3.4.2. ΔΕΥΤΕΡΗ ΠΕΡΙΠΤΩΣΗ ΒΑΡΩΝ

Όμοια με την πρώτη περίπτωση βαρών, εξετάζουμε τον έλεγχο υποθέσεων, σύμφωνα με τον οποίο, κάτω από τη μηδενική υπόθεση, παίρνουμε την κατανομή ***Weibull*** με παράμετρο $0.552$ ,έναντι μιας πληθώρας εναλλακτικών υποθέσεων. Στον **Πίνακα 3.6** και το **Σχήμα 3.6**, παρατίθενται τα αποτελέσματα και η γραφικές απεικονίσεις για αυτόν τον έλεγχο, σε επίπεδο σημαντικότητας $\alpha = 0.01$.

Στην αρχή εξετάζουμε το ***μέγεθος*** του ελέγχου, όπου κανείς μπορεί να παρατηρήσει ότι οι τιμές του είναι αρκετά ικανοποιητικές, πολύ κοντά δηλαδή στα αντίστοιχα επίπεδα σημαντικότητας. Ειδικά για τα διαστήματα $m = 4$ και $5$, ο έλεγχος της μέγιστης εντροπίας προσαρμόζεται πολύ καλά. Επίσης, παρατηρούμε όπως αναμένονταν, ότι όταν το μέγεθος του δείγματος είναι πολύ μικρό και ο αριθμός των κλάσεων είναι πολύ μεγάλος, οι τιμές είναι μικρότερες.

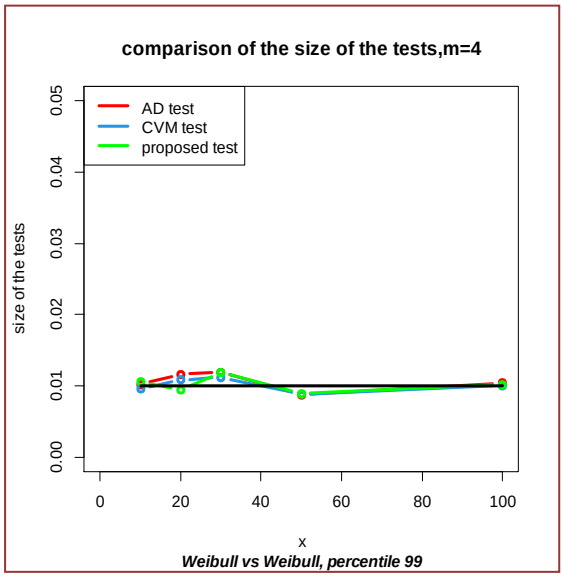
Παρατηρώντας τώρα την ***ισχύ*** του ελέγχου, γενικά μπορούμε να συμπεράνουμε ότι ο προτεινόμενος έλεγχος προσαρμόζεται ικανοποιητικά έναντι των εναλλακτικών υποθέσεων που έχουμε πάρει, όντας αρκετά κοντά στους εναλλακτικούς. Όταν η κατανομή *Inverse Gaussian* είναι στην εναλλακτική υπόθεση, η ισχύς ξεκινάει από χαμηλές τιμές (στα υπόλοιπα επίπεδα, ξεκινάει από υψηλότερες), τουλάχιστον για μικρά μεγέθη δείγματος, ωστόσο αυτή η διαφορά μειώνεται γρήγορα καθώς το μέγεθος του δείγματος αυξάνεται, μέχρις ότου όλες οι παρατηρούμενες τιμές ισχύος είναι ίσες με τη μονάδα. Ωστόσο, σε όλα τα επίπεδα, οι εναλλακτικοί έλεγχοι είναι καλύτεροι. Το ίδιο συμβαίνει και στην περίπτωση της κατανομής *Log-Normal*, με τη μόνη διαφορά ότι ο έλεγχος στο επίπεδο $\alpha = 0.05$, στα διαστήματα $m = 4, 5$ και $10$, προσαρμόζεται όμοια ή και καλύτερα από τον έλεγχο Anderson-Darling. Πολύ καλή προσαρμογή παρατηρεί κανείς και όταν η

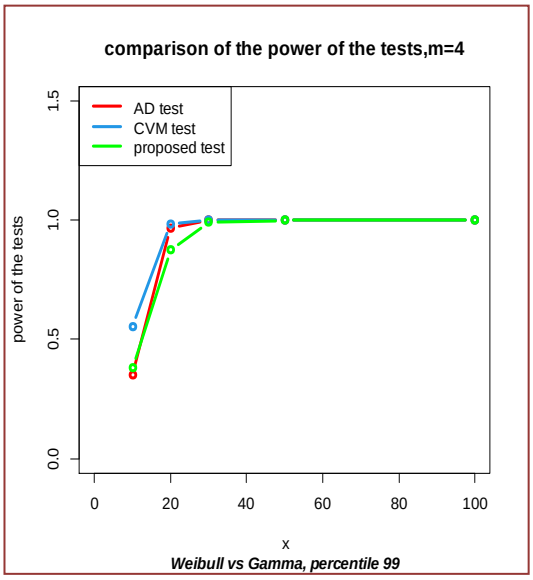
κατανομή *Pareto* είναι στην εναλλακτική υπόθεση, αφού έχουμε σταδιακή αύξηση της ισχύος, η οποία πλησιάζει τη μονάδα. Γενικά, σε αυτή την περίπτωση, ο έλεγχος των Anderson-Darling είναι ο καλύτερος, σε όλα τα επίπεδα σημαντικότητας, ενώ ο προτεινόμενος έλεγχος συμπεριφέρεται όπως των Cramér-von Mises στα επίπεδα $\alpha = 0.05$ και $0.10$ και σχετικά κοντά με αυτόν στο επίπεδο $\alpha = 0.01$ (**Σχήμα 3.6**). Στην περίπτωση της *Γάμμα* κατανομής ως εναλλακτικής, έχουμε πάλι μεγιστοποίηση της ισχύος, επομένως η εναλλακτική υπόθεση απομακρύνεται από τη μηδενική, ενώ οι τιμές του προτεινόμενου ελέγχου δεν διαφέρουν αρκετά και από εκείνες των εναλλακτικών. Πολύ καλή προσαρμογή, παρουσιάζεται στο επίπεδο $\alpha = 0.10$, με την ισχύ να παίρνει υψηλές τιμές από μικρά μεγέθη δείγματος.

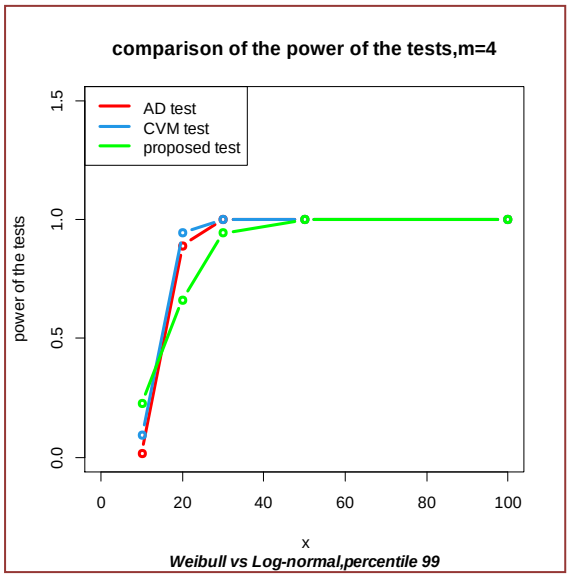
| | | **Maximum entropy test** | | | | | **Alternative test** $(n = 100)$ | |
|---|---|---|---|---|---|---|---|---|
| **m** | **shape** | **10** | **20** | **30** | **50** | **100** | **AD** | **CVM** |
| **3** | **Weibull 0.552** | 0.0187 | 0.0135 | 0.0086 | 0.0126 | 0.0121 | 0.0106 | 0.0107 |
| **4** | | 0.0106 | 0.0094 | 0.0119 | 0.0089 | 0.0102 | 0.0104 | 0.0100 |
| **5** | | 0.0092 | 0.0105 | 0.0107 | 0.0101 | 0.0119 | 0.0094 | 0.0086 |
| **10** | | 0.0080 | 0.0096 | 0.0108 | 0.0132 | 0.0203 | 0.0109 | 0.0103 |
| **3** | **Log-Normal 0.62** | 0.0577 | 0.1195 | 0.5339 | 0.9998 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.2283 | 0.6634 | 0.9440 | 0.9994 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.2421 | 0.6742 | 0.9537 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.0628 | 0.6160 | 0.9554 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Inverse Gaussian 1.059** | 0.0396 | 0.3079 | 0.6791 | 0.9652 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.0000 | 0.1395 | 0.7329 | 0.9998 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.0004 | 0.0170 | 0.1720 | 0.9022 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.0000 | 0.0004 | 0.0423 | 0.8986 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Gamma 2** | 0.5600 | 0.9009 | 0.9922 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.3829 | 0.8762 | 0.9918 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.4010 | 0.8547 | 0.9902 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.3121 | 0.8437 | 0.9886 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

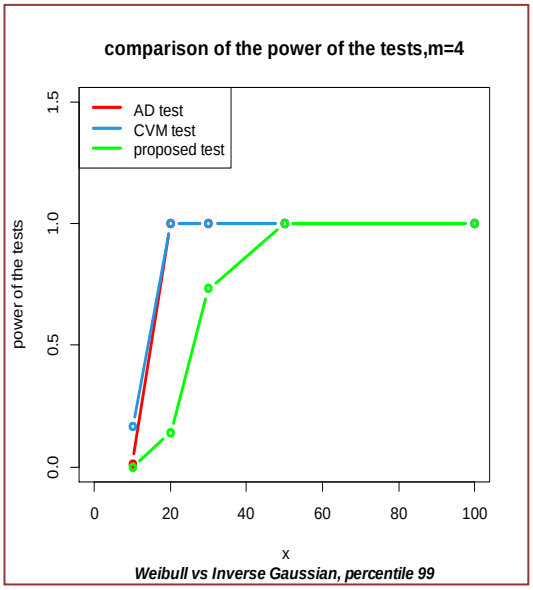
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **3** | **Pareto 0.78** | 0.2486 | 0.4335 | 0.6129 | 0.8875 | 0.9975 | 0.9999 | 0.9997 |
| **4** | | 0.1946 | 0.4343 | 0.6492 | 0.9025 | 0.9986 | 1.0000 | 0.9998 |
| **5** | | 0.1980 | 0.4430 | 0.6603 | 0.9117 | 0.9994 | 1.0000 | 0.9998 |
| **10** | | 0.2108 | 0.4838 | 0.7128 | 0.9435 | 0.9996 | 1.0000 | 0.9994 |

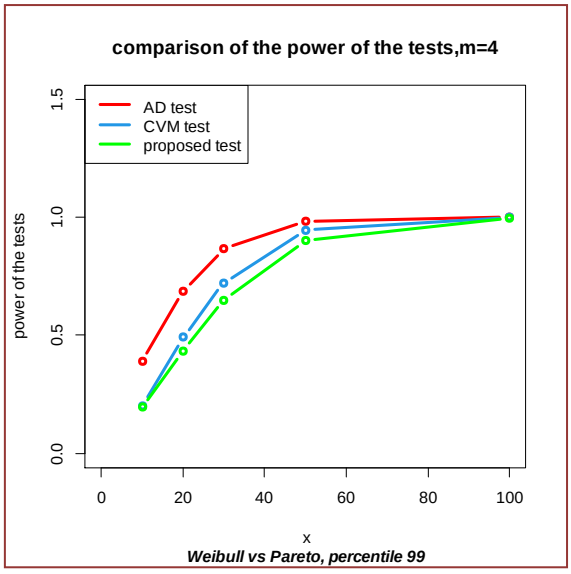
**Πίνακας 3.6 :** ***Weibull null model:* μέγεθος και ισχύ του ελέγχου της μέγιστης εντροπίας για $n = 10, 20, 30, 50, 100$ και οι εναλλακτικοί έλεγχοι για $n = 100$**






***Σχήμα 3.6 : Γραφική αναπαράσταση της κατανομής Weibull, κάτω από τη μηδενική υπόθεση, έναντι των κατανομών Weibull, Log-Normal, Gamma, Pareto και Inverse Gaussian ως εναλλακτικές υποθέσεις σε*** $\alpha = 0.01$

Τώρα κάτω από τη μηδενική υπόθεση, παίρνουμε την κατανομή ***Inverse Gaussian*** με παράμετρο $1.059$, ενώ τα αποτελέσματα από τις αναλύσεις παρατίθενται στον **Πίνακα 3.7** και στο **Σχήμα 3.7** .Όλα τα αποτελέσματα αφορούν το επίπεδο σημαντικότητας $\alpha = 0.01$, όμως παραπλήσια αποτελέσματα προκύπτουν και στα υπόλοιπα επίπεδα.

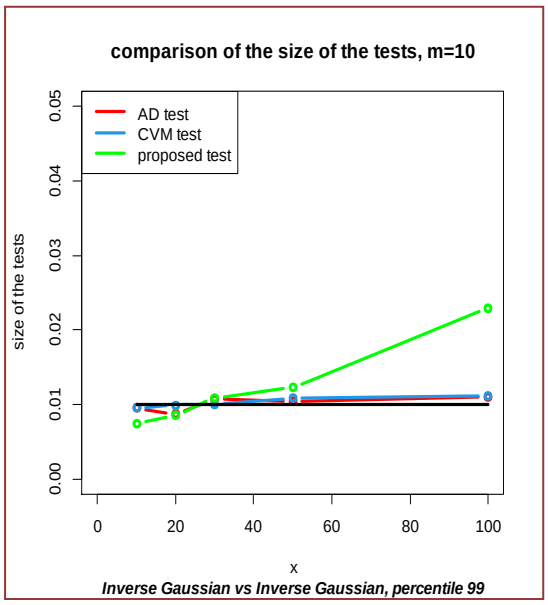
Αρχικά λοιπόν, μελετάμε το ***μέγεθος*** του ελέγχου, δηλαδή το σφάλμα τύπου I, όπου σε όλα τα επίπεδα σημαντικότητας είναι αρκετά ικανοποιητικό, με τιμές κοντά στο αντίστοιχο επίπεδο που παίρνουμε κάθε φορά. Ειδικά για τα διαστήματα $m = 4$ και $5$, η προσαρμογή του ελέγχου είναι πολύ καλή για όλα τα επίπεδα.

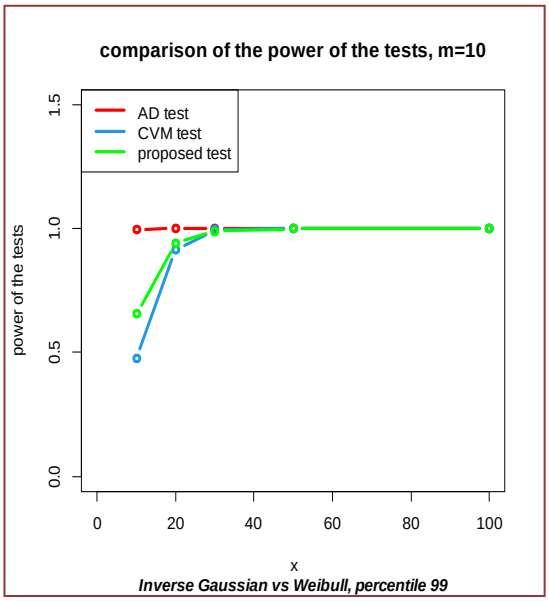
Για την ***ισχύ*** του ελέγχου, παρατηρούμε ότι όταν η κατανομή *Weibull* είναι στην εναλλακτική υπόθεση, οι παρατηρούμενες τιμές ισχύος αυξάνονται, καθώς αυξάνει το μέγεθος του δείγματος, φτάνοντας τη μέγιστη τιμή 1. Επομένως, αντιλαμβανόμαστε ότι υπάρχει σημαντική διαφοροποίηση μεταξύ της μηδενικής και της εναλλακτικής υπόθεσης, ελαχιστοποιώντας έτσι την πιθανότητα σφάλματος τύπου II . Σε όλα τα επίπεδα σημαντικότητας, ο προτεινόμενος έλεγχος προσαρμόζεται καλά, όχι όμως όπως οι εναλλακτικοί έλεγχοι και ειδικά τον έλεγχο Anderson-Darling, ενώ για το διάστημα $m = 10$, προσαρμόζεται όμοια ή και καλύτερα από τον έλεγχο Cramér-von Mises. Αντίθετα, στην περίπτωση της *Pareto*, όπου πάλι παρατηρείται αύξηση και μεγιστοποίηση της ισχύος, ο έλεγχος της μέγιστης εντροπίας προσαρμόζεται τέλεια, όπως και οι εναλλακτικοί, σε όλα τα επίπεδα σημαντικότητας. Παρόμοια, και όταν οι κατανομές *Log-Normal* και *Gamma* βρίσκονται στην εναλλακτική υπόθεση. Παρατηρούμε πολύ ικανοποιητική

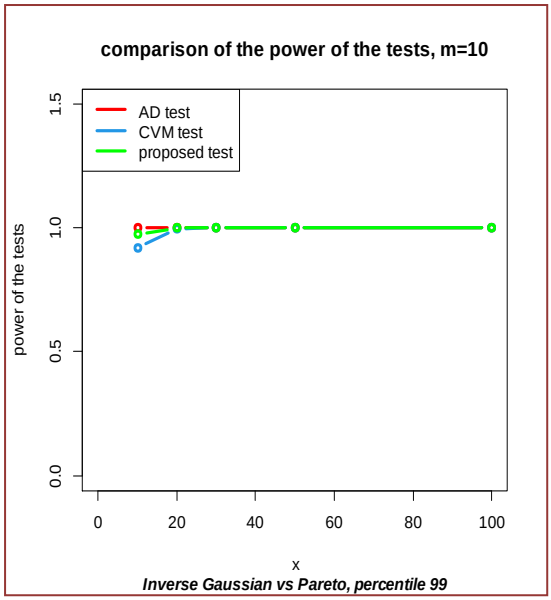
προσαρμογή του προτεινόμενου ελέγχου, σε όλα τα επίπεδα σημαντικότητας, διαφοροποιώντας πολύ τις πραγματικές κατανομές από την Inverse Gaussian.

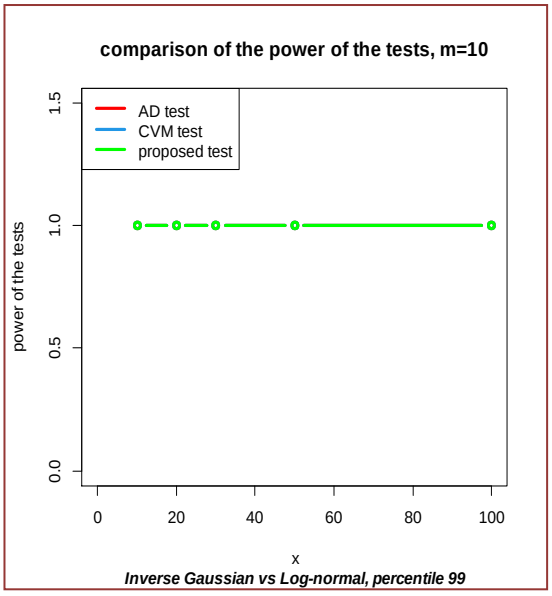
| | | Maximum entropy test | | | | | Alternative test $(n = 100)$ | |
|---|---|---|---|---|---|---|---|---|
| **m** | **shape** | **10** | **20** | **30** | **50** | **100** | **AD** | **CVM** |
| **3** | **Inverse Gaussian 1.059** | 0.0191 | 0.0122 | 0.0100 | 0.0142 | 0.0116 | 0.0093 | 0.0082 |
| **4** | | 0.0090 | 0.0115 | 0.0101 | 0.0104 | 0.0109 | 0.0091 | 0.0097 |
| **5** | | 0.0092 | 0.0109 | 0.0097 | 0.0113 | 0.0116 | 0.0082 | 0.0092 |
| **10** | | 0.0074 | 0.0085 | 0.0109 | 0.0123 | 0.0229 | 0.0110 | 0.0112 |
| **3** | **Log-Normal 0.62** | 0.9998 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.9996 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Weibull 0.552** | 0.3339 | 0.5341 | 0.6672 | 0.8793 | 0.9958 | 1.0000 | 1.0000 |
| **4** | | 0.4414 | 0.6929 | 0.8766 | 0.9783 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.4874 | 0.8112 | 0.9368 | 0.9944 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.6554 | 0.9421 | 0.9897 | 0.9996 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Pareto 0.78** | 0.8808 | 0.9903 | 0.9996 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.8996 | 0.9965 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.9459 | 0.9985 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.9749 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Gamma 2** | 0.9990 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

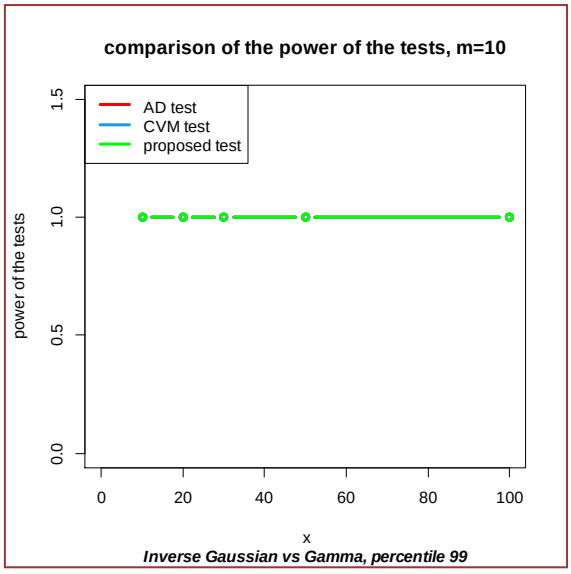
**Πίνακας 3.7 :** ***Inverse Gaussian null model:* μέγεθος και ισχύς του ελέγχου της μέγιστης εντροπίας για $n = 10, 20, 30, 50, 100$ και οι εναλλακτικοί έλεγχοι για $n = 100$**






***Σχήμα 3.7 : Γραφική αναπαράσταση της κατανομής Inverse Gaussian, κάτω από τη μηδενική υπόθεση, έναντι των κατανομών Inverse Gaussian,Weibull, Log-Normal, Gamma και Pareto ως εναλλακτικές υποθέσεις***

Στην τελευταία περίπτωση που εξετάζουμε, κάτω από τη μηδενική υπόθεση βρίσκεται η κατανομή ***Log-Normal*** με παράμετρο $0.62$. Όλα τα αποτελέσματα για αυτή την περίπτωση αφορούν το επίπεδο σημαντικότητας $\alpha = 0.01$, και παρατίθενται στον **Πίνακα 3.8** και το **Σχήμα 3.8**.

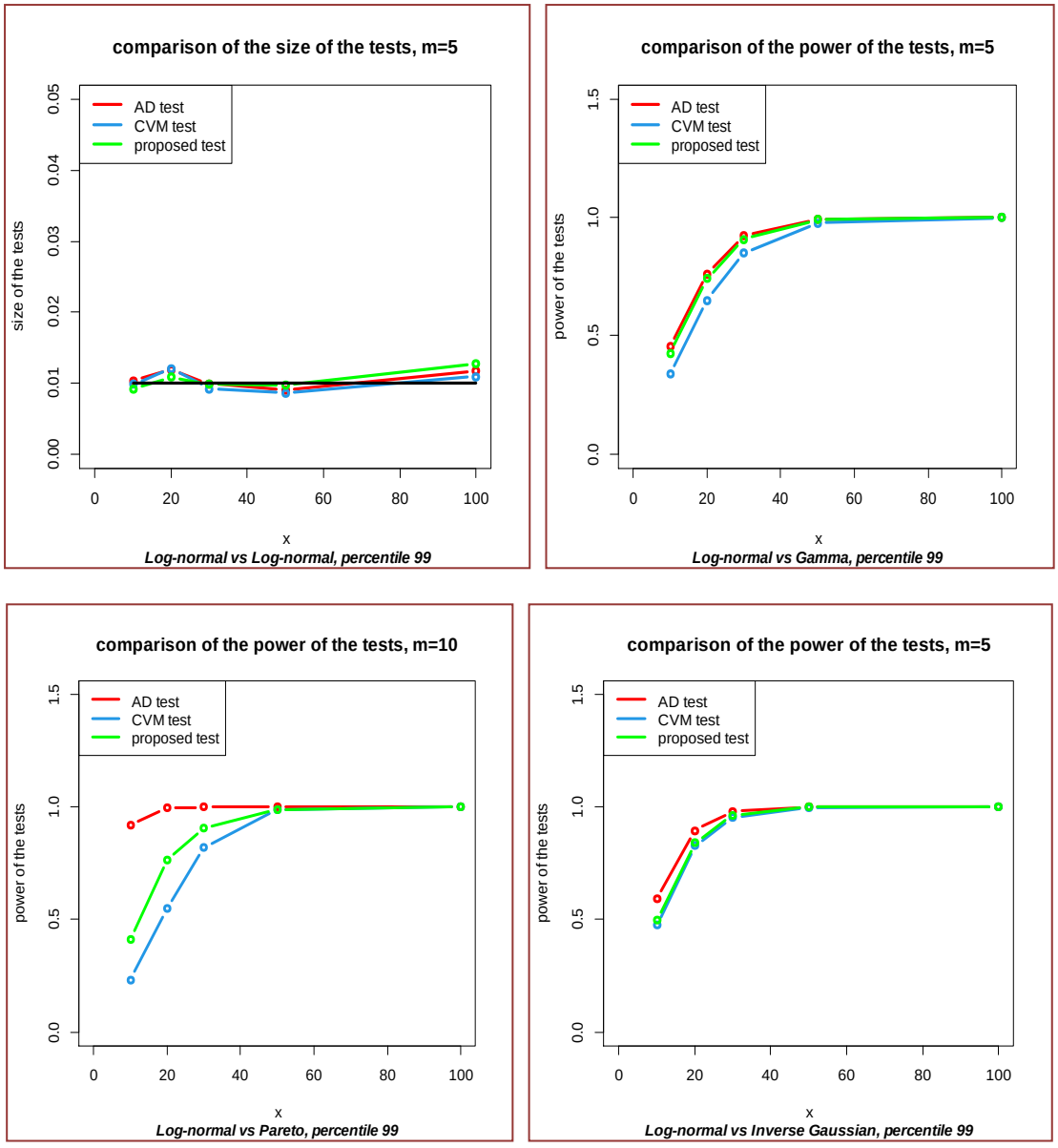
Όσον αφορά λοιπόν το ***μέγεθος*** του ελέγχου, παρατηρούμε ότι είναι αρκετά ικανοποιητικό, καθώς βρίσκεται κοντά στο επίπεδο που επιλέγουμε κάθε φορά, ακόμα και στα αρκετά μικρά μεγέθη δείγματος, ειδικά για τα διαστήματα $m = 4$ και $5$. Ελέγχοντας με αυτό τον τρόπο, το σφάλμα τύπου I, μπορούμε να προχωρήσουμε στη μελέτη της ισχύος, για να εξετάσουμε την προσαρμοστικότητα του ελέγχου.

Αναφορικά λοιπόν με την ***ισχύ*** του ελέγχου, όταν η κατανομή *Pareto* είναι στην εναλλακτική υπόθεση, η ισχύς αυξάνεται καθώς αυξάνεται το μέγεθος του δείγματος, πλησιάζοντας και φτάνοντας κιόλας τη μέγιστη τιμή της, εκφράζοντας με αυτό τον τρόπο την ικανοποιητική προσαρμοστικότητα του ελέγχου. Αυτό ισχύει σε όλα τα επίπεδα σημαντικότητας, ωστόσο οι εναλλακτικοί έλεγχοι είναι καλύτεροι, με πιο ισχυρό αυτόν των Anderson-Darling, ενώ παρατηρείται ότι στο διάστημα $m = 10$ σε όλα τα επίπεδα σημαντικότητας, ο προτεινόμενος έλεγχος συμπεριφέρεται καλύτερα από τον Cramér-von Mises (**Σχήμα 3.8**). Όταν η κατανομή *Gamma* βρίσκεται στην εναλλακτική υπόθεση, ο έλεγχος της μέγιστης εντροπίας, συμπεριφέρεται ικανοποιητικά, αφού υπάρχει αύξηση και μεγιστοποίηση της ισχύος, κάνοντας έτσι τις κατανομές να διαφέρουν πολύ. Σε όλα τα επίπεδα, ο προτεινόμενος έλεγχος είναι καλύτερος από τον Cramér-von Mises, ενώ αρκετές φορές συμπεριφέρεται όπως και ο Anderson-Darling. Η προσαρμοστικότητα του ελέγχου εξετάστηκε και για τις περιπτώσεις όπου οι κατανομές Weibull και Inverse

Gaussian είναι εναλλακτικές. Σε όλα τα επίπεδα σημαντικότητας, ο προτεινόμενος έλεγχος, συμπεριφέρεται είτε όμοια, είτε καλύτερα από τον έλεγχο των Cramér-von Mises, βέβαια με πάντα καλύτερο τον Anderson-Darling. Ωστόσο, στο επίπεδο $\alpha = 0.10$ σχεδόν σε όλα τα διαστήματα, έχει την ίδια προσαρμογή με αυτόν.

| | | **Maximum entropy test** | | | | | **Alternative test** $(n = 100)$ | |
|---|---|---|---|---|---|---|---|---|
| **m** | **shape** | **10** | **20** | **30** | **50** | **100** | **AD** | **CVM** |
| **3** | **Log-Normal 0.62** | 0.0210 | 0.0144 | 0.0090 | 0.0131 | 0.0116 | 0.0114 | 0.0104 |
| **4** | | 0.0110 | 0.0101 | 0.0116 | 0.0111 | 0.0131 | 0.0105 | 0.0101 |
| **5** | | 0.0092 | 0.0108 | 0.0098 | 0.0097 | 0.0127 | 0.0117 | 0.0109 |
| **10** | | 0.0067 | 0.0105 | 0.0102 | 0.0145 | 0.0183 | 0.0092 | 0.0091 |
| **3** | **Pareto 0.78** | 0.2012 | 0.3212 | 0.3914 | 0.6021 | 0.9044 | 1.0000 | 1.0000 |
| **4** | | 0.2620 | 0.4206 | 0.6270 | 0.8295 | 0.9876 | 1.0000 | 1.0000 |
| **5** | | 0.3008 | 0.5439 | 0.7122 | 0.9142 | 0.9972 | 1.0000 | 1.0000 |
| **10** | | 0.4115 | 0.7655 | 0.9068 | 0.9880 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Inverse Gaussian 3.57** | 0.5301 | 0.7940 | 0.9254 | 0.9958 | 1.0000 | 1.0000 | 1.0000 |
| **4** | | 0.4694 | 0.8217 | 0.9512 | 0.9981 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.4973 | 0.8418 | 0.9632 | 0.9993 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.5122 | 0.8653 | 0.9765 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |
| **3** | **Gamma 2** | 0.4361 | 0.6802 | 0.8333 | 0.9741 | 0.9997 | 1.0000 | 0.9998 |
| **4** | | 0.3870 | 0.7042 | 0.8882 | 0.9870 | 1.0000 | 1.0000 | 1.0000 |
| **5** | | 0.4258 | 0.7421 | 0.9079 | 0.9916 | 1.0000 | 1.0000 | 1.0000 |
| **10** | | 0.4290 | 0.7744 | 0.9341 | 0.9952 | 1.0000 | 1.0000 | 1.0000 |

**Πίνακας 3.8 :** ***Log-Normal null model:* μέγεθος και ισχύς του ελέγχου της μέγιστης εντροπίας για $n = 10, 20, 30, 50, 100$ και οι εναλλακτικοί έλεγχοι για $n = 100$**




***Σχήμα 3.8 : Γραφική αναπαράσταση της κατανομής Log-Normal, κάτω από τη μηδενική υπόθεση, έναντι των κατανομών Log-normal, Gamma, Inverse Gaussian και Pareto ως εναλλακτικές υποθέσεις***

## 3.5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Κύριο μέλημά μας γενικά, ήταν να προσπαθήσουμε να ελέγξουμε το σφάλμα τύπου I , διαλέγοντας ένα επίπεδο σημαντικότητας $\alpha$ και στη συνέχεια να ελαχιστοποιήσουμε την πιθανότητα του σφάλματος τύπου II ή αντίστοιχα να μεγιστοποιήσουμε την ισχύ. Το μέγεθος του ελέγχου, που εκφράζει τη μεγαλύτερη τιμή σφάλματος τύπου I, έπρεπε να είναι κοντά στο αντίστοιχο ποσοστημόριο $(90°, 95°, 99°)$ που επιλέγαμε κάθε φορά, αφού η πιθανότητα να γίνει σφάλμα τύπου I είναι ίση με το επίπεδο σημαντικότητας. Παράλληλα, μας ενδιέφερε η μεγάλη ισχύς του ελέγχου, ούτως ώστε να υπάρχει μεγάλη διαφορά μεταξύ της πραγματικής κατανομής, κάτω από την εναλλακτική υπόθεση και της υποθετικής

κατανομής (της κατανομής δηλαδή που βρίσκεται κάτω από τη μηδενική υπόθεση). Αφού, η ισχύς αποτελεί μια ένδειξη ευαισθησίας της στατιστικής διαδικασίας, με μέτρο την πιθανότητα απόρριψης της μηδενικής υπόθεσης, όταν αυτή είναι λανθασμένη και θα πρέπει πράγματι να απορριφθεί.

Λαμβάνοντας υπόψη λοιπόν τα αποτελέσματα για το μέγεθος και την ισχύ του προτεινόμενου ελέγχου μέγιστης εντροπίας, από τη μελέτη των προσομοιώσεων Monte-Carlo που εφαρμόστηκαν σε ένα φάσμα διαφορετικών κατανομών με παχιές ουρές και για διάφορες τιμές διαστημάτων $m$, και για τις δύο περιπτώσεις βαρών με έμφαση στις ουρές που εξετάσαμε, προκύπτει ότι ο έλεγχος έχει μια ικανοποιητική προσαρμογή για μικρά και μεσαία μεγέθη δείγματος. Στις περισσότερες περιπτώσεις, παρατηρείται ότι οι εναλλακτικοί έλεγχοι έχουν καλύτερη προσαρμογή, ειδικά ο έλεγχος των Anderson-Darling, ωστόσο, δεν είναι λίγες οι φορές, που ο προτεινόμενος έλεγχος συμπεριφέρεται όμοια ή και καλύτερα από αυτούς !

Συμπερασματικά, καταλήγουμε στο ότι ο τροποποιημένος έλεγχος με τις κατάλληλα επιλεγμένες κρίσιμες τιμές, που έχει προταθεί για μικρά και μεσαία μεγέθη δείγματος, αποτελεί έναν ικανοποιητικό έλεγχο καλής προσαρμογής, που είναι ικανός να χρησιμοποιηθεί στην ανάλυση αρκετών μοντέλων διάρκειας ζωής, περιγράφοντας πλήθος κριτηρίων που σχετίζονται με αυτή. Για παράδειγμα, όταν ο χρόνος ζωής $T$ μιας υπό μελέτης μονάδας έχει την ιδιότητα *increasing failure rate (IFR),* όταν δηλαδή ο ρυθμός αποτυχίας (*failure rate*) είναι αύξουσα, άρα η μονάδα "γερνά" (φθείρεται) με την πάροδο του χρόνου. Ή όταν ο χρόνος ζωής μιας υπό μελέτης μονάδας έχει την ιδιότητα *decreasing failure rate (DFR)*, όταν δηλαδή ο ρυθμός αποτυχίας είναι φθίνουσα, επομένως η μονάδα κατά κάποιο τρόπο λειτουργεί ομαλότερα όσο περνά ο χρόνος.

# ΒΙΒΛΙΟΓΡΑΦΙΑ ΞΕΝΗ


1. *Anderson T.W. and Darling D.A., (1954).* ***"A Test of Goodness of fit"****. Journal of the American Statistical Association, 49 (268), 765-769.*
2. *Conover W.J., (1999).* ***"Practical Nonparametric Statistics"****. 3rd edition. Wiley, New York.*
3. *Couallier V., Gerville-Réache L., Huber-Carol C., Limnios N. and Mesbah M. (2013)****."Statistical Models and Methods for Reliability and Survival Analysis"****, (chapter 3), Wiley, USA.*
4. *Cressie N. and Read T.R.C., (1984).* ***"Multinomial Goodness-of-fit Tests".*** *Journal of Royal Statistical Society, 46 (3), 440-464.*
5. *Cressie N., Pardo L. and Pardo M.C., (2013).* ***"Size and Power Considerations for Testing Linear Log Linear Models Using ϕ-Divergence Test Statistics".*** *Statistica Sinica, 13 (2), 555-570.*
6. *D'Agostino R.B. and Stephens M.A., (1986).* ***"Goodness-of-fit Techniques"****. Marcel Dekker, New York.*
7. *Durbin J. and Knott M., (1972).* ***"Components of Cramér-von Mises Statistics. I"****. Journal of the Royal Statistical Society, 34 (2), 290-307.*
8. *Forte B. and Hughes W. (1988).* ***"The maximum entropy principle: a tool to define new entropies"****. Reports on Mathematical Physics, 26 (2), 227-235.*
9. *Foss S., Korshunov D. and Zachary S. (2011).* ***"An Introduction to Heavy-Tailed and Subexponential Distributions"****. Springer, New York.*
10. *Gibbons J.D., (1976).* **"Nonparametric Methods for Quantitative Analysis".** *Holt, Rinehart and Winston, New York.*
11. *Gibbons J.D. and Chakraborti S.,(2003).* **"Nonparametric Statistical Inference".** *4th edition. Marcel Dekker, New York.*
12. *Gibson J.D. and Melsa J.L. (1975).* ***"Introduction to Nonparametric Tests Detection with Applications".*** *Academic Press, USA.*
13. *Gravetter F.J., Wallnau L.B., (2012).* **"Statistics for the Behavioral Sciences".** *9th edition. Wadsworth, USA.*
14. *Greenwood P.E. and Nikulin M.S., (1996).* ***"A Guide to Chi-squared Testing".*** *Wiley, USA.*
15. *Hollander M., Wolfe D.A. and Chicken E., (2014).* ***"Nonparametric Statistical Methods"*** *3rd edition. Wiley, New Jersey.*
16. *Lehmann E.L. and Romano J.P., (2005).* **"Testing Statistical Hypothesis".** *3rd edition. Springer, USA.*
17. *Lee S. , Vonta I. and Karagrigoriou A., (2011).* ***"A maximum entropy test of fit"****. Computational Statistics and Data Analysis, 55 (9), 2635-2643.*
18. *Lewis P.A.W., (1961).* ***"Distribution of the Anderson-Darling Statistic".*** *The Annals of Mathematical Statistics, 32 (4), 1118-1124.*

**19.** *Lilliefors H.W., (1967).* ***"On the Kolmogorov-Smirnov Test for Normality with Mean and Variance unknown"****. Journal of the American Statistical Association, 62 (318), 399-402.*

**20.** *Lilliefors H.W., (1969).* ***"On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean unknown"****. Journal of the American Statistical Association, 64 (325), 387-389.*

**21.** *Marsaglia G. and Marsaglia J.C.W. (2004)* ***"Evaluating the Anderson-Darling Distribution"****. Journal of Statistical Software, 9 (2), 1-5.*

**22.** *Massey F.J. (1951).* ***"The Kolmogorov-Smirnov Test for Goodness-of-fit".*** *Journal of the American Statistical Association, 46 (253), 68-78.*

**23.** *Maydeu-Olivares A. and Forero C.G., (2010).* **"Goodness-of-fit Testing".** *International encyclopedia of education,7 (1), 190-196.*

**24.** *Menéndez M., Morales D., Pardo L. and Vajda I., (2006).* ***"Approximations to powers of ϕ-Disparity Goodness-of-fit Tests".*** *Communications in Statistics (Theory and Methods), 30 (1), 105-134.*

**25.** *Pardo L., (2006).* ***"Statistical Inference Based on Divergence Measures"****. Chapman & Hall, New York.*

**26.** *Pardo L. and Martin N., (2010).* ***"Minimum Phi-Divergence Estimators and Phi-Divergence Test Statistics in Contingency Tables with Symmetry Structure: An Overview".*** *Symmetry, 2 (2), 1108-1120.*

**27.** *Pardo L. and Zografos K., (2000).* ***"Goodness of Fit Tests with Misclassified Data Based on ϕ-Divergences".*** *Biometrical Journal, 42 (2), 223-237.*

**28.** *Pardo J.A., Pardo L. and Zografos K., (2002). "****Minimum ϕ-Divergence estimators with constraints in multinomial populations".*** *Journal of Statistical Planning and Inference, 104 (1) 221-237.*

**29.** *Privitera G.J., (2017).* **"Statistics for the Behavioral Sciences".** *3rd edition, SAGE, USA.*

**30.** *Razali N.M. and Wah Y.B. (2011).* ***"Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests"****. Journal of Statistical Modeling and Analytics, 2 (I), 21-33.*

**31.** *Rayner J.C.W. and Best D.J., (1989).* ***"Smooth Tests of Goodness of Fit"****. Oxford University Press, New York.*

**32.** *Read T.R.C. and Cressie N., (1988).* ***"Goodness-of-fit Statistics for Discrete Multivariate Data".*** *Springer, New York.*

**33.** *Ross S.M.,(2009).* ***"Introduction to probability and statistics for engineers and scientists"****. 4th edition. Elsevier Academic Press, Canada.*

**34.** *Sen B., (2018)* ***"A Gentle Introduction to Empirical Process Theory and Applications"****, Columbia University (notes).*

**35.** *Singpurwalla D.,* (*2013).* **"A handbook of Statistics: An Overview of Statistical methods"**, *1st edition (notes).*

**36.** *Stephens M.A., (1965).* ***"The goodness-of-fit statistic $V_N$: distribution and significance points"****. Biometrika, 52 (3/4), 309-321.*

**37.** *Thas O. (2010).* ***"Comparing Distributions".*** *Springer-Verlag, New York.*

**38.** *Yap B.W. and Sim C.H., (2011).* ***"Comparisons of various types of normality tests"****. Journal of Statistical Computation and Simulation, 81 (12), 2141-2155.*

# ΒΙΒΛΙΟΓΡΑΦΙΑ ΕΛΛΗΝΙΚΗ

**39.** *Δαμιανού Χ., Κούτρας Μ., (2013).* **"Εισαγωγή στη Στατιστική: ΜΕΡΟΣ Ι"**. *Συμμετρία, Αθήνα.*

**40.** *Δαμιανού Χ. και Κούτρας Μ., (1998).* **"Εισαγωγή στη Στατιστική: ΜΕΡΟΣ ΙΙ"**. *Συμμετρία, Αθήνα.*

**41.** *Μπατσίδης Α., (2014).* **"Εισαγωγή στη μη Παραμετρική Στατιστική"**. *Πανεπιστήμιο Ιωαννίνων (σημειώσεις).*