ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΒΙΟΕΠΙΤΗΡΗΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΕ ΔΕΔΟΜΕΝΑ ΕΞΑΠΛΩΣΗΣ ΕΠΙΔΗΜΙΩΝ

Λουίζα Ι. Φακιολά

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική.

Πειραιάς
Σεπτέμβριος 2022

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Αναπληρωτής Καθηγητής Μπερσίμης Σωτήρης
- Επίκουρος Καθηγητής Ευαγγελάρας Χαράλαμπος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS
School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Statistical biosurveillance methods and their application in epidemic outbreak data

By

Louisa I. Fakiola

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2022

Στους γονείς μου,

Ιωάννη και Καλλιόπη

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κύριο Κούτρα Μάρκο, για την καθοριστική συμβολή του στην εκπόνηση της παρούσας διπλωματικής εργασίας, καθώς και την κατανόηση του απέναντι στα εμπόδια και τις δυσκολίες που προέκυψαν κατά την διάρκειά της.

Επιπρόσθετα, θα ήθελα να απευθύνω θερμές ευχαριστίες στους καθηγητές Μπερσίμη Σωτήρη και Ευαγγελάρα Χαράλαμπο, για τον χρόνο που διέθεσαν για την διόρθωση της διπλωματικής μου εργασίας.

Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου για την στήριξη, υλική και ψυχολογική, που μου παρείχαν κατά την διάρκεια των διπλωματικών μου σπουδών.

Περίληψη

Τον Δεκέμβριο του 2019, καταγράφηκε το πρώτο επιβεβαιωμένο κρούσμα κορονοϊού Covid–19 στην κωμόπολη *Wuhan* της Κίνας. Η ραγδαία αύξηση των κρουσμάτων σε παγκόσμια κλίμακα, ανέδειξε την ανάγκη για γρήγορη και έγκυρη ανίχνευση περιπτώσεων μεταδοτικών ασθενειών, προκειμένου να ληφθούν τα απαραίτητα μέτρα αναχαίτησης μιας επικείμενης υγειονομικής κρίσης.

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάλυση βασικών μεθόδων και μέσων Βιοεπιτήρησης. Πιο συγκεκριμένα στο Κεφάλαιο 2 εστιάζουμε στις Στατιστικές Συναρτήσεις Σάρωσης, στο Κεφάλαιο 3 στα Διαγράμματα Ελέγχου ενώ στο Κεφάλαιο 4 παρουσιάζονται οι Χρονοσειρές με χρήση μοντέλων *ARIMA* καθώς και *SIR*. Παράλληλα, παρουσιάζονται για κάθε μία από τις παραπάνω μεθόδους, παραδείγματα εφαρμογής πάνω σε επιδημιολογικά δεδομένα. Τέλος, στο τελευταίο κεφάλαιο, γίνεται ενδεικτική εφαρμογή των μεθόδων *ARIMA*, *p* – Διαγράμματος Ελέγχου και Χωρικής Σάρωσης σε δεδομένα που αφορούν την επιδημία του κορονοϊού στην Ελλάδα.

Abstract

In December 2019, the first confirmed case of coronavirus Covid – 19 was recorded in the Chinese city of Wuhan. The rapid increase in cases on a global scale has highlighted the need for fast and valid detection of infectious disease cases, in order to take the necessary measures to prevent an impending health crisis.

The purpose of this thesis is the analysis of basic methods and means of Biosurveillance. More specifically, in Chapter 2 we focus on Scan Statistics, in Chapter 3 on Statistical Process Control Charts, while in Chapter 4 we present Timeseries Methodology with use of ARIMA and SIR models. At the same time, for each of the above methods, we present examples on epidemiological and clinical data. Finally, in Chapter 5, an indicative application of methods of ARIMA, p – Control Chart and Space Scan Statistics is made in data concerning the coronavirus epidemic in Greece.

Περιεχόμενα

Ευχαριστίες

Περίληψη

Abstract

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1ο : Εισαγωγή

1.1 Επιδημιολογία και εξάπλωση επιδημιών στον σύγχρονο κόσμο 1

1.2 Στατιστική και Δημόσια Υγεία 2

1.3 Βιοεπιτήρηση 4

ΚΕΦΑΛΑΙΟ 2ο : Στατιστικές Συναρτήσεις Σάρωσης

2.1 Εισαγωγή 7

2.2 Μοντέλα *Poisson* και *Bernoulli* 9

2.3 Χωρική Μέθοδος Σάρωσης 11

2.4 Μέθοδοι Σάρωσης στον Χώρο και τον Χρόνο 14

2.5 Το λογισμικό *SaTScan* 17

2.6 Η μελέτη των *Desjardins et. al.* 18

ΚΕΦΑΛΑΙΟ 3ο : Διαγράμματα Ελέγχου

3.1 Εισαγωγή 25

3.2 Βασικές Έννοιες και Δομή Διαγραμμάτων Ελέγχου 27

3.3 Επιλογή κατάλληλου Διαγράμματος Ελέγχου στην Βιοεπιτήρηση 33

3.4 Διαγράμματα Ελέγχου Τύπου *Shewhart* για ιδιότητες 34

3.4.1 *p* Διάγραμμα Ελέγχου 34

3.4.2 *np* Διάγραμμα Ελέγχου 40

3.4.3 *c* Διάγραμμα Ελέγχου 44

3.4.4 *u* Διάγραμμα Ελέγχου 47

3.5 Διαγράμματα Ελέγχου Τύπου *Shewhart* για μεταβλητές 50

3.5.1 *X – Bar* Διάγραμμα Ελέγχου 51

3.5.2 Διάγραμμα Ελέγχου για την Διασπορά 53
3.5.3 Διάγραμμα Ελέγχου για Μεμονωμένες Παρατηρήσεις 57
3.6 Διαγράμματα Ελέγχου Με Μνήμη 61
3.6.1 Διάγραμμα Ελέγχου *CUSUM* 62
3.6.2 Διάγραμμα Ελέγχου *EWMA* 67
3.7 Διαγράμματα Ελέγχου για Πολυμεταβλητά Δεδομένα 70
3.7.1 Η περίπτωση των Υποομάδων 70
3.7.2 Η περίπτωση των μεμονωμένων παρατηρήσεων $n = 1$ 72
3.8 Διαγράμματα Ελέγχου Προσαρμοσμένα στον Κίνδυνο 75

ΚΕΦΑΛΑΙΟ 4ο : Χρονοσειρές

4.1 Εισαγωγή 79
4.2 Χρονοσειρές 80
4.2.1 Ορισμός Χρονοσειράς 80
4.2.2 Συνθετικά Στοιχεία Χρονοσειρών 81
4.2.1 Στασιμότητα 83
4.2.2 Λευκός Θόρυβος 86
4.3 Συντελεστές Αυτοσυσχέτισης και Μερικής Αυτοσυσχέτισης 86
4.4 Ανάλυση Μοντέλου Αυτοπαλινδρόμησης *AR*(p) 87
4.5 Ανάλυση Μοντέλου Κινητού Μέσου Όρου *MA*(q) 89
4.6 Ανάλυση Μοντέλου Αυτοπαλινδρόμησης Κινητού Μέσου Όρου *ARMA*(p,q) 91
4.7 Μεθοδολογία *Box – Jenkins* 92
4.8 Μοντέλα *SIR* 96
4.8.1 Παραλλαγές του Μοντέλου *SIR* 103

ΚΕΦΑΛΑΙΟ 5ο : Εφαρμογή των μεθόδων σε δεδομένα κορονοϊού

5.1 Εφαρμογή Μοντέλων *ARIMA* σε δεδομένα κορονοϊού 105
5.2 Εφαρμογή p – Διαγράμματος Ελέγχου σε δεδομένα κορονοϊού 115
5.3 Εφαρμογή Μεθόδου Στατιστικής Σάρωσης σε δεδομένα κορονοϊού 118

Βιβλιογραφία 122

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Επιδημιολογία και εξάπλωση επιδημιών στον σύγχρονο κόσμο

Τον Δεκέμβριο του 2019, στην κωμόπολη *Wuhan* της Κίνας, εμφανίστηκαν για πρώτη φορά κρούσματα της νόσου *Covid* – 19. Λόγω της εύκολης μεταδοτικότητας του ιού μέσω των αεραγωγών οδών του ανθρώπινου οργανισμού, επιβεβαιωμένα κρούσματα του ιού άρχισαν να κάνουν την εμφάνισή τους σε ολόκληρο τον πλανήτη, ενώ στις 11 Μαρτίου του 2020, ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) κηρύττει επίσημα την *Covid* - 19 ως πανδημία.

Η μετάδοση της νόσου χωρίζεται σε τρείς κατηγορίες. Την μετάδοση μέσα στη οικογένεια, όπου ο ιός περιορίζεται μέσα στα κλειστά πλαίσια των μελών της οικογένειας. Την μετάδοση στη κοινότητα, όπου σε αυτό το στάδιο δεν είναι εφικτό να βρεθεί ο ασθενής 0, από τον οποίο ξεκίνησε η νόσος, κι έτσι αυτή διαδίδεται μέσα στη κοινότητα. Και τέλος έχουμε την μετάδοση του ιού σε μεγάλα κομμάτια της επικράτειας, όπου πλέον ο ρυθμός μετάδοσης είναι ταχύτατος, λόγω της μετακίνησης του πληθυσμού.

Η ταχεία διάδοση της νόσου σε ολόκληρο τον κόσμο, οδήγησε τις κυβερνήσεις και τους αρμόδιους οργανισμούς υγείας στην λήψη μέτρων για την αναχαίτηση της μετάδοσης, τα οποία περιλάμβαναν την κοινωνική αποστασιοποίηση αλλά και το *Lockdown*. Ωστόσο, η λήψη μέτρων για την καταπολέμηση της νόσου καθώς και η αναζήτηση κατάλληλων εμβολίων και φαρμάκων, δεν ήταν τα μόνα όπλα στην φαρέτρα των ειδικών.

Ως επιδημιολογία ορίζουμε την επιστήμη η οποία μελετά την κατανομή και την εξέλιξη διάφορων νοσημάτων ή χαρακτηριστικών στον ανθρώπινο πληθυσμό και των παραγόντων που τον διαμορφώνουν ή επηρεάζουν (Ντζούφρας Ι., 2006) . Στόχος είναι ο έλεγχος της χρονικής εξέλιξης της νόσου, η κατανόηση των επιβαρυντικών παραγόντων, καθώς και η λήψη κατάλληλων υγειονομικών μέτρων για την πρόληψη ή και την αντιμετώπιση της επικείμενης νόσου.

Η επιδημιολογία βρίσκει τις ρίζες από τα πρώτα χρόνια της ανθρωπότητας. Ο Ιπποκράτης (460 – 357 π.Χ), μέσω του έργου του «Περί αέρων, υδάτων και τόπων», προσπάθησε να εντοπίσει τις αιτίες που προκαλούν τις διάφορες ασθένειες, μέσω των εμπειρικών παρατηρήσεων του ίδιου

σχετικά με τις σχέσεις των ασθενειών και παραγόντων όπως το νερό και ο αέρας. Το έργο του θεωρείται μία από τις πρώτες επιδημιολογικές μελέτες, ενώ ο ίδιος χαρακτηρίζεται από πολλούς ως ο πατέρας της επιδημιολογίας. Ο *John Graunt* (1620 – 1674), ένας από τους μεγαλύτερους επιστήμονες πάνω στην δημογραφία, θεωρείται ένας από τους πρώτους επιδημιολόγους, καθώς μελέτησε την θνησιμότητα βρεφών μέσω διάφορων ποσοτικών μεθόδων. Μετέπειτα ο *William Farr* (1807 – 1883) έθεσε τα θεμέλια της ιατρικής στατιστικής, χρησιμοποιώντας για πρώτη φορά πλήθος δημογραφικών δεδομένων για την μελέτη της δημόσιας υγείας αλλά και για τον έλεγχο επιδημιών που μάστιζαν την εποχή.

Πέρα από την πρόσφατη πανδημία του *Covid* – 19, πληθώρα επιδημιών έχουν σημαδέψει τον 20ο αιώνα. Μετά την λήξη του Α Παγκοσμίου Πολέμου και κατά την διάρκεια του Ρωσικού Εμφυλίου Πολέμου, έκανε την εμφάνιση του ο τύφος, ο οποίος έπληξε συνολικά πάνω από 30 εκατομμύρια ανθρώπους, οδηγώντας στον θάνατο σχεδόν το ένα δέκατο του πληθυσμού που νόσησε. Το 1968, κάνει την εμφάνισή της στην Κίνα, η γρίπη *H*3*N*2. Αν και στην αρχή δεν της δόθηκε η πρέπουσα σημασία από τους ειδικούς, εξελίχθηκε σε μία από της φονικότερες επιδημίες του σύγχρονου κόσμου, κάνοντας πλέον ορατή την ανάγκη επιδημιολογικής επιτήρησης. Τέλος από τις πιο πρόσφατες πανδημίες με τις οποίες ήρθε αντιμέτωπος ο σύγχρονος κόσμος είναι η πανδημία του *HIV/AIDS*. Από την αρχή της επιδημίας, αναφέρεται ότι σχεδόν 80 εκατομμύρια άνθρωποι έχουν μολυνθεί από τον ίο, ενώ έχουν πεθάνει 40 εκατομμύρια.

1.2 Στατιστική και Δημόσια Υγεία

Όπως αναφέρθηκε παραπάνω, οι στόχοι της επιδημιολογικής επιτήρησης των λοιμωδών νοσημάτων όπως αυτοί αναφέρονται στην ιστοσελίδα του Εθνικού Οργανισμού Δημόσιας Υγείας (ΕΟΔΥ) είναι η εκτίμηση της διασποράς ενός νοσήματος, ο προσδιορισμός των παραγόντων κινδύνου και των επιβαρυντικών παραγόντων που συμβάλουν στην μετάδοση της νόσου, η εκτίμηση των επιπτώσεων των νοσημάτων στον πληθυσμός, ο έγκαιρος εντοπισμός επιδημιών και ενδεχόμενων κινδύνων για την δημόσια υγεία καθώς και η αξιολόγηση των παρεμβάσεων και των δράσεων που πραγματοποιούνται για την διατήρηση της ποιότητας της δημόσιας υγείας.

Απαραίτητο εργαλείο ωστόσο αποτελεί η έγκαιρη και έγκυρη καταγραφή των επιβεβαιωμένων κρουσμάτων αλλά και των συμβάντων που τυχαίνουν υγειονομικού ενδιαφέροντος. Έτσι, η επιδημιολογική επιτήρηση μπορεί να πραγματοποιηθεί μέσω τριών συστημάτων:

- Συστήματα Υποχρεωτικής Δήλωσης, μέσω δελτίων τα οποία αποστέλλονται από τις Δομές Υγείας της χώρας και χρησιμοποιούνται για την επιδημιολογική επιτήρηση ΣΜΝ, Ηπατιτίδων καθώς και *HIV/AIDS*.
- Συστήματα Παρατηρητών Νοσηρότητας, μέσω των οποίων επιτηρούνται νοσήματα τα οποία παρουσιάζουν αυξημένη συχνότητα στον πληθυσμό, ωστόσο δεν κρίνεται η άμεση παρέμβαση των Υγειονομικών Αρχών, καθώς οι ασθενείς εμφανίζουν ήπια κλινική εικόνα. Μερικά από τα νοσήματα που καταγράφονται είναι η λοίμωξη του αναπνευστικού συστήματος με πυρετό, η γαστρεντερίτιδα, η ιλαρά και η ανεμοβλογιά. Ωστόσο, αν και όπως αναφέρθηκε, τα νοσήματα αυτά δεν απαιτούν άμεσες ενέργειες από τις υπεύθυνες υγειονομικές αρχές, η παρακολούθηση της εξέλιξής τους είναι ιδιαίτερα σημαντική, καθώς με τον τρόπο αυτό γίνεται εφικτή η έγκαιρη παρέμβαση σε περιόδους έξαρσης ή και επιτήρησης τυχόν μεταλλάξεων αυτών.
- Ειδικά Συστήματα Επιτήρησης / Δήλωση Κρουσμάτων, τα οποία χωρίζονται στην επιτήρηση μέσω εξειδικευμένων εργαστηρίων – κέντρων αναφοράς νόσων όπως σαλμονέλα, εντεροιών, μηνιγγίτιδας και στην επιτήρηση μέσω ειδικών κλινικοεργαστηριακών δικτύων, όπου γίνεται επιτήρηση γρίπης, γεγονός που καθιστά το σύστημα αυτό υψίστης σημασίας καθώς η επιτήρηση της γρίπης μπορεί να προλάβει πιθανές επιδημίες που ενδέχεται να οδηγήσουν σε αυξημένη νοσηρότητα και θνητότητα, ιδιαίτερα σε ομάδες υψηλού κινδύνου.

Τόσο στην Ευρωπαϊκή Ένωση, όσο και στην Ελλάδα έχουν δημιουργηθεί οργανισμοί που ασχολούνται με την καταγραφή λοιμωδών και όχι μόνο νοσημάτων, με στόχο την ανάλυση των δεδομένων που προέρχονται από τα καταγεγραμμένα κρούσματα, την εξαγωγή ασφαλών συμπερασμάτων από αυτά αλλά και την έγκαιρη ειδοποίηση των αρμόδιων φορέων Δημόσιας Υγείας και κυβερνήσεων για ενδεχόμενο ξέσπασμα επιδημίας ή άλλου υγειονομικού κινδύνου. Χαρακτηριστικά, το 2003, μετά την έξαρση του *SARS*, αλλά και την ταχεία εξάπλωσή του στην Ευρώπη, έγινε εμφανής η ανάγκη για την δημιουργία ενός ευρωπαϊκού οργανισμού, ο οποίος θα επιτηρεί την δημόσια υγεία μέσα στην Ένωση. Έτσι, το 2005 ιδρύθηκε το Ευρωπαϊκό Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (*ECDC*), το οποίο αναλύει και ερμηνεύει δεδομένα από τις χώρες της Ευρωπαϊκής Ένωσης, καταγράφοντας περιστατικά και επιβλέποντας 52 μεταδοτικές νόσους και παθήσεις, μέσω του Ευρωπαϊκού Συστήματος Επιτήρησης (*TESSy*), διασφαλίζοντας με τον τρόπο αυτό την έγκαιρη ανίχνευση νεοεμφανιζόμενων απειλών στην ΕΕ. Στην Ελλάδα, το 1992 ιδρύθηκε το Κέντρο Ελέγχου Ειδικών Λοιμώξεων (ΚΕΕΛ), το οποίο το 2005 μετονομάστηκε σε Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (ΚΕ.ΕΛ.Π.ΝΟ), ενώ στην συνέχεια έλαβε την σημερινή

του ονομασία ως Εθνικός Οργανισμός Δημόσιας Υγείας (Ε.Ο.Δ.Υ). Στόχος του οργανισμού είναι η προστασία και η διασφάλιση της ποιότητας της Δημόσιας Υγείας.

1.3 Βιοεπιτήρηση

Ως Βιοεπιτήρηση ορίζεται η συστηματική διαδικασία, κατά την οποία παρακολουθούνται και καταγράφονται τυχόν παράγοντες κινδύνου για την δημόσια υγεία, όπως βακτήρια, ιοί κλπ., και βάσει αυτών μπορούν να ανιχνευθούν και να προσδιοριστούν επικείμενα ξεσπάσματα επιδημιών ή άλλων υγειονομικών κρίσεων.

Η ανθρωπότητα από τις απαρχές της έχει έρθει αντιμέτωπη με πληθώρα επιδημιών, οι οποίες στην πλειοψηφία τους έπιασαν απροετοίμαστες της υγειονομικές αρχές. Οι υγειονομικές κρίσεις που παρουσιάστηκαν μέσα στην πορεία των χρόνων, όπως η έξαρση του *SARS* το 2004, αλλά και η επιδημία της γρίπης *H*1N*1*, έκαναν φανερό στις αρχές υγείας αλλά και τις κυβερνήσεις ανά τον κόσμο, ότι πρέπει να προετοιμαστούν με τα κατάλληλα εφόδια, προκειμένου να ανιχνεύσουν και να προλάβουν ένα επικείμενο ξέσπασμα μιας ακόμη επιδημίας, μειώνοντας έτσι τον αντίκτυπο που θα έχει αυτή στην κοινωνία. Ωστόσο, ιδιαίτερα μετά τις τρομοκρατικές επιθέσεις της 11ης Σεπτεμβρίου του 2001, καθώς και τις επιθέσεις με άνθρακα που ακολούθησαν, έγινε φανερή η ανάγκη δημιουργίας ενός συστήματος καταγραφής περιστατικών υγειονομικού και όχι μόνο ενδιαφέροντος, καθιστώντας έτσι τις μεθόδους Βιοεπιτήρησης ως τα βασικότερα όπλα αντιμετώπισης αλλά και πρόληψης των παραπάνω περιστατικών.

Μία διαδικασία Βιοεπιτήρησης μπορεί να χωριστεί σε 3 στάδια:

- Την ανίχνευση και καταγραφή επιβεβαιωμένων κρουσμάτων αλλά και περιστατικών που ενδεχομένως να αποτελούν κίνδυνο για την δημόσια υγεία.
- Την επεξεργασία και ανάλυση των διαθέσιμων δεδομένων για την εξαγωγή ασφαλών συμπερασμάτων σχετικά με την εκδήλωση κάποιας πιθανής υγειονομικής κρίσης.
- Την έγκαιρη ειδοποίηση των αρμόδιων αρχών υγείας, για την λήψη κατάλληλων μέτρων για την αντιμετώπιση της κρίσης, αλλά και τη χρήση των πληροφοριών που έχουν συλλεχθεί για καλύτερη μελλοντική διαχείριση της υγείας.

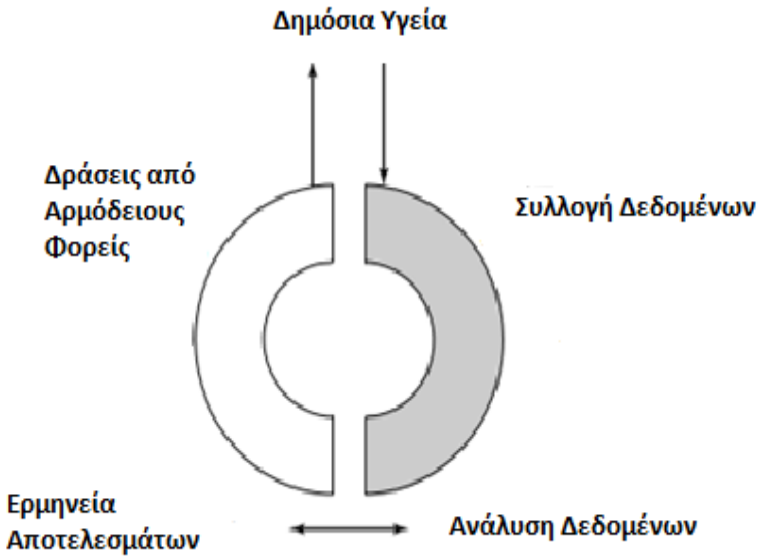
Τα δεδομένα που λαμβάνονται υπόψιν στην Βιοεπιτήρηση είναι συνήθως μετρήσεις οι οποίες αφορούν καταγεγραμμένα κρούσματα μιας νόσου, εισαγωγές στα νοσοκομεία, επισκέψεις ασθενών σε γιατρούς, πωλήσεις συγκεκριμένων φαρμακευτικών σκευασμάτων. Η αύξηση στις μετρήσεις

των παραπάνω αριθμών μπορεί να σηματοδοτεί την ύπαρξη κάποιας νόσου μέσα στην κοινότητα ή κάποιου άλλου υγειονομικού περιστατικού. Έτσι, τα παραπάνω δεδομένα θα πρέπει να αναλυθούν και να αξιολογηθούν ώστε να εξαχθούν ασφαλή συμπεράσματα σχετικά με την αυξητική πορεία των μετρήσεων, και αν αυτό είναι απαραίτητο να ληφθούν κατάλληλα μέτρα για την διασφάλιση της δημόσιας υγείας.

Υπάρχουν διάφορα είδη επιτήρησης της Δημόσιας Υγείας. Από τα κυριότερα είναι:

- Η παθητική επιτήρηση (*Passive Surveillance*), όπου πληροφορίες υγείας συλλέγονται από τις διάφορες δομές Υγείας όπως νοσοκομεία, κέντρα υγείας κλπ. Αν και αποτελεί το πιο διαδεδομένο μέσο επιτήρησης, καθώς είναι ιδιαίτερα απλό να αντληθούν οι απαιτούμενες πληροφορίες ενώ παράλληλα δεν απαιτούνται υψηλά χρηματικά ποσά για την διενέργειά της, η ποιότητα των δεδομένων που συλλέγονται εξαρτάται από τους ανθρώπους που στελεχώνουν τους διάφορους οργανισμούς υγείας.
- Η ενεργητική επιτήρηση (*Active Surveillance*). Οι αρμόδιοι φορείς και υπεύθυνοι Δημόσιας Υγείας, επικοινωνούν κατευθείαν με τους πολίτες προκειμένου να αντλήσουν τις απαιτούμενες πληροφορίες. Αν και οι πληροφορίες που συλλέγονται είναι ιδιαίτερα ακριβείς, το κόστος διενέργειας τέτοιων μορφών επιτήρησης είναι αρκετά υψηλό. Συνήθως συνδυάζεται με την παθητική επιτήρηση για τον σχηματισμό μιας πιο ολοκληρωμένης εικόνας σχετικά με την Δημόσια Υγεία.
- Η συνδρομική Επιτήρηση (*Syndromic Surveillance*). Αφορά την λήψη στοιχείων σχετικά με συγκεκριμένα εκδηλωμένα κλινικά σύνδρομα και όχι την δήλωση επιβεβαιωμένων κρουσμάτων. Για παράδειγμα, συλλέγει πληροφορίες για την εκδήλωση πυρετού και όχι για την διάγνωση λοίμωξης του αναπνευστικού. Έτσι, με την χρήση συνδρομικής επιτήρησης, φαινομενικά ασύνδετες περιπτώσεις, μπορούν να συνδεθούν με την χρήση κατάλληλων βιοστατιστικών μεθόδων. Ωστόσο, καθώς γίνεται λήψη μεγάλου όγκου δεδομένων, η ανάλυσή τους μπορεί να είναι ιδιαίτερα περίπλοκη και κοστοβόρα. Η χρήση συνδρομικών μεθόδων επιτήρησης πρέπει να γίνεται με προσοχή και συμπληρωματικά με τις δυο παραπάνω μεθόδους, καθώς επειδή βασίζεται στον εντοπισμό συνδρόμων και όχι επιβεβαιωμένων περιστατικών νόσων, είναι λιγότερο εξειδικευμένη και μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα.

Με βάση τα παραπάνω, ο κύκλος λειτουργίας μιας επιτυχημένης διεργασίας βιοεπιτήρησης μπορεί να συνοψιστεί στο παρακάτω διάγραμμα:



Διάγραμμα 1.3.1. Λειτουργία Επιτήρησης

.

Στην βιβλιογραφία έχουν προταθεί πληθώρα μεθόδων για την αποτελεσματική επιτήρηση επιδημιών. Μερικές από αυτές είναι μέσω Μαρκοβιανών Αλυσίδων, Μεθόδων Πολυμεταβλητής Ανάλυσης αλλά και μέσω Μπευζιανής Στατιστικής. Στην παρούσα διπλωματική εργασία, θα αναλυθούν οι Μέθοδοι Βιοεπιτήρησης:

- Μέσω Στατιστικών Συναρτήσεων Σάρωσης, όπου διεξάγωντας σάρωση σε μια υπό μελέτη περιοχής, μπορούμε να εντοπίσουμε σήμα επικίνδυνης αύξησης των κρουσμάτων ή επικίνδυνης μείωσης διάφορων παραγόντων που αποτελούν κίνδυνο για την Δημόσια Υγεία.
- Μέσω Διαγραμμάτων Ελέγχου, όπου η χρήση κατάλληλων διαγραμμάτων μπορούν να προειδοποιήσουν το χρήστη για το πότε η υπό μελέτη διεργασία βρίσκεται εκτός ελέγχου, και συνεπώς χρήζει περαιτέρω επιδημιολογικής επίβλεψης.
- Με χρήση χρονοσειρών, η οποία στοχεύει στην έγκαιρη ανίχνευση επιδημιών καθώς και την πρόβλεψη της πορείας τους.

ΚΕΦΑΛΑΙΟ 2

Στατιστικές Συναρτήσεις Σάρωσης

2.1 Εισαγωγή

Οι φορείς υγείας σε ολόκληρο τον κόσμο έχουν αναπτύξει συστήματα καταγραφής για τις διάφορες ασθένειες που εμφανίζονται, τόσο όσον αφορά την κλινική εικόνα και τα βιομετρικά χαρακτηριστικά των ασθενών, όσο και πληροφορίες σχετικά με τον χρόνο και την περιοχή εμφάνισης των κρουσμάτων. Λαμβάνοντας τα παραπάνω δεδομένα υπόψιν, ως προς τον χρόνο και την χωρική περιοχή εμφάνισης κρουσμάτων, οι επιδημιολόγοι εστιάζουν το ενδιαφέρον τους στην εύρεση και την ανάδειξη των συστάδων εκείνων που ενδεχομένως να αποτελούν κίνδυνο για τη δημόσια υγεία και πιθανόν να χρήζουν περαιτέρω επιδημιολογικής επιτήρησης. Για τον σκοπό αυτό έχουν προταθεί οι Στατιστικές Μέθοδοι Σάρωσης (*Scan Statistics*).

Στο σημείο αυτό, είναι αρκετά σημαντικό να τονιστεί ότι οι Στατιστικές Μέθοδοι Σάρωσης δεν θα πρέπει να συγχέονται με τις παραδοσιακές Μεθόδους Ανάλυσης σε Συστάδες. Η πρώτη εφαρμόζεται προκειμένου να ανιχνευθούν αν τυχόν υπάρχουν πιθανές συστάδες, για τις υπό μελέτη χρονικές περιόδους και περιοχές, ενώ η δεύτερη μέθοδος εξετάζει κατά πόσο είναι όμοιες κάποιες παρατηρήσεις ως προς ένα σύνολο μεταβλητών, με σκοπό να δημιουργηθεί συστάδα από όμοιες παρατηρήσεις. Επιπρόσθετα, οι κλασικές Μέθοδοι Ανάλυσης Συστάδων, μπορούν να λειτουργήσουν μόνο αναδρομικά και μόνο για ανάλυση μιας συγκεκριμένης περιοχής και χρονικής στιγμής, ενώ οι Συναρτήσεις Σάρωσης μπορούν να χρησιμοποιηθούν τόσο για αναδρομικές, όσο και για προοδευτικές μελέτες, οι οποίες μπορεί και να αποτελούνται από περιοδικά επαναλαμβανόμενες αναλύσεις (*Rogerson, 1997*).

Ως συστάδες ορίζουμε τις ομάδες εκείνες που αποτελούνται από παρατηρήσεις οι οποίες μοιάζουν μεταξύ τους ως προς κάποια χαρακτηριστικά. Για σκοπούς επιδημιολογικής επιτήρησης, εστιάζουμε το ενδιαφέρον μας στην ανάδειξη χωρικών ή/και χρονικών συστάδων. Μια χωρική συστάδα εντοπίζεται όταν σε αυτή συγκεντρώνεται ένα σύνολο περιοχών, στις οποίες υπάρχει ένδειξη ότι αυξήθηκε το ποσοστό εμφάνισης της υπό μελέτη νόσου καθώς και το ποσοστό κινδύνου.

Αντίστοιχα, χρονική συστάδα εντοπίζεται όταν παρατηρείται ένα σύνολο χρονικών στιγμών κατά τις οποίες ο κίνδυνος εμφάνισης της νόσου είναι αυξημένος.

Οι Στατιστικές Συναρτήσεις Σάρωσης μελετήθηκαν για πρώτη φορά εκτενώς από τον *Joseph Naus* το 1965, ο οποίος τις εφάρμοσε στον χρόνο (μονοδιάστατη περίπτωση) και στον χώρο (δισδιάστατη περίπτωση), ενώ σύμφωνα με τους *Tsui et al.* (2008), οι πιο γνωστές μέθοδοι επιτήρησης μπορούν να κατηγοριοποιηθούν σε 3 βασικές ομάδες : χρονική, χωρική και χωροχρονική επιτήρηση. Αρχικά, χρησιμοποιήθηκαν για την αναδρομική μελέτη και τον έλεγχο ύπαρξης συστάδων που αφορούσαν διάφορες μορφές καρκίνου και το σύνδρομο *Down*. Ωστόσο, μετά τα γεγονότα της 11ης Σεπτεμβρίου αλλά και την πληθώρα τρομοκρατικών επιθέσεων με χρήση επιστολών άνθρακα, έγινε επιτακτική η ανάγκη χρήσης των μεθόδων για προοδευτική ανάλυση με στόχο τον έγκαιρο εντοπισμό αλλά και πρόληψη τέτοιων περιπτώσεων.

Ως προς την μονοδιάστατη περίπτωση, οι Στατιστικές Συναρτήσεις Σάρωσης έχουν χρησιμοποιηθεί για εύρεση των χρονικών εκείνων συστάδων που ενδεχομένως να έχουν σχηματιστεί σε μία χρονική περίοδο. Ειδικότερα, ο *Naus*, θεώρησε μια ακολουθία από σημεία σε μία περίοδο $[a,b]$, καθώς κι ένα παράθυρο σάρωσης σταθερού μεγέθους $[t, t + w]$, όπου $w < b - a$, με το t να παίρνει τιμές από το a έως το $b - w$, και το οποίο θα κινείται πάνω στην δοθείσα χρονική περίοδο. Ως σημείο, θα θεωρούμε την καταγραφή του γεγονότος που μελετάμε. Για τις διάφορες τιμές που λαμβάνει το t, καταγράφουμε το μέγιστο αριθμό σημείων - περιστατικών που έχουν εμφανιστεί μέσα στο διάστημα $[t,t + w]$ του παραθύρου. Εν συνεχεία, ο αριθμός αυτός συγκρίνεται με την κατανομή των σημείων υπό την μηδενική υπόθεση, δηλαδή της υπόθεσης μη ύπαρξης συστάδων και αύξησης του ποσοστού κινδύνου.

Η χρονική σάρωση, η οποία αφορά την μονοδιάστατη περίπτωση, επικεντρώνεται στην αναγνώριση κάποιων τάσεων ή ακολουθιών όμοιων γεγονότων. Για παράδειγμα, έστω ότι διαθέτουμε ένα κινούμενο παράθυρο μήκους $m = 5$, για μία χρονική περίοδο $[a,b]$, όπου $a = 1$ και $b = 15$, το οποίο κινείται κατά μήκος μιας ακολουθίας δίτιμων αποτελεσμάτων (Επιτυχία – Αποτυχία), ΕΕΑΑΕΕΑΕΑΕΕΕΕΑΕ, μεγέθους n = 15. Τότε ο μέγιστος αριθμός επιτυχιών που μπορεί να παρατηρηθεί σε ένα παράθυρο θα είναι 4.

Οι Στατιστικές Συναρτήσεις Σάρωσης έχουν αναπτυχθεί πάνω σε τρεις περιπτώσεις χρόνου. Στην πρώτη μορφή, ένα παράθυρο μήκους w σαρώνει το χρονικό διάστημα $(0, T)$, ενώ παράλληλα καταγράφονται οι χρόνοι όπου σημειώθηκαν τα υπό μελέτη περιστατικά. Στην περίπτωση αυτή, η συνεχής στατιστική σάρωση είναι ο μέγιστος αριθμός

περιστατικών που έχουν λάβει χώρα στο παράθυρο. Στην δεύτερη περίπτωση, ο χρόνος χωρίζεται σε *T* ισομήκη διαστήματα, ενώ τα δεδομένα που λαμβάνουμε αφορούν την καταγραφή του πλήθους των περιστατικών που έχουν συμβεί σε καθένα από τα παραπάνω διαστήματα. Στην τρίτη και τελευταία περίπτωση, θεωρούμε ακολουθία που αποτελείται από *T* το πλήθος δοκιμές, όπου σε κάθε δοκιμή καταγράφεται αν έχει συμβεί ή όχι το υπό μελέτη περιστατικό. Στην περίπτωση αυτή, η στατιστική συνάρτηση σάρωσης είναι ο μέγιστος αριθμός περιστατικών που έχουν συμβεί για *w* συνεχόμενες δοκιμές.

Προκειμένου να προχωρήσουμε σε επιδημιολογική επιτήρηση μέσω Μεθόδων Στατιστικής Σάρωσης, είναι απαραίτητο να έχουν οριστεί σαφώς 3 βασικές ιδιότητες των μεθόδων. Αρχικά, θα πρέπει να δοθεί με ακρίβεια η περιοχή όπου θα πραγματοποιηθεί η σάρωση. Στη συνέχεια, ανάλογα με τον τύπο της ανάλυσης που θέλουμε να πραγματοποιήσουμε, αν θέλουμε δηλαδή χωρική ή και χρονική σάρωση, θα πρέπει να ορίσουμε το κατάλληλο παράθυρο, ως προς το μέγεθος και το σχήμα του. Επιπλέον, θα πρέπει να ορίσουμε τη συνάρτηση πιθανότητας όπου υπό τη μηδενική υπόθεση εμφανίζονται τα υπό μελέτη περιστατικά. Καθώς μας ενδιαφέρει η καταγραφή και μελέτη του ρυθμού εμφάνισης νόσου, προτείνεται η χρήση μοντέλων με βάση την κατανομή *Bernoulli* και *Poisson* για τον αριθμό των συμβάντων. Κριτήριο για την επιλογή του βέλτιστου από τα δύο μοντέλα, είναι ο παρονομαστής που χρησιμοποιείται για την εύρεση του ρυθμού εμφάνισης της υπό μελέτη νόσου. Για παράδειγμα, χρησιμοποιούμε μοντέλο *Poisson*, όταν μελετάμε την γεωγραφική κατανομή των περιστατικών καρκίνου του μαστού για μια περίοδο 5 χρόνων, και ο παρονομαστής του ρυθμού εμφάνισης αφορά τα χρόνια που έζησε το άτομο τα 5 υπό μελέτη χρόνια. Ωστόσο, όταν ενδιαφερόμαστε για τη γεωγραφική κατανομή των περιστατικών καρκίνου του μαστού αρχικού σταδίου, και ο παρονομαστής είναι το σύνολο των ασθενών με καρκίνο του μαστού, είναι προτιμότερο να προχωρήσουμε με μοντέλο *Bernoulli*. Τέλος, στην περίπτωση όπου έχουμε στη διάθεση μας δεδομένα που αφορούν χρόνους επιβίωσης, προτείνεται η χρήση εκθετικού μοντέλου.

2.2 Μοντέλα Poisson και Bernoulli

Τα μοντέλα *Poisson* και *Bernoulli* έχουν μελετηθεί εκτενώς από τον *Kulldorff* (1997). Έστω μία ακολουθία σημείων, η οποία σχετίζεται με τον αριθμό των καταγεγραμμένων συμβάντων στον χώρο, και ας συμβολίσουμε με *N(A)* τον αριθμό των υπό μελέτη

γεγονότων στην υποπεριοχή $A \subset G$, όπου G η συνολική γεωγραφική περιοχή που μελετάμε. Καθώς το παράθυρο σάρωσης κινείται πάνω στην περιοχή που μελετάμε, δημιουργεί ένα σύνολο Z ζωνών $Z \subset G$.

Για το μοντέλο *Bernoulli*, θεωρούμε μέτρο μ για κάθε μία από τις περιοχές $A \subset G$, τέτοια ώστε $\mu(A)$ να είναι ένας ακέραιος αριθμός για όλα τα υποσύνολα A. Κάθε μονάδα του μ θα αντιστοιχεί σε ένα άτομο, το οποίο μπορεί να βρίσκεται σε δυο καταστάσεις, να έχει ή να μην έχει το χαρακτηριστικό που μας ενδιαφέρει να μελετήσουμε. Καθένα από τα άτομα θα αναφέρεται ως σημείο, ενώ η γεωγραφική θέση του ατόμου θα συνθέτει μία ακολουθία σημείων. Βάσει του μοντέλου *Bernoulli*, υπάρχει μία μοναδική ζώνη $Z \subset G$, τέτοια ώστε κάθε άτομο εντός της έχει πιθανότητα p να αποτελεί σημείο, δηλαδή να έχει το υπό μελέτη χαρακτηριστικό, ενώ η πιθανότητα εκτός της ζώνης είναι q. Ο έλεγχος υποθέσεων διαμορφώνεται ως εξής: υπό την μηδενική υπόθεση, θεωρούμε ότι $H_0 : p = q$ ή διαφορετικά $H_0 : N(A) \sim Bin\ (\mu(A),p), \forall A$, έναντι της εναλλακτικής $H_1 : p > q$ ή διαφορετικά $H_1 : N(A) \sim Bin\ (\mu(A),p), \ \forall A \subset Z$, και $N(A) \sim Bin\ (\mu(A),q), \ \forall A \subset Z^c$.

Για την περίπτωση όπου τα δεδομένα μας προέρχονται από μία μη ομογενή διαδικασία *Poisson*, δηλαδή όταν τα γεγονότα συμβαίνουν με τυχαίο τρόπο, θεωρούμε ότι υπάρχει μία μοναδική ζώνη $Z \subset G$, τέτοια ώστε $N(A) \sim P\ (p\mu(AZ) + q\mu(AZ^c))$. Υπό τη μηδενική υπόθεση, θεωρούμε ότι $H_0 : p = q$ ή διαφορετικά $H_0 : N(A) \sim P\ (p\mu(A))$, , έναντι της εναλλακτικής $H_1 : p > q$.

Η επιλογή του καταλληλότερου μοντέλου εξαρτάται όπως προαναφέρθηκε από τον ρυθμό εμφάνισης της νόσου που θέλουμε να μελετήσουμε. Ωστόσο, σημαντικό κριτήριο αποτελεί και η φύση των δεδομένων που έχουμε στη διάθεσή μας. Αν διαθέτουμε δίτιμα δεδομένα (θετικός στην νόσο – όχι θετικός στην νόσο), τότε χρησιμοποιούμε μοντέλο *Bernoulli*, ενώ όταν έχουμε μετρήσεις που σχετίζονται με κάποιο παράγοντα κινδύνου, όπως για παράδειγμα μετρήσεις ραδιενέργειας *Geiger*, τότε θα πρέπει να προχωρήσουμε στη χρήση μοντέλου *Poisson* (*Kulldorff, 1997*). Ωστόσο, στην περίπτωση που ο συνολικός αριθμός σημείων - περιστατικών είναι αρκετά μικρότερος συγκριτικά με το $\mu(G)$, τότε τα δύο μοντέλα προσεγγίζουν ικανοποιητικά το ένα το άλλο.

2.3 Χωρική Μέθοδος Σάρωσης

Όπως αναφέρθηκε παραπάνω, μια από τις σημαντικότερες εφαρμογές των Μεθόδων Στατιστικής Σάρωσης αφορά την πρόληψη στη δημόσια υγεία, και πιο συγκεκριμένα τον έγκαιρο εντοπισμό των συστάδων εκείνων που ενδεχομένως να αποτελούν απόρροια κάποιου παράγοντα κινδύνου. Ιδιαίτερα, στο κομμάτι της επιδημιολογικής επιτήρησης, μας ενδιαφέρει η εύρεση αναδυόμενων συστάδων νόσων, οι οποίες είναι πιθανό να αποτελούν ένδειξη ξεσπάσματος κάποιας επιδημίας. Συνεπώς, είναι επιτακτική η ανάγκη επιδημιολογικής επιτήρησης αναλύοντας τα δεδομένα υγείας σε τακτά χρονικά διαστήματα. Με την έγκαιρη ανίχνευση συστάδων, προσπαθούμε να επιτύχουμε έγκαιρη αντιμετώπιση του προβλήματος πχ. μέσω εμβολιασμών, καθώς και να μειώσουμε τα ποσοστά θνησιμότητας.

Οι χωρικές συναρτήσεις σάρωσης παρουσιάστηκαν για πρώτη φορά από τους *Kulldorff* και *Nagarwalla*. Αναπτύχθηκαν με σκοπό την ανίχνευση πιθανών γεωγραφικών συστάδων καθώς και για τον εντοπισμό της θέσης τους (*Kulldorff*, 1997). Ανάλογα με τα δεδομένα που έχουμε στη διάθεσή μας, η χωρική σάρωση μπορεί να εφαρμοστεί είτε σε συγκεντρωτικά δεδομένα για την υπό μελέτη περιοχή, είτε στην ειδική περίπτωση όπου κάθε περιοχή αντιστοιχεί σε ένα άτομο σε κίνδυνο, και πιο συγκεκριμένα η γεωγραφική περιοχή είναι η ακριβής θέση του ατόμου (*Kulldorff*, 2000). Σύμφωνα με τους *Chen et al* (2010), η μέθοδος χωρικής σάρωσης που προτάθηκε από τον *Kulldorff*, χρησιμοποιείται κυρίως για αναδρομική επιτήρηση, δηλαδή, για να ελέγξουμε αν ένα σύνολο γεγονότων είναι τυχαία κατανεμημένα στον χώρο για μια συγκεκριμένη γεωγραφική περιοχή και χρονική περίοδο.

Με βάση την μέθοδο της χωρικής σάρωσης, ένα κυκλικό παράθυρο σαρώνει την υπό μελέτη περιοχή στον χάρτη, ενώ παράλληλα το κέντρο του κινείται με τρόπο τέτοιο ώστε κάθε διαφορετική θέση που λαμβάνει να περιέχει διαφορετικό συνδυασμό γειτονικών περιοχών. Όταν στο παράθυρο περιέχεται το κέντρο μίας περιοχής θα θεωρούμε ότι ολόκληρη η περιοχή περιέχεται μέσα σε αυτό. Επιπρόσθετα, για κάθε κέντρο του κύκλου, η ακτίνα του λαμβάνει τιμές από το μηδέν έως ένα ανώτατο όριο, το οποίο ορίζεται από τον ερευνητή, με την προϋπόθεση ότι κάθε φορά δεν θα περιλαμβάνεται στον κύκλο παραπάνω από τον μισό πληθυσμό σε κίνδυνο. Με τον τρόπο αυτό, δημιουργείται ένα πλήθος διακεκριμένων παραθύρων, καθένα από τα οποία περιλαμβάνει διαφορετικό συνδυασμό γειτονικών περιοχών, ενώ παράλληλα είναι πιθανό να περιέχει μία συστάδα.

Έστω ότι ο συνολικός αριθμός των γεγονότων που παρατηρούνται είναι *N*, τότε η στατιστική συνάρτηση *S* που χρησιμοποιείται ορίζεται ως ο λόγος της μέγιστης πιθανοφάνειας για όλους τους κύκλους *Z*, που σχηματίστηκαν από το παράθυρο. Πιο συγκεκριμένα:

$$\mathrm{S} = \frac{max_Z\{L(Z)\}}{L_0} = max_Z\left\{\frac{L(Z)}{L_0}\right\}$$

όπου το *L(Z)* αντιστοιχεί στη συνάρτηση μέγιστης πιθανοφάνειας του κύκλου *Z*, ενώ το L_0 είναι η συνάρτηση πιθανοφάνειας υπό τη μηδενική υπόθεση. Ως συνάρτηση πιθανοφάνειας για τον κύκλο *Z*, ορίζεται η πιθανότητα το ποσοστό των υπό μελέτη γεγονότων να είναι διαφορετικό εντός και εκτός του κύκλου.

Θεωρούμε ότι τα δεδομένα που έχουμε στη διάθεσή μας προέρχονται από μια διαδικασία Poisson. Έστω n_Z το πλήθος των γεγονότων στον κύκλο *Z* και *μ(Z)* ο αναμενόμενος αριθμός γεγονότων που περιμένουμε να συμβούν υπό τη μηδενική υπόθεση, τέτοιο ώστε *μ(A) = N*, όπου *A* το σύνολο της περιοχής που μας ενδιαφέρει να μελετήσουμε. Τότε, αποδεικνύεται ότι:

$$\frac{L(Z)}{L_0} = \begin{cases} \left\{\frac{n_z}{\mu(Z)}\right\}^{n_z}\left\{\frac{N-n_z}{N-\mu(Z)}\right\}^{N-n_z} & ,για\ n_z > \mu(Z) \\ 1 & ,για\ n_z \le \mu(Z) \end{cases}, \qquad (Kulldorff,\ 1997)$$

Κατά τη διάρκεια της σάρωσης, ο παραπάνω λόγος πιθανοφάνειας μεγιστοποιείται για όλους τους κύκλους, ενώ παράλληλα μας δίνει τον κύκλο που είναι πιο πιθανό να αποτελεί συστάδα. Το *p – value* για τον έλεγχο υποθέσεων λαμβάνεται μέσω της διαδικασίας *Monte Carlo*.

Είναι σημαντικό ωστόσο να επισημανθεί πως στην περίπτωση που έχουμε επιλέξει παράθυρο σταθερού μεγέθους, η στατιστική ελέγχου που θα χρησιμοποιηθεί θα είναι ο μέγιστος αριθμός περιστατικών που παρατηρήθηκαν στο παράθυρο, σε οποιαδήποτε χρονική στιγμή της σάρωσης, ενώ στην περίπτωση που επιλέξουμε ένα παράθυρο σάρωσης μεταβλητού μεγέθους, η στατιστική ελέγχου θα γίνει μέσω του ελέγχου πιθανοφανειών όπως περιγράφηκε παραπάνω.

Η τεχνική Monte Carlo αναπτύχθηκε από τον *Dwass* (1957), ενώ χρησιμοποιήθηκε για πρώτη φορά για την εύρεση *p – value* στις Μεθόδους Σάρωσης από τους *Turnbull et al.*

(1990). Με βάση τη διαδικασία *Monte Carlo*, καθώς είναι ήδη γνωστό το μέτρο μ, μπορούμε να παράγουμε «αντίγραφα» των δεδομένων μας υπό την H_0. Με την δημιουργία 9999 τέτοιων «αντιγράφων», η στατιστική συνάρτηση ελέγχου είναι στατιστικά σημαντική σε επίπεδο στατιστικής σημαντικότητας 5%, όταν η τιμή της στατιστικής συνάρτησης των πραγματικών μας δεδομένων, είναι ανάμεσα στις 500 υψηλότερες τιμές των στατιστικών συναρτήσεων που προκύπτουν από τα δεδομένα που έχουν παραχθεί μέσω αντιγράφων (*Kulldorff, 1997*).

Η μέθοδος της χωρικής σάρωσης, πέρα από την πιο πιθανή συστάδα, είναι σε θέση να εντοπίσει και δευτερεύουσες συστάδες, ενώ παράλληλα μπορεί να τις ταξινομήσει με βάσει τον λόγο πιθανοφάνειας της καθεμίας. Η μέθοδος όπως περιγράφηκε παραπάνω είναι ιδιαίτερα εύχρηστη ενώ επιπλέον αποτελεί πολύ σημαντικό εργαλείο για όσους επιθυμούν να πραγματοποιήσουν επιδημιολογική επιτήρηση. Πιο συγκεκριμένα, προσαρμόζεται τόσο για πυκνότητα ανομοιογενούς πληθυσμού, όσο και για κάθε αριθμό συμμεταβλητών. Επιπρόσθετα, ένα πολύ σημαντικό πλεονέκτημα της μεθόδου είναι ότι μηδενίζει τη μεροληψία που θα προέκυπτε αν ορίζαμε εκ των προτέρων τόσο το μέγεθος όσο και την τοποθεσία της συστάδας. Τέλος, ο έλεγχος του λόγου πιθανοφάνειας, λαμβάνει υπόψιν τους πολλαπλούς ελέγχους που γίνονται, αλλά δίνει ένα $p - value$ για τον έλεγχο της μηδενικής υπόθεσης. Στην περίπτωση της απόρριψης της μηδενικής υπόθεσης, η μέθοδος μπορεί να μας δώσει την κατά προσέγγιση θέση της συστάδας. Σύμφωνα με τους *Wanger et al* (2011) «Ο κύριος στόχος των χωρικών μεθόδων σάρωσης είναι η αναγνώριση της περιοχής, του μεγέθους και του σχήματος των πιθανών συστάδων, και στην συνέχεια ο έλεγχος για το αν η πιθανές συστάδες αποτελούν πραγματικές συστάδες, και συνεπώς χρήζουν περαιτέρω επιδημιολογικής μελέτης, ή αν αυτές έχουν σχηματιστεί κατά τύχη».

Ωστόσο, η επιδημιολογική επιτήρηση με χρήση αμιγώς χωρικών μεθόδων σάρωσης, παρουσιάζει σημαντικά προβλήματα. Πιο συγκεκριμένα, αν κάποιο αιφνίδιο ξέσπασμα κάποιας επιδημίας συνέβαινε σε μία μικρή περιοχή, ενδεχομένως να μην έπεφτε στην αντίληψή μας, καθώς υπάρχει πιθανότητα η περιοχή να αποτελεί μέρος ομαδοποιημένων δεδομένων μιας μεγαλύτερης και ευρύτερης περιοχής. Μια λύση στο παραπάνω θα ήταν να διαμερίζαμε τις περιοχές που μας ενδιαφέρει, και να τις μελετήσουμε σε μικρότερες υποπεριοχές, προκειμένου να μπορέσουμε να εποπτεύσουμε κάθε περιοχή ξεχωριστά. Κάτι τέτοιο όμως, θα αύξανε σημαντικά τον αριθμό των περιοχών παρακολούθησης, με αποτέλεσμα πραγματοποίηση πολλαπλών ελέγχων, οι οποίοι με τη σειρά τους θα μας

έδιναν αρκετούς λανθασμένους συναγερμούς, δηλαδή θα είχαμε σημαντική αύξηση της πιθανότητας Σφάλματος Τύπου *I*. Επιπρόσθετα, η μέθοδος της χωρικής σάρωσης, δεν είναι σε θέση να εντοπίσει ξέσπασμα το οποίο συμβαίνει στα σύνορα δυο υπό μελέτη περιοχών.

Τα συστήματα καταγραφής των αρμόδιων φορέων, ανανεώνονται διαρκώς με την εισαγωγή των νέων κρουσμάτων, σε μία προσπάθεια επικαιροποίησης των αρχείων αυτών. Έτσι, προκειμένου να είμαστε σε θέση να εντοπίσουμε τις νέες συστάδες που ενδεχομένως να σχηματιστούν με την καταγραφή νέων περιπτώσεων νόσου, θα έπρεπε να επαναλάβουμε τη χωρική σάρωση σε τακτά χρονικά διαστήματα. Ωστόσο, επαναλαμβανόμενες χωρικές σαρώσεις μπορεί να δημιουργήσουν σοβαρά προβλήματα στη μελέτη. Στην προσπάθειά μας να εντοπίσουμε όσο το δυνατόν γρηγορότερα τις πιθανές συστάδες, η ισχύς του ελέγχου θα ελαττώνεται. Πιο συγκεκριμένα, αν κάποιος παράγοντας κινδύνου έχει εμφανιστεί τα τελευταία χρόνια, τότε πραγματοποιώντας χωρική σάρωση είναι πιθανό να μην μπορέσουμε να εντοπίσουμε τις συστάδες που θα δημιουργηθούν με την εμφάνιση του νέου παράγονται κινδύνου. Αυτό μπορεί να προκύψει καθώς η μέθοδος λαμβάνει υπόψιν και τις παλαιότερες περιόδους, όπου ο παράγοντας κινδύνου δεν υπήρχε, έχοντας σαν αποτέλεσμα την εξασθένιση της δυναμικότητας του παράγοντα αυτού. Έτσι, οι ερευνητές οδηγήθηκαν στην ανάπτυξη μίας νέας μεθόδου σάρωσης, η οποία επιλύει τα προβλήματα που προαναφέρθηκαν, την Μέθοδο Σάρωσης στον Χώρο – Χρόνο (*Space – Time Scan Statistics*).

2.4 Μέθοδοι Σάρωσης στον Χώρο και τον Χρόνο

Όπως επισημάνθηκε παραπάνω, η χρήση αμιγώς χωρικής σάρωσης για αρκετά μεγάλες χρονικές περιόδους, έχει σαν αποτέλεσμα τον μη εντοπισμό των συστάδων εκείνων που ενδεχομένως να έχουν σχηματιστεί τις πιο πρόσφατες χρονικές στιγμές της μελέτης μας. Μία λύση στο παραπάνω πρόβλημα θα ήταν να χρησιμοποιούνται στη μελέτη μόνο οι πιο πρόσφατες χρονικές περίοδοι. Καθώς, όμως, δεν είναι γνωστός ο επαρκής αριθμός περιόδων που πρέπει να χρησιμοποιηθούν για την πραγματοποίηση μιας επιτυχούς χωρικής σάρωσης, θα έπρεπε να είμαστε πολύ προσεκτικοί στην επιλογή του πλήθους των περιόδων που θα λάβουμε υπόψιν στην επιτήρηση. Η χρήση μικρού αριθμού περιόδων, θα οδηγούσε σε έλεγχο μικρότερης ισχύος, με αποτέλεσμα να μην μπορεί να εντοπιστεί κάποιος παράγοντας ρίσκου μέτριου κινδύνου, ενώ αντίθετα, η χρήση μεγάλου αριθμού περιόδων

θα οδηγούσε και πάλι σε έλεγχο χαμηλότερης ισχύς, με αποτέλεσμα να μην είναι δυνατός ο εντοπισμός των συστάδων εκείνων που σχηματίστηκαν από παράγοντα κινδύνου, ο οποίος εμφανίστηκε τις πιο πρόσφατες χρονικές στιγμές. Επιπρόσθετα, η εφαρμογή χωρικής σάρωσης, θέτει περιορισμούς και ως προς τον τύπο των συστάδων που μπορεί να εντοπίσει. Ενώ, αναδεικνύει έγκαιρα τις πιο συμπαγείς συστάδες, υστερεί στον εντοπισμό επιμηκών συστάδων, οι οποίες χρειάζονται περισσότερο χρόνο για τη διερεύνηση τους, γεγονός ασύμφορο, καθώς, ιδιαίτερα σε περιπτώσεις μεγάλου επιδημιολογικού συναγερμού, ο έγκαιρος εντοπισμός των πιθανών συστάδων είναι προαπαιτούμενος. Λαμβάνοντας υπόψιν τους παραπάνω κινδύνους χρήσης μόνο χωρικής σάρωσης, προτάθηκε η χρήση Μεθόδων Σάρωσης στον Χώρο και τον Χρόνο ταυτόχρονα (*Space – Time Scan Statistics*).

Στο σημείο αυτό είναι πολύ σημαντικό να επισημανθεί η προσθήκη του χρόνου στη σάρωση μας, και πώς η προσθήκη αυτή διαφοροποιεί τη μέθοδο σάρωσης στον χώρο και τον χρόνο ταυτόχρονα από τη μέθοδο χωρικής σάρωσης. Ο χρόνος έχει ένα πολύ σημαντικό σημείο αναφοράς, το παρόν. Ιδιαίτερα στην επιδημιολογία, το ενδιαφέρον μας επικεντρώνεται στις συστάδες εκείνες που εξακολουθούν να είναι ενεργές στον παρόντα χρόνο, συνεπώς η χρήση προοδευτικών μεθόδων για την εύρεση συστάδων που τελειώνουν στο παρόν είναι απαραίτητη, σε αντίθεση με την χρήση αναδρομικών μεθόδων. Επιπλέον, ο χρόνος έχει μια συγκεκριμένη κατεύθυνση, από το παρελθόν, στο παρόν και στην συνέχεια στο μέλλον. Στόχος της βιοεπιτήρησης, είναι ο εντοπισμός των συστάδων εκείνων που αναδύονται στην πορεία του χρόνου. Για παράδειγμα, στην επιδημιολογία, και πιο συγκεκριμένα στην περίπτωση της *Covid – 19*, ενώ η επιδημία ξεκίνησε με την εμφάνιση πολύ μικρού αριθμού κρουσμάτων, με την πάροδο του χρόνου, ο αριθμός αυτός σημείωσε ραγδαία αύξηση. Στόχος της Σάρωσης στον χρόνο και στον χώρο, είναι η ανάδειξη χωρο - χρονικών συστάδων, που εμφανίζουν μετρήσεις (π.χ. αριθμό κρουσμάτων) σημαντικά υψηλότερες από ό,τι οι αναμενόμενες.

Σε αντίθεση με τις χωρικές μεθόδους σάρωσης, στις οποίες γίνεται χρήση κυκλικού παραθύρου, οι μέθοδοι στον χώρο και χρόνο χρησιμοποιούν ένα κυλινδρικό παράθυρο. Η βάση του κυλίνδρου εκφράζει τον χώρο, ενώ το ύψος του τον χρόνο. Ωστόσο, το μέγεθος της βάσης, δεν εξαρτάται από το ύψος, δηλαδή ο χρόνος και ο χώρος μεταβάλλονται ανεξάρτητα μεταξύ τους, ενώ παράλληλα ο κύλινδρος είναι ευέλικτος ως προς το μέγεθος της βάσης αλλά και ως προς τον χρόνο που επιλέγεται ως έναρξη της επιτήρησης. Κάθε

κύλινδρος που σχηματίζεται αποτελεί και μία πιθανή συστάδα. Στα πλεονεκτήματα της μεθόδου συγκαταλέγεται το γεγονός ότι λαμβάνουμε υπόψη μόνο τους κυλίνδρους εκείνους που φτάνουν μέχρι τη λήξη της επιτήρησης. Συγκεκριμένα, στο τέλος, επιλέγονται οι συστάδες εκείνες που είναι ενεργές, δηλαδή που εξακολουθούν να βρίσκονται σε κίνδυνο, κατά τη λήξη της επιτήρησης, ενώ παράλληλα παραλείπονται εκείνες οι οποίες ενδεχομένως να υπήρχαν στο παρελθόν, αλλά πλέον δεν αποτελούν πρόβλημα για τη δημόσια υγεία. Υποθέτοντας λοιπόν ότι, $[Y_1, Y_2]$ η περίοδος η οποία επιλέγεται για μελέτη, και για την οποία έχουμε στην διάθεσή μας δεδομένα, s η ημερομηνία έναρξης του κυλίνδρου και t η ημερομηνία λήξης του, στο τέλος επιλέγονται οι κύλινδροι εκείνοι για τους οποίους ισχύει: $Y_1 \leq s \leq t = Y_2$.

Εν συνεχεία, σε αντιστοιχία με τις μεθόδους χωρικής σάρωσης, υπολογίζονται οι λόγοι πιθανοφάνειας, και πραγματοποιείται έλεγχος λόγου πιθανοφανειών, με τον ίδιο τρόπο, όπως αναφέρθηκε παραπάνω. Ωστόσο σημαντική διαφορά αποτελεί το γεγονός ότι η πιθανοφάνεια που υπολογίζεται σε αυτή την περίπτωση είναι πιο ευαίσθητη στους υπολογισμούς καθώς αυτή δεν εφαρμόζεται μόνο σε δύο διαστάσεις αλλά σε τρεις.

Ωστόσο, ένα σημαντικό μειονέκτημα της μεθόδου είναι ότι τα υπολογιζόμενα $p - values$ που χρησιμοποιούνται για τον έλεγχο της στατιστικής σημαντικότητας των συστάδων που έχουν αναδειχθεί, ενώ υπολογίζονται για πολλά πιθανά μεγέθη και θέσεις συστάδων αλλά και πολλών χρονικών περιόδων, αποτελούν $p - value$ ενός μόνο ελέγχου, και όχι πολλαπλών περιοδικών ελέγχων που πραγματοποιούνται κατά την διάρκεια των χρόνων της μελέτης μας.

Το παραπάνω αποτελεί πρόβλημα, καθώς, ιδιαίτερα στον τομέα της υγείας, τα συστήματα καταγραφής νέων κρουσμάτων, ανανεώνονται ανά τακτά χρονικά διαστήματα, ενώ έχει ιδιαίτερη βαρύτητα για την επιδημιολογική επιτήρηση, τα κρούσματα να καταγράφονται όσο το δυνατόν πιο έγκαιρα. Έτσι, με την προσθήκη των νεών περιπτώσεων στα συστήματα καταγραφής των αρμόδιων υπηρεσιών, είναι απαραίτητο να προσαρμόζουμε τα διαρκώς ανανεωμένα δεδομένα μας σε πολλαπλούς ελέγχους, λαμβάνοντας υπόψιν όχι μόνο τις πιθανές θέσεις και τα μεγέθη των συστάδων στη διάρκεια του χρόνου επιτήρησης, αλλά συνεκτιμώντας και τις αναλύσεις που έχουν προηγηθεί σε προγενέστερους χρόνους.

Και σε αυτή την περίπτωση, θεωρούμε τους κυλίνδρους εκείνους για τους οποίους ισχύει $Y_1 \leq s \leq t = Y_2$, για μία περίοδο μελέτης $[Y_1, Y_2]$. Η διαφοροποίηση παρουσιάζεται

στο ότι, προκειμένου να προσαρμοστούν τα δεδομένα μας στους πολλαπλούς ελέγχους, ενώ η πιθανοφάνεια για τα πραγματικά δεδομένα μας ορίζεται όπως και πριν ως ο μεγαλύτερος κύλινδρος, ο οποίος εξακολουθεί να υφίσταται κατά τη λήξη της επιτήρησης, για δεδομένα που προέρχονται από τυχαία δειγματοληψία, η πιθανοφάνεια ορίζεται ως ο μέγιστος κύλινδρος ανάμεσα στους κυλίνδρους που έχουν παρουσιαστεί σε προηγούμενες αναλύσεις. Συγκεκριμένα, επιλέγονται οι κύλινδροι για τους οποίους ισχύει $Y_1 \leq s \leq t \leq Y_2$ και $t \geq Y_m$, , όπου Y_m η περίοδος όπου ξεκίνησε και πάλι η επιτήρηση. Στη συνέχεια, η συστάδα που ανιχνεύεται είναι στατιστικά σημαντική σε επίπεδο στατιστικής σημαντικότητας α, αν η πιθανότητα να έχει ανιχνευθεί συστάδα με μεγαλύτερη πιθανοφάνεια κατά τη διάρκεια όλων των προηγούμενων αναλύσεων ή των πιο πρόσφατων είναι το πολύ α (*Kulldorff,* 2001). Τα *p – value* των συστάδων ωστόσο, θα είναι υψηλότερα σε σχέση με την ανάλυση που έχει προηγηθεί, καθώς προσαρμόζονται πλέον σε χρονικά περιοδικές αναλύσεις.

2.5 Το Λογισμικό SaTScan

Το *SaTScan* είναι ένα δωρεάν λογισμικό πακέτο, το οποίο χρησιμοποιείται για την ανάλυση χωρικών, χρονικών και χωρο – χρονικών δεδομένων. Αναπτύχθηκε από τον *Matrin Kulldorf* σε συνεργασία με την *Information Management Services Inc,* ενώ παράλληλα χρηματοδοτήθηκε από πολλούς οργανισμούς όπως το Εθνικό Ινστιτούτο Καρκίνου καθώς και το Εθνικό Ινστιτούτο Παιδικής Υγείας της Αμερικής.

Η ανάπτυξη του πακέτου βασίστηκε στις χωρικές, χρονικές και χωρο – χρονικές μεθόδους σάρωσης, ενώ η χρήση του αποσκοπεί στην διενέργεια 4 βασικών μορφών βιοεπιτήρησης:

- Την διενέργεια γεωγραφικής επιτήρησης μιας ασθένειας, προκειμένου να βρεθούν πιθανές χωρικές και χωρο – χρονικές συστάδες, και να εξεταστεί αν αυτές είναι στατιστικά σημαντικές.
- Τον έλεγχο για το αν μια ασθένεια είναι τυχαία κατανεμημένη στον χώρο, τον χρόνο ή τον χώρο και τον χρόνο.
- Την αξιολόγηση της στατιστικής σημαντικότητας των συστάδων μιας ασθένειας.

- Την διεξαγωγή επαναλαμβανόμενων χρονο – περιοδικών παρακολουθήσεων ασθενειών, προκειμένου να ανιχνευτούν έγκαιρα επικείμενα ξεσπάσματα επιδημιών.

Το πακέτο μπορεί να χρησιμοποιηθεί για διακριτές αλλά και συνεχείς συναρτήσεις σάρωσης. Στην διακριτή περίπτωση, η τοποθεσία των παρατηρήσεων που λαμβάνουμε πρέπει να είναι γνωστές και να δηλώνονται από τον χρήστη, ενώ στην συνεχή περίπτωση, η τοποθεσία των δεδομένων μας είναι τυχαία ενώ τα συμβάντα μπορούν να συμβούν οπουδήποτε μέσα στην προκαθορισμένη περιοχή έρευνας.

Επίσης, για την διακριτή περίπτωση σάρωσης, το λογισμικό χρησιμοποιεί είτε μοντέλα που βασίζονται στην διακριτή κατανομή *Poisson*, όπου τα καταγεγραμμένα γεγονότα στην υπό μελέτη περιοχή ακολουθεί την κατανομή *Poisson*, βάσει ενός γνωστού πληθυσμού σε κίνδυνο, είτε μοντέλα *Bernoulli*, όπου τα δεδομένα έχουν την μορφή συμβάντων (*cases*) και μη – συμβάντων (*control*). Επίσης το πακέτο δίνει την δυνατότητα χρήσης και άλλων μοντέλων όπως το μοντέλο χωρο – χρονικής αντιμετάθεσης, το οποίο χρησιμοποιεί μόνο το πλήθος των καταγεγραμμένων συμβάντων, το πολυωνυμικό μοντέλο για κατηγορικά δεδομένα αλλά και το εκθετικό μοντέλο για δεδομένα που αφορούν χρόνους επιβίωσης.

2.6 Η μελέτη των *Desjardins et al.*

Η έρευνα που θα παρουσιαστεί στο παρόν κεφάλαιο πραγματοποιήθηκε από τους *Desjardins, Hohl, and Delmelle* (*2020*). Το πρώτο διαγνωσμένο κρούσμα της *Covid* – 19 εμφανίστηκε πρώτη φορά στην πόλη Ουχάν, πρωτεύουσα της επαρχίας Χουπέι της Κίνας, τον Δεκέμβριο του 2019. Στη συνέχεια, κρούσματα της νόσου έκαναν την εμφάνιση τους σε ολόκληρη την υφήλιο, καθιστώντας την *Covid* – 19 τη μεγαλύτερη πανδημία των τελευταίων ετών. Μέχρι τις 28 Μαρτίου 2020, είχαν σημειωθεί παγκοσμίως 649.000 κρούσματα και 115.500 θάνατοι. Αντίστοιχα, για την ίδια χρονική περίοδο, στις Ηνωμένες Πολιτείες Αμερικής, καταγράφηκαν 115.000 επιβεβαιωμένα κρούσματα της νόσου και 1891 θάνατοι. Η συμπτωματολογία της *Covid* – 19 για το 80% των περιπτώσεων που παρατηρήθηκαν ήταν ήπια, ενώ σε αυτή περιλαμβάνονται πυρετός, βήχας και δυσκολία στην αναπνοή.

Τα δεδομένα που αναλύθηκαν παρακάτω, προέρχονται το πανεπιστήμιο *Johns Hopkins* και μπορούν να βρεθούν στην ιστοσελίδα του *GitHub* (https://github.com/CSSEGISandData/COVID-19), ενώ αφορούν το χρονικό διάστημα από τις 22 Γενάρη του 2020 έως τις 27 Μαρτίου του 2020. Επιπρόσθετα, τα δεδομένα περιέχουν χωρικές και χρονικές πληροφορίες των κρουσμάτων, ενώ παράλληλα ομαδοποιούνται σε επίπεδο καταγραφής κρουσμάτων ανά πολιτεία. Στην μελέτη συμπεριλαμβάνονται 48 όμορες πολιτείες καθώς και η πολιτεία της Ουάσιγκτον, ενώ δεν λαμβάνονται υπόψη κρούσματα αγνώστου προελεύσεως.

Στόχος της μελέτης είναι να βρεθούν συστάδες χώρου και χρόνου, οι οποίες είναι ακόμη ενεργές, δηλαδή τις συστάδες εκείνες για τις οποίες ο κίνδυνος παραμένει μέχρι και το τέλος της σάρωσης, ενώ θέλουμε να απορρίψουμε εκείνες για τις οποίες ο κίνδυνος δεν είναι στατιστικά σημαντικός. Για τον σκοπό αυτό, θα γίνει χρήση μοντέλου *Poisson* για χωρική και χρονική σάρωση, μέσω της εφαρμογής *SatScan*. Κατά τη διάρκεια της χωροχρονικής σάρωσης, κινούμενοι κύλινδροι ποικίλου μεγέθους θα σαρώνουν τις υπό μελέτη περιοχές στον χάρτη, ενώ το κέντρο κάθε κυλίνδρου θα είναι το κέντρο κάθε πολιτείας. Σε καθέναν από τους κυλίνδρους θα υπολογίζονται τόσο τα αναμενόμενα κρούσματα της *Covid* – 19 , όσο και τα πραγματικά. Κάθε κύλινδρος θα επεκτείνεται μέχρι μία μέγιστη ακτίνα, η οποία ορίζεται από τον χρήστη, ενώ ο κύλινδρος που θα σχηματίζεται κάθε φορά θα αποτελεί και μία πιθανή συστάδα. Στην παρούσα εργασία έχει οριστεί πως κάθε παράθυρο δεν θα περιλαμβάνει περισσότερο από το 10% του πληθυσμού σε κίνδυνο, προκειμένου να αποφευχθούν υπερβολικά μεγάλες συστάδες, ενώ παράλληλα θα περιλαμβάνει το 50% του χρόνου της μελέτης μας. Κάθε συστάδα η οποία θα λαμβάνεται υπόψη, θα πρέπει να έχει διάρκεια το λιγότερο 2 μέρες, καθώς επίσης θα πρέπει να έχουν καταγραφεί σε αυτή τουλάχιστον 5 επιβεβαιωμένα κρούσματα της νόσου.

Θεωρώντας ότι τα δεδομένα μας προέρχονται από την κατανομή Poisson, ο έλεγχος που πραγματοποιείται αφορά την μηδενική υπόθεση H_0 , ότι κάθε πολιτεία έχει ένα σταθερό κίνδυνο έντασης μ, ο οποίος είναι ανάλογος του πληθυσμού σε κίνδυνο, έναντι της εναλλακτικής H_1, όπου οι καταγεγραμμένες περιπτώσεις της *Covid* – 19 είναι σημαντικά αυξημένες σε σχέση με τον αναμενόμενο αριθμό περιπτώσεων που προκύπτει υπό την H_0. Υπό την H_0, ο αναμενόμενος αριθμός κρουσμάτων της *Covid* – 19 προκύπτει από τον τύπο:

$$\mu = p\frac{C}{P}$$

όπου p ο πληθυσμός στην υπό μελέτη πολιτεία, C ο συνολικός αριθμός κρουσμάτων *Covid* – 19 στις Ηνωμένες Πολιτείες Αμερικής, και P ο συνολικός εκτιμώμενος πληθυσμός στις Ηνωμένες Πολιτείες.

Στην συνέχεια, ο έλεγχος λόγου πιθανοφανειών πραγματοποιείται μέσω της χρήσης του τύπου:

$$\frac{L(z)}{L_0} = \frac{\left(\frac{n_z}{\mu(z)}\right)^{n_z}\left(\frac{N-n_z}{N-\mu(z)}\right)^{N-n_z}}{\left(\frac{N}{\mu(T)}\right)}$$

όπου $L(z)$ η συνάρτηση πιθανοφάνειας του κυλίνδρου *Z*, L_0 η συνάρτηση πιθανοφάνειας υπό την H_0, n_z ο αριθμός κρουσμάτων που παρατηρούνται στον κύλινδρο Z, $\mu(z)$ ο αναμενόμενος αριθμός κρουσμάτων στον κύλινδρο *Z*, N ο συνολικός αριθμός παρατηρούμενων κρουσμάτων στις Ηνωμένες Πολιτείες καθ' όλη την διάρκεια της μελέτης και $\mu(T)$ ο συνολικός αναμενόμενος αριθμός περιπτώσεων $Covid - 19$.

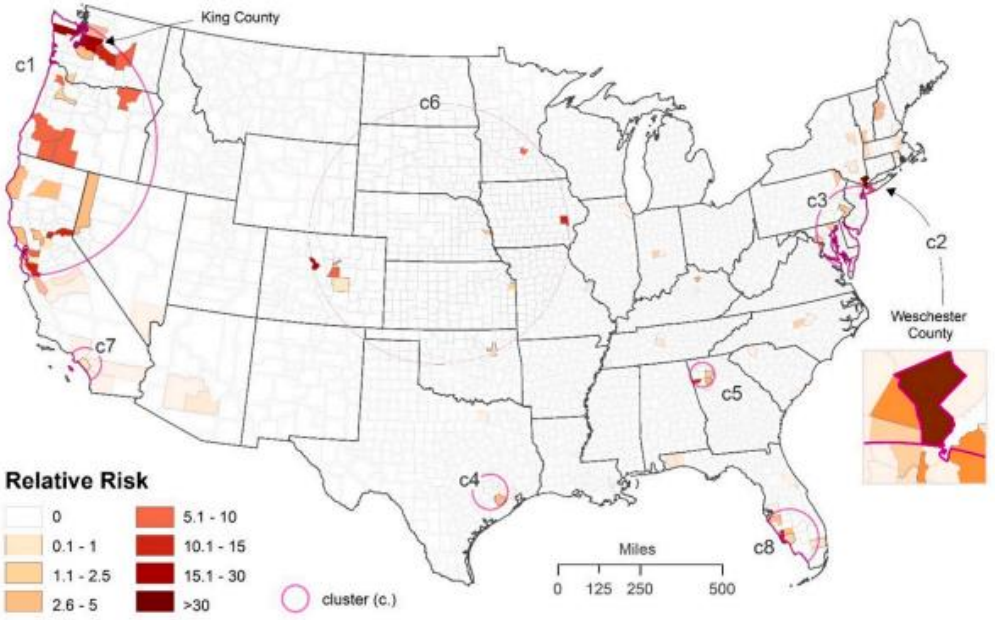
Θα θεωρούμε ότι σε έναν κύλινδρο παρατηρείται αυξημένος κίνδυνος, όταν ο λόγος πιθανοφανειών είναι μεγαλύτερος της μονάδας, και πιο συγκεκριμένα $\frac{n_z}{\mu(z)} > \frac{N-n_z}{N-\mu(z)}$. Ο κύλινδρος για τον οποίο προκύπτει ο μεγαλύτερος λόγος πιθανοφανειών, θα θεωρείται η πιο πιθανή συστάδα. Επιπλέον, χρησιμοποιείται η τεχνική *Monte Carlo*, με 999 προσομοιώσεις, προκειμένου να εκτιμηθεί η στατιστική σημαντικότητα των συστάδων που έχουν προκύψει. Για κάθε προσομοίωση, προσαρμόζεται ο ίδιος αριθμός περιπτώσεων $Covid$ – 19, και στην συνέχεια υπολογίζεται η πιθανοφάνεια, με αποτέλεσμα να λαμβάνουμε 999 λόγους πιθανοφανειών για κάθε υποψήφια συστάδα, υπό την H_0. Επιπλέον, λαμβάνονται υπόψη και δευτερεύουσες συστάδες, οι οποίες θα θεωρούνται στατιστικά σημαντικές εάν $p - value < 0.05$.

Η ανάλυση έχει γίνει για δύο περιόδους, Η πρώτη αφορά την περίοδο από 22 Γενάρη 2020 έως 9 Μαρτίου 2020, ενώ η δεύτερη έγινε για την περίοδο 22 Γενάρη 2020 έως 27 Μαρτίου 2020, προκειμένου να ληφθούν υπόψη τα νέα κρούσματα τα οποία καταγράφηκαν από τις 10 Μάρτη έως τις 27.

Παρακάτω δίνεται ο πίνακας ανάλυσης της περιόδου 22 Γενάρη 2020 με 9 Μαρτίου 2020 καθώς και ο χάρτης των περιοχών σάρωσης.

Cluster	Duration (days)	*p*	Observed	Expected	RR	# of counties	# of counties with RR > 1
1	Feb 29th - Mar 9th	<0.001	207	7.9	43.2	107	23
2	Mar 4th - Mar 9th	<0.001	97	1.5	639	1	1
3	Mar 5th - Mar 9th	<0.001	53	5.1	11.3	66	9
4	Mar 5th - Mar 9th	<0.001	12	0.9	13.1	12	2
5	Mar 3rd - Mar 9th	<0.001	10	0.6	16.3	12	4
6	Mar 6th - Mar 9th	0.001	17	2.8	6.3	552	10
7	Mar 4th - Mar 9th	0.002	16	2.5	6.4	2	2
8	Mar 7th - Mar 9th	0.017	8	0.5	14.4	13	4

Πίνακας 2.5.1. Αναδυόμενες χωρο – χρονικές συστάδες περιόδου 22/01/2020 έως 09/03/2020
Πηγή: *Desjardins M.R. et al.* (*2020*)



Διάγραμμα 2.5.1. Χωρική κατανομή των αναδυόμενων χωρο – χρονικών συστάδων Covid – 19
Πηγή: *Desjardins M.R. et al.* (*2020*)

Βάσει των παραπάνω αποτελεσμάτων παρατηρούμε ότι έχουμε 8 αναδυόμενες συστάδες κρουσμάτων της *Covid* – 19, για την περίοδο από 22 Γενάρη έως 9 Μαρτίου Η πρώτη συστάδα (*c*1) βρίσκεται στο βορειοδυτικό μέρος των Ηνωμένων Πολιτείων και περιλαμβάνει 107 κομητείες. Για την συστάδα αυτή, παρατηρούμε ότι για την περίοδο 29 Φεβρουαρίου έως 9 Μαρτίου, τα κρούσματα που έχουν καταγραφεί είναι 207, ενώ ο αναμενόμενος αριθμός είναι 7.9. Ο σχετικός κίνδυνος της συστάδας είναι 43.2, δηλαδή άτομα τα οποία βρίσκονται εντός της συγκεκριμένης συστάδας, έχουν 43.2 φορές μεγαλύτερη πιθανότητα να νοσήσουν σε σχέση με άτομα που βρίσκονται εκτός αυτής. Μέσα στην συστάδα, υπάρχουν 23 κομητείες για τις οποίες ο σχετικός κίνδυνος είναι μεγαλύτερος από την μονάδα. Μέσα στην συστάδα αυτή, περιλαμβάνεται και η

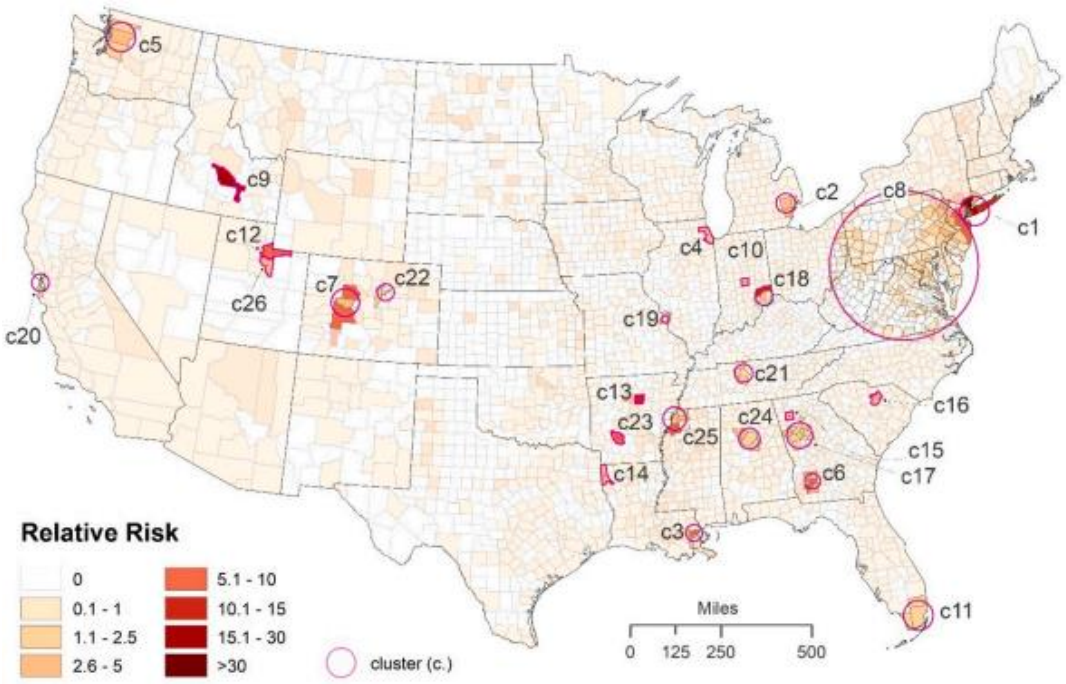
κομητεία *King County*, στην οποία είχε καταγραφεί και η πρώτη περίπτωση *Covid* -19 στην Αμερική, προερχόμενο από ταξιδιώτες που είχαν επισκεφτεί την Κίνα.

Η δεύτερη συστάδα (*c*2) αφορά την περίοδο 4 Μαρτίου έως 9 Μαρτίου, και περιλαμβάνει μόνο την κομητεία του *Westchester,* στην οποία παρατηρούνται 97 κρούσματα της νόσου, με αναμενόμενο αριθμό περιπτώσεων στο 1.5, καθώς επίσης και με σχετικό κίνδυνο 639. Οι κομητείες *King County* και *Westchester* είναι γνωστές ως τα πρώτα δυο *hotspots* της νόσου στην Αμερική. Η τρίτη συστάδα (*c*3) αφορά την περίοδο 5 Μαρτίου έως 9 Μαρτίου, και περιλαμβάνει 66 κομητείες, όπου ο παρατηρούμενος αριθμός κρουσμάτων είναι 53, ο αναμενόμενος αριθμός 5.1, με σχετικό κίνδυνο 11.3, ενώ από το σύνολο των 66 κομητειών, 9 από αυτές έχουν σχετικό κίνδυνο μεγαλύτερο της μονάδας. Στον πίνακα 1 εμφανίζονται επιπλέον 5 ακόμη συστάδες των οποίων τα *p – value* είναι μικρότερα του 0.05.

Μία δεύτερη ανάλυση πραγματοποιείται για την περίοδο 22 Γενάρη 2020 έως 27 Μαρτίου 2020, καθώς προστέθηκαν τα νέα κρούσματα που καταγράφηκαν την περίοδο 10 έως 27 Μαρτίου.

Cluster	Duration (days)	*p*	Observed	Expected	RR	# of counties	# of counties with RR > 1
1	Mar 19th - Mar 27th	<0.001	56,189	3343.8	33.1	14	14
2	Mar 21st - Mar 27th	<0.001	3036	835.8	3.7	3	3
3	Mar 19th - Mar 27th	<0.001	1477	228.0	6.5	2	2
4	Mar 24th - Mar 27th	<0.001	1953	636.4	3.1	1	1
5	Mar 17th - Mar 27th	<0.001	1929	1032.9	1.9	4	4
6	Mar 20th - Mar 27th	<0.001	251	35.3	7.1	5	5
7	Mar 11th - Mar 27th	<0.001	218	30.5	7.2	4	3
8	Mar 13th - Mar 27th	<0.001	3214	2173.1	1.5	273	43
9	Mar 8th - Mar 27th	<0.001	93	4.8	19.1	1	1
10	Mar 25th - Mar 27th	<0.001	323	87.9	3.7	1	1
11	Mar 26th - Mar 27th	<0.001	630	294.0	2.1	3	3
12	Mar 19th - Mar 27th	<0.001	95	11.6	8.2	1	1
13	Mar 23rd - Mar 27th	<0.001	49	3.8	12.8	1	1
14	Mar 25th - Mar 27th	<0.001	100	22.2	4.5	1	1
15	Mar 20th - Mar 27th	<0.001	98	26.1	3.7	1	1
16	Mar 21st - Mar 27th	<0.001	63	14.1	4.5	1	1
17	Mar 26th - Mar 27th	<0.001	294	189.7	1.5	14	11
18	Mar 26th - Mar 27th	<0.001	44	12.5	3.5	8	4
19	Mar 26th - Mar 27th	<0.001	146	79.8	1.8	2	2
20	Mar 26th - Mar 27th	<0.001	175	101.5	1.7	2	2
21	Mar 24th - Mar 27th	<0.001	205	127.2	1.6	4	3
22	Mar 25th - Mar 27th	<0.001	198	125.8	1.5	3	3
23	Mar 23rd - Mar 27th	0.003	18	3.4	5.3	1	1
24	Mar 25th - Mar 27th	0.003	143	86.4	1.6	3	3
25	Mar 26th - Mar 27th	0.004	48	19.1	2.5	8	5
26	Mar 23rd - Mar 27th	0.019	21	5.1	4.1	1	1

Πίνακας 2.5.2. Αναδυόμενες χωρο – χρονικές συστάδες περιόδου 22/01/2020 έως 27/03/2020
Πηγή: *Desjardins M.R. et al.* (2020)



Διάγραμμα 2.5.2. Χωρική κατανομή των αναδυόμενων χωρο – χρονικών συστάδων *Covid* – 19
Πηγή: *Desjardins M.R. et al.* (2020)

Με την προσθήκη 18 επιπλέον ημερών και των αντίστοιχων κρουσμάτων που σημειώθηκαν, παρατηρούμε ότι βρέθηκαν 26 στατιστικά σημαντικές συστάδες που αφορούν περιπτώσεις της νόσου *Covid* – 19. Η πρώτη πιο πιθανή συστάδα (*c*1), αφορά την περίοδο 19 με 27 Μαρτίου, και εντοπίζεται στο βορειοανατολικό κομμάτι του χάρτη, περιλαμβάνονται 14 κομητείες στην Νέα Υόρκη, Κονέκτικατ και Νιου Τζέρσεϋ ενώ η περιοχή του Μανχάταν εμφανίζει το μεγαλύτερο σχετικό κίνδυνο, ο οποίος είναι 96.8. Στο σύνολό της, η πρώτη συστάδα καταγράφει 56,189 της νόσου, με αναμενόμενο αριθμό περιπτώσεων 3343.8 και σχετικό κίνδυνο στην συστάδα 33.1. Η δεύτερη συστάδα (*c*2) αφορά την περίοδο 21 με 27 Μάρτη και περιέχει τρεις κομητείες στο Μίσιγκαν και Γουέιν, όπου καταγράφονται 3036 κρούσματα και σχετικό κίνδυνο 3.7. Ακολούθως, έχουν βρεθεί ακόμη 24 στατιστικά σημαντικές συστάδες.

Στο σημείο αυτό είναι σημαντικό να σημειωθεί πως με την προσθήκη 18 επιπλέον ημερών στην ανάλυση μας, άλλαξαν σημαντικά τα αποτελέσματα που λάβαμε. Παρατηρούμε ότι ο αριθμός των στατιστικά σημαντικών συστάδων έχει αυξηθεί, ενώ παράλληλα οι συστάδες έγιναν πιο πυκνές και μικρότερες σε μέγεθος. Ωστόσο, για την συνέχεια της επιτήρησης, είναι αναγκαίο να ερευνηθεί ο λόγος της αύξησης του αριθμού των συστάδων, καθώς αυτό μπορεί να οφείλεται σε αύξηση των τεστ ανίχνευσης της *Covid* – 19 στις συστάδες που ελέγχθηκαν ενώ αντίστοιχα σε μειωμένη χρήση τεστ σε περιοχές που εμφανίζουν λίγα έως και μηδενικά κρούσματα. Καταληκτικά θα πρέπει να

επισημανθεί πως για τον έγκυρο και έγκαιρο εντοπισμό των χωρικών ή και χρονικών συστάδων που ενδεχομένως να αποτελούν κίνδυνο για την δημόσια υγεία θα πρέπει τα αποτελέσματα των σαρώσεων που πραγματοποιούμε, να αξιολογούνται διαρκώς και η ανάλυση να επαναλαμβάνεται σε τακτά χρονικά διαστήματα με την καταγραφή νέων περιπτώσεων.

ΚΕΦΑΛΑΙΟ 3

Διαγράμματα Ελέγχου

3.1 Εισαγωγή

Ο Στατιστικός Έλεγχος Ποιότητας αποτελεί την παλαιότερη και πιο γνωστή μέθοδο ελέγχου μίας διεργασίας, και στόχο έχει την βελτίωση της ποιότητας των παραγόμενων προϊόντων ή και υπηρεσιών, προκειμένου αυτά να ανταποκρίνονται στα ποιοτικά πρότυπα που έχουν οριστεί (Ταγαράς, 2001). Αποτελείται από ένα σύνολο στατιστικών τεχνικών, το οποίο μπορεί να χωριστεί σε τρία βασικά υποσύνολα:

- Σχεδιασμός και Ανάλυση Πειραμάτων
- Στατιστικός Έλεγχος Διεργασιών
- Δειγματοληψία Αποδοχής

Ειδικότερα, ο Στατιστικός Έλεγχος Διεργασιών (*Statistical Process Control – SPC*), με τον οποίο θα ασχοληθούμε και στο παρόν κεφάλαιο, αποτελεί ένα από τα πιο σημαντικά εργαλεία του Στατιστικού Ελέγχου Ποιότητας. Περιέχει το σύνολο των στατιστικών τεχνικών εκείνων, οι οποίες είναι απαραίτητες για τον έλεγχο της παραγωγικής διεργασίας κατά την διάρκεια παραγωγής των προϊόντων.

Όπως αναφέρθηκε και παραπάνω, ο Στατιστικός Έλεγχος Διεργασιών χρησιμοποιεί ένα σύνολο εργαλείων για την επίβλεψη της σταθερότητας μιας παραγωγικής διεργασίας, με τα 7 κυριότερα από αυτά ή αλλιώς “ *The magnificent seven* ”, να είναι:

- Το Ιστόγραμμα (*Histogram)*
- Διάγραμμα Μίσχου-Φύλλων (*Stem- and -Leaf Plot*)
- Το Φύλλο Ελέγχου (*Check Sheet*)
- Το Διάγραμμα Pareto (*Pareto Chart*)
- Το Διάγραμμα Αιτίας - Αποτελέσματος (*Cause - and - Effect Diagram*)
- Το Διάγραμμα Συγκέντρωσης Ελαττωμάτων (*Defect Concentration Diagram*)

- Το Διάγραμμα Διασποράς ή Διασκόρπισης (*Scatter Plot*)
- Το Διάγραμμα Ελέγχου (*Control Chart*).

Ένα από τα σημαντικότερα πλεονεκτήματα του Στατιστικού Ελέγχου Διεργασιών, το οποίο το κατέστησε ιδιαίτερα χρήσιμο εργαλείο στον κλάδο της Βιομηχανίας και όχι μόνο, αποτελεί το γεγονός πως βασίζεται στον έλεγχο της λειτουργίας της παραγωγής κατά την διάρκεια της παραγωγικής διαδικασίας, και όχι απλά στον έλεγχο του τελικού προϊόντος στο τέλος της παραγωγής, με αποτέλεσμα την έγκαιρη διασφάλιση της ποιότητας του παραγόμενου αγαθού ή υπηρεσίας, πριν από τον τελικό έλεγχο. Έτσι, εισάγει για πρώτη φορά την έννοια της πρόληψης.

Καθώς η πρόληψη αποτελεί τον θεμέλιο λίθο για τον έγκαιρο εντοπισμό και αντιμετώπιση μια πιθανής επικείμενης επιδημίας, τα τελευταία χρόνια ο Στατιστικός Έλεγχος Διεργασιών, έχει μπει στην φαρέτρα των εργαλείων των αρμόδιων φορέων υγείας, στην προσπάθεια τους να εντοπίσουν και να χειριστούν κατάλληλα οτιδήποτε αποτελεί κίνδυνο για την δημόσια υγεία. Η δυνατότητα που δίνεται μέσω του Στατιστικού Ελέγχου Διεργασιών, να προλαμβάνει και να εξασφαλίζει την ποιότητα των παραγόμενων προϊόντων, το καθιστά ένα πολύ σημαντικό εργαλείο στα χέρια όσων ασχολούνται με την ποιότητα της Δημόσιας Υγείας αλλά και με την Βιοεπιτήρηση. Συγκεκριμένα, τα Διαγράμματα Ελέγχου, προσφέρουν την δυνατότητα αξιολόγησης της ποιότητας υγείας, αλλά και της ενεργοποίησης «συναγερμού» όταν αυτή παρουσιάζει αποκλίσεις από τα επιθυμητά όρια που έχουν ορίσει οι υπεύθυνοι υγείας, με τρόπο που θα παρουσιαστεί εκτενώς παρακάτω.

Τα Διαγράμματα Ελέγχου αναπτύχθηκαν για πρώτη φορά από τον φυσικό *Walter. A. Shewhart* στα εργαστήρια της *Bell* το 1924. Αφορμή στάθηκε το γεγονός πως ο *Shewhart* παρατήρησε πως, όσο καλά σχεδιασμένη κι αν είναι μια παραγωγική διαδικασία ή όσο καλά κι αν αυτή συντηρείται, τα προϊόντα που παράγονται δεν είναι πανομοιότυπα το ένα με το άλλο, δηλαδή θα υπάρχει πάντα μια διασπορά στις τιμές των ποιοτικών χαρακτηριστικών των προϊόντων, η οποία οφείλεται σε ένα σύνολο πολλών μικρών και αναπόφευκτων αιτιών, τις οποίες ονόμασε φυσική μεταβλητότητα. Ωστόσο, τόνισε πως μπορούν να υπάρξουν κι άλλες αιτίες μεταβλητότητας, μη – φυσικές, όπως κακοσυντηρημένες μηχανές ή χαμηλής ποιότητας πρώτη ύλη. Η μεταβλητότητα η οποία οφείλεται στους παραπάνω παράγοντες είναι στις περισσότερες περιπτώσεις μεγαλύτερη

της φυσικής, και έχει ως αποτέλεσμα να παράγονται μη αποδεκτά προϊόντα. Έτσι, όταν αυτή εμφανίζεται, είναι απαραίτητο να γίνεται αντιληπτή όσο το δυνατόν πιο γρήγορα και να λαμβάνονται τα κατάλληλα μέτρα για αντιμετώπισή της και την διασφάλιση της ποιότητας των προϊόντων.

Η χρήση των Διαγραμμάτων Ελέγχου στην Βιοεπιτήρηση, δεν διαφέρει καθόλου από την χρήση τους στην Βιομηχανία. Στον παρακάτω πίνακα, παρουσιάζεται ένας παραλληλισμός βασικών εννοιών των Διαγραμμάτων Ελέγχου από την Βιομηχανία στην Δημόσια Υγεία

ΣΤΑΤΙΣΤΙΚΟΣ ΈΛΕΓΧΟΣ ΠΟΙΟΤΗΤΑΣ	**ΕΠΙΔΗΜΙΟΛΟΓΙΑ - ΒΙΟΕΠΙΤΗΡΗΣΗ**
Επιτήρηση Διεργασίας	Επιδημιολογική Επίβλεψη
Φυσική Μεταβλητότητα	Ενδημική Νόσος
Ειδική Μεταβλητότητα	Πιθανή Πανδημική Νόσος
Όρια Ελέγχου	Όρια Λήψης Μέτρων
Εκτός Ορίων Ελέγχου Σημεία	Κατάσταση Επιτήρησης Νοσηρότητας

Πίνακας 3.1.1. Ορολογία των Διαγραμμάτων Ελέγχου

3.2 Βασικές Έννοιες και Δομή Διαγραμμάτων Ελέγχου

Όπως επισημάνθηκε παραπάνω, μια παραγωγική διεργασία, όσο καλά σχεδιασμένη ή συντηρημένη κι αν είναι, περιέχει μια μορφή φυσικής μεταβλητότητας. Αυτή η μορφή μεταβλητότητας αποτελείται από ένα σύνολο πολλών αιτιών, οι οποίες δεν μπορούν να ελεγχθούν, και γι' αυτό τον λόγο αναφέρονται ως κοινές αιτίες μεταβλητότητας. Η φυσική μεταβλητότητα είναι συνήθως μικρή σε μέγεθος, κι έτσι όταν μια παραγωγική διεργασία λειτουργεί μόνο με την παρουσία των κοινών αιτιών μεταβλητότητας, λέμε ότι είναι εντός στατιστικού ελέγχου διεργασία ή διαφορετικά αναφέρεται ως σταθερή διεργασία.

Ωστόσο, πέρα από την φυσική μεταβλητότητα, σε μια διεργασία μπορούν να εμφανιστούν και άλλες μορφές μεταβλητότητας, οι οποίες δεν σχετίζονται με την φυσική. Τέτοιου είδους μεταβλητότητα μπορεί να οφείλεται σε λανθασμένο χειρισμό των μηχανών παραγωγής ή κατώτερης ποιότητας πρώτη ύλη. Σε αντίθεση με την μεταβλητότητα που οφείλεται στις κοινές αιτίες, η μεταβλητότητα αυτή είναι αρκετά μεγαλύτερη σε μέγεθος,

και η παρουσία της οδηγεί σε μη – αποδεκτά παραγόμενα προϊόντα ή υπηρεσίες, με αποτέλεσμα η διεργασία να είναι εκτός στατιστικού ελέγχου και να χαρακτηρίζεται ως μη – σταθερή διεργασία. Η παραπάνω μορφή μεταβλητότητας ονομάζεται ειδική μεταβλητότητα και οι αιτίες που την παράγουν, ειδικές ή συστηματικές αιτίες μεταβλητότητας. Σε αντίθεση με την φυσική μεταβλητότητα, η ύπαρξη της οποίας δεν μπορεί να ελεγχθεί αλλά παράλληλα δεν επηρεάζει την ποιότητα της διεργασίας, η ειδική μεταβλητότητα είναι απαραίτητο να εντοπιστεί και να ληφθούν τα κατάλληλα μέτρα έτσι ώστε αν όχι να εξαλειφθεί τελείως, να μειωθεί σε σημαντικό βαθμό.

Τα παραπάνω, μπορούν εύκολα να εφαρμοσθούν σε τομείς Υγείας. Για παράδειγμα, αν θέλαμε να επιτηρήσουμε τα περιστατικά αναπνευστικής λοίμωξης σε μία κοινότητα, μια μορφή φυσικής μεταβλητότητας η οποία θα επηρέαζε αυξητικά τον αριθμό των λοιμώξεων, αλλά όχι σε ανησυχητικό βαθμό, θα μπορούσε να θεωρηθεί η αλλαγή θερμοκρασίας και καιρού, αιτία που δεν μπορεί να ελεγχθεί από τους αρμόδιους φορείς. Αντιθέτως, μια τεράστια αύξηση των περιστατικών λοιμώξεων του αναπνευστικού, η οποία δεν θα μπορούσε να αποδοθεί σε αιτίες όπως ο καιρός, μπορεί να σημαίνει την ύπαρξη στην κοινότητα ενός παράγοντα κινδύνου (παρουσία νέου ιού - *Covid* – 19), ο οποίος θα πρέπει να εντοπισθεί, να μελετηθεί και να αντιμετωπισθεί όσο το δυνατόν γρηγορότερα.

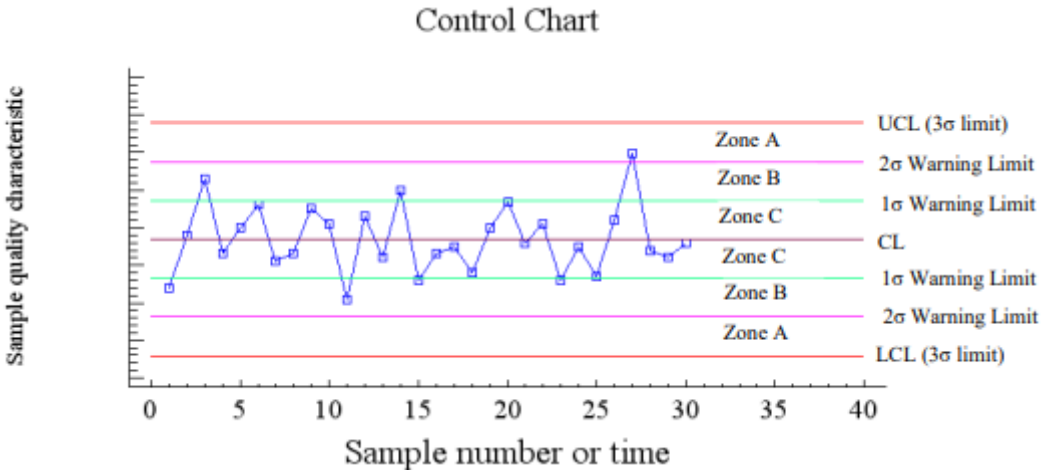
Τα διαγράμματα ελέγχου χαρακτηρίζονται ως ένα από τα βασικότερα εργαλεία του Στατιστικού Ελέγχου Διεργασιών. Αποτελούν την γραφική απεικόνιση των διακυμάνσεων των τιμών της κρίσιμης ποσότητας που μας ενδιαφέρει να μελετήσουμε. Έστω Χ το υπό μελέτη χαρακτηριστικό. Τότε, λαμβάνουμε τυχαία δείγματα x_1, x_2,…, x_n από ένα πληθυσμό σε διαφορετικές χρονικές στιγμές. Επιπλέον, λαμβάνοντας υπόψιν τα τυχαία δείγματα, υπολογίζουμε την τιμή της $g(x_i) = W_i$, $i = 1,2,\ldots,n$, όπου g κατάλληλη στατιστική συνάρτηση, η οποία εκτίμα, στην πλειοψηφία των περιπτώσεων αμερόληπτα, την κρίσιμη ποσότητα που μελετάμε. Έτσι, μέσω της παρακολούθησης των τιμών που παίρνει η στατιστική συνάρτηση για καθένα από τα n δείγματα, μπορούμε εύκολα να μελετήσουμε την διαχρονική συμπεριφορά των τιμών της κρίσιμης ποσότητας.

Στην πιο απλή τους μορφή, ένα διάγραμμα ελέγχου αποτελείται από μια κεντρική γραμμή (*CL – Center Line*), η οποία αντιπροσωπεύει στην πλειοψηφία των περιπτώσεων την αναμενόμενη ή επιθυμητή τιμή της στατιστικής συνάρτησης W. Εκατέρωθεν αυτής, τοποθετούνται τα άνω και κάτω όρια ελέγχου *(UCL – Upper Control Limit / LCL – Lower Control Limit*), τα οποία ορίζουν τα επιθυμητά όρια μέσα στα οποία θέλουμε να λειτουργεί

η διεργασία μας και επέχουν από την κεντρική γραμμή απόσταση L τυπικών αποκλίσεων. Ανάλογα, με την θέση των σημείων στο διάγραμμα, μπορούμε να αποφασίσουμε αν μια διεργασία βρίσκεται εντός ή εκτός στατιστικού ελέγχου. Ειδικότερα, αν όλα τα σημεία βρίσκονται ανάμεσα στα δυο όρια ελέγχου, τότε η διεργασία βρίσκεται εντός ελέγχου, ενώ αν ένα σημείο βρεθεί εκτός αυτών, τότε λαμβάνουμε ένδειξη ότι η διεργασία είναι μη – σταθερή και συνεπώς πέρα από την φυσική μεταβλητότητα, η διεργασία επηρεάζεται και από ειδικές αιτίες μεταβλητότητας. Ωστόσο, το παραπάνω δεν αποτελεί πανάκεια για τον χαρακτηρισμό της σταθερότητας μίας διεργασίας.

Ακόμη και στην περίπτωση που όλα τα σημεία της στατιστικής συνάρτησης βρεθούν εντός των ορίων ελέγχου, θα πρέπει να ελέγξουμε το τρόπο με τον οποίο αυτά συμπεριφέρονται και τοποθετούνται στο διάγραμμα. Αν αυτά συμπεριφέρονται με ένα συστηματικό τρόπο, τότε και πάλι λαμβάνουμε ένδειξη ότι η διεργασία μας βρίσκεται εκτός ελέγχου. Για τον λόγο αυτό, η χρήση των διαγραμμάτων ελέγχου επεκτείνεται και στην μελέτη του τύπου της μεταβλητότητας της διεργασίας. Έτσι, μια διεργασία λέμε ότι έχει στάσιμη συμπεριφορά (*stationary behavior*), όταν τα σημεία του διαγράμματος κινούνται γύρω από μία συγκεκριμένη τιμή με προβλέψιμο τρόπο. Στις στάσιμες διεργασίες, τα δεδομένα που χρησιμοποιούμε μπορούν να είναι είτε ασυσχέτιστα (*uncorrelated*) είτε αυτοσυσχετιζόμενα (*autocorrelated*). Ως ασυσχέτιστα ορίζουμε τα δεδομένα όπου η τιμή που λαμβάνουν την χρονική στιγμή t, δεν επηρεάζει την τιμή που θα λάβουμε την χρονική στιγμή $t + 1$. Αντίθετα, στα αυτοσυσχετιζόμενα δεδομένα, μπορούμε βάσει της συμπεριφοράς των τιμών των παρελθοντικών χρονικών στιγμών να προβλέψουμε τον τρόπο με τον οποίο θα κυμανθούν οι τιμές στο μέλλον. Ωστόσο, στην περίπτωση όπου τα δεδομένα τοποθετούνται στο διάγραμμα χωρίς κάποια λογική, λέμε πως η διεργασία είναι μη – στάσιμη (*nonstationary*).

Ιδιαίτερα στον τομέα της Υγείας, είναι επιτακτική η ανάγκη σήμανσης γρήγορου συναγερμού προκειμένου να ληφθούν έγκαιρα τα κατάλληλα μέτρα, όταν παρουσιάζεται στην κοινότητα ένας ιός. Για τον σκοπό αυτό, τα διαγράμματα ελέγχου μπορούν να γίνουν πιο ευαίσθητα ως προς την ικανότητα τους να εντοπίζουν γρηγορότερα μια διεργασία εκτός ελέγχου, με χρήση, πέρα από του άνω και κάτω ορίου, προειδοποιητικών ορίων (*warning limits*), τα οποία τοποθετούνται εντός αυτών.



Διάγραμμα 3.2.1. Προειδοποιητικά όρια Διαγράμματος Ελέγχου

Παραπάνω δίνεται η γενική μορφή ενός διαγράμματος ελέγχου με χρήση άνω και κάτω ορίων ελέγχου ($UCL - LCL$), τα οποία απέχουν από την κεντρική γραμμή (CL) απόσταση 3 τυπικών αποκλίσεων, καθώς και χρήση προειδοποιητικών ορίων ($1\sigma - 2\sigma$ *Warning Limits*). Προκειμένου να αποφασίσουμε αν η διεργασία είναι εντός ή εκτός στατιστικού ελέγχου, έχει αναπτυχθεί ένα σύνολο κανόνων που μας βοηθούν. Παρακάτω, δίνονται οι δέκα πιο σημαντικοί κανόνες, οι οποίοι όταν ενεργοποιηθούν, λαμβάνουμε ένδειξη εκτός ελέγχου διεργασία:

- Ένα ή περισσότερα σημεία εκτός των ορίων ελέγχου
- Δύο από τρία συνεχόμενα σημεία στην Ζώνη Α (σε μια από τις δύο ζώνες Α)
- Τέσσερα από πέντε συνεχόμενα σημεία πέραν της Ζώνης C (σε μια από τις δύο περιοχές)
- Οκτώ συνεχόμενα σημεία στην ίδια μεριά (επάνω ή κάτω) της κεντρικής γραμμής
- Έξι συνεχόμενα σημεία σε αύξουσα ή φθίνουσα διάταξη
- Δεκαπέντε συνεχόμενα σημεία στην ολική ζώνη C
- Δεκατέσσερα συνεχόμενα σημεία σε εναλλασσόμενη μορφή “πάνω-κάτω”
- Οκτώ συνεχόμενα σημεία εκτός της ολικής Ζώνης C
- Οποιαδήποτε ασυνήθιστη ή μη τυχαία ακολουθία σημείων
- Ένα ή περισσότερα σημεία κοντά στα προειδοποιητικά όρια ή τα όρια ελέγχου.

Ωστόσο, θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην ερμηνεία των αποτελεσμάτων όταν γίνεται συνδυαστική χρήση των παραπάνω κανόνων, καθώς υπάρχει κίνδυνος να λάβουμε λανθασμένους συναγερμούς, λόγω της αυξημένης ευαισθησίας του διαγράμματος. Επιπλέον, μεγάλη προσοχή θα πρέπει να δίνεται στην απόσταση που θα έχουν τα όρια ελέγχου από την κεντρική γραμμή, καθώς αν επιλέξουμε μικρό *L*, προκύπτουν πολύ στενά όρια ελέγχου, με αποτέλεσμα να οδηγούμαστε σε αυξημένη πιθανότητα σφάλματος τύπου *I*, δηλαδή το διάγραμμα θα δίνει ένδειξη ότι η διαδικασία είναι εκτός στατιστικού ελέγχου μέσω λανθασμένων θετικών συναγερμών, ενώ στην πραγματικότητα κάτι τέτοιο δεν θα ισχύει. Σε αντίθετη περίπτωση, επιλέγοντας μεγάλο *L*, τότε δημιουργούμε πολύ πλατιά όρια ελέγχου, με αποτέλεσμα να αυξάνουμε την πιθανότητα σφάλματος τύπου *II*, δηλαδή το διάγραμμα δεν θα μας δίνει ένδειξη ότι η διαδικασία βρίσκεται εκτός στατιστικού ελέγχου.

Στο σημείο αυτό, αξίζει να αναφέρουμε τις φάσεις στις οποίες χωρίζεται ο έλεγχος μιας παραγωγικής διεργασίας, την φάση *I* (*Phase I*) και φάση *II* (*Phase II)*. Η φάση *I* ή αλλιώς αναδρομική φάση, ως στόχο έχει την διαμόρφωση των κατάλληλων ορίων ελέγχου. Συγκεκριμένα, ο χρήστης εξετάζοντας δείγματα που έχουν ληφθεί στο παρελθόν, πρέπει να ορίσει αν η διεργασία ήταν ή όχι εντός στατιστικού ελέγχου. Στην περίπτωση που η διεργασία ήταν μη – σταθερή, τότε με κατάλληλους χειρισμούς θα πρέπει να την φέρει εντός ελέγχου, και τα όρια που θα προκύψουν (κεντρική γραμμή / άνω και κάτω όριο) θα χρησιμοποιούνται πλέον ως όρια για μελλοντική παρακολούθηση της διεργασίας. Εν συνεχεία, αξιοποιώντας τα όρια που έχουμε λάβει στην φάση *I*, περνάμε στην φάση *II*, όπου λαμβάνοντας πλέον συνεχώς νέα δεδομένα από την παραγωγική διαδικασία, ο χρήστης θα πρέπει να κρίνει αν η διαδικασία εξακολουθεί να παραμένει εντός στατιστικού ελέγχου.

Τα κυριότερα Διαγράμματα Ελέγχου χωρίζονται σε 3 βασικές κατηγορίες. Αρχικά, έχουμε τα Διαγράμματα Ελέγχου τύπου *Shewhart*, η χρήση των οποίων γίνεται όταν παρατηρείται μεγάλη αλλαγή στο μέσο επίπεδο της διεργασίας και παράλληλα το γνωρίζουμε την κατανομή των αρχικών δεδομένων. Στην συνέχεια, υπάρχουν τα Διαγράμματα Ελέγχου τύπου *CUSUM*, όπου η χρήση τους συνίσταται όταν θέλουμε να ανιχνεύσουμε μικρές μετατοπίσεις του μέσου επιπέδου της διεργασίας και η κατανομή των αρχικών δεδομένων είναι γνωστή. Τέλος, έχουμε τα Διαγράμματα Ελέγχου τύπου *EWMA*, όπου και αυτά χρησιμοποιούνται όταν θέλουμε να ανιχνεύσουμε μικρές μετατοπίσεις του

μέσου επιπέδου της διεργασίας, ενώ στην περίπτωση αυτή η κατανομή των αρχικών δεδομένων είναι άγνωστη. Στην βιβλιογραφία, προτείνεται ότι τα καταλληλότερα διαγράμματα για την φάση *I* είναι τα διαγράμματα Shewhart, ενώ τα διαγράμματα τύπου *CUSUM* και *EWMA* προτιμώνται για την παρακολούθηση της διαδικασίας στην φάση *II*, λόγω της ικανότητας τους να ανιχνεύουν καλύτερα τις μικρές μετατοπίσεις της διεργασίας (*Mohammed et al., 2007*).

Τέλος, ένας επιπλέον διαχωρισμός των διαγραμμάτων ελέγχου γίνεται με βάση το είδος της υπό μελέτη μεταβλητής ή χαρακτηριστικού που μας ενδιαφέρει. Όταν το χαρακτηριστικό που μας ενδιαφέρει να μελετήσουμε μπορεί να μετρηθεί βάσει μιας συνεχούς αριθμητικής κλίμακας (συστολική πίεση), τότε έχουμε τα διαγράμματα ελέγχου για μεταβλητές (*control charts for variables*), ενώ, όταν το υπό μελέτη χαρακτηριστικό αναφέρεται σε κάποια ιδιότητα (ελλαττωματικό ή όχι προϊόν / θετικός για *Covid* – 19 ή όχι) έχουμε τα διαγράμματα ελέγχου για ιδιότητες (*control charts for attributes*). Έτσι, με βάση τον παραπάνω διαχωρισμό των διαγραμμάτων ελέγχου, τα διαγράμματα για μεταβλητές είναι:

- $\bar{x}$ και *R* (Διάγραμμα Ελέγχου για την Μέση τιμή και Εύρος)
- $\tilde{x}$ και *R* (Διάγραμμα Ελέγχου για την Διάμεσο και Εύρος)
- $\bar{x}$ και *S* (Διάγραμμα Ελέγχου για την Μέση τιμή και Τυπική Απόκλιση)
- *EWMA* (Διάγραμμα Ελέγχου για τον Μέσο μιας Διεργασίας)
- *CUSUM* (Αθροιστικό Διάγραμμα Ελέγχου)

ενώ τα διαγράμματα ελέγχου για ιδιότητες είναι:

- *u* Διάγραμμα Ελέγχου (Διάγραμμα Ελέγχου για το μέσο αριθμό των ελαττωμάτων)
- *c* Διάγραμμα Ελέγχου (Διάγραμμα Ελέγχου για το μέσο αριθμό των ελαττωμάτων)
- *np* Διάγραμμα Ελέγχου (Διάγραμμα Ελέγχου για τον αριθμό των ελαττωματικών προϊόντων)
- *p* Διάγραμμα Ελέγχου (Διάγραμμα Ελέγχου για το Ποσοστό των ελαττωματικών προϊόντων).

3.3 Επιλογή κατάλληλου Διαγράμματος Ελέγχου στην Βιοεπιτήρηση

Τα διαγράμματα ελέγχου όπως αναφέραμε παραπάνω, αν και χρησιμοποιήθηκαν αρχικά για την βελτίωση της ποιότητας των προϊόντων που παράγονται σε διάφορους τομείς της Βιομηχανίας, γρήγορα λόγω της εύκολης εφαρμογής τους αλλά και της δυνατότητας που προσέφεραν για ενεργοποίηση έγκαιρων συναγερμών, εφαρμόσθηκαν και στον τομέα της Υγείας, τόσο σε ατομικό επίπεδο ασθενούς, όσο και σε επίπεδο οργανισμού Υγείας αλλά και σε επίπεδο κοινότητας. Ωστόσο, η βασική διαφορά ανάμεσα στην χρήση των διαγραμμάτων στην Βιομηχανία και στην Υγεία, έγκειται στο γεγονός πως σε αντίθεση με τις μεταβλητές που συναντώνται στην βιομηχανία, οι μεταβλητές που μας ενδιαφέρουν στην υγεία, είναι συνήθως είτε ποσοτικές που δεν ακολουθούν την κανονική κατανομή, είτε ποιοτικές (Μπερσίμης, 2019). Για παράδειγμα, ο αριθμός των λοιμώξεων του αναπνευστικού σε μία κοινότητα ακολουθεί κατανομή *Poisson*. Συνεπώς, με βάση την κατανομή της πιθανότητας που ακολουθούν τα δεδομένα που μελετώνται, χρησιμοποιούμε τα διαγράμματα ελέγχου με τον ακόλουθο τρόπο. Τα διαγράμματα ελέγχου *p* και *np* χρησιμοποιούνται όταν τα δεδομένα ακολουθούν διωνυμική κατανομή. Στην περίπτωση δεδομένων που προέρχονται από την κατανομή *Poisson* χρησιμοποιούμε τα διαγράμματα ελέγχου *c* και *u*, ενώ όταν τα δεδομένα ακολουθούν την κανονική κατανομή, τότε τα καταλληλότερα διαγράμματα είναι τα διαγράμματα ελέγχου για την Μέση τιμή και το Εύρος ή την Τυπική Απόκλιση. Παρακάτω δίνεται ένας συνοπτικός πίνακας, με την κατάλληλη επιλογή διαγράμματος ανά περίπτωση μελέτης:

Τύπος Διαγράμματος	**Κατανομή Πιθανότητας**	**Παράδειγμα**
$\bar{X}$ και S	Κανονική	Διάρκεια Επέμβασης, Συγκέντρωση Ζαχάρου στο Αίμα
np	Διωνυμική	Αριθμός Κρουσμάτων Λοιμώδης Νόσου
p	Διωνυμική	Αριθμός Ασθενών που Παρουσιάζουν επιπλοκή από Χορήγηση Θεραπείας
c	Poisson	Αριθμός Επιπλοκών Ασθενών
u	Poisson	Ρυθμός Θανάτου Ασθενών μετά από Χειρουργική Επέμβαση

Πίνακας 3.4.1. Τύποι Διαγραμμάτων για χρήση στην Βιοεπιτήρηση

3.4 Διαγράμματα Ελέγχου Τύπου *Shewhart* για ιδιότητες

Τα διαγράμματα ελέγχου τύπου *Shewhart* για ιδιότητες βασίζονται στην έννοια του ελαττωματικού προϊόντος. Ως ελαττωματικό χαρακτηρίζεται ένα προϊόν το οποίο δεν ανταποκρίνεται στις τεχνικές προδιαγραφές που έχουν οριστεί από τον χρήστη. Μέσω αυτού του τύπου τα διαγράμματα, στοχεύουμε στον έγκαιρο εντοπισμό αυτών των προϊόντων, την απομάκρυνσή τους, αλλά και την διεξαγωγή κατάλληλης έρευνας προκειμένου να βρεθούν καθώς και να αντιμετωπιστούν οι αιτίες ειδικής μεταβλητότητας που οδήγησαν την διεργασία εκτός στατιστικού ελέγχου.

Ωστόσο, μιας και στην Βιοεπιτήρηση, δεν μπορεί να γίνεται λόγος για ελαττωματικά προϊόντα, τα διαγράμματα ελέγχου στην περίπτωση αυτή, έχουν σαν στόχο τον έλεγχο του αριθμού αλλά και του ποσοστού εμφανίσεως του γεγονότος που μας ενδιαφέρει. Παραδείγματος χάριν, σε μία νοσοκομειακή μονάδα μας ενδιαφέρει να μελετήσουμε τον αριθμό των ενδονοσοκομειακών λοιμώξεων. Σε επίπεδο κοινότητας, μας ενδιαφέρει να μελετήσουμε τον αριθμό ή και ποσοστό κρουσμάτων ηπατίτιδας σε παιδιά.

3.4.1 *p* Διάγραμμα Ελέγχου

Στην περίπτωση που μας ενδιαφέρει να μελετήσουμε το ποσοστό των περιπτώσεων ενός γεγονότος που αφορά την Δημόσια Υγεία, όπως το ποσοστό εμφάνισης λοιμώξεων του αναπνευστικού ή κρουσμάτων ευλογιάς των πιθήκων στην Ελλάδα, χρησιμοποιούμε το *p* Διάγραμμα Ελέγχου. Έστω, ότι σε μία κοινότητα είναι γνωστό ότι το ποσοστό περιστατικών της υπό μελέτης νόσου είναι ίσο με *p*. Από την κοινότητα, επιλέγονται με τυχαίο τρόπο *m* ανεξάρτητα δείγματα, το καθένα από αυτά αποτελούμενο από *n* άτομα. Για καθένα από αυτά τα άτομα, καταγράφουμε αν νοσεί ή όχι. Ορίζουμε έτσι την τυχαία μεταβλητή X_{ij}, για $i = 1,2,\ldots,m$ και $j = 1,2,\ldots,n$, η οποία λαμβάνει την τιμή 0 αν το j άτομο του i δεν νοσεί, ενώ αν νοσεί, λαμβάνει την τιμή 1. Επιπλέον, γνωρίζουμε ότι το κάθε άτομο έχει p πιθανότητα να νοσήσει, δηλαδή $X_{ij} \sim B(1,p)$. Έτσι, ορίζουμε την τυχαία μεταβλητή $X_i = X_{i1} + X_{i2} + \ldots + X_{in}$, η οποία αντιπροσωπεύει τον αριθμό των περιστατικών νόσησης στο δείγμα i και $X_i \sim B(n,p)$.

Με βάση τα παραπάνω, ορίζουμε την τυχαία μεταβλητή:

$$W_i = p_i = \frac{x_i}{n}, \text{ με } 1 \leq i \leq m,$$

η οποία αντιπροσωπεύει το ποσοστό των ασθενών στο δείγμα i, και για την οποία ισχύει:

$$\mu_{w_i} = p \text{ και } \sigma^2_{w_i} = \frac{p(1-p)}{n}.$$

Στην περίπτωση που το ποσοστό νόσησης δεν είναι γνωστό, τότε πρέπει να το εκτιμήσουμε. Θεωρούμε ότι έχουμε m ανεξάρτητα προκαταρτικά τυχαία δείγματα, μεγέθους n το κάθε ένα, δηλαδή έχουμε $X_i = (X_{i1}, X_{i2}, \ldots, X_{in})$ για $1 \leq i \leq m$. Επιπλέον, ορίζουμε το ποσοστό:

$$p_i = \frac{X_i}{n} = \frac{X_1+X_2+\ldots+X_n}{n} \text{ , για } 1 \leq i \leq m,$$

καθώς και

$$\hat{p} = \frac{p_1+p_2+\ldots+p_m}{m} = \frac{X_1+X_2+\ldots+X_m}{mn} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{mn}.$$

Ο $\hat{p}$ είναι αμερόληπτός εκτιμητής του p, καθώς $E(\hat{p}) = p$, αφού $\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij} \sim B(mn,p)$. Για τον λόγο αυτό, ο $\hat{p}$ χρησιμοποιείται ως εκτιμητής του ποσοστού p.

Λαμβάνοντας τα παραπάνω υπόψιν, προκύπτουν τα παρακάτω L όρια ελέγχου για την παρακολούθηση του ποσοστού περιστατικών νόσησης, τόσο για την Φάση I, όπου χρησιμοποιούμε τον εκτιμητή $\hat{p}$, όσο και για την Φάση II:

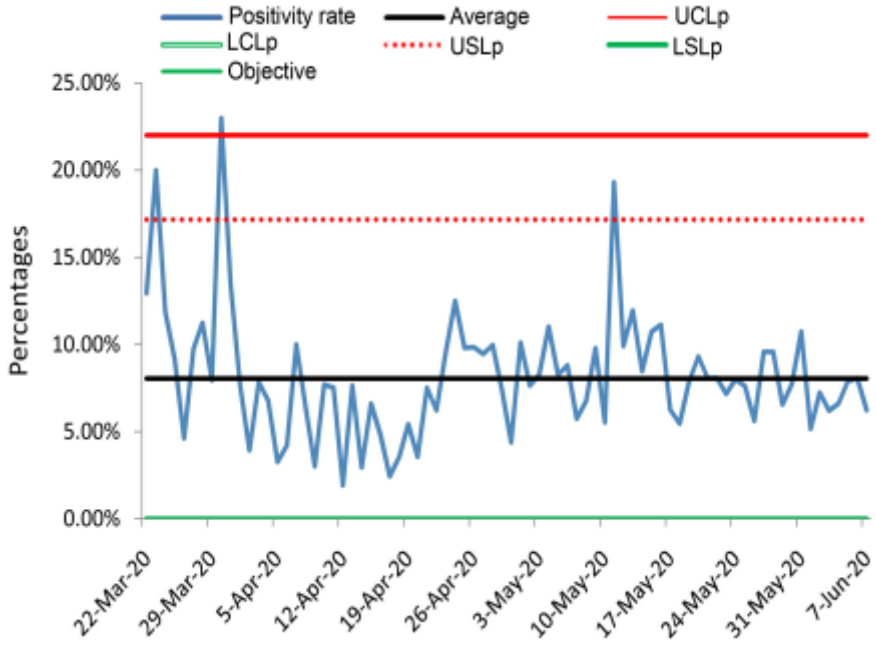
***p* Διάγραμμα Ελέγχου**	***p* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *I	***L* Όρια Ελέγχου Φάσης *II***
$UCL = \hat{p} + L\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $CL = \hat{p}$ $LCL = \hat{p} - L\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$UCL = p + L\sqrt{\frac{p(1-p)}{n}}$ $CL = p$ $LCL = p - L\sqrt{\frac{p(1-p)}{n}}$

Πίνακας 3.4.1.1. *L* Όρια ελέγχου *p* - διαγράμματος

Πολλές φορές δεν μπορούμε να διατηρήσουμε το μέγεθος των δειγμάτων μας σταθερό. Στις περιπτώσεις όπου έχουμε m προκαταρκτικά ανεξάρτητα δείγματα μεγέθους n_i το καθένα, δηλαδή $X_i = (X_{i1}, X_{i2,\dots}, X_{in})$ και $Xi \sim B(n_i,p)$, ο εκτιμητής του ποσοστού δίνεται από την σχέση

$$\hat{p} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{\sum_{i=1}^{m} n_i}.$$

Η παρακάτω μελέτη πραγματοποιήθηκε από τους *Mbaye, Sgar. and Ngom* (2021), οι οποίοι κατέγραψαν την ημερήσια θετικότητας κρουσμάτων *Covid* – 19, για την περίοδο 22/03/2020 έως 07/06/2020. Ως δείκτη θετικότητας ορίζεται ο αριθμός των θετικών τεστ της ημέρας προς τον συνολικό αριθμό τεστ που πραγματοποιήθηκαν. Καθώς, ενδιαφερόμαστε για την παρακολούθηση του ποσοστού των θετικών κρουσμάτων, το p – διάγραμμα θεωρήθηκε καταλληλότερο. Όπως φαίνεται στο παρακάτω διάγραμμα, η κεντρική τιμή του διαγράμματος βρίσκεται στο 8%, ενώ το άνω όριο ελέγχου στο 22%. Ωστόσο το διάγραμμα μας παρέχει ένδειξη ότι η διεργασία βρίσκεται εκτός ελέγχου το διάστημα ανάμεσα στις 29/03/2020 έως τις 05/04/2020, όπου για πρώτη φορά βρίσκεται σημείο εκτός του άνω ορίου ελέγχου. Αυτό, αποτελεί ένδειξη ύπαρξης ειδικής αιτίας μεταβλητότητας (μεγάλη διασπορά του ιού στην κοινότητα), η οποία θα πρέπει να ερευνηθεί και να αντιμετωπιστεί κατάλληλα από τους αρμόδιους φορείς υγείας.



Διάγραμμα 3.4.1.1. *p* – Διάγραμμα ελέγχου θετικών κρουσμάτων *Covid* - 19

Πηγή*: Mbaye et al. (2021)*

Ωστόσο, ιδιαίτερα όσον αφορά ιατρικές μετρήσεις και μελέτες, μπορεί να βρεθούμε αντιμέτωποι με δυο περιπτώσεις. Είτε να έχουμε στην διάθεσή μας υποομάδες με μεγάλο αριθμό συμμετεχόντων, με αποτέλεσμα τα δεδομένα μας να παρουσιάζουν μεγάλη διασπορά, είτε σε αντίθετη περίπτωση, οι ομάδες που συμμετέχουν στην επιτήρηση να αποτελούνται από πολύ μικρό αριθμό ατόμων, όπως για παράδειγμα στην μελέτη σπάνιων συνδρόμων (οζώδη σκλήρυνση), με αποτέλεσμα τα δεδομένα μας να παρουσιάζουν πολύ μικρή διασπορά. Και στις δυο περιπτώσεις, θα λαμβάναμε ένα *p* διάγραμμα ελέγχου, αποτελούμενο είτε από πολύ στενά είτε αρκετά πλατιά όρια ελέγχου. Τα παραπάνω θα είχαν ως αποτέλεσμα πολλούς εσφαλμένους θετικούς ή αρνητικούς συναγερμούς.

Η λύση στο πρόβλημα της διασποράς των δεδομένων δίνεται με την χρήση του *Laney p* διαγράμματος ελέγχου. Στην βασική του μορφή δεν διαφέρει από το κλασικό *p* διάγραμμα καθώς και τα δυο καταγράφουν τον αριθμό των ελαττωματικών προϊόντων, ή στην περίπτωση μας, ασθενών. Η διαφορά τους έγκειται στην χρήση διαφορετικών ορίων ελέγχου.

Με βάση την συγκεκριμένη προσέγγιση, γίνεται υπολογισμός των z – *scores* για κάθε ένα από τα δείγματά μας, δηλαδή:

$$z_i = \frac{p_i - \bar{p}}{\sigma_{p_i}}$$

και

$$\sigma_{pi} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}$$

Στην συνέχεια υπολογίζεται το :

$$\sigma_z = \frac{\bar{R}}{k} \text{ όπου } R_i = |z_i - z_{i-1}|$$

όπου οι τιμές του k εξαρτώνται από το $\bar{R}$.

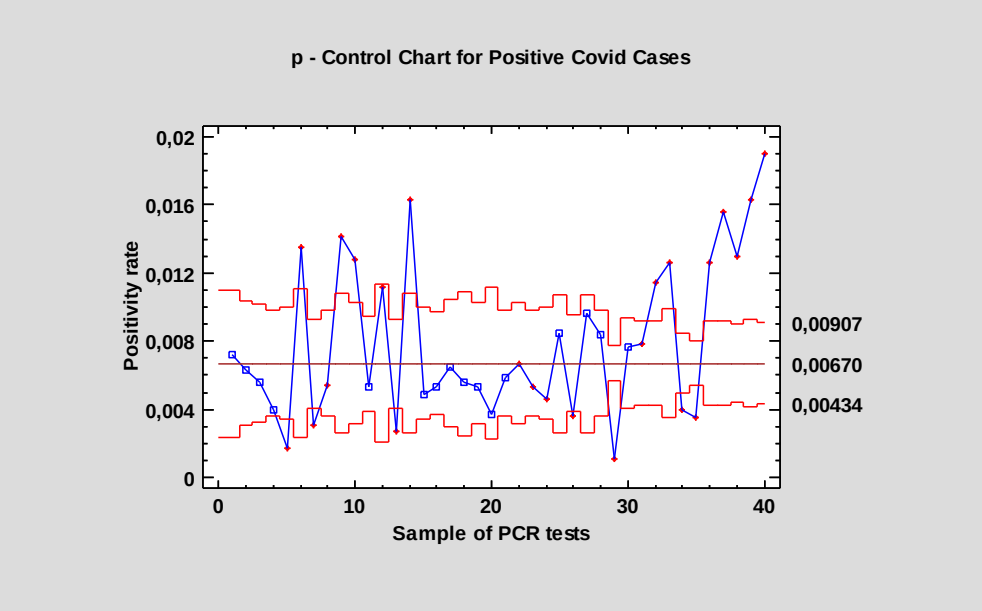
Έτσι, τα όρια που διαμορφώνονται είναι τα ακόλουθα:

Laney p΄Διάγραμμα Ελέγχου
***L* Όρια Ελέγχου**
UCL = $\bar{p}$ + L$\sigma_{p_i}\sigma_z$
CL = $\bar{p}$
LCL = $\bar{p}$ - L$\sigma_{p_i}\sigma_z$

Πίνακας 3.4.1.2. *L* Όρια ελέγχου *Laney p΄* - διαγράμματος

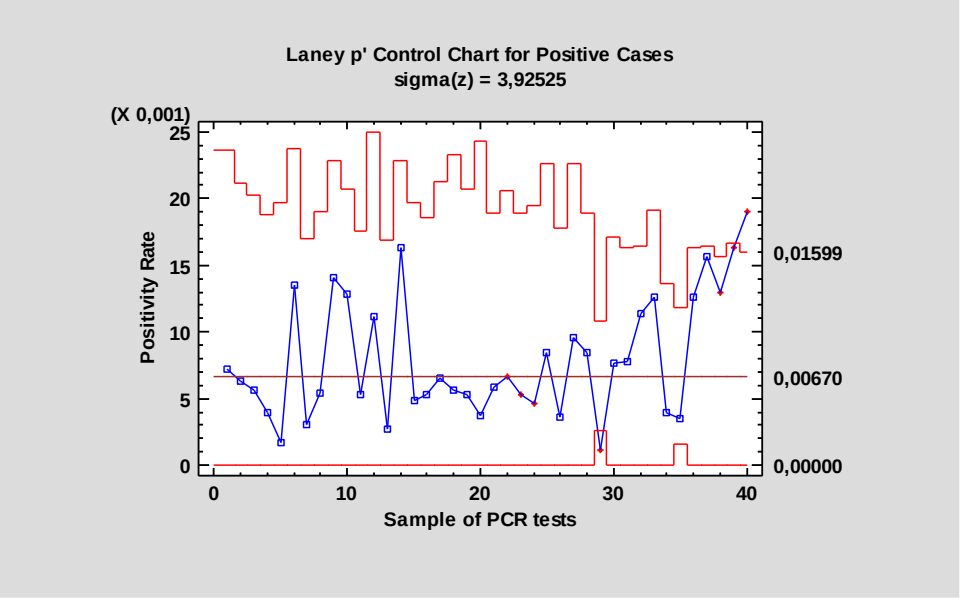
Παρακάτω πραγματοποιήσαμε μελέτη για την περίοδο 01/07/2020 έως 09/08/2020. Καταγράφηκαν τα ημερήσια επιβεβαιωμένα κρούσματα *Covid* - 19 όπως αυτά ανακοινώθηκαν από την επίσημη σελίδα του ΕΟΔΥ (*https://eody.gov.gr/epidimiologika-statistika-dedomena/ektheseis-epidimiologikis-epitirisis-loimoxis-apo-ton-sars-cov-2/imerisies-ektheseis-covid-19-2020/*), αλλά και το πλήθος των εργαστηριακών τεστ (*PCR*) που πραγματοποιήθηκαν κατά την παραπάνω περίοδο. Καθώς την περίοδο εκείνη το πλήθος των θετικών τεστ ήταν σημαντικά μικρότερο από το σύνολο των εργαστηριακών

ελέγχων που είχαν διεξαχθεί, αν επιλέγαμε την χρήση ενός p – διαγράμματος ελέγχου, θα λαμβάναμε το παρακάτω διάγραμμα:



Διάγραμμα 3.4.1.2. *p* – Διάγραμμα ελέγχου θετικών κρουσμάτων *Covid - 19*

Από το παραπάνω διάγραμμα, γίνεται αντιληπτό πως λαμβάνουμε ένδειξη εκτός ελέγχου διεργασίας, καθώς παρατηρούμε πληθώρα σημείων να βρίσκονται εκτός των ορίων ελέγχου. Ωστόσο, κάτι τέτοιο μπορεί να μην αντιπροσωπεύει την πραγματικότητα, και να οφείλεται στο σχετικά μικρό αριθμό θετικών τεστ σε σχέση με τον υψηλό αριθμό τεστ που έχουν διενεργηθεί. Η παραπάνω εικόνα φαίνεται να αλλάζει όταν γίνεται χρήση του *Laney p΄* διαγράμματος ελέγχου.



Διάγραμμα 3.4.1.3. *Laney p΄* – Διάγραμμα ελέγχου θετικών κρουσμάτων *Covid - 19*

Από το παραπάνω *Laney p΄* διαγράμματος ελέγχου, λαμβάνουμε για πρώτη φορά ένδειξη ότι η διεργασία μας ξεπέρασε για πρώτη φορά το άνω όριο ελέγχου στις 09/08/2020. Η χρήση αυτού του τύπου διαγράμματός μας προφυλάσσει από το πρόβλημα της υπερδιασποράς και της σήμανσης λανθασμένων συναγερμών, που ενδεχομένως να οδηγούσαν στην λήψη μη απαραίτητων υγειονομικών μέτρων και περιορισμών.

3.4.2 *np* Διάγραμμα Ελέγχου

Στην περίπτωση που μας ενδιαφέρει να μελετήσουμε τον αριθμό των περιπτώσεων ενός γεγονότος που αφορά την Δημόσια Υγεία, χρησιμοποιούμε το *np* Διάγραμμα Ελέγχου. Η διαδικασία κατασκευής που διαγράμματος, είναι η ίδια με αυτή που περιεγράφηκε παραπάνω, με την διαφορά ότι η κρίσιμη ποσότητα που απεικονίζεται στο διάγραμμα είναι η $W_i = X_i$, και για την οποία ισχύει ότι:

$$\mu_{w_i} = np \text{ και } \sigma^2_{w_i} = np(1-p).$$

Έτσι, τα L όρια ελέγχου για την παρακολούθηση του αριθμού των περιστατικών που μελετάμε, για την Φάση *I* και Φάση *II* είναι τα ακόλουθα:

***np* Διάγραμμα Ελέγχου**	***np* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *I	***L* Όρια Ελέγχου Φάσης *II***
UCL = $n\hat{P}$ + L$\sqrt{n\hat{P}(1-\hat{P})}$ CL = $n\hat{P}$ LCL = $n\hat{P}$ - L$\sqrt{n\hat{P}(1-\hat{P})}$	UCL = np + L$\sqrt{np(1-p)}$ CL = np LCL = np - L$\sqrt{np(1-p)}$

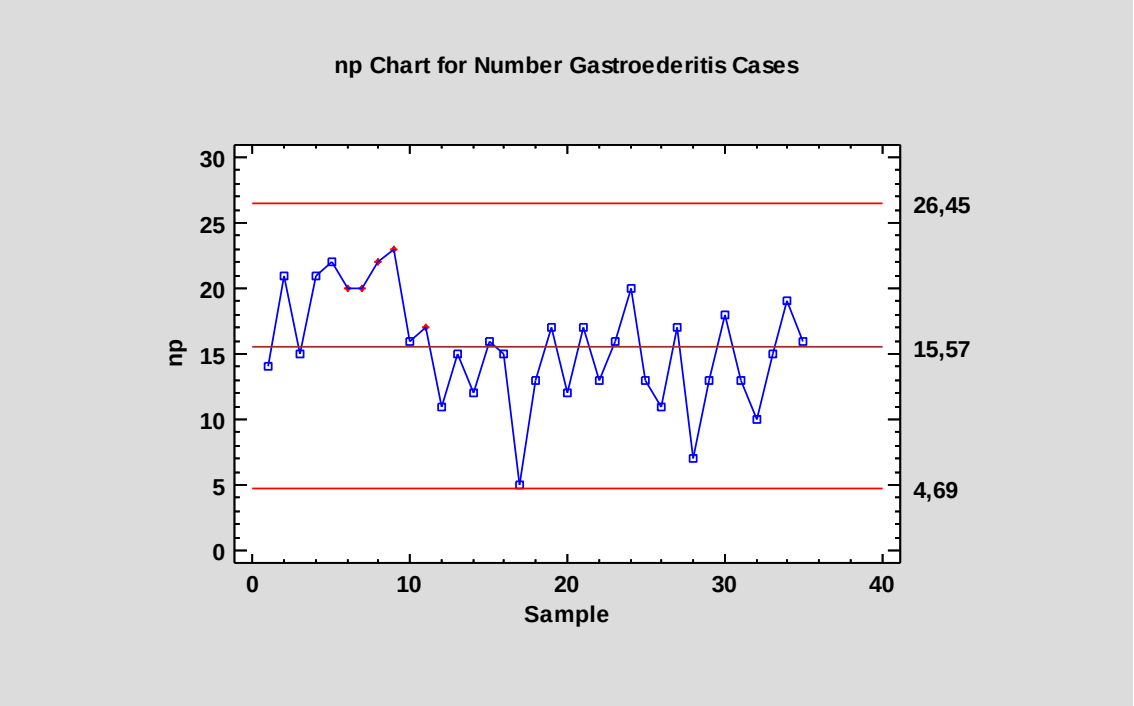
Πίνακας 3.4.2.1. *L* Όρια ελέγχου *np* - διαγράμματος

Ας υποθέσουμε ότι θέλουμε να παρακολουθήσουμε τα κρούσματα γαστρεντερίτιδας. Σε ένα νοσοκομείο, εξετάζεται δείγμα 100 ασθενών για 35 ημέρες, και καταγράφονται τα επιβεβαιωμένα κρούσματα γαστρεντερίτιδας και τα οποία θα χρησιμοποιηθούν για ανάλυση φάσης *I*, όπως φαίνονται στον παρακάτω πίνακα:

Αριθμός Δείγματος	Μέγεθος Δείγματος	Κρούσματα Γαστρεντερίτιδας
1	100	14
2	100	21
3	100	15
4	100	21
5	100	22
6	100	20
7	100	20
8	100	22
9	100	23
10	100	16
11	100	17
12	100	11
13	100	15
14	100	12
15	100	16
16	100	15
17	100	5
18	100	13
19	100	17
20	100	12
21	100	17
22	100	13
23	100	16
24	100	20
25	100	13
26	100	11

27	100	17
28	100	7
29	100	13
30	100	18
31	100	13
32	100	10
33	100	15
34	100	19
35	100	16

Πίνακας 3.4.2.2. Καταγεγραμμένα περιστατικά Γαστρεντερίτιδας



Διάγραμμα 3.4.2.1. *np* – Διάγραμμα ελέγχου κρουσμάτων γαστρεντερίτιδας Φάσης *I*

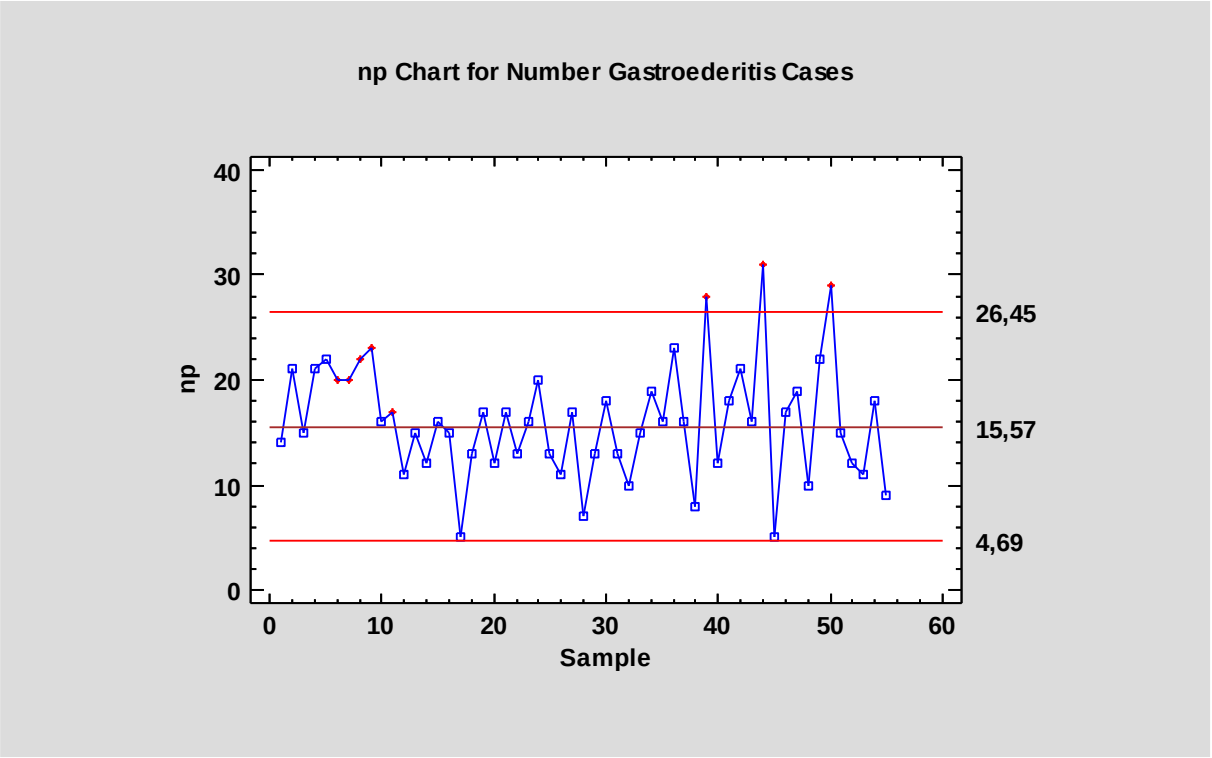
Όπως φαίνεται από το παραπάνω διάγραμμα, κανένα σημείο δεν βρίσκεται εκτός των ορίων. Ο αναμενόμενος αριθμός κρουσμάτων γαστρεντερίτιδας είναι 16, ενώ τα 3σ - όρια ελέγχου που θα χρησιμοποιηθούν πλέον για μελλοντική παρακολούθηση φάσης *II* θα είναι από 4.69 έως 26.45.

Στην συνέχεια, καταγράφηκαν τα κρούσματα για τις επόμενες 20 ημέρες, όπως φαίνεται στον παρακάτω πίνακα, για την πραγματοποίηση ανάλυσης φάσης *II*:

Αριθμός Δείγματος	Μέγεθος Δείγματος	Κρούσματα Γαστρεντερίτιδας
36	100	23
37	100	16
38	100	8
39	100	28
40	100	12
41	100	18
42	100	21
43	100	16
44	100	31
45	100	5
46	100	17
47	100	19
48	100	10
49	100	22
50	100	29
51	100	15
52	100	12
53	100	11
54	100	18
55	100	9

Πίνακας 3.4.2.3. Καταγεγραμμένα περιστατικά Γαστρεντερίτιδας

Όπως φαίνεται από το παρακάτω διάγραμμα, 3 σημεία βρίσκονται εκτός των ορίων ελέγχου, όπως αυτά ορίστηκαν κατά την ανάλυση φάσης *I*. Παρατηρείται επίσης αύξηση των κρουσμάτων, καθώς τα περισσότερα από τα είκοσι καινούργια δείγματα που προστέθηκαν βρίσκονται πάνω από την κεντρική γραμμή του διαγράμματος. Συνεπώς το νοσοκομείο θα πρέπει να τεθεί σε επαγρύπνηση για πιθανή έξαρση γαστρεντερίτιδας στην κοινότητα.



Διάγραμμα 3.4.2.2. *np* – Διάγραμμα ελέγχου κρουσμάτων γαστρεντερίτιδας Φάσης *II*

3.4.3 *c* Διάγραμμα Ελέγχου

Στα διαγράμματα ελέγχου *p* και *np* ήταν φανερό ότι τα δεδομένα μας προέρχονταν από την διωνυμική κατανομή, όπου η μεταβλητή μας ήταν δίτιμη, με τους συμμετέχοντες στις ομάδες να έχουν ή όχι την υπό μελέτη νόσο. Ωστόσο, για τις περιπτώσεις εκείνες που μας ενδιαφέρει να μελετήσουμε τον αριθμό των μολύνσεων ή ιών που μπορεί να εμφανίσει ένας ασθενής, δεν μπορούμε να προχωρήσουμε σε έλεγχο μέσω χρήσης των παραπάνω διαγραμμάτων. Έτσι, όταν τα υπό μελέτη περιστατικά καταγράφονται με ένα γνωστό και σταθερό ρυθμό, και η εμφάνιση του προηγούμενου δεν επηρεάζει την εμφάνιση του επόμενου, δηλαδή προέρχονται από την κατανομή *Poisson*, τότε προχωράμε με σε έλεγχο με χρήση των *c* διαγραμμάτων ελέγχου.

Ας υποθέσουμε ότι X είναι ο αριθμός που εκφράζει το πλήθος των ενδονοσοκομειακών λοιμώξεων από τις οποίες μολύνεται ένας ασθενής εντός της ΜΕΘ, ο οποίος ακολουθεί κατανομή *Poisson* με παράμετρο *c*, δηλαδή $X \sim P(c)$ και

$$P(X = x) = e^{-c}\frac{c^x}{x!}\text{, x = 0,1,…}$$

Τότε θα ισχύει:

$$\mu_X = c \text{ και } \sigma_X^2 = c$$

Στα c διαγράμματα ελέγχου απεικονίζεται η στατιστική συνάρτηση $W_i = X_i$, και έστω ότι αυτή δηλώνει τον αριθμό των μολύνσεων που έχουν βρεθεί στον i ασθενή.

Ωστόσο, στην περίπτωση που δεν είναι γνωστή η παράμετρος c της κατανομής, τότε αυτή θα πρέπει να εκτιμηθεί. Έστω, ότι έχουμε δείγμα m ατόμων, και X_i ο αριθμός των μολύνσεων που εμφανίζει ο i ασθενής. Τότε, ο αμερόληπτος εκτιμητής του c θα είναι:

$$\hat{C} = \frac{X_1+X_2+\cdots+X_m}{m}\text{, με } 1 \le i \le m$$

Έτσι, τα L όρια ελέγχου για την φάση I και φάση II είναι τα ακόλουθα:

***c* Διάγραμμα Ελέγχου**	***c* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *I	***L* Όρια Ελέγχου Φάσης *II***
UCL = $\hat{c}$ + L$\sqrt{\hat{c}}$ CL = $\hat{c}$ LCL = $\hat{c}$ - L$\sqrt{\hat{c}}$	UCL = c + L$\sqrt{c}$ CL = c LCL = c - L$\sqrt{c}$

Πίνακας 3.4.3.1. L Όρια ελέγχου c - διαγράμματος

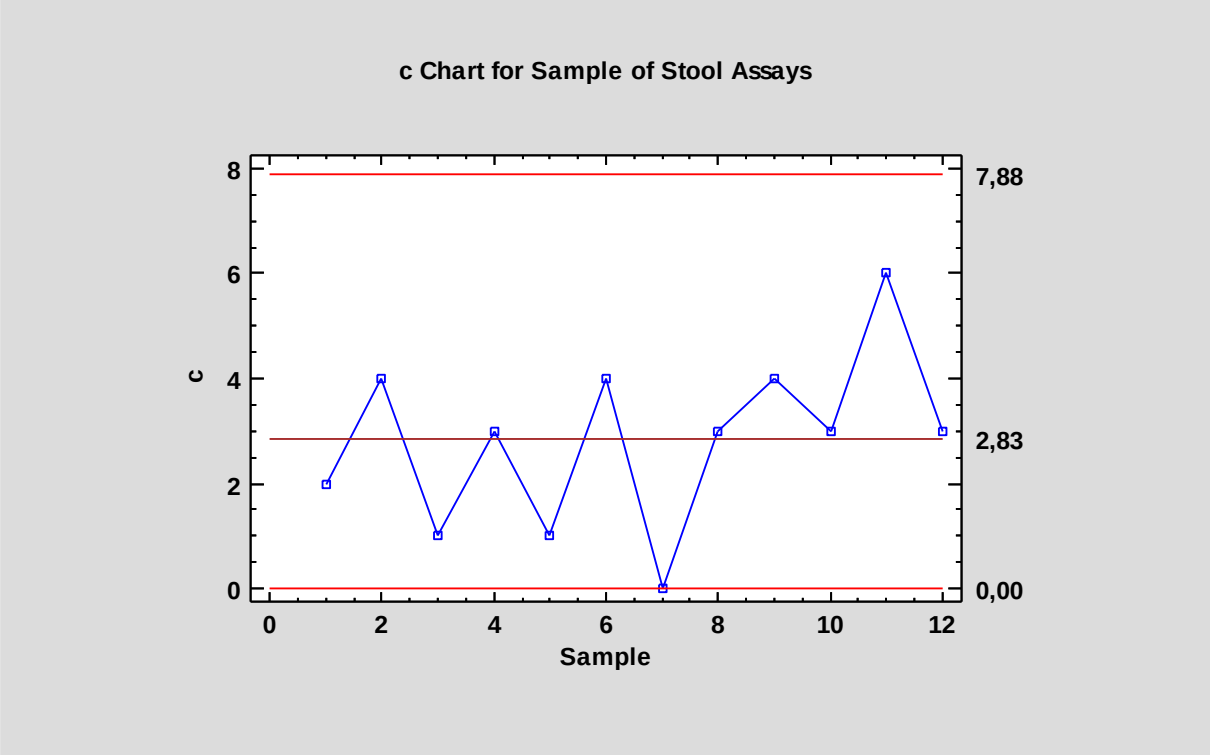
Σημαντικό να σημειωθεί πως αν το LCL είναι αρνητικό, τότε χωρίς βλάβη της γενικότητας το θέτουμε ίσο με το 0.

Τα παρακάτω δεδομένα αφορούν την εξέταση δείγματος κοπράνων ασθενών σε νοσοκομείο με σκοπό τον έλεγχο ύπαρξης τοξινών Α και Β του βακτηρίου *Clostridium difficile,* οι οποίες προκαλούν νέκρωση του βλεννογόνου του παχέος εντέρου. Οι καταγραφές αφορούν μηνιαία περιστατικά, και η έρευνα διήρκησε 12 μήνες.

Μήνας	Θετικά Δείγματα
Ιανουάριος	2
Φεβρουάριος	4
Μάρτιος	1
Απρίλιος	3
Μάιος	1
Ιούνιος	4
Ιούλιος	0
Αύγουστος	3
Σεπτέμβριος	4
Οκτώβριος	3
Νοέμβριος	6
Δεκέμβριος	3

Πίνακας 3.4.3.2. Θετικά Δείγματα σε Τοξίνη *A* και *B*
Πηγή: *Sellick Jr. J.* (1993)

Προκειμένου να προχωρήσουμε σε Στατιστικό Έλεγχο Ποιότητας των δειγμάτων μέσω *c* – διαγράμματος ελέγχου, θα υποθέσουμε ότι το πλήθος των ασθενών στο νοσοκομείο παραμένει σταθερό. Λαμβάνουμε το ακόλουθο διάγραμμα ελέγχου:



Διάγραμμα 3.4.3.1. *c* – Διάγραμμα ελέγχου για ύπαρξη τοξινών *A* και *B*

Από τα αποτελέσματα του διαγράμματος παρατηρούμε πως δεν υπάρχουν σημεία εκτός των ορίων ελέγχου, συνεπώς Τα 3σ - όρια ελέγχου που θα χρησιμοποιηθούν πλέον για μελλοντική παρακολούθηση – ανάλυση φάσης *II*, θα είναι από 0.0 έως 7.88, δηλαδή ανά δειγματοληψία είναι επιτρεπτό να υπάρχουν από 0 έως 8 δείγματα θετικά. Η κεντρική γραμμή είναι 2.83, που σημαίνει ότι όταν η διεργασία θεωρείται εντός στατιστικού ελέγχου επιτρέπεται να έχουμε κατά μέσο όρο 3 θετικά δείγματα ανά μονάδα ελέγχου. Ωστόσο, έχει ενδιαφέρον να παρατηρήσουμε πως από τον Αύγουστο και μετά, τα σημεία βρίσκονται πάνω από την κεντρική γραμμή, που σημαίνει ότι τα ευρήματα θα πρέπει να θέσουν τις αρχές σε ετοιμότητα για την αντιμετώπιση των αυξημένων θετικών δειγμάτων.

3.4.4 *u* Διάγραμμα Ελέγχου

Όπως στην περίπτωση των *c* διαγραμμάτων ελέγχου, έτσι κι εδώ βασική υπόθεση είναι ότι τα δεδομένα μας προέρχονται από κατανομή *Poisson* με παράμετρο *c*. Για την κατασκευή των *u* διαγραμμάτων ελέγχου, μας ενδιαφέρει η μελέτη του μέσου αριθμού μολύνσεων σε έναν ασθενή. Ωστόσο, ειδοποιός διαφορά ανάμεσα σε αυτά τα δυο διαγράμματα, είναι πως στην περίπτωση των *u* διαγραμμάτων ελέγχου, τα δείγματά μας μπορούν να είναι μεγαλύτερα του ενός ασθενούς το καθένα.

Έστω ότι έχουμε στην διάθεσή μας m ανεξάρτητα τυχαία δείγματα, μεγέθους n το καθένα, και X_{ij} η τυχαία μεταβλητή που συμβολίζει τον αριθμό των μολύνσεων του j ασθενούς στο i δείγμα, με $1 \leq i \leq m$ και $1 \leq j \leq n$. Όπως και παραπάνω, η τυχαία μεταβλητή X_{ij} ακολουθεί *Poisson* με παράμετρο c. Επιπλέον, ορίζουμε $X_i = X_{i1} + X_{i2} + \ldots + X_{in}$, τυχαία μεταβλητή η οποία δηλώνει το συνολικό αριθμό μολύνσεων στο i δείγμα και η οποία ακολουθεί κι αυτή Poisson με παράμετρο nc.

Επιπλέον, θεωρούμε την τυχαία μεταβλητή $U_i = \frac{X_i}{n}$, η οποία εκφράζει το μέσο αριθμό μολύνσεων ανά ασθενή στο i δείγμα, και για την οποία ισχύει:

$$\mu_{U_i} = c \quad \text{και} \quad \sigma^2_{U_i} = \frac{c}{n}, \; i \geq 1$$

Στην περίπτωση που το c δεν είναι γνωστό εκ των προτέρων και θα χρειαστεί να το εκτιμήσουμε, θεωρούμε m ομάδες ασθενών, μεγέθους n η κάθε μία, και $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$ για $1 \leq i \leq m$. Επιπρόσθετα, θέτουμε

$$U_i = \frac{X_i}{n} = \frac{X_{i1}+X_{i2}+\cdots+X_{in}}{n}$$

και

$$\overline{U} = \frac{U_1+U_2+\cdots+U_m}{m} = \frac{X_1+X_2+\cdots+X_m}{mn} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{mn}$$

Ο τελευταίος αποτελεί αμερόληπτο εκτιμητή του c, καθώς $\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij} \sim P(mnc)$.

Έτσι, τα L όρια ελέγχου για την φάση I και φάση II είναι τα ακόλουθα:

***u* Διάγραμμα Ελέγχου**	***u* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *I	***L* Όρια Ελέγχου Φάσης *II***
UCL = $\overline{U} + L\sqrt{\frac{\overline{U}}{n}}$ CL = $\overline{U}$ LCL = $\overline{U} - L\sqrt{\frac{\overline{U}}{n}}$	UCL = $U + L\sqrt{\frac{U}{n}}$ CL = U LCL = $U - L\sqrt{\frac{U}{n}}$

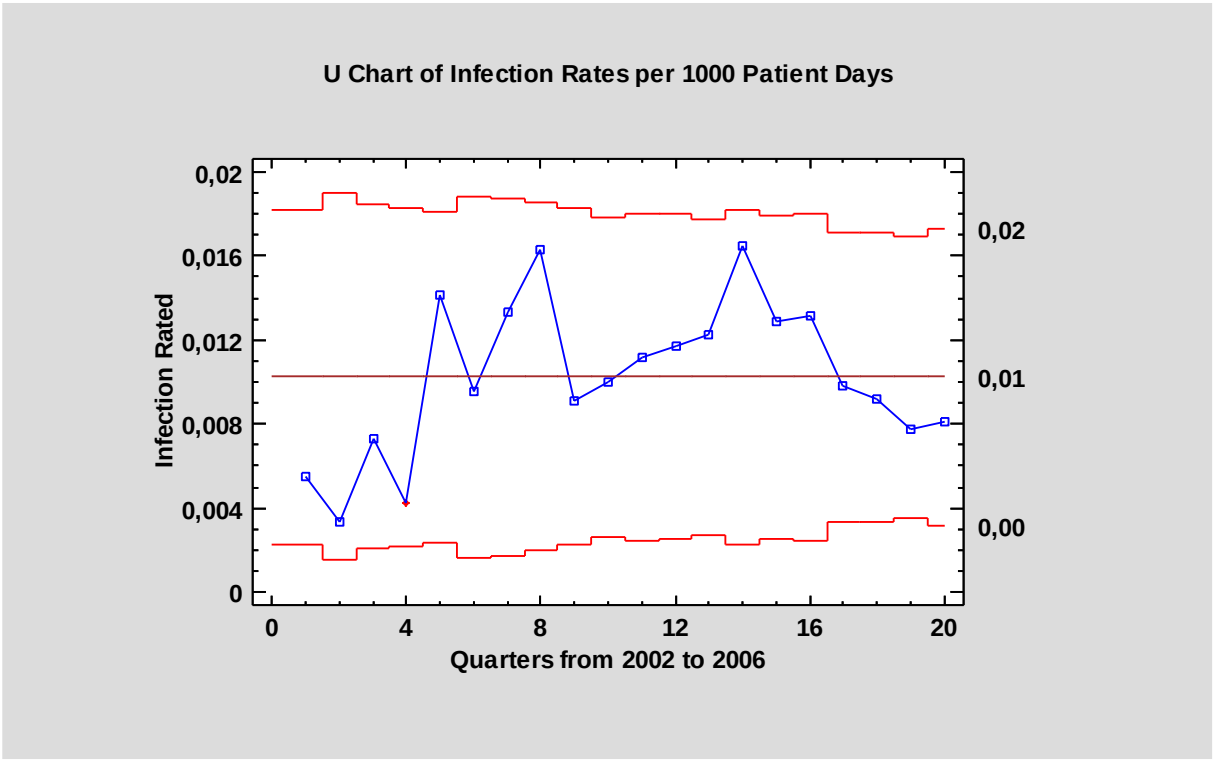
Πίνακας 3.4.4.1. *L* Όρια ελέγχου *u* – διαγράμματος

Τα παρακάτω δεδομένα αφορούν περιστατικά λοιμώξεων που εισήχθησαν στο Παιδιατρικό Νοσοκομείο του *Seattle*, την περίοδο από το 2002 έως το 2006 (*Limaye et al.*, 2008).

Τρίμηνο	Πλήθος Λοιμώξεων	Συνολικές Ημέρες Νοσηλείας	Πλήθος Λοιμώξεων ανα Ημέρα Νοσηλείας
Q1 of 2002	8	1451	0,005513439
Q2 of 2002	4	1207	0,003314002
Q3 of 2002	10	1372	0,00728863
Q4 of 2002	6	1412	0,004249292
Q1 of 2003	21	1481	0,014179608
Q2 of 2003	12	1256	0,00955414
Q3 of 2003	17	1275	0,013333333
Q4 of 2003	22	1348	0,016320475
Q1 of 2004	13	1432	0,009078212
Q2 of 2004	16	1596	0,010025063
Q3 of 2004	17	1526	0,011140236
Q4 of 2004	18	1543	0,011665587
Q1 of 2005	20	1629	0,012277471
Q2 of 2005	24	1453	0,01651755
Q3 of 2005	20	1552	0,012886598
Q4 of 2005	20	1519	0,013166557
Q1 of 2006	19	1934	0,009824199
Q2 of 2006	18	1950	0,009230769
Q3 of 2006	16	2066	0,007744434
Q4 of 2006	15	1857	0,008077544

Πίνακας 3.4.4.2. Πλήθος Λοιμώξεων Παιδιατρικού Νοσοκομείου, *Seattle*
Πηγή: *Limaye et al.* (2008)

Από το παρακάτω διάγραμμα παρατηρούμε πως δεν υπάρχουν σημεία εκτός των ορίων ελέγχου, και συνεπώς δεν υπάρχει κάποια ένδειξη ύπαρξης ειδικής αιτίας μεταβλητότητας.



Διάγραμμα 3.4.4.1. *c* - Διάγραμμα Ελέγχου

3.5 Διαγράμματα Ελέγχου Τύπου *Shewhart* για μεταβλητές

Σε αντίθεση με τα διαγράμματα ελέγχου για ιδιότητες, τα διαγράμματα ελέγχου τύπου *Shewhart* για μεταβλητές χρησιμοποιούνται όταν το υπό μελέτη χαρακτηριστικό εκφράζεται μέσω κάποιας αριθμητικής κλίμακας, όπως ο αριθμός των καρκινικών κυττάρων μετά από έναν κύκλο χημειοθεραπειών, ή οι τιμές χοληστερόλης στο αίμα, δηλαδή αφορούν κάποια συνεχή μεταβλητή.

Έστω, X το υπό μελέτη χαρακτηριστικό. Τα διαγράμματα τύπου *Shewhart* για μεταβλητές μελετούν την μέση τιμή και την διασπορά των τιμών του X ταυτόχρονα, προκειμένου να παρακολουθήσουν την διεργασία.

3.5.1 X – *Bar* Διάγραμμα Ελέγχου

Έστω $X_i = (X_{i1}, X_{i2,\dots} X_{in})$, $i \geq 1$, τυχαίο δείγμα μεγέθους n, το οποίο προέρχεται από κανονικό πληθυσμό με γνωστή μέση τιμή μ και διασπορά σ^2. Τότε, ο δειγματικός μέσος

$$\bar{X}_i = \frac{X_{in} + X_{in} + \cdots + X_{in}}{n}$$

ακολουθεί την κατανομή $N\left(\mu, \frac{\sigma^2}{n}\right)$, και μάλιστα αποτελεί και αμερόληπτο εκτιμητή της μέσης τιμής του χαρακτηριστικού Χ που μας ενδιαφέρει.

Θέτουμε $W_i = \bar{X}_i$, τον δειγματικό μέσο του δείγματος n, ο οποίος, με πιθανότητα $1 - \alpha$, θα παίρνει τιμές στο διάστημα

$$\left[\mu - z_{a/2}\frac{\sigma}{\sqrt{n}}, \mu + z_{a/2}\frac{\sigma}{\sqrt{n}}\right]$$

Έτσι, για την παρακολούθηση της μέσης τιμής της διεργασίας, προκύπτουν τα ακόλουθα όρια ελέγχου:

$\bar{X}$ **Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *II
UCL = $\mu + L\frac{\sigma}{\sqrt{n}}$
CL = μ
LCL = $\mu + L\frac{\sigma}{\sqrt{n}}$

Πίνακας 3.5.1.1. *L* Όρια ελέγχου $\bar{X}$ – διαγράμματος Φάσης *II*

Επιπρόσθετα, το αντίστοιχο διάγραμμα με όρια πιθανότητας α, θα είχε τα παρακάτω όρια ελέγχου:

$\bar{X}$ **Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Πιθανότητας α Φάσης *II
UCL = $\mu + za_{/2}\frac{\sigma}{\sqrt{n}}$ CL = μ LCL = $\mu + za_{/2}\frac{\sigma}{\sqrt{n}}$

Πίνακας 3.5.1.2. *L* Όρια ελέγχου $\bar{X}$ – διαγράμματος Φάσης *I*

Για παράδειγμα, αν α = 0.0027, τότε θα προκύπταν τα όρια $\left[\mu - 3\frac{\sigma}{\sqrt{n}}, \mu + 3\frac{\sigma}{\sqrt{n}}\right]$, δηλαδή, ο δειγματικός μέσος W_i θα είχε πιθανότητα 99.73% να βρεθεί στο παραπάνω διάστημα και πιθανότητα 0.27% να βρεθεί εκτός αυτού.

Ωστόσο, πολλές φορές οι παράμετροι της κατανομής του χαρακτηριστικού που μας ενδιαφέρει δεν είναι εκ των προτέρων γνωστές, κι γι' αυτό θα χρειαστεί να τις εκτιμήσουμε.

Έστω $X_i = (X_{i1}, X_{i2,\dots} X_{in})$, $1 \leq i \leq m$, τυχαίο δείγμα μεγέθους n, το οποίο προέρχεται από κανονικό πληθυσμό με άγνωστη μέση τιμή μ και διασπορά σ^2. Τότε, για την εκτίμηση της μέσης τιμής χρησιμοποιούμε την στατιστική συνάρτηση:

$$\hat{\mu} = \bar{\bar{X}}$$

Για την εκτίμηση της τυπικής απόκλισης μπορούν να χρησιμοποιηθούν οι τρεις εκτιμητές:

- **Μέθοδος *R*** :

$\hat{\sigma} = \frac{\bar{R}}{d_2}$, όπου $R_i = max(x_{ij}, j = 1,2,\dots,n) - min(x_{ij}, j = 1,2,\dots,n)$ και

$$\bar{R} = \frac{1}{m}\sum_{i=1} R_i$$

- **Μέθοδος S** :

$\hat{\sigma} = \frac{\bar{S}}{c_4}$, όπου $S = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2}$ και $\bar{S} = \frac{1}{m}\sum_{i=1}^{m} S_i$

- **Μέθοδος S^2**:

$\hat{\sigma} = \sqrt{\overline{S^2}}$, όπου $\overline{s^2} = \frac{1}{m}\sum_{i=1}^{m} S_i^2$

Σημειώνεται ότι οι δυο πρώτες εκτιμήσεις της τυπικής απόκλισης αποτελούν και αμερόληπτους εκτιμητές της, καθώς επίσης ότι τα d_2 και c_4 είναι σταθερές οι οποίες εξαρτώνται από το μέγεθος του δείγματος. Έτσι, ανάλογα με τον εκτιμητή της τυπικής απόκλισης που επιλέγουμε κάθε φορά, λαμβάνουμε τα παρακάτω όρια ελέγχου:

$\bar{X}$ Διάγραμμα Ελέγχου	**$\bar{X}$ Διάγραμμα Ελέγχου**	**$\bar{X}$ Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *I* Μέθοδος *R	***L* Όρια Ελέγχου Φάσης *I* Μέθοδος *S***	***L* Όρια Ελέγχου Φάσης *I* Μέθοδος S^2**
UCL = $\bar{\bar{X}} + A_2\bar{R}$ CL = $\bar{\bar{X}}$ LCL = $\bar{\bar{X}} - A_2\bar{R}$	UCL = $\bar{\bar{X}} + A_3\bar{S}$ CL = $\bar{\bar{X}}$ LCL = $\bar{\bar{X}} - A_3\bar{S}$	UCL = $\bar{\bar{X}} + A\sqrt{\overline{S^2}}$ CL = $\bar{\bar{X}}$ LCL = $\bar{\bar{X}} - A\sqrt{\overline{S^2}}$

Πίνακας 3.5.1.3. *L* Όρια ελέγχου $\bar{X}$ – διαγράμματος Φάσης *I* βάσει μεθόδου *R*, *S* και S^2

όπου $A_2 = \frac{L}{d_2\sqrt{n}}$, $A_3 = \frac{L}{c_4\sqrt{n}}$ $και$ $A = \frac{L}{\sqrt{n}}$.

3.5.2 Διάγραμμα Ελέγχου για την Διασπορά

Παραπάνω αναφερθήκαμε στην παρακολούθηση των τιμών ενός χαρακτηριστικού μέσω του διαγράμματος ελέγχου για την μέση τιμή, το οποίο μπορεί να μας δώσει ένδειξη εντός ελέγχου διεργασίας όταν όλα τα σημεία βρεθούν εντός των ορίων ελέγχου, με την

προϋπόθεση ωστόσο ότι η διασπορά του χαρακτηριστικού X παραμένει σταθερή (Αντζουλάκος, 2020). Ωστόσο, σε μία διεργασία, αν και ο μέσος της μπορεί να μην μετατοπίζεται, εντούτοις αυτή να βρίσκεται εκτός στατιστικού ελέγχου λόγω μη – σταθερότητας της διασποράς. Για τον σκοπό αυτό, έχουν αναπτυχθεί τα διαγράμματα ελέγχου για την διασπορά.

Έστω, Χ το υπό μελέτη χαρακτηριστικό το οποίο ακολουθεί κανονική κατανομή με γνωστή μέση τιμή και τυπική απόκλιση, $X \sim N(\mu, \sigma^2)$. Επίσης, έχουμε τυχαίο δείγμα μεγέθους n, $X_i = (X_{i1}, X_{i2,\dots} X_{in})$, $1 \leq i \leq m$. Θεωρούμε την στατιστική συνάρτηση

$$W_i = R_i = X_{i(n)} - X_{i(1)}$$

για την οποία ισχύει

$$\mu_{R_i} = \sigma d_2 \text{ και } \sigma_{R_i} = \sigma d_3$$

όπου d_2 και d_3 σταθερές οι οποίες εξαρτώνται από το μέγεθος n του δείγματος.

Έτσι, για την παρακολούθηση της διασποράς της διεργασίας μέσω του εύρους των δειγμάτων, προκύπτουν τα ακόλουθα όρια ελέγχου:

***R* Διάγραμμα Ελέγχου**
***L* Όρια Ελέγχου Φάσης II**
UCL = $D_2\sigma$
CL = $d_2\sigma$
LCL = $D_1\sigma$

Πίνακας 3.5.2.1. *L* Όρια ελέγχου *R* – διαγράμματος Φάσης *II*

Όπου $D_1 = d_2 - Ld_3$ και $D_2 = d_2 + Ld_3$, ενώ όταν για το μέγεθος του δείγματος ισχύει ότι $n \leq 6$

τότε θέτουμε $D_1 = 0$.

Επιπλέον, αντί της στατιστικής συνάρτησης που ορίστηκε παραπάνω, θα μπορούσε να χρησιμοποιηθεί η στατιστική συνάρτηση:

$$W_i = S_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}\right)^2}$$

για την οποία ισχύει:

$$\mu_{S_i} = \sigma c_4 \text{ και } \sigma_{S_i} = \sigma\sqrt{1 - c_4^2}$$

Έτσι, για την παρακολούθηση της διασποράς της διεργασίας μέσω των δειγματικών αποκλίσεων, προκύπτουν τα ακόλουθα όρια ελέγχου:

***S* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *II
UCL = $B_6\sigma$
CL = $c_4\sigma$
LCL = $B_5\sigma$

Πίνακας 3.5.2.2. *L* Όρια ελέγχου S – διαγράμματος Φάσης *II*

όπου $B_5 = c_4 - L\sqrt{1 - c_4^2}$ και $B_5 = c_4 + L\sqrt{1 - c_4^2}$, ενώ όταν για το μέγεθος του δείγματος ισχύει ότι $n \leq 5$ τότε θέτουμε $B_5 = 0$.

Τέλος, μπορεί η παρακολούθηση της διασποράς να γίνει μέσω της στατιστικής συνάρτησης:

$$W_i = S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2, 1 \leq i \leq m$$

για την οποία ισχύει:

$$\mu_{S_i^2} = \sigma^2 \text{ και } \frac{(n-1)S_i^2}{\sigma^2} \sim X_{n-1}^2 - 1$$

Βάσει των παραπάνω προκύπτει ότι:

$$P(X_{n-1;1-a/2}^2 \leq \frac{(n-1)S_i^2}{\sigma^2} \leq X_{n-1;a/2}^2) = 1\text{-}\alpha$$

και

$$P(\frac{\sigma^2}{x-1}X_{n-1;1-a/2}^2 \leq S^2 \leq \frac{\sigma^2}{x-1}X_{n-1;a/2}^2) = 1\text{-}\alpha$$

Έτσι, για την παρακολούθηση της διασποράς της διεργασίας μέσω της δειγματικής διακύμανσης, προκύπτουν τα ακόλουθα όρια ελέγχου πιθανότητας α:

S^2 Διάγραμμα Ελέγχου
Όρια Ελέγχου Πιθανότητας α Φάσης *II*
UCL = $\frac{\sigma^2}{n-1}X^2_{n-1;a/2}$ CL = σ^2 LCL = $\frac{\sigma^2}{n-1}X^2_{n-1;1-a/2}$

Πίνακας 3.5.2.3. *L* Όρια ελέγχου S^2 – διαγράμματος Φάσης *II*

Ωστόσο, όπως αναφέραμε και παραπάνω, στις περισσότερες περιπτώσεις το σ είναι άγνωστο και γι' αυτό πρέπει να εκτιμηθεί. Οι εκτιμήσεις γίνονται μέσω των μεθόδων *R*, *S* και S^2, οι οποίες αναπτύχθηκαν στο προηγούμενο κεφάλαιο. Αναλόγως λοιπόν με την μέθοδο εκτίμησης που θα χρησιμοποιήσουμε, παίρνουμε τα αντίστοιχα όρια ελέγχου Φάσης *I*:

R Διάγραμμα Ελέγχου	**S Διάγραμμα Ελέγχου**	**S^2 Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *I* Μέθοδος *R	***L* Όρια Ελέγχου Φάσης *I* Μέθοδος *S***	***L* Όρια Ελέγχου Φάσης *I* Μέθοδος S^2**
UCL = $D_4\bar{R}$ CL = $\bar{R}$ LCL = $D_3\bar{R}$	UCL = $B_4\bar{S}$ CL = $\bar{S}$ LCL = $B_3\bar{S}$	UCL = $\frac{\overline{S^2}}{n-1}X^2_{n-1;a/2}$ CL = $\overline{S^2}$ LCL = $\frac{\overline{S^2}}{n-1}X^2_{n-1;1-a/2}$

Πίνακας 3.5.2.3. *L* Όρια ελέγχου Φάσης *I* βάσει μεθόδου *R*, *S* και S^2

όπου

- $D_3 = 1 - L\frac{d_3}{d_2}$ και $D_4 = 1 + L\frac{d_3}{d_2}$, ενώ όταν για το μέγεθος του δείγματος ισχύει ότι $n \leq 6$ τότε θέτουμε $D_3 = 0$
- Όπου $B_3 = 1 - \frac{L}{C_4}\sqrt{1 - c_4^2}$ και $B_4 = 1 + \frac{L}{C_4}\sqrt{1 - c_4^2}$, ενώ όταν για το μέγεθος του δείγματος ισχύει ότι $n \leq 5$ τότε θέτουμε $B_3 = 0$.

3.5.3 Διάγραμμα Ελέγχου για Μεμονωμένες Παρατηρήσεις

Σε περιπτώσεις ωστόσο, που το υπό μελέτη χαρακτηριστικό αφορά κάποιο σπάνιο σύνδρομο, δεν είναι εφικτό να έχουμε στην διάθεσή μας μεγάλα δείγματα. Επίσης, πολλές φορές μας ενδιαφέρουν μετρήσεις που αφορούν έναν ασθενή, δηλαδή $n = 1$. Έτσι, προτάθηκαν τα διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις.

Έστω X το υπό μελέτη χαρακτηριστικό, για το οποίο ισχύει $X \sim N(\mu, \sigma^2)$, όπου μ, σ^2 γνωστά και $W_i = X_i, i \geq 1$ η κατάλληλη στατιστική συνάρτηση. Τότε τα όρια του ελέγχου για το X – διάγραμμα είναι τα:

***X* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *II
UCL = $\mu + L\sigma$
CL = μ
LCL = $\mu - L\sigma$

Πίνακας 3.5.3.1. *L* Όρια ελέγχου *X* διαγράμματος Φάσης *II*

Για την παρακολούθηση της διασποράς, θα χρησιμοποιήσουμε το κινούμενο εύρος μήκους 2 (*Moving Range*) των μεμονωμένων παρατηρήσεων:

$$MR_i = |X_i - X_{i-1}| = max(X_{i-1}, X_i) - min(X_{i-1}, X_i),\ i \geq 2$$

για το οποίο ισχύει ότι $\mu_{MR_i} = \sigma d_2$ και $\sigma_{MR_i} = \sigma d_3$.

Έτσι, τα όρια ελέγχου που προκύπτουν είναι τα ακόλουθα:

***MR* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *II
UCL = $D_2\sigma$
CL = $d_2\ \sigma$
LCL = $D_1\sigma$

Πίνακας 3.5.3.2. *L* Όρια ελέγχου *MR* διαγράμματος Φάσης *II*

όπου $D_1 = d_2 - Ld_3$ και $D_1 = d_2 + Ld_3$. $Τα\ d_2\ και\ d_3$ είναι σταθερές που εξαρτώνται από το μέγεθος του δείγματος. Στην περίπτωσή μας, υπολογίζονται για n = 2.

Στην περίπτωση όπου τα μ και σ^2 δεν είναι εκ των προτέρων γνωστά, τότε θα χρειαστεί να τα εκτιμήσουμε. Έστω $X = (X_1, X_2, \ldots X_m)$ τυχαίο δείγμα μεγέθους m. Τότε, ένας εκτιμητής του μ είναι ο:

$$\hat{\mu} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_m}{m}$$

και για τον οποίο ισχύει $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{m}\right)$.

Για την εκτίμηση του σ, ορίζουμε τον αμερόληπτο εκτιμητή:

$$\hat{\sigma} = \frac{\bar{MR}}{d_2} = \frac{\frac{MR_2 + MR + \cdots + MR_m}{m-1}}{d_2}$$

Έτσι, τα όρια ελέγχου που προκύπτουν είναι τα ακόλουθα:

$\boldsymbol{X}$ **Διάγραμμα Ελέγχου**	$\boldsymbol{MR}$ **Διάγραμμα Ελέγχου**
$\boldsymbol{L}$ **Όρια Ελέγχου Φάσης** $\boldsymbol{I}$	$\boldsymbol{L}$ **Όρια Ελέγχου Φάσης** $\boldsymbol{I}$
UCL = $\bar{X} + L\frac{\overline{MR}}{d_2}$ CL = $\bar{X}$ LCL = $\bar{X} - L\frac{\overline{MR}}{d_2}$	UCL = $D_4\overline{MR}$ CL = $\overline{MR}$ LCL = $D_3\overline{MR}$

Πίνακας 3.5.3.3. *L* Όρια ελέγχου *X* και *MR* διαγραμμάτων Φάσης *I*

όπου $D_3 = 1 - L\frac{d_3}{d_2}$ και $D_4 = 1 + L\frac{d_3}{d_2}$. Τα D_4 *και* D_3 είναι σταθερές που εξαρτώνται από το μέγεθος του δείγματος. Στην περίπτωσή μας, υπολογίζονται για $n = 2$.

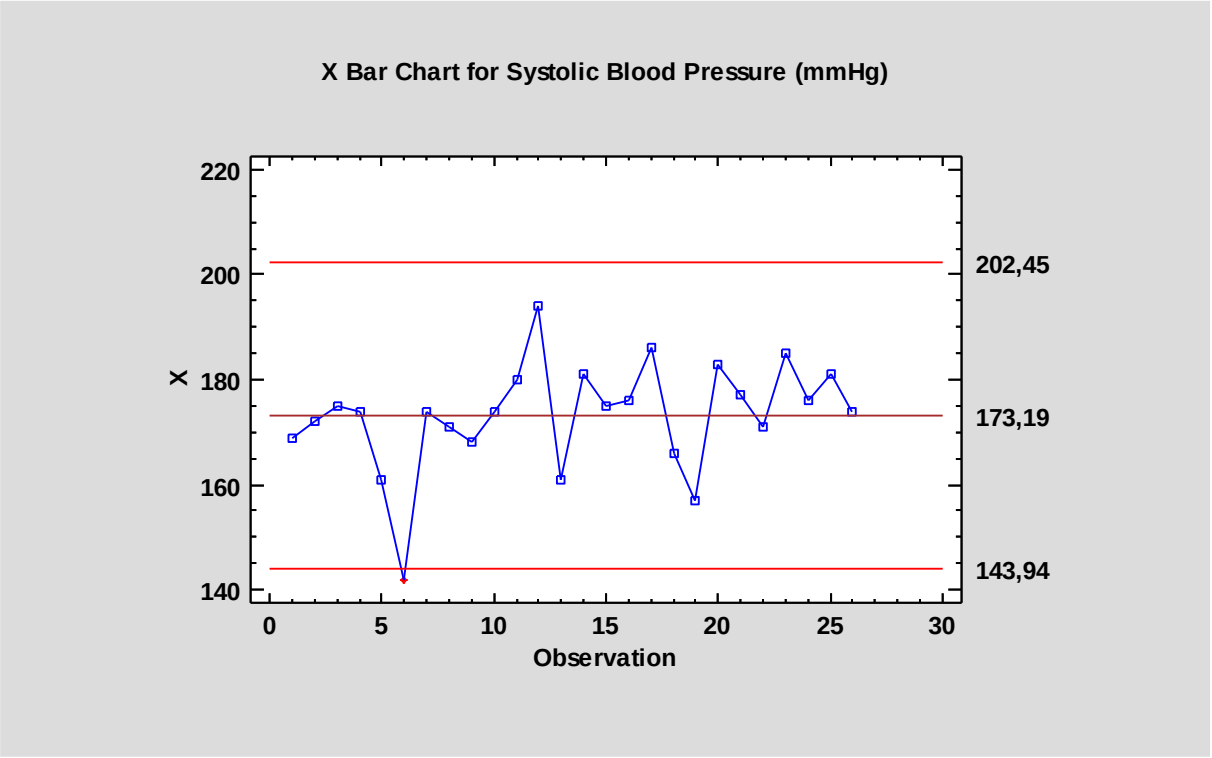
Ενδεικτικό της χρήσης των διαγραμμάτων ελέγχου X και MR για μεμονωμένες παρατηρήσεις, είναι το παράδειγμα των *Mohammed et al* (2007). Στον παρακάτω πίνακα, δίνονται οι μετρήσεις της συστολικής πίεσης ασθενούς, όπως μετρήθηκε καταγράφηκαν για 26 συνεχόμενα πρωινά.

Ημέρα	Συστολική Πίεση (mmHg)
1η	169
2η	172
3η	175
4η	174
5η	161
6η	142
7η	174
8η	171
9η	168
10η	174
11η	180
12η	194
13η	161
14η	181
15η	175
16η	176
17η	186
18η	166
19η	157
20η	183

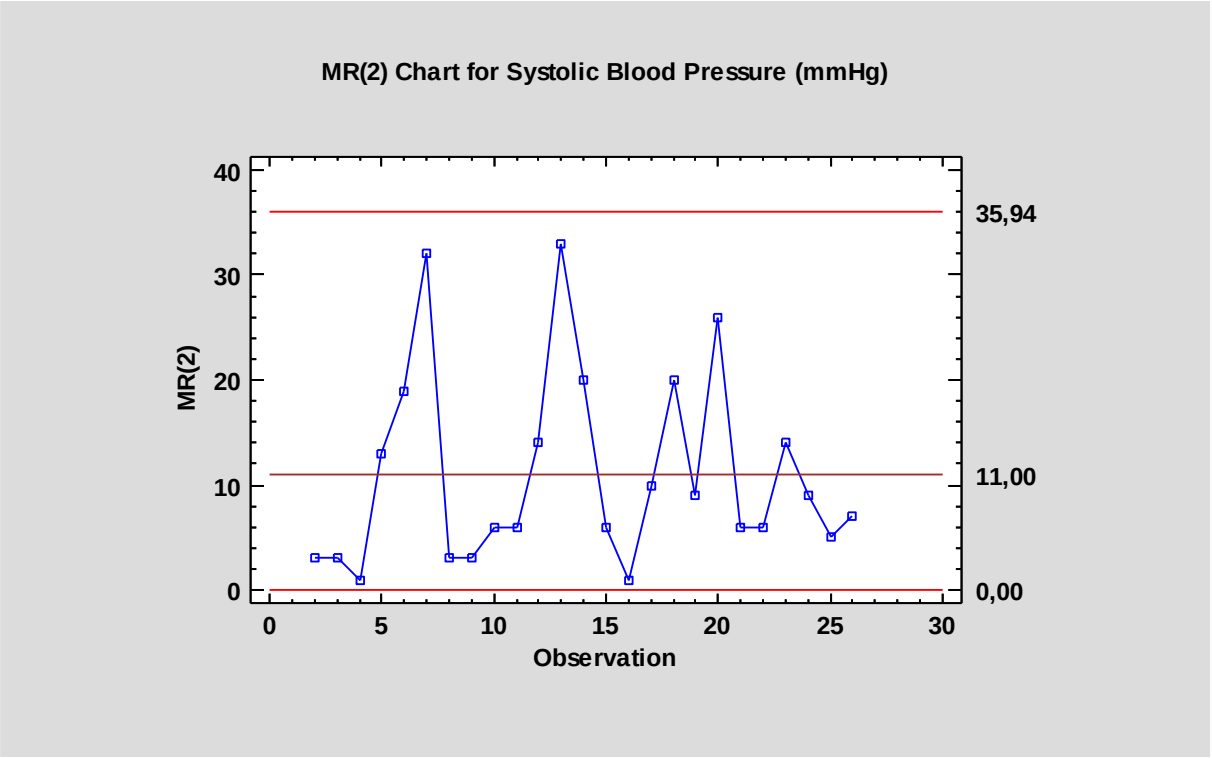
21η	177
22η	171
23η	185
24η	176
25η	181
26η	174

Πίνακας 3.5.3.4. Μετρήσεις Συστολικής Πίεσης Ασθενούς
Πηγή*: Mohammed et al.* (2007)

Από τα παρακάτω διαγράμματα παρατηρούμε πως στο *X – Bar* διάγραμμα λαμβάνουμε ένδειξη εκτός ελέγχου διεργασίας, καθώς υπάρχει ένα σημείο όπου βρίσκεται εκτός των ορίων ελέγχου, και συγκεκριμένα η μέτρηση όπου έγινε την 6^{η} ημέρα. Αντιθέτως, στο *MR* διάγραμμα, όλα τα σημεία βρίσκονται εντός των ορίων. Ωστόσο, για να μπορέσουμε να χρησιμοποιήσουμε τα όρια ελέγχου για ανάλυση φάσης *II*, θα πρέπει να αφαιρέσουμε το σημείο που βρίσκεται εκτός, και να προχωρήσουμε στον επαναυπολογισμό τους.



Διάγραμμα 3.5.3.1. *X bar* - Διάγραμμα Ελέγχου για την Συστολική Πίεση Ασθενούς



Διάγραμμα 3.5.3.2. *MR* - Διάγραμμα Ελέγχου για την Συστολική Πίεση Ασθενούς

3.6 Διαγράμματα ελέγχου με μνήμη

Τα διαγράμματα τύπου *Shewhart,* αν και είναι ιδιαίτερα εύχρηστα, έχουν ένα σημαντικό μειονέκτημα. Για την κατασκευή τους, λαμβάνουμε υπόψιν την πληροφορία που δίνεται μόνο στο τελευταίο χρονικά δείγμα, ενώ αγνοούν οποιαδήποτε πληροφορία μπορούμε να λάβουμε από το σύνολο των προγενέστερων δειγμάτων. Έτσι, τα διαγράμματα τύπου *Shewhart,* αν και μπορούν να εντοπίσουν εξαιρετικά γρήγορα μεγάλες μετατοπίσεις της διεργασίας, εντούτοις, δεν μπορούν να ανιχνεύσουν με την ίδια αποτελεσματικότητα μετατοπίσεις της διεργασίας της τάξης των 1.5σ ή και χαμηλότερων. Συνεπώς, αν και είναι ιδιαίτερα χρήσιμα για τον έλεγχο της διεργασίας στην φάση *I*, η εφαρμογή τους δεν συνίσταται για την φάση *II*. Επίσης, στην πλειονότητα τους τα διαγράμματα τύπου *Shewhart,* υποθέτουν ότι τα δεδομένα προέρχονται από την κανονική κατανομή. Ωστόσο, ιδιαίτερα όσο αναφορά δεδομένα Υγείας, η υπόθεση της κανονικότητας πολύ συχνά παραβιάζεται. Έτσι, τα δεδομένα μας συχνά προέρχονται από κατανομές οι οποίες εμφανίζουν ιδιαίτερα μεγάλη λοξότητα. Συνεπώς, η χρήση διαγραμμάτων *Shewhart* σε

αυτές τις περιπτώσεις, θα οδηγούσε σε μεγάλη αύξηση της συχνότητας των λανθασμένων συναγερμών.

Για την αντιμετώπιση των παραπάνω προβλημάτων, προτάθηκαν τα διαγράμματα ελέγχου με μνήμη, με τα πιο γνωστά από αυτά να είναι τα διαγράμματα *CUSUM* και *EWMA*.

3.6.1 Διάγραμμα CUSUM

Τα διαγράμματα *CUSUM* (*Cumulative Sum Chart*) ή αλλιώς αθροιστικά διαγράμματα ελέγχου, εισήχθησαν για πρώτη φορά από τον *Page* (1954) και προτάθηκαν λόγω της ικανότητάς τους να ανιχνεύουν μικρές μετατοπίσεις της διεργασίας από την μέση τιμή της στην φάση *II*. Βασίζονται στον υπολογισμό των συσωρευτικών αθροισμάτων (*Cumulative Sums*), δηλαδή στις διαφορές των δειγμάτων που λαμβάνουμε από την μέση τιμή της διεργασίας μ_0 , ενώ διαθέτουν μη περιορισμένη και ομοιόμορφη μνήμη, καθώς για την δημιουργία τους γίνεται προσαρμογή της πληροφορίας που λάβαμε για το σύνολο της ακολουθίας των παρατηρήσεων που έχουν προηγηθεί, έχοντας κάθε μία από αυτές τον ίδιο συντελεστή βαρύτητας.

Για την κατασκευή διαγραμμάτων *CUSUM* πρέπει να υπολογιστεί το συσσωρευμένο άθροισμα των αποκλίσεων των παρατηρήσεων που λαμβάνουμε από την μέση τιμή μ_0 της διεργασίας. Έστω $W_i = g(X_i)$ η στατιστική συνάρτηση που αντιστοιχεί σε κάθε τυχαίο δείγμα X_i, $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$, $n \geq 1$. Στα διαγράμματα ελέγχου *CUSUM* απεικονίζεται η στατιστική συνάρτηση:

$$Y_t = \sum_{i=1}^{t}[g(X_i) - \mu_0] = -tY_0 + \sum_{i=1}^{t} g(X_i)$$

όπου Y_0 είναι μια τιμή στόχος.

Στα διαγράμματα τύπου *CUSUM* για τον μέσο μίας διεργασίας, απεικονίζονται δυο ποσότητες με σκοπό την ανίχνευση των μετατοπίσεων της μέσης τιμής της διεργασίας σε υψηλότερο ή χαμηλότερο επίπεδο, δηλαδή μετατοπίσεις της μορφής $\mu_1 = \mu_0 \pm \delta\sigma,\ \delta > 0$, και συγκεκριμένα οι:

$$S_i^+ = \max\{0, X_i - (\mu_0 + K) + S_{i-1}^+\}\ , S_0^+ = 0$$

$$S_i^- = \min\{0, X_i - (\mu_0 - K) + S_{i-1}^-\}\ , S_0^- = 0$$

όπου X_i η τιμή που υπό μελέτη χαρακτηριστικού, την χρονική στιγμή i, η οποία προέρχεται από κατανομή με μέση τιμή μ_0 και τυπική απόκλιση σ. Η παραπάνω μέθοδος ονομάζεται αλγοριθμική και εισήχθη για πρώτη από φορά από τον *Page* (1954).

Η *K* ονομάζεται τιμή αναφοράς, και δίνεται από την σχέση:

$$K = \frac{\sigma\delta}{2} = \frac{|\mu_1 - \mu_0|}{2} = k\sigma,\ k = \frac{\delta}{2}$$

Επιπρόσθετα, πάνω σε ένα διάγραμμα *CUSUM* σχεδιάζονται δυο ευθείες παράλληλες προς την μέση τιμή της διεργασίας μ_0, οι H^+ και H^-, οι οποίες αποτελούν τα όρια ελέγχου του διαγράμματος. Το *H* καλείται διάστημα απόφασης και δίνεται από την σχέση

$$H = h\sigma$$

Για τις περισσότερες εφαρμογές χρησιμοποιείται το $h = 4$ ή $h = 5$. Το διάγραμμα μας δίνει ένδειξη ότι η διεργασία είναι εκτός στατιστικού ελέγχου όταν $S_i^+ > H^+$ ή $S_i^- < H^-$. Επιπρόσθετα, είναι σημαντικό να τονιστεί ότι σε αντίθεση με τα διαγράμματα τύπου *Shewhart*, τα διαγράμματα *CUSUM* μπορούν να ερμηνευτούν μόνο ποιοτικά, με αποτέλεσμα να μην έχουμε ακριβής εικόνα για τις τιμές που διέπουν την διεργασία μας.

Ωστόσο, ένα από τα βασικότερα μειονεκτήματα τέτοιου τύπου διαγραμμάτων είναι πως δεν δίνουν εγκαίρως ειδοποίηση ότι η διεργασία επανήλθε εντός στατιστικού ελέγχου.

Παρακάτω, δίνονται δεδομένα που αφορούν την ανίχνευση περιστατικών *HIV/AIDS*, από το *Oyo state Hospital Management Board*, για την περίοδο Γενάρης 2001 έως Δεκέμβριος 2004.

Year/Month	2001	2002	2003	2004
January	31	33	33	36
February	45	48	48	28
March	41	47	26	24
April	40	25	26	43
May	53	28	30	32
June	48	18	41	26
July	55	36	40	44
August	71	23	27	25
September	56	31	49	48
October	64	14	41	51
November	47	6	51	41
December	47	16	13	46

Πίνακας 3.6.1.1. Καταγεγραμμένες περιπτώσεις *HIV/AIDS*
Πηγή: *Adeoti O.* (2013)

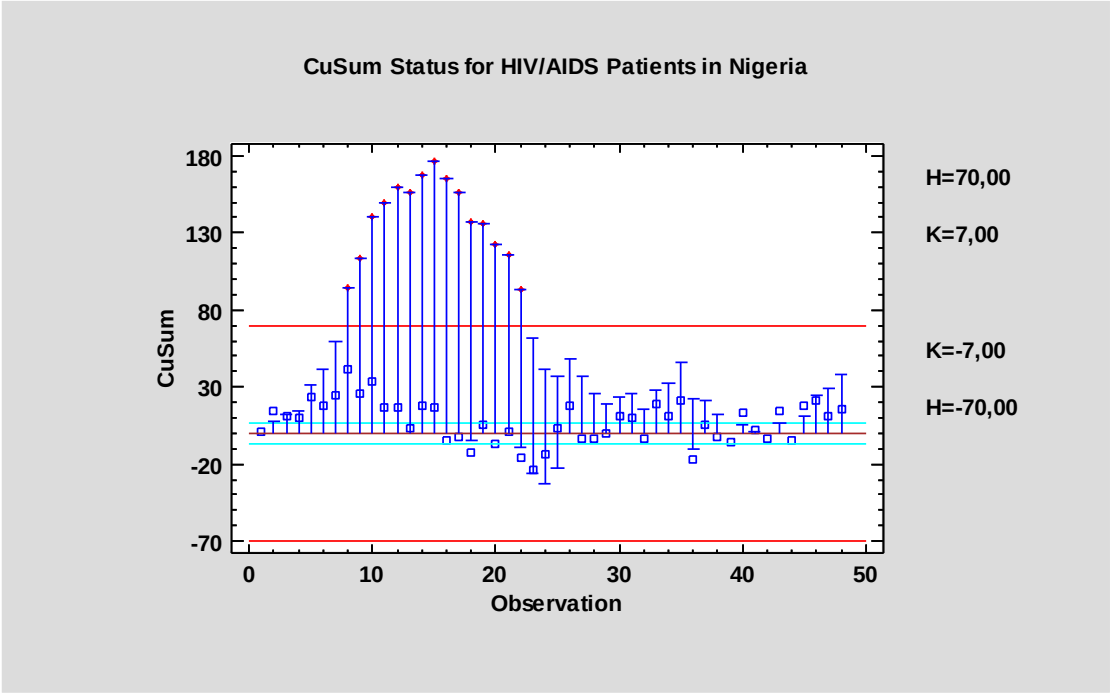
Για την κατασκευή του διαγράμματος *CUSUM* για την μελέτη φάσης *II* επιλέχθηκε H = 70, h = 5, k = 0.5, K = 7, μ_0 = 30 και std = 14. Μας ενδιαφέρει να εντοπίσουμε μετατόπιση του μέσου της διεργασίας προς τα πάνω, και πιο συγκεκριμένα $S_i^+ > H$.

ΔΕΙΓΜΑ	Xi	Xi - 37	Si	Ni
1	31	-6	0	0
2	45	8	8	1
3	41	4	12	2
4	40	3	15	3
5	53	16	31	4
6	48	11	41	5
7	55	18	59	6
8	71	34	93	7
9	56	19	112	8
10	64	27	139	9
11	47	10	149	10
12	47	10	159	11
13	33	-4	155	12
14	48	11	165	13
15	47	10	175	14
16	25	-12	163	15
17	28	-9	154	16
18	18	-19	135	17
19	36	-1	134	18
20	23	-14	120	19

21	31	-6	114	20
22	14	-23	91	21
23	6	-31	60	22
24	16	-21	39	23
25	33	-4	35	24
26	48	11	46	25
27	26	-11	35	26
28	26	-11	24	27
29	30	-7	17	28
30	41	4	21	29
31	40	3	24	30
32	27	-10	14	31
33	49	12	26	32
34	41	4	30	33
35	51	14	44	34
36	13	-24	20	35
37	36	-1	19	36
38	28	-9	10	37
39	24	-13	0	0
40	43	6	6	1
41	32	-5	1	2
42	26	-11	0	0
43	44	7	7	1
44	25	-12	0	0
45	48	11	11	1
46	51	14	25	2
47	41	4	29	3
48	46	9	38	4

Πίνακας 3.6.1.2. Αθροίσματα *CUSUM* για τις περιπτώσεις *HIV/AIDS*

Όπως, διακρίνεται στο παρακάτω διάγραμμα, λαμβάνουμε ένδειξη εκτός ελέγχου διεργασίας, καθώς παρατηρούμε ότι το διάγραμμα βγαίνει εκτός του άνω ορίου που θέσαμε και ο μέσος της διεργασίας μετατοπίζεται σε υψηλότερο επίπεδο, γεγονός που ενδέχεται να αποτελεί απόρροια ύπαρξης παράγοντα κινδύνου.



Διάγραμμα 3.6.1.1. Διάγραμμα *CUSUM* για τις περιπτώσεις *HIV/AIDS*

Ωστόσο, πολλές μελέτες που περιστρέφονται γύρω από την επιδημιολογία, αφορούν την αύξηση ή μείωση του αριθμού κρουσμάτων μιας μολυσματικής νόσου. Για τον σκοπό αυτό, έχουν προταθεί από τον *Lucas*, το 1985, τα διαγράμματα *CUSUM* για ιδιότητες, και συγκεκριμένα το διάγραμμα *Poisson CUSUM*.

Οι στατιστικές συναρτήσεις που χρησιμοποιούνται σε ένα τέτοιο διάγραμμα είναι οι:

.

$$S_i^+ = \max\{0, X_i - K + S_{i-1}^+\}$$

$$S_i^- = \min\{0, X_i + K + S_{i-1}^-\}$$

ενώ για την έγκαιρη ένδειξη ότι η διεργασία είναι εκτός ελέγχου στην αρχή της επιτήρησης, ο *Lucas* (1985) προτείνει την χρήση του $S_0 = \frac{h}{2}$.

Πολύ σημαντική στην κατασκευή του διαγράμματος *Poisson CUSUM* αποτελεί η επιλογή των κατάλληλων τιμών K και h. Η τιμή της K, επιλέγεται με τρόπο τέτοιο ώστε να είναι ανάμεσα στην αποδεκτή τιμή του μέσου της διεργασίας και την μέση τιμή των μετρήσεων των περιστατικών που θέλουμε να ανιχνεύσουμε. Ωστόσο, παρόλο που τις

περισσότερες φορές, η αποδεκτή μέση τιμή της διεργασίας επιθυμούμε να είναι μηδέν (μηδενικός αριθμός κρουσμάτων), κάτι τέτοιο δεν χρησιμοποιείται για τον σχεδιασμό ενός *Poisson CUSUM*. Στην πράξη, η αποδεκτή μέση τιμή, επιλέγεται με τρόπο τέτοιο ώστε να βρίσκεται κοντά στην μέση τιμή της διεργασίας μας (*Lucas,* 1985). Έτσι, η τιμή αναφοράς ενός διαγράμματος *Poisson CUSUM*, επιλέγεται βάσει του παρακάτω τύπου:

$$K_p = \frac{\mu_d - \mu_a}{\ln \mu_d - \ln \mu_a}$$

όπου, μ_a η αποδεκτή μέση τιμή του μέσου της διεργασίας και μ_d η μέση τιμή των περιστατικών που μελετάμε. Στην συνέχεια, και μετά τον υπολογισμό της K, μέσω κατάλληλων πινάκων, επιλέγεται και η τιμή της h.

3.6.2 Διάγραμμα EWMA

Τα διαγράμματα ελέγχου εκθετικά κινούμενου μέσου (*EWMA)*, εισήχθησαν για πρώτη φορά από τον *Roberts* το 1959, και αποτελούν άριστη εναλλακτική των διαγραμμάτων ελέγχου τύπου *Shewhart*, καθώς είναι ιδιαίτερα αποτελεσματικά στην ανίχνευση μικρών μετατοπίσεων του μέσου επιπέδου της διεργασίας. Τα συγκεκριμένα διαγράμματα έχουν μη περιορισμένη και μη ομοιόμορφη μνήμη καθώς κατά την δημιουργία τους γίνεται χρήση της πληροφορίας που λήφθηκε από όλα τα προγενέστερα δείγματα, με το καθένα από αυτά να έχει διαφορετικό συντελεστή βαρύτητας.

Για την κατασκευή τους, χρησιμοποιείται η στατιστική συνάρτηση:

$$Z_t = (1 - \lambda)Z_{t-1} + \lambda X_t,\ 0 < \lambda \leq 1 \qquad (3.6.1)$$

Όπου X_t η τιμή του υπό μελέτη χαρακτηριστικού την χρονική στιγμή t, και λ συντελεστής εξομάλυνσης ή η σταθερά βάσει της οποίας καθορίζεται η βαρύτητα που δίνεται στο κάθε δείγμα. Συγκεκριμένα, το λ καθορίζει την εξάρτηση που θα έχει κάθε δείγμα από τις προηγούμενες τιμές. Για παράδειγμα, για λ = 0.2, Z_t εξαρτάται κατά 80% από τις προηγούμενες τιμές που έχουμε λάβει από την διεργασία και κατά 20% από την τιμή της διεργασίας την χρονική στιγμή t. Για $t = 0$, το Z_0 είναι ίσο με μ_0, δηλαδή με το

μέσο της εντός ελέγχου διεργασίας ή με την τιμή στόχο. Είναι φανερό, πως η τιμή του Z_t δεν εξαρτάται μονάχα από την τιμή του χαρακτηριστικού X την χρονική στιγμή t, αλλά και από τις προηγούμενες τιμές Z_{t-1}, δηλαδή γίνεται λαμβάνονται υπόψιν και οι παρελθοντικές τιμές.

Αν αναπτύξουμε την σχέση (3.6.1) παίρνουμε:

$$Z_t = (1-\lambda)Z_{t-1} + \lambda X_t = (1-\lambda)[(1-\lambda)Z_{t-2} + \lambda X_{t-1}] + \lambda X_t = \ldots = (1-\lambda)^t Z_0 + \lambda \sum_{i=1}^{t} (1-\lambda)^{t-i} X_i$$

Από την παραπάνω σχέση προκύπτει ότι το Z_t αποτελεί τον σταθμισμένο μέσο των παρατηρήσεων $Z_0, X_1, \ldots, X_t$ με βάρη $(1-\lambda)^t$, $\lambda(1-\lambda)^{t-1}, \ldots, \lambda$ αντίστοιχα.

Για την στατιστική συνάρτηση Z_t ισχύει ότι:

$$\mu_{z_t} = \mu_0$$

και

$$\sigma_{z_t}^2 = \sigma^2 \left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2t}]$$

Έτσι προκύπτουν τα παρακάτω όρια ελέγχου για το διάγραμμα EWMA:

***EWMA* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *II
UCL = $\mu_{z_t} + L\sigma_{z_t} = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]}$
CL = μ_0
LCL = $\mu_{z_t} - L\sigma_{z_t} = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]}$

Πίνακας 3.6.2.1. *L* Όρια ελέγχου *EWMA* διαγράμματος Φάσης *II*

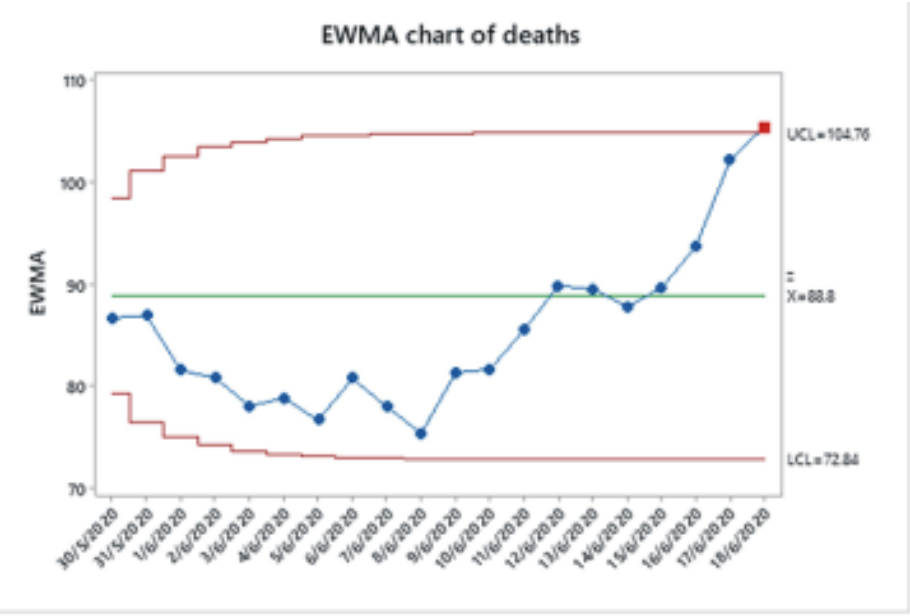
Λόγω της χρήσης της παραμέτρου t, τα παραπάνω όρια ελέγχου δεν είναι σταθερά αλλά μεταβλητά, αν και όσο αυξάνεται το t, η ποσότητα $(1-\lambda)^{2t}$ τείνει στο μηδέν, οπότε και τα όρια σταθεροποιούνται. Έτσι, τα σταθερά όρια ελέγχου δίνονται από τον παρακάτω πίνακα:

***EWMA* Διάγραμμα Ελέγχου**
L* Όρια Ελέγχου Φάσης *II
$\text{UCL} = \mu_{z_t} + L\sigma_{z_t} = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}}$ $\text{CL} = \mu_0$ $\text{LCL} = \mu_{z_t} - L\sigma_{z_t} = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}}$

Πίνακας 3.6.2.2. *L* Όρια ελέγχου *EWMA* διαγράμματος Φάσης *II*

Τέλος, αξίζει να σημειωθεί ένα βασικό πλεονέκτημα των διαγραμμάτων *EWMA*, ιδιαίτερα έναντι των διαγραμμάτων τύπου *Shewhart*. Καθώς, όπως προαναφέρθηκε τα διαγράμματα *EWMA* χρησιμοποιούν τον σταθμισμένο μέσο της πιο πρόσφατης παρατήρησης, αλλά και των προγενέστερων αυτής, είναι αρκετά ανθεκτικά στην παραβίαση της κανονικότητας των δεδομένων και γι' αυτό τον λόγο αναφέρονται πολύ συχνά ως *distribution free charts*.

Στο παρακάτω διάγραμμα, πραγματοποιήθηκε παρακολούθηση των θανάτων από Covid – 19, στην Ινδία, για την περίοδο 10/05/2020 έως 18/06/2020. Η αναμενόμενη μέση τιμή για την διεργασία έχει οριστεί στο 88.8, ενώ ως ανώτατο όριο 104.7. Παρατηρούμε, ότι ενώ από την αρχή της παρακολούθησης οι ημερήσιοι θάνατοι κυμαίνονται κάτω από το μέσο επίπεδο της διεργασίας, λαμβάνουμε για πρώτη φορά ένδειξη εκτός ελέγχου διεργασίας την στις 18/06 όπου παρατηρείται μεγάλη αύξηση των ημερήσιων καταγραφών θανάτου λόγω του ιού, δηλαδή ο μέσος της διεργασίας μετατοπίζεται πάνω από το ανώτατο όριο. Έτσι, θα έπρεπε οι ειδικοί υγείας την περίοδο αυτή, να ερευνήσουν για τυχόν ύπαρξη ειδικών αιτιών μεταβλητότητας, οι οποίοι έχουν οδηγήσει στην αλματώδη αύξηση των θανάτων λόγω *Covid - 19*.



Διάγραμμα 3.6.2.1. *EWMA* Διάγραμμα Ελέγχου για τους καταγεγραμμένους θανάτους στο Πακιστάν

Πηγή*: Mahmood. et al.* (2020)

3.7 Διαγράμματα Ελέγχου για Πολυμεταβλητά Δεδομένα

Παραπάνω μελετήσαμε διαγράμματα που αφορούν δεδομένα τα οποία αντιπροσωπεύουν μόνο μία μεταβλητή. Ωστόσο, πολλές φορές βρισκόμαστε αντιμέτωποι με δεδομένα που σχετίζονται με παραπάνω από μία μεταβλητές. Το 1947, ο *Harold Hotelling*, εισήγαγε τα διαγράμματα ελέγχου για πολυμεταβλητά δεδομένα, τα T^2 διαγράμματα ελέγχου ή *Hotelling* T^2 *Control Charts*.

3.7.1 Η περίπτωση των Υποομάδων

Έστω ότι μας ενδιαφέρει η μελέτη ενός γεγονότος σε m ανεξάρτητα δείγματα μεγέθους n το καθένα, n > 1, μέσω p συσχετιζόμενων ποιοτικών χαρακτηριστικών, $X_1, X_{2,\dots}, X_p$. Θεωρούμε ότι τα X_i προέρχονται από την p – διάστατη κανονική κατανομή.

Για την κατασκευή του διαγράμματος Φάσης *I*, η παράμετρος μ της μέσης τιμής καθώς και ο πίνακας *S* των διακυμάνσεων και συνδιακυμάνσεων δεν είναι γνωστοί και θα πρέπει να εκτιμηθούν.

Για τον σκοπό αυτό, η στατιστική συνάρτηση η οποία χρησιμοποιείται για την κατασκευή του T^2 χρησιμοποιείται η στατιστική συνάρτηση:

$$T^2 = n\left(\overline{\mathrm{X}} - \bar{\bar{X}}\right)' S^{-1}\left(\overline{\mathrm{X}} - \overline{\overline{\mathrm{X}}}\right)$$

Τα όρια ελέγχου που προκύπτουν για την παρακολούθηση της διεργασίας στη Φάση *I* είναι τα ακόλουθα:

T^2**Διάγραμμα Ελέγχου**
Όρια Ελέγχου Φάσης *I*
UCL = $\frac{p(m-1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-p+1}$ LCL = 0

Πίνακας 3.7.1.1. Όρια ελέγχου T^2 διαγράμματος Φάσης *I*

Για την κατασκευή του διαγράμματος Φάσης *II*, η παράμετρος μ της μέσης τιμής καθώς και ο πίνακας Σ των διακυμάνσεων και συνδιακυμάνσεων είναι γνωστές. Τα όρια ελέγχου που προκύπτουν για την παρακολούθηση της διεργασίας στη Φάση *II* είναι τα ακόλουθα:

T^2**Διάγραμμα Ελέγχου**
Όρια Ελέγχου Φάσης *II*
UCL = $\frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,\mathrm{p},\mathrm{mn-p+1}}$ LCL = 0

Πίνακας 3.7.1.2. Όρια ελέγχου T^2 διαγράμματος Φάσης *II*

3.7.2 Η περίπτωση των μεμονωμένων παρατηρήσεων $n = 1$

Στην περίπτωση όπου το κάθε δείγμα αποτελείται από $n = 1$ παρατηρήσεις, τότε η στατιστική συνάρτηση που χρησιμοποιείται σε ένα διάγραμμα *Hotelling* T^2 είναι η:

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X})$$

Έτσι, τα όρια ελέγχου που διαμορφώνονται για την Φάση *II* είναι τα παρακάτω:

T^2**Διάγραμμα Ελέγχου**
Όρια Ελέγχου Φάσης *II*
UCL = $p\frac{(m+1)(m-1)}{m^2-mp} F_{\alpha,p,m-p+1}$ LCL = 0

Πίνακας 3.7.2.1. Όρια ελέγχου T^2 διαγράμματος Φάσης *II* για $n = 1$

Ωστόσο, σημειώνεται πως όταν το πλήθος των δειγμάτων είναι τέτοιο ώστε $m > 100$, τότε προτείνεται να χρησιμοποιηθούν τα παρακάτω όρια ελέγχου:

T^2**Διάγραμμα Ελέγχου**
Όρια Ελέγχου Φάσης *II* ($m > 100$)
UCL = $p\frac{(m-1)}{m-p} F_{\alpha,p,m-p}$ ή $\chi^2_{\alpha,\rho}$ LCL = 0

Πίνακας 3.7.2.2. Όρια ελέγχου T^2 διαγράμματος Φάσης *II* για πλήθος δειγμάτων $m > 100$

Σημειώνεται, πως κατά τους *Tracy et al.* (2018), στην Φάση *I*, τα όρια ελέγχου θα πρέπει να υπολογίζονται με βάση την κατανομή *Βήτα*. Έτσι, τα όρια ελέγχου διαμορφώνονται ως εξής:

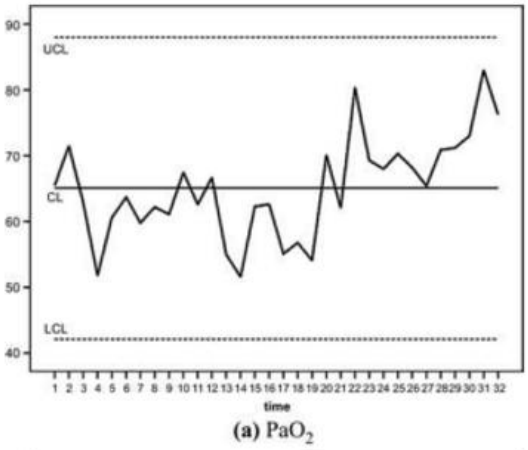
T^2**Διάγραμμα Ελέγχου**
Όρια Ελέγχου Φάσης *I*
UCL = $\frac{(m-1)^2}{m}\beta_{\alpha,p/2,(m-p-1)/2}$ LCL = 0

Πίνακας 3.7.2.3. Όρια ελέγχου T^2 διαγράμματος Φάσης *I*

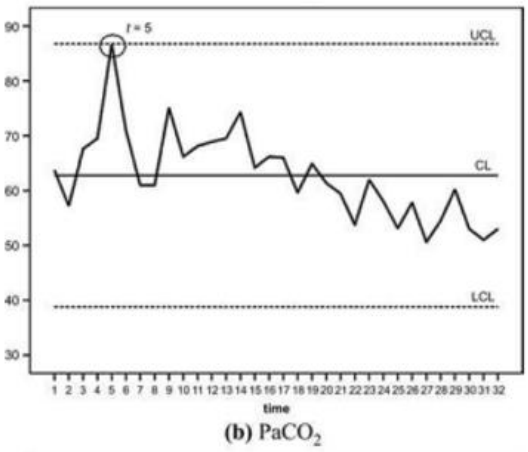
όπου $\beta_{\alpha,p/2,(m-p-1)/2}$ είναι το άνω α – ποσοστιαίο σημείο της κατανομής *Βήτα*, με παραμέτρους *p/2* και *(m-p-1)/2*.

Η μελέτη των *Correia et al*. (2011) αποτελεί χαρακτηριστικό παράδειγμα της χρήσης των T^2 διαγραμμάτων ελέγχου έναντι των κλασικών διαγραμμάτων *Shewhart* για μεταβλητές. Ειδικότερα μελέτησαν τους δείκτες PaO_2, $PaCO_2$ και *BMI* για δυο ασθενείς με χρόνια αποφρακτική πνευμονοπάθεια, καθώς οι δείκτες αυτοί μπορούν να αξιολογήσουν την κατάσταση υγείας του ασθενή.

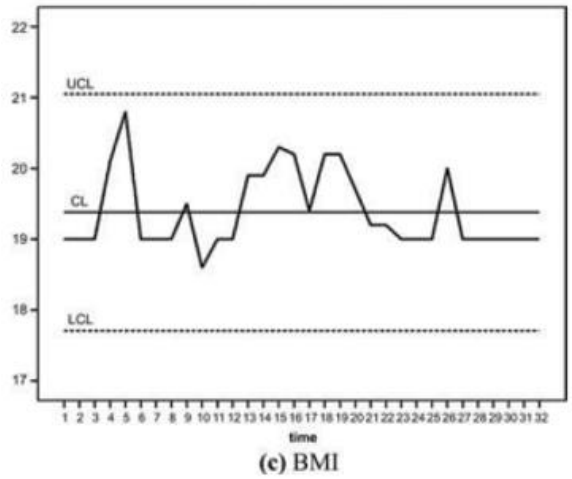
Αρχικά, χρησιμοποιήθηκαν διαγράμματα Shewhart για τους τρεις δείκτες ξεχωριστά.



(a) PaO_2



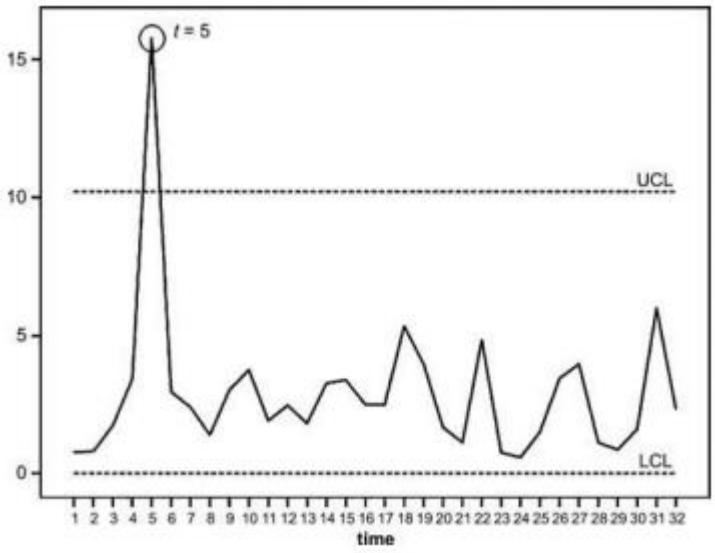
(b) $PaCO_2$



Πίνακας 3.7.2.1. Διαγράμματα *Shewhart* για την παρακολούθηση δεικτών PaO_2, $PaCO_2$ και *BMI*
Πηγή: *Correia et al.* (2011)

Από τα παραπάνω διαγράμματα είναι φανερό πως αν και για τους δείκτες PaO_2 και *BMI* δεν υπάρχουν σημεία εκτός των ορίων ελέγχου, για τον δείκτη $PaCO_2$ το σημείο για $t = 5$ είναι πάνω στο άνω όριο ελέγχου, χωρίς να μας δίνεται ξεκάθαρη εικόνα για το αν η διεργασία είναι εκτός στατιστικού ελέγχου.

Η παραπάνω εικόνα ωστόσο φαίνεται να αλλάζει όταν γίνεται χρήση του διαγράμματος T^2 για την ταυτόχρονη παρακολούθηση και των τριών δεικτών.



Πίνακας 3.7.2.2. T^2 Διάγραμμα για την ταυτόχρονη παρακολούθηση δεικτών PaO_2, $PaCO_2$ και *BMI*
Πηγή: *Correia et al.* (2011)

Είναι φανερό πως με την ταυτόχρονη απεικόνιση και των τριών δεικτών, την χρονική στιγμή $t = 5$ υπάρχει σημείο εκτός του άνω ορίου ελέγχου. Λαμβάνουμε έτσι γρηγορότερα ένδειξη εκτός ελέγχου διεργασίας, συμπέρασμα όπου δεν θα μπορούσαμε να εξάγουμε αν μελετούσαμε τους δείκτες μεμονωμένα.

3.8 Διαγράμματα Ελέγχου Προσαρμοσμένα στον Κίνδυνο

Παραπάνω παρουσιάστηκαν τα διαγράμματα ελέγχου που μπορούν να αξιοποιηθούν για χρήση στην Βιοεπιτήρηση. Ωστόσο, σε κανένα από αυτά δεν γίνεται λόγος, ούτε λαμβάνονται υπόψιν οι επιβαρυντικοί παράγοντες που επηρεάζουν την υγεία των ασθενών αλλά και τις πιθανότητες να ξεπεράσουν την υπό μελέτη ασθένεια. Κάτι τέτοιο είναι υψίστης σημασία, καθώς σε αντίθεση με τα βιομηχανικά προϊόντα, τα δεδομένα υγείας που έχουμε στην διάθεσή μας, διαφοροποιούνται ως προς κάποια πολύ βασικά χαρακτηριστικά. Για παράδειγμα, ο τύπος της νόσησης από *Covid* – 19, διαφέρει από ασθενή σε ασθενή, καθώς εξαρτάται από την ηλικία των νοσούντων, τα προϋπάρχοντα προβλήματα υγείας κλπ. Έτσι, είναι απαραίτητο σε κάποιες περιπτώσεις μελέτης, να λαμβάνονται υπόψιν οι διάφοροι επιβαρυντικοί παράγοντες.

Η μελέτη των *Alemi et al.* (1990) για τα διαγράμματα ελέγχου προσαρμοσμένα στον κίνδυνο, στηρίχθηκε στο ποσοστό της θνησιμότητας. Ως ποσοστό θνησιμότητας ορίζουμε τον αριθμό των καταγεγραμμένων θανάτων προς τον συνολικό αριθμό των νοσούντων. Η βασική διαφορά με τα διαγράμματα ελέγχου που έχουν ήδη προταθεί, είναι πως τα διαγράμματα ελέγχου προσαρμοσμένα στον κίνδυνο, δεν στηρίζονται στον παρατηρούμενο αριθμό θανάτων, δηλαδή στο παρατηρούμενο ποσοστό θνησιμότητας, αλλά στο αναμενόμενο ποσοστό.

Για την κατασκευή των διαγραμμάτων, ακολουθούνται τα παρακάτω βήματα. Αρχικά, γίνεται υπολογισμός των αναμενόμενων θανάτων μετά την προσαρμογή των διάφορων επιβαρυντικών παραγόντων κινδύνου. Συγκεκριμένα, αυτό πετυχαίνετε μέσω κατασκευής κατάλληλου μοντέλου παλινδρόμησης, στο οποίο, η εξαρτημένη μεταβλητή θα είναι αν ο ασθενής επιβιώνει ή όχι (0 αν επιβιώνει – 1 αν αποβιώνει), ενώ ανεξάρτητες μεταβλητές θα είναι οι παράγοντες κινδύνου για την ασθένεια, όπως αυτοί προκύπτουν από προγενέστερες ιατρικές μελέτες. Στην συνέχεια, υπολογίζονται το αναμενόμενο ποσοστό

θνησιμότητας αλλά η τυπική απόκλιση του αναμενόμενου ποσοστού, μέσω των παρακάτω τύπων:

$$\hat{P}_i = \sum_{j=1}^{n_i} \frac{\hat{P}_{ij}}{n_i}$$

και

$$\hat{S}_i = \frac{\sqrt{\sum_{j=1}^{n_i} \hat{P}_{ij}(1-\hat{P}_{ij})}}{n_i}$$

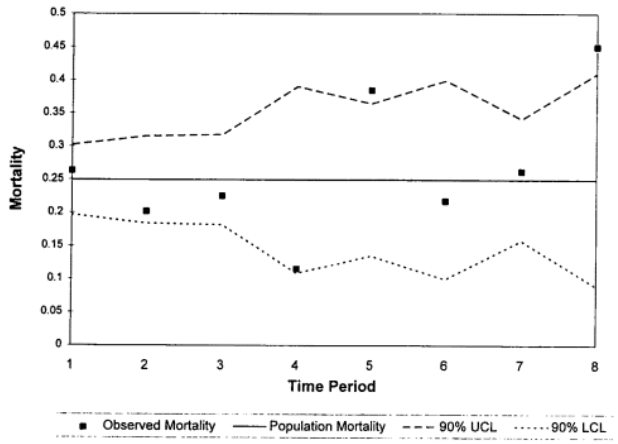
Όπου $\hat{P}_{ij}$ η πιθανότητα να αποβιώσει ο j ασθενής του δείγματος i και $\hat{P}_i$ η πιθανότητα να αποβιώσει ασθενής στο δείγμα i.

Με βάση τα παραπάνω, τα μεταβλητά όρια ελέγχου είναι τα παρακάτω:

$$UCL_i = \hat{P}_i + t_{\frac{\alpha}{2}}(\hat{s}_i)$$

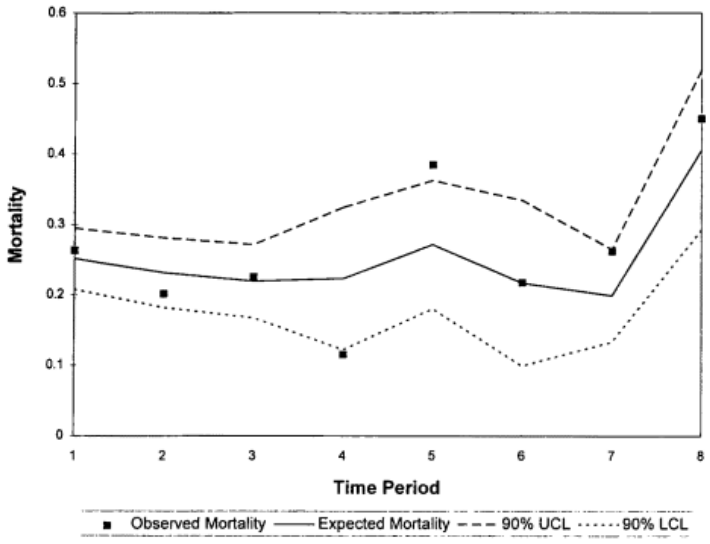
$$LCL_i = \hat{P}_i - t_{\frac{\alpha}{2}}(\hat{s}_i)$$

Παρακάτω δίνονται δυο διαγράμματα ελέγχου, ένα μη – προσαρμοσμένο στον κίνδυνο και ένα προσαρμοσμένο.



Διάγραμμα 3.8.1. Μη – προσαρμοσμένο στον κίνδυνο Διάγραμμα Ελέγχου για το ποσοστό θνησιμότητας

Πηγή: *Alemi et al.* (1996)



Διάγραμμα 3.8.2. Προσαρμοσμένο στον κίνδυνο Διάγραμμα Ελέγχου για το ποσοστό θνησιμότητας

Πηγή: *Alemi et al.* (1996)

Το μη – προσαρμοσμένο στον κίνδυνο διάγραμμα 3.8.1. κατασκευάστηκε με βάση τον παρατηρούμενο αριθμό θανάτων, και συνεπώς τον παρατηρούμενο ποσοστό θνητότητας. Σε αντίθεση, το προσαρμοσμένο στον κίνδυνο διάγραμμα 3.8.2. κατασκευάστηκε χρησιμοποιώντας τον αναμενόμενο αριθμό θανάτων. Ενώ στο διάγραμμα 3.8.1. παρατηρούμε ότι για την $5^{η}$ και $8^{η}$ περίοδο, τα σημεία βρίσκονται εκτός των ορίων, δίνοντας μας ένδειξη ότι η διεργασία βρίσκεται εκτός ελέγχου, στο διάγραμμα 3.8.2. κατά την $8^{η}$ περίοδο δεν λαμβάνουμε ένδειξη εκτός ελέγχου διεργασίας. Πιθανότατα, κατά την περίοδο αυτή κάποιοι ασθενείς να νοσήσαν πολύ βαριά συγκριτικά με άλλες περιόδους, αιτιολογώντας το μεγάλο ποσοστό θνητότητας αυτό το χρονικό διάστημα. Επιπλέον, στο διάγραμμα 3.8.2, στην $4^{η}$ περίοδο το σημείο βρίσκεται εκτός του κάτω ορίου ελέγχου, και καθώς αναφερόμαστε σε αριθμό θανάτων, αυτό μας δίνει ένδειξη ότι ενδεχομένως κατά την περίοδο αυτή, οι υπηρεσίες υγείας είναι πολύ καλές.

Τέλος, μελέτες σχετικά με τα διαγράμματα ελέγχου προσαρμοσμένα στο κίνδυνο έχουν προταθεί από πολλούς ερευνητές, με κάποιους από αυτούς να είναι οι *Hendryx et al.* (1999), οι οποίοι προσάρμοσαν τα διαγράμματα σε ασθενείς ψυχιατρικών νόσων, καθώς και οι *Koetsier et al.* (2012), οι οποίοι με βάση τα διαγράμματα ελέγχου προσαρμοσμένα

στον κίνδυνο προσπάθησαν να επιτηρήσουν ποσοστά θνησιμότητας ασθενών σε μονάδες εντατικής θεραπείας.

ΚΕΦΑΛΑΙΟ 4

Χρονοσειρές

4.1 Εισαγωγή

Το 2019 σημαδεύτηκε από την έναρξη της επιδημίας του *Covid* – 19, όταν το πρώτο γνωστό κρούσμα σημειώθηκε στην Κίνα. Αν και στην αρχή, οι κυβερνήσεις ανά τον κόσμο, καθώς και οι οργανισμοί υγείας αντέδρασαν με ιδιαίτερα γρήγορα αντανακλαστικά, κανείς δεν περίμενε την τόσο γρήγορη εξάπλωση της νόσου σε παγκόσμιο επίπεδο, έως ότου το 11 Μαρτίου , ο Παγκόσμιος Οργανισμός Υγείας απέδωσε στον κορονοϊό τον χαρακτηρισμό πανδημία. Έτσι, έγινε επιτακτική η ανάγκη αποτελεσματικής διαχείρισης της πανδημίας μέσω της πρόβλεψης της εξέλιξής της, προκειμένου να μπορούν να λαμβάνονται μέτρα που θα μας φέρνουν ένα βήμα πιο μπροστά από αυτή.

Σύμφωνα με τους *Montgomery et al.*(2015) ως πρόβλεψη ορίζεται η ικανότητα να μπορούμε να διακρίνουμε τι θα γίνει στο μέλλον. Μέσω των στατιστικών μεθόδων πρόβλεψης, γίνεται προσπάθεια πρόγνωσης των παραγόντων που επηρεάζουν την εξέλιξη του γεγονότος που μας ενδιαφέρει να μελετήσουμε. Η χρήση τους, επεκτείνεται πίσω στον 19ο αιώνα, μέσω της ανάλυσης παλινδρόμησης, ενώ πιο πρόσφατα, τον 20ο αιώνα αναπτύχθηκε και η μεθοδολογία των *Box – Jenkins* με την οποία θα ασχοληθούμε εκτενώς στο παρόν κεφάλαιο.

Φορείς σχετικοί με την δημόσια υγεία αλλά και ανεξάρτητοι οργανισμοί καταγράφουν ανά τακτά χρονικά διαστήματα επιβεβαιωμένα κρούσματα και δεδομένα που αφορούν περιστατικά που μας ενδιαφέρουν, όπως κρούσματα *Covid – 19, H*1*N*1 κλπ.. Τα δεδομένα αυτά συχνά έχουν την μορφή χρονοσειρών. Η χρήση και η ανάλυσή τους αποτελούν κλειδί στην κατανόηση της εξέλιξης του υπό μελέτη φαινομένου και συνεπώς στην πρόβλεψη της πορείας του.

4.2 Χρονοσειρές

4.2.1 Ορισμός Χρονοσειράς

Με τον όρο χρονοσειρά ή χρονική σειρά (*time series*) ορίζεται η ακολουθία παρατηρήσεων ενός χαρακτηριστικού που μας ενδιαφέρει, οι οποίες λαμβάνονται ανά τακτά χρονικά διαστήματα (ωριαίες παρατηρήσεις, ημερήσιες, μηνιαίες κλπ). Ειδικότερα, μια χρονοσειρά συμβολίζεται ως $\{X_t, t\epsilon T\}$, όπου X_t η τιμή την χρονική στιγμή t της υπό μελέτη μεταβλητής X του συστήματος που μας ενδιαφέρει, ενώ ως T ορίζεται το σύνολο των χρονικών στιγμών που λαμβάνονται οι μετρήσεις.

Ο μηχανισμός παραγωγής μιας χρονοσειράς μπορεί να θεωρηθεί ως μια στοχαστική διαδικασία, δηλαδή ως ένα σύνολο τυχαίων τιμών $\{X_t, t = 1,2,3, \cdots\}$. Ως στοχαστική διαδικασία, μία χρονοσειρά μπορεί να προσδιοριστεί από την συνάρτηση κατανομής της. Ωστόσο, πρακτικά δεν είναι δυνατόν να βρεθεί η συνάρτηση αυτή. Για τον λόγο αυτό, μία χρονοσειρά ορίζεται μέσω των πρώτων και δεύτερων ροπών της για κάθε χρονική στιγμή, δηλαδή μέσω της μέσης τιμής της $E(X_t) = \mu_t$, της διακύμανσης $Var(X_t) = \sigma_t^2$ αλλά και της αυτοδιακύμνασης $Cov(X_t, X_s) = \gamma_{t,s}, t \neq s$.

Ανάλογα με το είδος του συνόλου T, οι χρονοσειρές κατηγοριοποιούνται σε δυο μεγάλες ομάδες: τις χρονοσειρές διακριτού χρόνου, όπου τα δεδομένα καταγράφονται σε ορισμένες και διακριτές χρονικές στιγμές και το σύνολο T είναι της μορφής $T = \{0,1,2, \dots\}$, και σε χρονοσειρές συνεχούς χρόνου, όπου υπάρχει συνεχής καταγραφή των μετρήσεων μέσα στον χρόνο, δηλαδή το σύνολο T είναι της μορφής $T = (0, +\infty)$. Μια επιπλέον κατηγοριοποίηση των χρονοσειρών βασίζεται στο πλήθος των υπό μελέτη χαρακτηριστικών. Συγκεκριμένα, όταν μας ενδιαφέρει η μελέτη μιας μόνο μεταβλητής του υπό μελέτη χαρακτηριστικού, τότε η χρονοσειρά καλείται μονοδιάστατη, ενώ όταν καταγράφεται η τιμή περισσότερων της μίας μεταβλητών, η χρονοσειρά ονομάζεται πολυδιάστατη.

4.2.2 Συνθετικά Στοιχεία Χρονοσειρών

Η ανάλυση χρονοσειρών μελετά τη συμπεριφορά ενός ή περισσοτέρων χαρακτηριστικών ενός συστήματος, αγνοώντας την παρουσία άλλων μεταβλητών που ενδεχομένως να επηρεάζουν τη συμπεριφορά του, καθώς θεωρείται πως το σύνολο της επίδρασης τους, εμπεριέχεται με έμμεσο τρόπο μέσα στις τιμές τις χρονοσειράς. Κύριος στόχος της ανάλυσης χρονοσειρών αποτελεί η αναζήτηση ενός μοντέλου, το οποίο μπορεί να περιγράψει ικανοποιητικά τον μηχανισμό της στοχαστικής διαδικασίας που δημιουργεί την χρονοσειρά. Συγκεκριμένα, γίνεται προσπάθεια αναγνώρισης του μηχανισμού εκείνου που παράγει τις διάφορες τιμές της χρονοσειράς στο παρελθόν, προκειμένου να χρησιμοποιηθεί για την μελέτη και την πρόβλεψη των μελλοντικών τιμών της.

Υπάρχουν τρεις βασικές μέθοδοι ανάλυσης, οι οποίες χρησιμοποιούνται για σκοπούς πρόβλεψης:

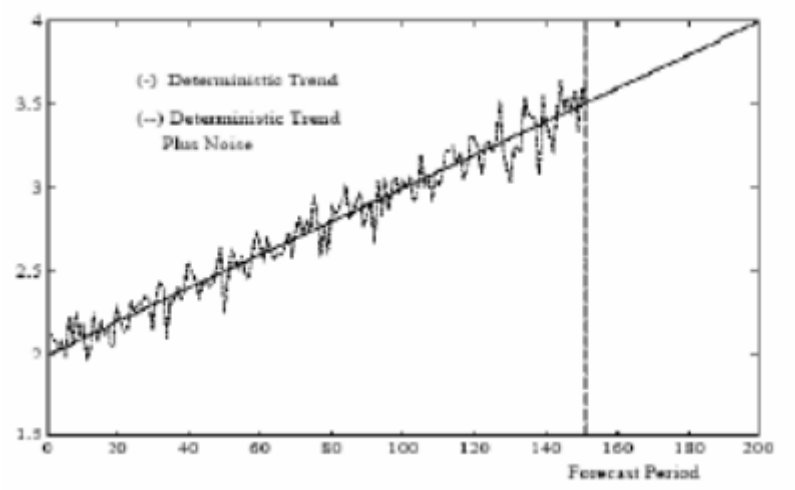
- Μέθοδοι Εξομάλυνσης
- Διάσπαση Χρονοσειρών
- Ανάλυση *ARIMA*

Οι διαφορές των τριών αυτών μεθόδων εντοπίζονται στις διαφορετικές συναρτησιακές σχέσεις που χρησιμοποιούν για να περιγράψουν το μηχανισμό παραγωγής των τιμών της χρονοσειράς. Οι πρώτες δυο αποτελούν ντετερμινιστικές – προσδιοριστικές μέθοδοι ανάλυσης, ενώ η τελευταία είναι στοχαστική διαδικασία, η οποία θα αναλυθεί περαιτέρω στο παρόν κεφάλαιο για τους σκοπούς της διπλωματικής εργασίας.

Πριν προχωρήσουμε στην μοντελοποίηση και ανάλυση των χρονοσειρών, είναι αναγκαίο να περιγραφούν τα συνθετικά στοιχεία και χαρακτηριστικά της, τα οποία είναι: η τάση, η εποχικότητα, η κυκλικότητα και η μη – κανονικότητα. Όσο καλύτερα γνωρίζουμε τα στοιχεία αυτά, τόσο ευκολότερο είναι να αντιληφθούμε τον τρόπο και μηχανισμό παραγωγής των τιμών του χαρακτηριστικού που μελετάμε, καθώς και να εξάγουμε ασφαλέστερες αλλά και πιο τεκμηριωμένες προβλέψεις.

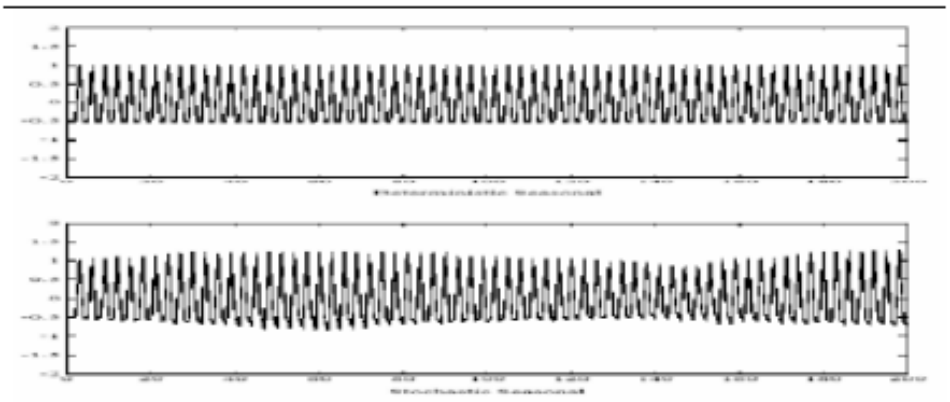
Ως τάση (*Tr -Trend*) ορίζεται η μακροχρόνια γενική κίνηση που ακολουθεί η χρονοσειρά, και αναπαριστά την αύξηση ή πτώση των τιμών της σε μία εκτεταμένη χρονική περίοδο. Προκειμένου να μετρηθεί η τάση χρησιμοποιούνται υποδείγματα που

χρησιμοποιούν ως εξαρτημένη μεταβλητή τις τιμές της χρονοσειράς ενώ ως ανεξάρτητη μεταβλητή χρησιμοποιούν τον χρόνο.



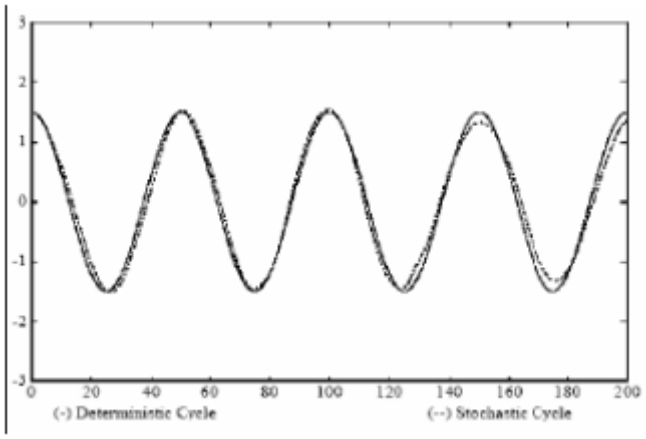
Διάγραμμα 4.2.2.1. Διάγραμμα Χρονοσειράς με Τάση
Πηγή: Κυριακίδης Μ. (2017)

Ως εποχικότητα (*S - Seasonality*) χαρακτηρίζονται οι διακυμάνσεις στις τιμές της χρονοσειράς, οι οποίες εμφανίζονται κατά την διάρκεια του έτους και επαναλαμβάνονται με την ίδια ή περίπου με την ίδια μορφή από έτος σε έτος. Η εμφάνιση εποχικότητας, είναι πιθανόν να επιφέρει μεγάλη μεταβλητότητα στη μέση τιμή της χρονοσειράς, ανάλογα και με την περίοδο του έτους που αντιστοιχεί. Επειδή η εποχικότητα σε μία χρονοσειρά είναι σταθερή και έχει γνωστή συχνότητα εμφάνισης, είναι εύκολο να προσδιοριστεί.



Διάγραμμα 4.2.2.2. Διάγραμμα Χρονοσειράς με Εποχικότητα
Πηγή: Κυριακίδης Μ. (2017)

Η κυκλικότητα (*C - Cyclic*) αντιπροσωπεύει το σύνολο των επαναλαμβανόμενων κυμάνσεων γύρω από την τάση. Συγκεκριμένα, αυτό το συνθετικό στοιχείο της χρονοσειράς εμφανίζεται ακανόνιστα σε κυματοειδή μορφή και έχει διάρκεια αρκετά μεγαλύτερη του ενός έτους. Στο σημείο αυτό αξίζει να σημειωθεί πως, όταν οι διακυμάνσεις μιας χρονοσειράς δεν είναι σταθερής περιόδου και έχουν διάρκεια μεγαλύτερη του ενός έτους, τότε χαρακτηρίζονται ως κυκλικές, ενώ όταν έχουν σταθερή περιοδικότητα, η οποία φαίνεται να σχετίζεται με κάποιο ημερολογιακό πρότυπο, τότε στην χρονοσειρά υπάρχει εποχικότητα.



Διάγραμμα 4.2.2.3. Διάγραμμα Χρονοσειράς με Κυκλικότητα
Πηγή: Κυριακίδης Μ. (2017)

Τέλος, η μη – κανονικότητα (*I*) σε μία χρονοσειρά, αφορά τις επιδράσεις πάνω στις τιμές της χρονοσειράς, οι οποίες συμβαίνουν κατά τυχαίο, μη- συστηματικό και μη – επαναλαμβανόμενο τρόπο, και επομένως είναι αδύνατον να προσδιοριστούν και να αποδοθούν στην τάση, την εποχικότητα ή την κυκλικότητα.

4.2.3 Στασιμότητα

Μια χρονοσειρά καλείται αυστηρώς στάσιμη (*strictly stationary*) όταν το σύνολο των στατιστικών της ιδιοτήτων παραμένουν σταθερές και δεν επηρεάζονται από τις αλλαγές στον χρόνο. Ειδικότερα, η αυστηρή στασιμότητα υποδηλώνει ότι η από κοινού κατανομή

της χρονοσειράς για τις παρατηρήσεις $X_1, X_{2,\dots}, X_n$ θα είναι ίδια με την από κοινού κατανομή των παρατηρήσεων $X_{k+1}, X_{k+2,\dots}, X_{k+n}$, δηλαδή:

$$f(x_1, x_2, \dots, x_n) = f\left(x_{k+1}, x_{k+2,\dots}, x_{k+n}\right)$$

Ωστόσο, καθώς είναι πρακτικά αδύνατο να εξασφαλιστούν οι συνθήκες σταθερότητας για το σύνολο των ροπών της χρονοσειράς, για την εξασφάλιση της στασιμότητας περιοριζόμαστε στην σταθερότητα και την ανεξαρτησία από τον χρόνο των δυο πρώτων ροπών της χρονοσειράς. Συγκεκριμένα, μια χρονοσειρά καλείται ασθενώς στάσιμη (*weakly stationary*) όταν σε κάθε χρονική στιγμή η μέση τιμή είναι σταθερή, η διακύμανση είναι σταθερή και πεπερασμένη και η αυτοδιακύμανση εξαρτάται μόνο από την χρονική υστέρηση και όχι από το συγκεκριμένο χρονικό σημείο που αυτές υπολογίζονται, δηλαδή:

$$E(X_t) = E(X_{t-k}) = \mu, \forall\, t, k \in T,$$

$$Var(X_t) = Var(X_{t-k}) = \sigma_X^2 < +\infty,$$

και

$$Cov(X_t, X_{t-k}) = E[(X_t - \mu)(X_{t-k} - \mu)] = \gamma_k \text{ , } k = 1, 2, \dots$$

Η εξασφάλιση της συνθήκης στασιμότητας αποτελεί ένα από τα βασικότερα βήματα, πριν προχωρήσουμε στην ανάλυση της χρονοσειράς, καθώς για να προχωρήσουμε σε πραγματοποίηση προβλέψεων, πρέπει να είμαστε σε θέση να μπορούμε να παράγουμε τις παρελθοντικές τιμές με σχετική ακρίβεια. Γενικότερα, η ύπαρξη μη – στασιμότητας αποτελεί βασικό πρόβλημα στην δημιουργία προβλέψιμων προτύπων συμπεριφοράς. Η ύπαρξη τάσης ή εποχικότητας στις τιμές της χρονοσειράς φανερώνει μη στάσιμη διαδικασία.

Προκειμένου να ελεγχθεί αν μία χρονοσειρά είναι στάσιμη, χρησιμοποιείται αρχικά το κορελόγραμμα, δηλαδή το διάγραμμα των συσχετίσεων και των μερικών συσχετίσεων των τιμών της χρονοσειράς. Ένδειξη ότι η χρονοσειρά που εξετάζεται δεν είναι στάσιμη αποτελεί το γεγονός πως στο διάγραμμα οι αυτοσυσχετίσεις και μερικές αυτοσυσχετίσεις δεν σβήνουν παρά μόνο πολύ αργά (Αγιακλόγλου, 2018). Άλλοι έλεγχοι που μπορούν να

χρησιμοποιηθούν για τον έλεγχο της στασιμότητας της χρονοσειράς είναι μέσω του τεστ ύπαρξης μοναδιαίας ρίζας των Dickey – Fuller. Για παράδειγμα, έστω η χρονοσειρά :

$$X_t = a + \beta t + \rho X_{t-1} + e_t$$

Με βάση του τεστ μοναδιαίας ρίζας, γίνεται ο έλεγχος υποθέσεων:

H_0: Αν $|\rho| < 1$ και $\beta = 0$, τότε η χρονοσειρά είναι στάσιμη
H_1: Αν $|\rho| \geq 1$ και $\beta \neq 0$, τότε η χρονοσειρά δεν είναι στάσιμη

Αν διαπιστωθεί η μη – στασιμότητα της χρονοσειράς, θα πρέπει, πριν προχωρήσουμε στην δημιουργία κάποιου μοντέλου πρόβλεψης, να την καταστήσουμε στάσιμη. Αυτό πραγματοποιείται μέσω της χρήσης των πρώτων διαφορών, όπου τότε η σειρά ονομάζεται ολοκληρωμένη πρώτης τάξης (*Integrated first order – I*(1)).

Η σειρά που προκύπτει μέσω των πρώτων διαφορών είναι η *I*(1*)*:

$$X'_t = X_t - X_{t-1}$$

Αν και μετά την χρήση των πρώτων διαφορών, η σειρά εξακολουθεί να μην είναι στάσιμη επαναλαμβάνουμε την διαδικασία. Ωστόσο, η χρήση της διαφόρισης πρέπει να γίνεται με προσοχή, καθώς σε κάθε επανάληψη της διαδικασία μειώνεται το πλήθος των δεδομένων που έχουμε διαθέσιμα.

Σε περιπτώσεις όπου η εποχικότητα είναι ιδιαίτερα έντονη, μπορούμε να προχωρήσουμε και σε εποχική διαφόριση. Και πάλι ανάλογα με την τάξη της διαφόρισης, η νέα χρονοσειρά θα είναι της μορφής:

$$X'_t = X_t - X_{t-m}$$

ή

$$X''_t = X'_t - X'_{t-m}.$$

4.2.4 Λευκός Θόρυβος

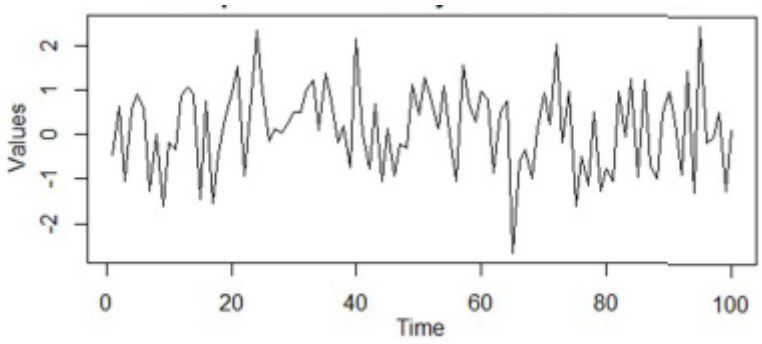
Για να έχει νόημα η οποιαδήποτε ανάλυση διεξάγουμε πάνω στις παρατηρούμενες τιμές μιας χρονοσειράς, θα πρέπει πρώτα να απορρίψουμε την υπόθεση ότι οι τιμές που έχουμε παράγονται με εντελώς τυχαίο τρόπο. Συγκεκριμένα, μια στοχαστική διαδικασία $\{X_t, t = 1,2,3,\cdots\}$ καλείται λευκός θόρυβος (*white noise*) όταν οι παρατηρούμενες τιμές είναι ανεξάρτητες μεταξύ τους και ισχύει:

$$E(X_t) = 0$$

και

$$Cov(X_t, X_S) = \begin{cases} \sigma^2, & t = s \\ 0 \ , & t \neq s. \end{cases}$$

Αξίζει να σημειωθεί πως, ο λευκός θόρυβος θεωρείται στάσιμη στοχαστική διαδικασία καθώς έχει μηδενική μέση τιμή – δηλαδή δεν εμφανίζεται τάση ή εποχικότητα – αλλά και σταθερή διασπορά.



Διάγραμμα 4.2.4.1. Διάγραμμα Χρονοσειράς Λευκού Θορύβου
Πηγή: *Jebb et. al.* (2015)

4.3 Συντελεστές Αυτοσυσχέτισης και Μερικής Αυτοσυσχέτισης

Η συνάρτηση αυτοσυσχέτισης, όπως ο μέσος και η διακύμανση, παίζει πολύ σημαντικό ρόλο στο στάδιο της αναγνώρισης μιας χρονοσειράς, καθώς μας δίνει πληροφορίες σχετικά με την στοχαστική διαδικασία που περιγράφει την χρονοσειρά αλλά φανερώνει και την χρονική διάρκεια της μνήμης της.

Η συνάρτηση αυτοσυσχέτισης (*Autocorrelation Function – ACF*) μιας χρονοσειράς ορίζεται ως:

$$\rho_k = \frac{\mathrm{COV}(X_t, X_{t-k})}{\sqrt{V(X_t)V(X_{t-k})}} = \frac{\gamma_k}{\gamma_0} \ , \ k = 1, 2, \ldots$$

Όπως φαίνεται από τον παραπάνω τύπου, η αυτοσυσχέτιση μας δείχνει κατά πόσο η τιμή της χρονοσειράς την χρονική περίοδο *t*, εξαρτάται από την τιμή αυτής *k* υστερήσεων πίσω. Λαμβάνει τιμές στο διάστημα [-1,1], ενώ όταν *ACF = 0*, τότε οι δυο παρατηρήσεις δεν έχουν καμία σχέση μεταξύ τους και τα δεδομένα μας είναι τυχαία. Τέλος, όταν στα δεδομένα υπάρχει τάση, τότε η τιμή του *ACF* για μία υστέρηση είναι συνήθως κοντά 1, ενώ όταν υπάρχει εποχικότητα, η αυτοσυσχέτιση παρουσιάζει απότομη συμπεριφορά κοντά στην 4 υστέρηση για δεδομένα που αφορούν τρίμηνο.

Στην μελέτη των χρονοσειρών, υπολογίζεται επίσης και ο συντελεστής μερικής αυτοσυσχέτισης (*Partial Autocorrelation Function – PACF*). Συγκεκριμένα, μέσω του *PACF* υπολογίζεται η συσχέτιση ανάμεσα στις τιμές της χρονοσειράς X_t *και* X_{t-k}, χωρίς να λαμβάνονται υπόψιν οι τυχόν επιδράσεις που μπορεί να έχουν πάνω σε αυτές οι ενδιάμεσες τιμές της $X_{t-1,}\, X_{t-2,}\ldots,\, X_{t-(k-1)}$.

Τέλος, σημειώνεται πως η γραφική απεικόνιση των συντελεστών αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, δίνεται μέσω του κορρελογράμματος. Στο διάγραμμα αυτό, χαρακτηριστικό γνώρισμα των μη - στάσιμων χρονοσειρών είναι πως ο *ACF* και *PACF* δεν μηδενίζονται, αλλά σβήνουν με πολύ αργούς ρυθμούς προς το μηδέν.

4.4 Ανάλυση Μοντέλου Αυτοπαλινδρόμησης *AR(p)*

Σε αντίθεση με τα συνήθη μοντέλα παλινδρόμησης, τα οποία προσπαθούν να περιγράψουν την σχέση μιας εξαρτημένης μεταβλητής με ένα σύνολο παραγόντων που ενδέχεται να επηρεάζουν τις τιμές της, τα μοντέλα αυτοπαλινδρόμησης τάξης *p* (*Autoregressive Models – AR*(*p*)), προσπαθούν να περιγράψουν τις τιμές της υπό μελέτη μεταβλητής μέσω της γραμμικής σχέσης των τιμών της μεταβλητής με τις τιμές που αυτή έχει λάβει στο παρελθόν. Ο όρος αυτοπαλίνδρομο, υποδηλώνει ότι η τιμή της εξαρτημένης μεταβλητής την χρονική στιγμή t, εξαρτάται και παλινδρομείτε ανάμεσα στις τιμές της μεταβλητής τις παρελθοντικές χρονικές στιγμές.

Συγκεκριμένα, ένα μοντέλο αυτοπαλινδρόμησης, δίνεται μέσω της σχέσης:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + e_t$$

Όπου ϕ_0 σταθερός όρος, ϕ_i συντελεστές αυτοσυσχέτισης της χρονοσειράς ενώ ο e_t είναι λευκός θόρυβος. Τα $X_{t-1}, X_{t-2}, \dots, X_{t-p}$ συμβολίζουν τις τιμές της χρονοσειράς για *t -1, t – 2,..., t – p* χρονικές υστερήσεις (*lags*). Αν γίνει χρήση και του τελεστή υστέρησης B, τότε η παραπάνω σχέση γίνεται:

$$(1 - \varphi_1 B - \cdots - \varphi_P B^p) X_t = \varphi_0 + e_t$$

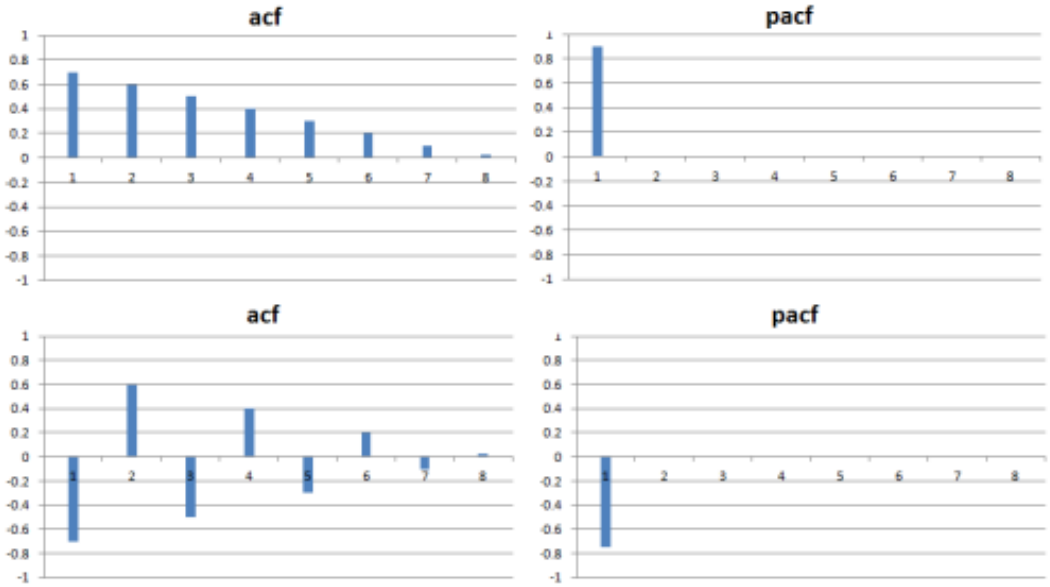
ή

$$\Phi(B) X_t = \varphi_0 + e_t$$

Όπου $\Phi(B)$ το χαρακτηριστικό πολυώνυμο του μοντέλου, και το οποίο μπορεί να χρησιμοποιηθεί για τον έλεγχο της στασιμότητας του $AR(p)$. Συγκεκριμένα, η χρονοσειρά θα θεωρείται στάσιμη, αν οι ρίζες του χαρακτηριστικού πολυωνύμου, βρίσκονται εκτός του μοναδιαίου κύκλου.

Σύμφωνα με το παραπάνω μοντέλο, παρατηρούμε ότι η τιμή της χρονοσειράς την χρονική στιγμή *t*, εξαρτάται από την τιμή της χρονοσειράς τη την χρονική στιγμή $t - i$, πολλαπλασιασμένη κατά ϕ_i. Ο αριθμός p δηλώνει την τάξη του μοντέλου αυτοπαλινδρόμησης, και δηλώνει το μήκος της χρονικής υστέρησης που χρησιμοποιείται για τον υπολογισμό της X_t. Το πιο απλό αυτοπαλινδρομούμενο μοντέλο είναι το $AR(1)$, όπου σε αυτό κάθε παρατήρηση εξαρτάται μόνο από την τιμή που έλαβε η προηγούμενη της.

Παρακάτω δίνονται τα γραφήματα των αυτοσυχετίσεων και μερικών αυτοσυσχετίσεων μοντέλων $AR(1)$ και $AR(2)$ αντίστοιχα. Παρατηρούμε ότι στα μοντέλα αυτοπαλινδρόμησης, οι συντελεστές αυτοσυσχέτισης φθίνουν σταδιακά στο μηδέν, σε αντίθεση με τον συντελεστή μερικής αυτοσυσχέτισης, ο οποίος μηδενίζεται ακαριαία μετά από 1 και 2 αντίστοιχα χρονικές υστερήσεις.



Διάγραμμα 4.4.1. Διαγράμματα Συσχετίσεων και Μερικών Αυτοσυσχετίσεων μοντέλων $AR(1)$ και $AR(2)$
Πηγή: Ασημακόπουλος Β.(2020)

Η επιλογή της παραμέτρου p, βασίζεται στα διαγράμματα των αυτοσυχετίσεων και μερικών αυτοσυσχετίσεων της χρονοσειράς. Συγκεκριμένα, για ένα στάσιμο μοντέλο $AR(p)$ θα ισχύει ότι ο συντελεστής αυτοσυσχέτισης θα φθίνει σταδιακά στο μηδέν, ακολουθώντας εκθετική ή ημιτονοειδή πορεία, σε αντίθεση με τον συντελεστή μερικής αυτοσυσχέτισης, ο οποίος θα μηδενίζεται ακαριαία μετά από p περιόδους υστέρησης.

4.5 Ανάλυση Μοντέλου Κινητού Μέσου Όρου *MA(q)*

Όπως αναφέρθηκε παραπάνω, τα μοντέλα αυτοπαλινδρόμησης θεωρούν την γραμμική σχέση ανάμεσα στην τιμή της χρονοσειράς την χρονική στιγμή t, και των p παρελθοντικών τιμών αυτής. Ωστόσο, τα μοντέλα κινητού μέσου όρου, θεωρούν την ύπαρξη γραμμικής σχέσης ανάμεσα στην χρονοσειρά την χρονική στιγμή t και των παρελθοντικών σφαλμάτων που παρουσίασε η χρονοσειρά.

Ειδικότερα, ένα μοντέλο κινητού μέσου όρου, δίνεται μέσω της σχέσης

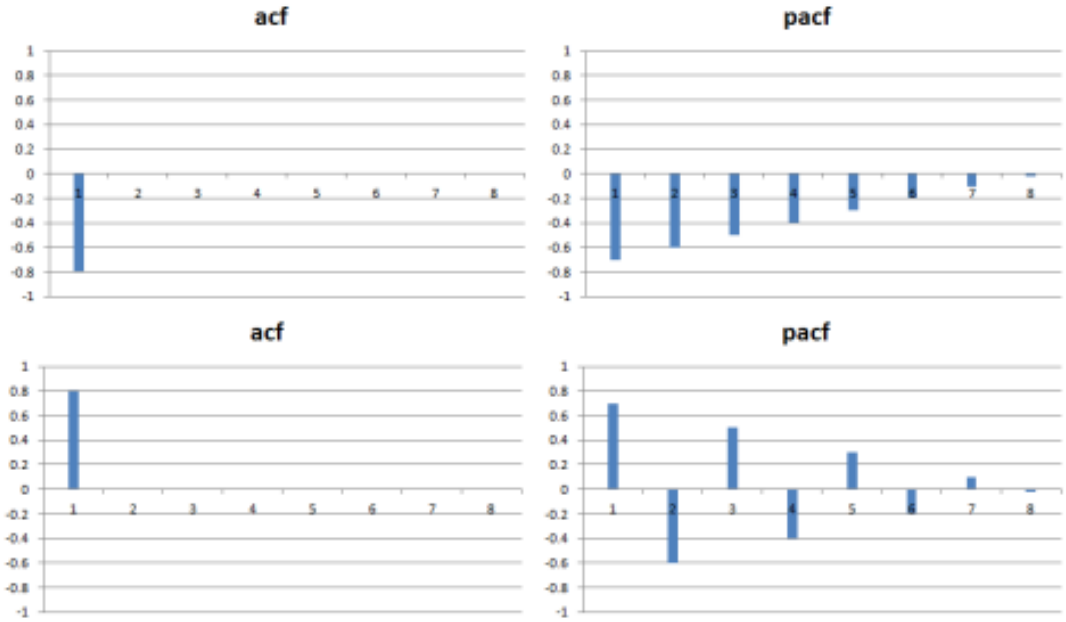
$$X_t = \theta_0 - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t$$

όπου θ_0 σταθερός όρος, θ_i συντελεστές μερικής αυτοσυσχέτισης της χρονοσειράς για i χρονικές υστερήσεις, ενώ ο e_t είναι λευκός θόρυβος. Αν γίνει χρήση και του τελεστή υστέρησης B, τότε η παραπάνω σχέση γίνεται:

$$\bar{X}_t = \left(1 - \theta B - \theta_2 B^2 - \cdots - \theta_q B^q\right) e_t$$

Σημειώνεται πως κάθε μοντέλο AR μπορεί να γραφτεί ως μοντέλο MA άπειρων όρων.

Παρακάτω δίνονται τα γραφήματα των αυτοσυχετίσεων και μερικών αυτοσυσχετίσεων μοντέλων $MA(1)$ και $MA(2)$ αντίστοιχα. Παρατηρούμε ότι στα μοντέλα κινητού μέσου όρου, οι συντελεστές μερικής αυτοσυσχέτισης φθίνουν σταδιακά στο μηδέν, σε αντίθεση με τον συντελεστή αυτοσυσχέτισης, ο οποίος μηδενίζεται ακαριαία μετά από 1 και 2 αντίστοιχα χρονικές υστερήσεις.



Διάγραμμα 4.4.2. Διαγράμματα Συσχετίσεων και Μερικών Αυτοσυσχετίσεων μοντέλων $MA(1)$ και $MA(2)$
Πηγή: Ασημακόπουλος Β. (2020)

Η επιλογή της παραμέτρου q, βασίζεται στα διαγράμματα των αυτοσυχετίσεων και μερικών αυτοσυσχετίσεων της χρονοσειράς. Συγκεκριμένα, για ένα στάσιμο μοντέλο $MA(q)$ θα ισχύει ότι ο συντελεστής μερικής αυτοσυσχέτισης θα φθίνει σταδιακά στο μηδέν,

ακολουθώντας εκθετική ή ημιτονοειδή πορεία, σε αντίθεση με τον συντελεστή αυτοσυσχέτισης, ο οποίος θα μηδενίζεται ακαριαία μετά από q περιόδους υστέρησης.

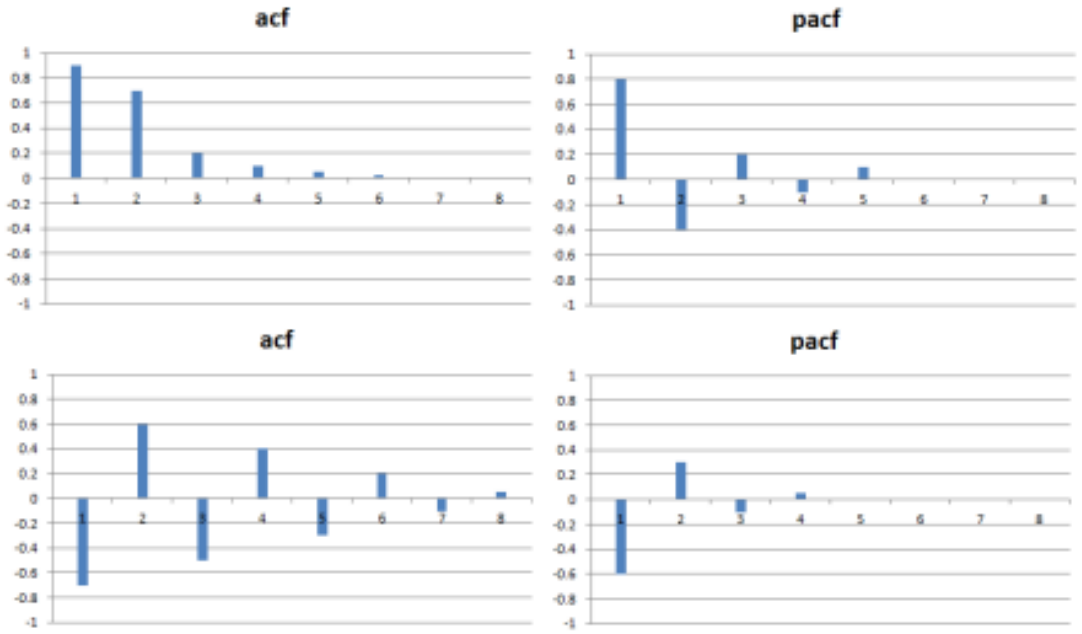
4.6 Ανάλυση Μοντέλου Αυτοπαλινδρόμησης Κινητού Μέσου Όρου *ARMA(p,q)*

Όταν η χρονοσειρά είναι στάσιμη, τα μοντέλα αυτοπαλινδρόμησης και κινητού μέσου όρου, μπορούν να συνδυαστούν κατάλληλα για την πρόβλεψη των τιμών της χρονοσειράς. Ο συνδυασμός των δυο παραπάνω μοντέλων έχει ιδιαίτερη σημασία, καθώς αποδεικνύεται χρήσιμος για τις περιπτώσεις όπου η χρήση των μοντέλων μεμονωμένα, θα απαιτούσε την χρήση μεγάλου αριθμού παραμέτρων.

Στην γενική του μορφή, ένα μοντέλο *ARMA(p,q)* δίνεται από την σχέση:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \theta_0 + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

Παρακάτω δίνονται τα γραφήματα των αυτοσυχετίσεων και μερικών αυτοσυσχετίσεων μοντέλων *ARMA(1,1)*



Διάγραμμα 4.6.1. Διαγράμματα Συσχετίσεων και Μερικών Αυτοσυσχετίσεων μοντέλου *ARMA*(1,1)

Πηγή: Ασημακόπουλος Β. (2020)

Αξίζει να παρατηρήσουμε ότι σε αντίθεση με τα μοντέλα αυτοπαλινδρόμησης και κινητού μέσου όρου, στα μοντέλα *ARMA*, τόσο η αυτοσυσχέτιση, όσο και η μερική αυτοσυσχέτιση δεν μηδενίζονται ακαριαία μετά από *k* χρονικές υστερήσεις, αλλά φθίνουν σταδιακά εκθετικά ή με ημιτονοειδή τρόπο στο μηδέν.

Οι *Reis et al*. (2003) μελετώντας ένα πλήθος μοντέλων *ARMA*, απέδειξαν πως το *ARMA*(2,1) περιγράφει ικανοποιητικά το πλήθος εισαγωγών στα επείγοντα ενός νοσοκομείου. Αντίθετα, το μοντέλο *ARMA*(1,1) επιλέχθηκε για τις επισκέψεις στα επείγοντα που σχετίζονται με το περιστατικά νόσησης του αναπνευστικού συστήματος. Τέλος, οι *Earnest et al*. (2005), μελετώντας τον αριθμό των κλινών που χρησιμοποιούνται σε ένα νοσοκομείο της Σιγκαπούρης, κατά την διάρκεια της επιδημίας του *SARS*, απέδειξαν πως το μοντέλο *ARMA*(1,3) είναι το καταλληλότερο, ενώ ο *Lai* (2005) χρησιμοποίησε τα μοντέλα *AR*(1) και *ARMA*(1,1) για την παρακολούθηση του *SARS* στην Κίνα.

4.7 Μεθοδολογία *Box – Jenkins*

Για τον σχηματισμό των μοντέλων που αναφέρθηκαν παραπάνω, απαραίτητη προϋπόθεση αποτελεί η χρήση στάσιμων χρονοσειρών. Ωστόσο, και ιδιαιτέρα όσον αφορά δεδομένα υγείας για χρήση στην επιδημιολογική επιτήρηση, η στασιμότητα της διαθέσιμης χρονοσειράς δεν είναι σχεδόν ποτέ εξασφαλισμένη. Έτσι, προκειμένου να γίνει η κατάλληλη προσαρμογή μοντέλου στα διαθέσιμα δεδομένα, θα πρέπει να επιτευχθεί η δημιουργία στάσιμης χρονοσειράς και στην συνέχεια να προχωρήσουμε σε δημιουργία μοντέλου πρόβλεψης.

Το 1970 οι μαθηματικοί *George Box* και *Gwilym Jenkins* ανέπτυξαν την ομώνυμη μεθοδολογία για την μελέτη μη – στάσιμων χρονοσειρών, στην προσπάθειά τους να μελετήσουν τις αλλαγές στην συμπεριφορά μιας χρονοσειράς. Βασίστηκαν στις αρχές της αυτοπαλινδρόμησης, του κινούμενου μέσου όρου αλλά και των χρήση των πρώτων διαφορών, μέσω της οποίας επιτυγχάνεται και η ζητούμενη στασιμότητα. Το εξαγόμενο μοντέλο λαμβάνει την ονομασία *ARIMA* (*p,d,q*) όπου *p* και *q* οι συντελεστές που αναφέρθηκαν παραπάνω, ενώ *d* το πλήθος των διαφορών που πρέπει να παραχθούν για την εξάλειψη της τάσης και την επίτευξη της στασιμότητας. Η ύπαρξη στασιμότητας εξασφαλίζει ότι η μέση τιμή, η διακύμανση και η αυτοσυσχέτιση παραμένουν σταθερές και ανεξάρτητες του χρόνου, καθώς μόνο έτσι μπορεί να γίνει η στοχαστική μελέτη της

χρονοσειράς. Έτσι, τα μοντέλα *ARIMA* ανήκουν στην κατηγορία των μοντέλων εκείνων που μπορούν να προσαρμοστούν τόσο σε στάσιμες, όσο και σε μη – στάσιμες χρονοσειρές.

Στην γενική τους μορφή, τα μοντέλα *ARIMA* είναι ο γραμμικός συνδυασμός των παρελθοντικών τιμών της χρονοσειράς, πολλαπλασιασμένων με κατάλληλους στοχαστικούς παράγοντες, και κάποιου τυχαία παράγοντα – σφάλμα πρόβλεψης. Ωστόσο, προτού προχωρήσουμε στην μεθοδολογία των *Box – Jenkins*, θα πρέπει να επισημάνουμε πως αν και τα μοντέλα *ARIMA* είναι ιδιαίτερα αποδοτικά στην δημιουργία βραχυπρόθεσμων προβλέψεων, δεν είναι ιδιαίτερα αξιόπιστα για μακροπρόθεσμες προβλέψεις. Αυτό συμβαίνει καθώς όσο προχωράμε στο μέλλον, οι προβλέψεις που παράγονται δεν βασίζονται σε πραγματικά δεδομένα της χρονοσειράς που διαθέτουμε, αλλά σε προβλέψεις που έχουν προηγηθεί.

Η μεθοδολογία που πρότειναν οι *Box – Jenkins* βασίζεται σε 4 βήματα:

- Της αναγνώρισης
- Της εκτίμησης
- Του διαγνωστικού ελέγχου
- Της πρόβλεψης

Στο στάδιο της αναγνώρισης, γίνεται χρήση των διαγραμμάτων της αυτοσυσχέτισης και μερικής αυτοσυσχέτισης προκειμένου να επιλεχθούν ένα ή και περισσότερα μοντέλα *ARIMA*, τα οποία περιγράφουν ικανοποιητικά την χρονοσειρά μας. Συγκεκριμένα, παρατηρείται ότι η τιμή της αυτοσυσχέτισης στο κορελόγραμμα φθίνει στο μηδέν έπειτα από q - p υστερήσεις, σε αντίθεση με την μερική αυτοσυσχέτιση, η οποία φθίνει στο μηδέν μετά από $p – q$ υστερήσεις. Στην συνέχεια, γίνεται έλεγχος για την στασιμότητα της σειράς. Χαρακτηριστικό μη - στάσιμων χρονοσειρών αποτελεί το ότι στο διάγραμμα των αυτοσυσχετίσεων, η αυτοσυσχέτιση δεν εξασθενεί γρήγορα. Έτσι, αφού πρώτα καθορίσουμε το κατάλληλο μοντέλο *ARMA (p,q)*, χρησιμοποιούμε τις d πρώτες διαφορές, προκειμένου να καταστήσουμε την χρονοσειρά στάσιμη. Τα μοντέλα που θα προκύψουν θα είναι τα *ARIMA (p,d,q),* όπου το p θα συμβολίζει την τάξη του αυτοπαλινδρούμενου μοντέλου, το d τον αριθμό των διαφορών που χρησιμοποιήθηκαν και το q την τάξη του μοντέλου κινούμενου μέσου.

Μετά την αναγνώριση του πιθανού ή πιθανών μοντέλων, προχωράμε στην εκτίμηση των παραμέτρων του. Ο υπολογισμός τους μπορεί να γίνει με ποικίλους τρόπους, ωστόσο οι

πιο γνωστοί είναι μέσω της Μεθόδου Ελαχίστων Τετραγώνων αλλά και της Μεθόδου των Ροπών. Η πρώτης βασίζεται στην ελαχιστοποίησή των σφαλμάτων προσαρμογής, ενώ η δεύτερη στην εκτίμηση της αυτοσυσχέτισης.

Εφόσον εκτιμηθούν οι κατάλληλες παράμετροι για τα μοντέλα μας, πρέπει να ελεγχθεί αν τα παραγόμενα μοντέλα είναι στατιστικά σημαντικά και περιγράφουν κατάλληλα την χρονοσειρά μας και αν ναι, να επιλεχθεί το βέλτιστο αυτών. Ο έλεγχος πραγματοποιείται με την χρήση του *Akaike's Information Criterion* (*AIC*) ή *Bayesian Information Criterion* (*BIC*).

Με βάση το κριτήριο του *Akaike*, υπολογίζεται η τιμή του:

$$AIC = -2\,log\,L + 2(p + q + k + 1)$$

όπου $k = 0$ αν η σταθερά του εξεταζόμενου μοντέλου είναι ίση με μηδέν, ενώ σε αντίθετη περίπτωση $k = 1$. Καταλληλότερο μοντέλο είναι αυτό που ελαχιστοποιεί την τιμή του *AIC*.

Αν γίνει χρήση του κριτηρίου *BIC*, τότε επιλέγεται το μοντέλο εκείνο που ελαχιστοποιεί την τιμή του:

$$BIC = AIC + log(n)\,(p + q + k + 1)$$

Τέλος, αφού γίνει η επιλογή του καταλληλότερου μοντέλου, προχωράμε στο στάδιο της πρόβλεψης, όπου γίνονται προβλέψεις για μία ή περισσότερες περιόδους, ενώ συνίσταται να υπολογίζονται και τα κατάλληλα διαστήματα εμπιστοσύνης. Επίσης, θα πρέπει να γίνεται και διαρκής αξιολόγηση των σφαλμάτων των προβλέψεων, καθώς αν αυτά αυξάνονται διαρκώς έχουμε ένδειξη ότι το μοντέλο μας θα πρέπει να επαναξιολογηθεί.

Προκειμένου να γίνει έλεγχος της προβλεπτικής ικανότητας του μοντέλου *ARIMA* που καταλήξαμε, συγκρίνουμε τα αποτελέσματα των προβλέψεων με τα πραγματικά δεδομένα της χρονοσειράς. Για την αξιολόγηση των προβλέψεων, θέτοντας X_t τις πραγματικές τιμές της χρονοσειράς και Y_t τις προβλέψεις που λάβαμε από το επιλεγμένο μοντέλο, έχουν προταθεί τα ακόλουθα μέτρα:

- Μέσο Απόλυτο Ποσοστιαίο Σφάλμα – *Mean Absolute Percentage Error* (*MAPE*):

$$MAPE = \frac{1}{N}\Sigma\left|\frac{X_t - Y_t}{X_t}\right| = \frac{1}{N}\Sigma\left|\frac{e_t}{X_t}\right|,$$

- Μέσο Απόλυτο Σφάλμα – *Mean Absolute Error* (*MAE*):

$$MAE = \frac{1}{N}\Sigma|X_t - Y_t| = \frac{1}{N}\Sigma|e_t| \ ,$$

- Μέσο Τετραγωνικό Σφάλμα – *Mean Squared Error* (*MSE*)

$$MSE = \frac{1}{N}\Sigma(X_t - Y_t)^2 = \frac{1}{N}\Sigma(e_t)^2,$$

- Τετραγωνική Ρίζα Μέσου Τετραγωνισμένου Σφάλματος – *Root Mean Squared Error* (*RMSE*)

$$RMSE = \sqrt{\frac{1}{N}\Sigma(X_t - Y_t)^2} = \sqrt{\frac{1}{N}\Sigma(e_t)^2},$$

- Συντελεστής Ανισότητας του *Theil – Theil's Inequality Coefficient:*

$$U = \sqrt{\frac{\frac{1}{N}\Sigma(X_t - Y_t)^2}{\frac{1}{N}\Sigma X_t^2}}.$$

Σε αντίθεση με τους πρώτους τέσσερις δείκτες, η χρήση των οποίων πρέπει να γίνεται με ιδιαίτερη προσοχή καθώς επηρεάζονται από τις μονάδες μέτρησης των δεδομένων μας, ο συντελεστής *U* είναι ανεξάρτητος των μονάδων. Αν $U > 1$, τότε το μοντέλο δεν παράγει καλές προβλέψεις, συνεπώς δεν μπορεί να χρησιμοποιηθεί, ενώ αν $U = 0$, τότε οι προβλέψεις συμπίπτουν με τις πραγματικές τιμές της χρονοσειράς.

Σε ερευνητικό επίπεδο, πολλές εργασίες δημοσιεύθηκαν την περίοδο του *Covid* – 19, χρησιμοποιώντας την μεθοδολογία των *Box – Jenkins* για την μελέτη της εξέλιξης της πανδημίας του κορονοϊού. Ενδεικτικά, η εργασία του *Ramesh* (2020) χρησιμοποιώντας την μέθοδο των *Box – Jenkins,* προσπάθησε να προβλέψει την εξέλιξη του αριθμού των κρουσμάτων Covid – 19, για την περίοδο 12 Ιουλίου – 11 Σεπτεμβρίου 2020, λαμβάνοντας δεδομένα από 7 χώρες, όπως Ιταλία, Ισπανία, Κίνα.

4.8 Μοντέλα *SIR*

Όπως αναφέρθηκε παραπάνω, μέσω των χρονοσειρών και των μοντέλων *ARIMA* μπορεί να επιτευχθεί με έγκαιρο τρόπο η μελέτη και η πρόβλεψη της πορείας της νόσου που μας ενδιαφέρει. Ωστόσο, μια πανδημία είναι ένα δυναμικό φαινόμενο, το οποίο εξελίσσεται κατά την διάρκεια μιας περιόδου, ενώ η εξέλιξη αυτή εξαρτάται σε μεγάλο βαθμό από την συμπεριφορά του υπό μελέτη πληθυσμού, όπως για παράδειγμα από τις κοινωνικές επαφές των κρουσμάτων.

Το 1927, οι *Kermack* και *McKendrik*, στην προσπάθεια μελέτης της εξέλιξης επιδημιών, διαμέρισαν για πρώτη φορά τον πληθυσμό σε 3 κατηγορίες:

- Τα υγιή άτομα, τα οποία δεν διαθέτουν ανοσία και μπορούν να νοσήσουν (*Susceptibles* – Ευπαθή άτομα)
- Τα άτομα τα οποία έχουν μολυνθεί και συνεπώς μεταδίδουν την ασθένεια (I*nfectious* – Μολυσματικά άτομα)
- Τα άτομα τα οποία έχουν νοσήσει και πλέον είτε έχουν ανοσία είτε έχουν αποβιώσει (*Removed* – Διαγεγραμμένα άτομα)

Έτσι, θεωρώντας έναν πληθυσμό μεγέθους $N = \{1, 2, \ldots, n\}$, έχουμε την σχέση:

$$S(t) + I(t) + R(t) = n,\ t \geq 0$$

όπου για την χρονική στιγμή t, $S(t)$ το πλήθος των ευπαθή ατόμων , $I(t)$ το πλήθος των ατόμων που νοσούν και μεταδίδουν την ασθένεια και $R(t)$ το πλήθος των ατόμων που έχουν πλέον αναρρώσει. Το μοντέλο που αναπτύχθηκε βάσει του παραπάνω διαχωρισμού του πληθυσμού έλαβε την ονομασία *SIR*. Στην επιδημιολογία, το μοντέλο *SIR* θεωρείται το πιο ρεαλιστικό αλλά και βασικότερο μέσο μελέτης της εξέλιξης της πορείας ενός εν δυνάμει πανδημικού φαινομένου.

Το μοντέλο των Kermack και McKendrik έχει την κάτωθι χρονική εξέλιξη, η οποία διαγραμματικά περιγράφεται ως:

$$S \overset{\beta SI}{\rightarrow} I \overset{aI}{\rightarrow} R$$

Ένα άτομο το οποίο ανήκει στην ομάδα των ευπαθή ατόμων, νοσεί με ρυθμό ανάλογο του β και συνεπώς περνάει στην ομάδα των μολυσματικών ατόμων, και στην συνέχεια είτε αποκτά ανοσία είτε πεθαίνει. Το μοντέλο δεν διαφοροποιεί όσους πεθαίνουν και όσους αποκτούν ανοσία, και τους κατατάσσει σε μία ομάδα. Αν θεωρήσουμε ότι β είναι ρυθμός μετάδοσης και μόλυνσης, α το αντίστοιχο ποσοστό ανάρρωσης και $\frac{1}{\alpha}$ ο μέσος χρόνος όπου κάποιος ασθενεί και μεταδίδει τον ιό, τότε το μοντέλο που προτείνεται από τους Kermack και McKendrik περιγράφεται από τις παρακάτω τρεις διαφορικές εξισώσεις:

$$\frac{dS}{dt} = -\frac{\beta}{N}SI \quad , \beta > 0 \tag{4.8.1}$$

$$\frac{dI}{dt} = \frac{\beta}{N}SI - aI \quad , \alpha, \beta > 0 \tag{4.8.2}$$

$$\frac{dR}{dt} = aI \ , \quad \alpha > 0 \tag{4.8.3}$$

Για το παραπάνω σύστημα διαφορικών εξισώσεων ισχύουν οι αρχικές συνθήκες:

$$S(0) = S_0 > 0, I(0) = I_0 > 0, R(0) = 0$$

Το μοντέλο SIR στηρίζεται σε τρείς βασικές υποθέσεις. Αρχικά, τα μολυσμένα άτομα μεταδίδουν τον ιό σε άτομα που ανήκουν στην ευπαθή ομάδα με ρυθμό ανάλογο των επαφών που υπάρχουν ανάμεσα στις δυο ομάδες. Στην συνέχεια, όσοι μολυνθούν αποκτούν ανοσία, με ρυθμό ανάλογο του αριθμού των νοσήσαντων. Τέλος, το μοντέλο θεωρεί ότι ο πληθυσμός κατά την διάρκεια της μελέτης παραμένει σταθερός, συνεπώς δεν λαμβάνει υπόψη τυχόν γεννήσεις και θανάτους.

Με χρήση χρονοσειρών, το μοντέλο μπορεί να γραφτεί ως:

$$S_{t+1} = S_t - \frac{\beta S_t}{N} I_t \tag{4.8.4}$$

$$I_{t+1} = I_t - aI_t + \frac{\beta S_t}{N} I_t \tag{4.8.5}$$

$$R_{t+1} = R_t + aI_t \qquad (4.8.6)$$

Από τις παραπάνω εξισώσεις, ο όρος $\frac{\beta S_t}{N}$ δηλώνει πόσους μπορεί να κολλήσει ένα άτομο που νοσεί, σε κάθε χρονική στιγμή t, ενώ ο όρος aI_t περιγράφει πόσοι γίνονται καλά (αποκτώντας ανοσία ή πεθαίνοντας) σε κάθε χρονική στιγμή.

Οι *Mohtashemi et al*. (2006), κάνουν λόγο για χρήση μόνο των δυο πρώτων σχέσεων, καθώς κατά την διάρκεια ενός εν δυνάμει επιδημικού κύματος, δεν είναι σημαντικοί αλλά ούτε εύκολο να καταγραφούν οι χρόνοι ανάρρωσης. Επιπλέον, τόσο ο ευπαθής πληθυσμός S, όσο και η διαγεγραμμένη ομάδα R, δεν είναι δυνατόν να καταγράφονται και να ενημερώνονται συστηματικά σε αντίθεση με τον μολυσμένο πληθυσμό I, ο οποίος αποτελεί τον σημαντικότερο δείκτη κατά την διάρκεια μιας επιδημιολογικής επιτήρησης.

Έτσι, είναι σημαντικό να καταφέρουμε να υπολογίσουμε το I_n χωρίς να γίνεται χρήση του S_n και R_n. Χρησιμοποιώντας τις σχέσεις (4.8.4) και (4.8.5), καταλήγουμε στον τύπο:

$$I_{n+2} = \frac{I_{n+1}^2}{I_n} - \beta I_{n+1}(I_{n+1} - (1-\alpha)I_n) \qquad (4.8.7)$$

Στην παραπάνω σχέση, έχουμε να εκτιμήσουμε μόνο μια παράμετρο, την β. Οι *Mohtashemi et al*. (2006) εκτίμησαν την παράμετρο μέσω παλινδρόμησης με χρήση της μεθόδου ελαχίστων τετραγώνων, για κάθε χρονολογικό παράθυρο, προκαθορισμένου μεγέθους, και μάλιστα θεωρούν ότι το β παραμένει σταθερό κατά την διάρκεια μικρών χρονικών περιόδων. Για κάθε μέρα του χρόνου, επιλέγουμε παράθυρο μεγέθους Τ (από το σήμερα στο παρελθόν), και με τον τρόπο αυτό σχηματίζονται Τ – 2 το πλήθος εξισώσεις για την εκτίμηση του β μέσω παλινδρόμησης με χρήση της μεθόδου ελαχίστων τετραγώνων. Οι *Mohtashemi et al*. κάνουν χρήση παραθύρου μεγέθους 7, καθώς εξισορροπούν την επίδραση της μέρας της εβδομάδας, ωστόσο μπορούν να γίνουν και χρήση παραθύρων μήκους 14. Εν συνεχεία, για κάθε β που εκτιμάται καθημερινά, γίνεται σύγκρισή του με το μέσο ποσοστό μετάδοσης που αντιστοιχεί στην ίδια μέρα του έτους από τα ιστορικά δεδομένα που έχουν στην διάθεσή τους. Αν η εκτίμηση που έχουν κάνει, υπερβαίνει το όριο που υπάρχει με βάση τα ιστορικά δεδομένα τότε σημάνει συναγερμός για ενδεχόμενο επιδημιολογικό κύμα (*Shtatland E*. (2008)).

Στην εργασίας του οι *Mohtashemi et al.* (2006), έκαναν χρήση του αυτοπαλίνδρομου μοντέλου $AR(p)$ με την βοήθεια του *SIR*, για την έγκαιρη ανίχνευση ξεσπάσματος μίας επιδημίας. Συγκεκριμένα, χρησιμοποιώντας τις σχέσεις (4.8.4), (4.8.5), (4.8.6) και (4.8.7), καταλήγουν στην εξίσωση:

$$I_{n+2} = k_1 I_{n+1} + k_2 I_n \qquad (4.8.8)$$

όπου

$$k_1 = \frac{I_{n+1}}{I_n} - \beta(I_{n+1} - I_n) \; και \; k_2 = k_1 I_n.$$

Στην έναρξη μιας επιδημίας ωστόσο τα κρούσματα είναι ιδιαίτερα χαμηλά, συνεπώς το k_2 λαμβάνει πολύ μικρές τιμές, και γι' αυτό τον λόγο μπορεί να παραληφθεί. Έτσι, η σχέση (4.8.8) μπορεί να γραφτεί ως:

$$I_{n+2} = k_1 I_{n+1} \quad ή \quad I_{n+1} = k_1 I_n.$$

Η παραπάνω σχέση μας παρέχει ένδειξη εκθετικής αύξησης των κρουσμάτων. Επίσης, καθώς στην αρχή τα καταγεγραμμένα κρούσματα είναι πολύ λίγα σε αριθμό, χωρίς βλάβη της γενικότητας μπορούμε να υποθέσουμε ότι $S_n \approx S_0 \approx N$. Έτσι , μπορούμε να καταλήξουμε στην σχέση:

$$I_{n+1} = (1 + \beta^* - \alpha) I_n \;, \quad \text{όπου } \beta^* = \beta N.$$

Η παραπάνω εξίσωση είναι γραμμική, και καταλήξαμε σε αυτή μέσω προσεγγίσεων από την αμιγώς μη – γραμμική εξίσωση της σχέσης (4.8.7). Γίνεται κατανοητό ότι οι παραπάνω προσεγγίσεις παρήγαγαν πολλά μικρά σφάλματα. Επιπλέον, υπάρχει έντονη στοχαστική διακύμανση στα β και α. Αν λάβουμε υπόψη τις παρατηρήσεις που αναφέραμε, καταλήγουμε σε ένα μοντέλο $AR(1)$ επιδημιολογικής επιτήρησης, το οποίο είναι:

$$I_{n+1} = \frac{(1 - \beta^* - a)}{I_n} + w_n$$

όπου w_n λευκός θόρυβος. Με βάση την παραπάνω σχέση, όταν $\beta^* - a < 0$, δηλαδή όταν ο ρυθμός ανάρρωσης είναι μεγαλύτερος από τον ρυθμό νόσησης, τότε δεν έχουμε έξαρση επιδημίας. Σε αντίθετη περίπτωση, όταν $\beta^* - a > 0$, παρατηρείται εκθετική αύξηση των κρουσμάτων, δηλαδή λαμβάνουμε ένδειξη για επικείμενο ξέσπασμα. Αν $\beta^* - a = 0$, έχουμε μη – στάσιμη διαδικασία.

Είναι σημαντικό να σημειωθεί, ότι σχεδόν όλα τα μοντέλα *ARMA* και *ARIMA* που χρησιμοποιούνται για επιδημιολογική επιτήρηση, χρησιμοποιούν το αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης.

Όπως αναφέρθηκε παραπάνω, στην έναρξη της επιδημίας, χωρίς βλάβη της γενικότητας μπορούμε να υποθέσουμε ότι $S \approx N$. Έτσι, η σχέση (4.8.2) γράφεται ως:

$$\frac{dI}{dt} \sim (\beta - \alpha)I \qquad (4.8.9)$$

από όπου ολοκληρώνοντας κατά μέλη έχουμε:

$$I = I_0 e^{(\beta-\alpha)t}.$$

Αν θεωρήσουμε ότι το I είναι μια σταθερά, και συγκεκριμένα $I = I_0$, τότε η σχέση (4.8.3) γίνεται:

$$\frac{dR}{dt} = a\, I_0$$

από όπου ολοκληρώνοντας κατά μέλη έχουμε:

$$R_t = a\, t\, I_0.$$

Αν υποθέσουμε ότι ο χρόνος από την στιγμή νόσησης μέχρι την απόκτηση ανοσίας είναι T, και $R(T) = I_0$, τότε έχουμε ότι ο ρυθμός ανάρρωσης είναι ίσος με:

$$a \approx \frac{1}{T}$$

Επιπρόσθετα, όσο ο χρόνος αλλάζει, δηλαδή $dt = k$, τότε η σχέση (4.8.3) γίνεται:

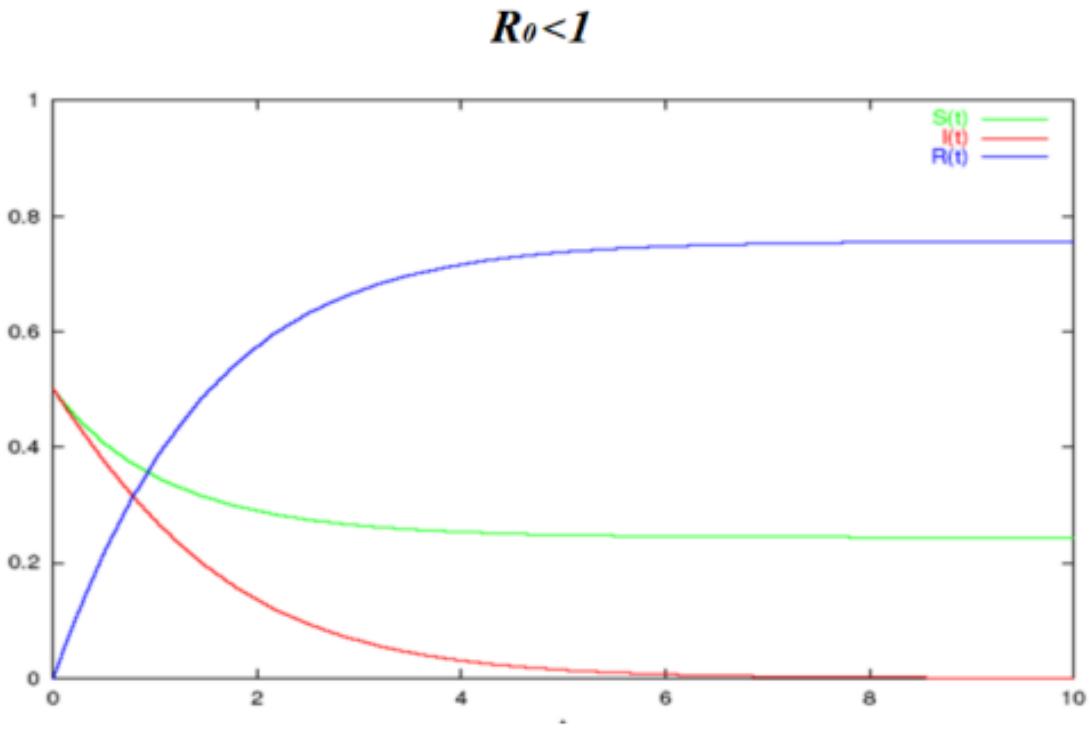
$$\frac{dR(t+k)}{k} = aI$$

Όπου ο αντίστοιχος ρυθμός ανάρρωσης είναι ίσος με:

$$\alpha = \frac{R(t+1)-R(t)}{I(t)}.$$

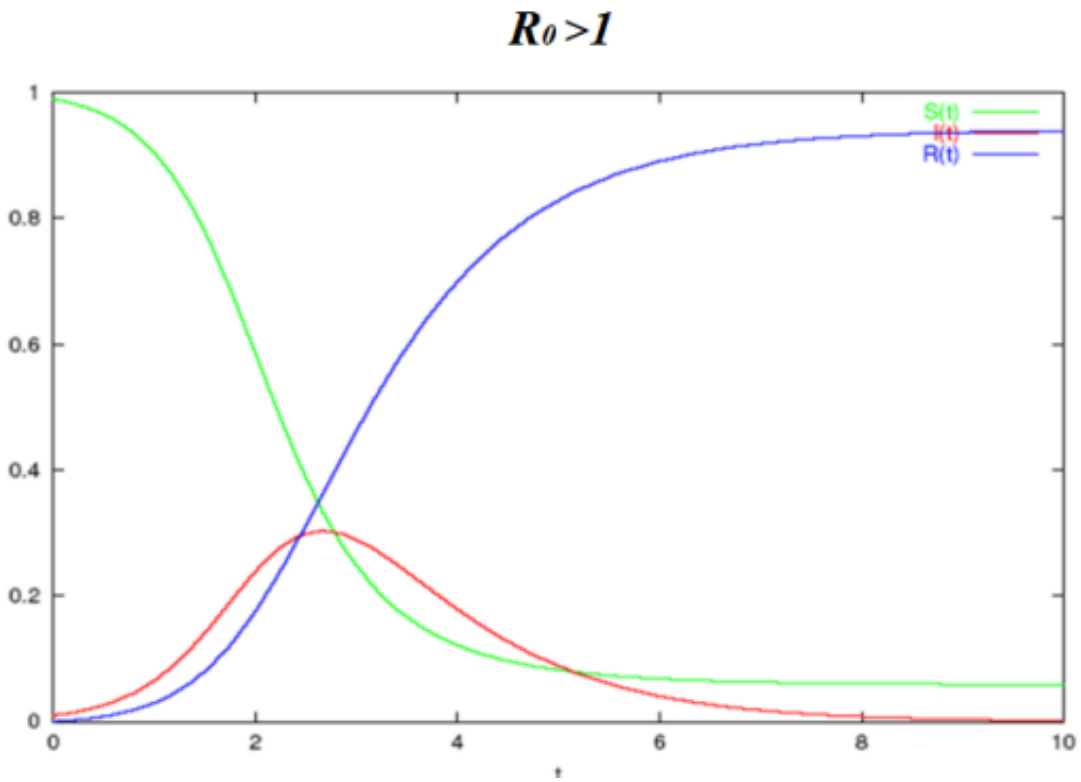
Πολύ σημαντικός δείκτης για τα παραπάνω μοντέλα αποτελεί ο βασικός ρυθμός αναπαραγωγής της ασθένειας, ο οποίος συμβολίζεται με R_0 και ο οποίος είναι ίσος με $\frac{\beta}{\alpha}$. Η ερμηνεία του βασικού ρυθμού αναπαραγωγής της ασθένειας είναι πως για κάθε άτομο το οποίο έχει μολυνθεί, αναμένεται να μεταφέρει την νόσο σε R_0 άτομα. Καθώς το δυναμικό σύστημα που περιγράφει την εν δυνάμει πανδημία αναπτύσσεται στον χρόνο παρατηρούμε πως αν $R_0 > 1$, τότε η επιδημία εξαπλώνεται και εξελίσσεται και πιθανόν να αποτελεί κίνδυνο για την δημόσια υγεία, ενώ αν $R_0 < 1$, τότε η επιδημία φθίνει, καθώς από την σχέση (4.8.9) προκύπτει ότι $\frac{dI}{dt} < 0$, αφού $\beta < \alpha$. Αξίζει να σημειωθεί πως το β μπορεί να μειωθεί αποτελεσματικά με πολιτικές όπως η κοινωνική αποστασιοποίηση και καραντίνα, εμβολιασμός, σωστή υγιεινή, ενώ το α εξαρτάται από τα βιολογικά χαρακτηριστικά της ασθένειας.

Στο παρακάτω διάγραμμα, απεικονίζονται οι συμπεριφορές των $S(t)$, $I(t)$, $R(t)$ όταν ο βασικός ρυθμός αναπαραγωγής είναι μικρότερος της μονάδας. Στην περίπτωση αυτή, κάθε μολυσμένο άτομο μολύνει λιγότερα του ενός άτομα, με αποτέλεσμα $\frac{dI}{dt} < 0$. Έτσι, η ασθένεια έχει πτωτική πορεία, με αποτέλεσμα αν και στην αρχή ο αριθμός των ατόμων που βρίσκονται στην ευπαθή ομάδα να μειώνεται, σε πολύ μικρό χρονικό διάστημα σταθεροποιείται.



Διάγραμμα 4.8.1. Γραφική Απεικόνιση *SIR* μοντέλου με $R_0 < 1$

Στο επόμενο διάγραμμα, παρουσιάζεται η περίπτωση όπου ο βασικός ρυθμός αναπαραγωγής της ασθένειας είναι μεγαλύτερος της μονάδας. Σε αντιστοιχία με ό,τι αναφέρθηκε παραπάνω, τώρα κάθε μολυσμένο άτομο είναι ικανό να μεταδώσει τον ιό σε περισσότερα του ενός άτομα. Ο αριθμός των ευπαθή ατόμων μειώνεται σημαντικά στην έναρξη της επιδημίας και στην συνέχεια σταθεροποιείται, όπου και μηδενίζεται ή τείνει στο μηδέν. Το ίδιο χρονικό διάστημα, ο αριθμός των μολυσμένων αυξάνεται, μέχρι να φτάσει το υψηλότερη τιμή της, και στην συνέχεια, καθώς έχει μολυνθεί η μεγαλύτερη μερίδα του υπό μελέτη πληθυσμού μειώνεται μέχρι ως ότου μηδενιστεί.



Διάγραμμα 4.8.2. Γραφική Απεικόνιση *SIR* μοντέλου με $R_0 > 1$

4.8.1 Παραλλαγές του μοντέλου *SIR*

Τα μοντέλα *SIR* αποτελούν βασικό μέσο της προσπάθειας επιτήρησης ασθενειών. Ωστόσο, τα ιδιαίτερα βιολογικά χαρακτηριστικά που παρουσιάζει κάθε ιός, τόνισε την ανάγκη προσαρμογής του βασικού μοντέλου σε αυτά. Έτσι προέκυψαν οι ακόλουθες παραλλαγές:

- Μοντέλο *SIS*: Με βάση το συγκεκριμένο μοντέλο, ο πληθυσμός χωρίζεται σε δυο ομάδες, την ευπαθή και τους μολυσμένους. Το άτομο που νοσεί, δεν αποκτά ανοσία, με αποτέλεσμα κατά την λήξη της περιόδου νόσησης, να επιστρέφει ξανά στην ομάδα των ευπαθών. Το παραπάνω μοντέλο μπορεί να εφαρμοστεί με επιτυχία στην περίπτωση του Covid – 19, και ιδιαίτερα στις υπομεταλλάξεις του στελέχους Όμικρον, καθώς η προηγούμενη νόσηση δεν προσφέρει ανοσία σε μία επικείμενη επαναμόλυνση. Ενδιαφέρον παρουσιάζει η εργασία του *Otunuga* (2021), ο οποίος προσάρμοσε το μοντέλο *SIS* για τα

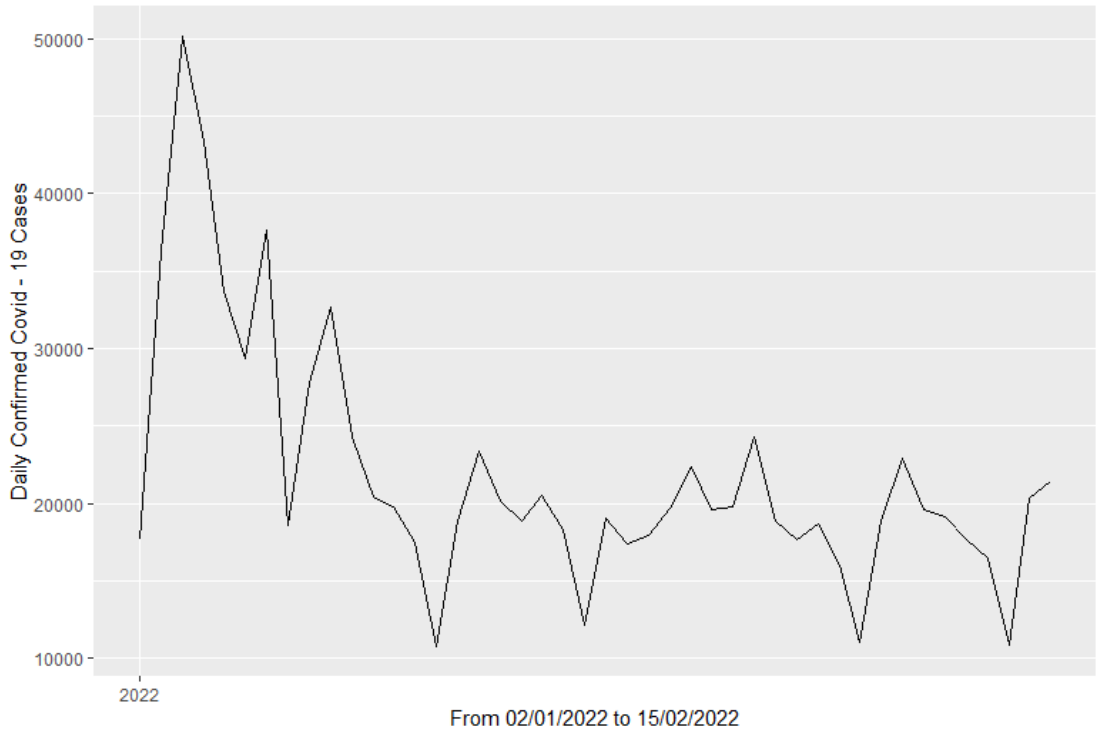
επιβεβαιωμένα κρούσματα κορονοϊού των Ηνωμένων Πολιτειών την περίοδο από 22 Ιανουάριο 2020 – 23 Μαρτίου 2021.

- Μοντέλο *SIRS*: Ο πληθυσμός χωρίζεται σε τρεις κατηγορίες, όπως και στο κλασικό μοντέλο *SIR*. Η διαφοροποίηση των δύο μοντέλων είναι πως, τώρα τα μολυσμένα άτομα αποκτούν προσωρινή ανοσία, κι έτσι μετά από σύντομο χρονικό διάστημα επιστρέφουν στην ομάδα των ευπαθών και μπορούν και πάλι να νοσήσουν.
- Μοντέλο *SIRD*: Στο μοντέλο αυτό, όσοι νοσούν, είτε αναρρώνουν αποκτώντας ανοσία, είτε πεθαίνουν. Έτσι, πέρα από την ομάδα των ευπαθών, των μολυσμένων και των διαγεγραμμένων, εισέρχεται μία επιπλέον κατηγοριοποίηση, η ομάδα των νεκρών. Υπενθυμίζεται ότι στο βασικό μοντέλο, δεν υπάρχει διαχωρισμός ανάμεσα σε όσους αποκτούν ανοσία και σε όσους πεθαίνουν, και κατατάσσονται στην ίδια κατηγορία. Στην πρόσφατα δημοσιοποιημένοι εργασία τους, οι *Sen et. al* (2021) προτείνουν την χρήση ενός τροποποιημένου μοντέλου *SIRD,* το οποίο λαμβάνει υπόψιν την επίδραση που έχει το μέτρο της καραντίνας καθώς και η ύπαρξη ασυμπτωματικών φορέων της νόσου, στην εξέλιξη της επιδημίας του κορονοϊού.
- Μοντέλο *SEIS - SEID* : Στο συγκεκριμένο μοντέλο, ένα άτομο πριν νοσήσει, διανύει μια περίοδο κατά την οποία ενώ έχει έρθει σε επαφή με τον ίο δεν νοσεί και συνεπώς δεν μπορεί να μεταδώσει τον ιό. Στην συνέχεια, περνά στην ομάδα των μολυσμένων, και έπειτα στην ομάδα των διαγεγραμμένων ή ανάλογα με τα χαρακτηριστικά της ασθένειας, στην ομάδα των ευπαθών. Η ομάδα στην οποία ο ιός επωάζεται και δεν μεταδίδεται ονομάζεται Εκτεθειμένη (*Exposed*).

Κ Ε Φ Α Λ Α Ι Ο 5

5.1 Εφαρμογή Μοντέλων ARIMA σε δεδομένα κορονοϊού

Στο παρόν κεφάλαιο, γίνεται προσπάθεια πρόβλεψης των ημερήσιων κρουσμάτων *Covid* – 19 για την ελληνική επικράτεια. Η ανάλυση έγινε μέσω της *R – Studio,* ενώ χρησιμοποιήθηκαν οι βιβλιοθήκες *forecast, tseries* και *TSstudio*.Τα δεδομένα που χρησιμοποιήθηκαν αφορούν την χρονική περίοδο από τις 02/01/2022 έως τις 15/02/2022. Η συλλογή των δεδομένων πραγματοποιήθηκε μέσω των ημερήσιων εκθέσεων *Covid* – 19 του Εθνικού Οργανισμού Δημόσιας Υγείας (ΕΟΔΥ) και μπορούν να βρεθούν στην ιστοσελίδα https://eody.gov.gr/epidimiologika-statistika-dedomena/imerisies-ektheseis-covid-19/.

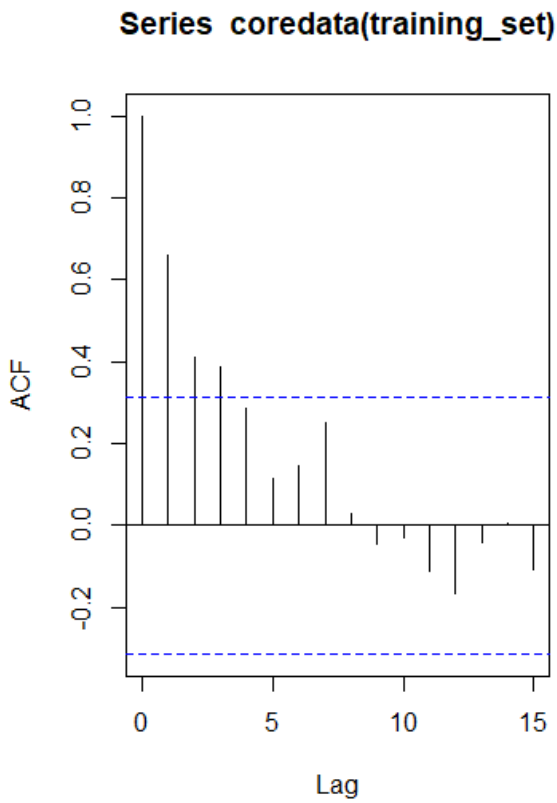


Διάγραμμα 5.1.1.
Γραφική παρουσίαση των επιβεβαιωμένων κρουσμάτων *Covid* – 19 για την περίοδο 02/01/2022 έως 15/02/2022.

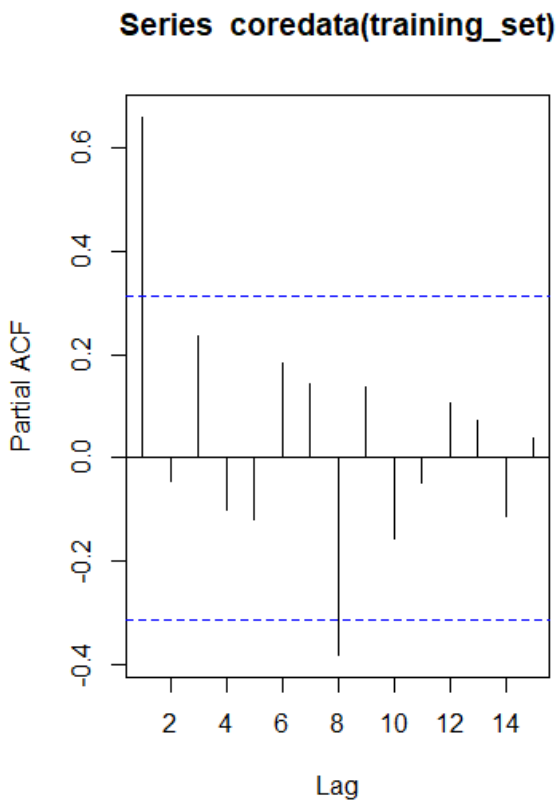
Βάσει του παραπάνω διαγράμματος παρατηρούμε ότι στην αρχή της περιόδου καταγραφής μας, τα καταγεγραμμένα κρούσματα παρουσιάζουν έντονη αυξητική τάση ενώ στην συνέχεια αν και παρατηρείται πτώση των κρουσμάτων, εξακολουθούν να κυμαίνονται σε υψηλά επίπεδα με περιόδους έξαρσης.

Πριν προχωρήσουμε στην ανάλυση των δεδομένων μας, θα χρειαστεί να χωρίσουμε τα δεδομένα μας σε δύο ομάδες, το *training set* και το *test set*. Το *training set* θα περιέχει τις παρατηρήσεις από τις 02/01/2022 έως και τις 10/02/2022, ενώ το *test set* θα περιέχει τις παρατηρήσεις από τις 11/02/2022 έως και τις 15/02/2022.

Στην συνέχεια, για το *training set* θα μελετήσουμε τα γραφήματα των αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων της χρονοσειράς, για να ελέγξουμε αν αυτή είναι στάσιμη.



Διάγραμμα 5.1.2.
Συνάρτηση αυτοσυσχέτισης (*ACF*) των δεδομένων κορονοϊού



Διάγραμμα 5.1.3.
Συνάρτηση μερικής αυτοσυσχέτισης (*PACF*) των δεδομένων κορονοϊού

Από τα παραπάνω διαγράμματα των συσχετίσεων και μερικών αυτοσυσχετίσεων είναι φανερό, τουλάχιστον οπτικά, ότι η χρονοσειρά μας δεν είναι στάσιμη. Ιδιαίτερα, από το διάγραμμα της συνάρτησης αυτοσυσχέτισης (Διάγραμμα 5.1.2.) φαίνεται πως αν και οι τιμές της συνάρτησης φθίνουν με αργό ρυθμό στο μηδέν, εν αντιθέσει με μία στάσιμη χρονοσειρά, αυτό δεν πραγματοποιείται με εκθετικό τρόπο. Το παραπάνω συμπέρασμα επιβεβαιώνεται και μέσω του ελέγχου *Augmented Dickey-Fuller,* βάσει του οποίου πραγματοποιείται ο ακόλουθος έλεγχος υποθέσεων:

H_0: Η χρονοσειρά είναι μη – στάσιμη
H_1: Η χρονοσειρά είναι στάσιμη

Τα αποτελέσματα τα οποία λάβαμε από τον έλεγχο είναι τα ακόλουθα:

```
> adf.test(training_set)

        Augmented Dickey-Fuller Test

data:  training_set
Dickey-Fuller = -2.7146, Lag order = 3, p-value =
0.2936
alternative hypothesis: stationary
```

Παρατηρούμε ότι σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$, το p – value του ελέγχου είναι 0.2936, δηλαδή δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση, συνεπώς, επιβεβαιώνοντας και τα αποτελέσματα των γραφημάτων της αυτοσυσχέτισης και μερικής αυτοσυσχέτισης, η χρονοσειρά είναι μη – στάσιμη.

Καθώς η χρονοσειρά είναι μη στάσιμη, για να προχωρήσουμε στην μελέτη, θα πρέπει να βρούμε την τάξη των διαφορών που χρειάζεται να χρησιμοποιηθούν προκειμένου η χρονοσειρά να καταστεί στάσιμη. Μέσω της *R*, χρησιμοποιούμε την εντολή *ndiffs*, η οποία μας δίνει την τάξη της διαφόρισης. Συγκεκριμένα, όπως θα δούμε παρακάτω θα χρειαστεί να χρησιμοποιήσουμε την διαφόριση $1^{ης}$ τάξης.

```
> covid_data_ts1<-diff(training_set,differences = 1)
> covid_data_ts1
```

Πραγματοποιώντας εκ νέου τον έλεγχο *Augmented Dickey-Fuller*, για την στασιμότητα της χρονοσειράς με την χρήση των πρώτων διαφορών, διαπιστώνουμε πως πλέον ο χρονοσειρά είναι στάσιμη. Πράγματι, έχουμε:

```
> adf.test(covid_data_ts1)

        Augmented Dickey-Fuller Test

data:  covid_data_ts1
Dickey-Fuller = -4.1123, Lag order = 3, p-value = 0.01603
alternative hypothesis: stationary
```

Παρατηρούμε ότι σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$, το p – value του ελέγχου είναι 0.01603, δηλαδή απορρίπτουμε την μηδενική υπόθεση της μη – στασιμότητας, και πλέον μπορούμε να προχωρήσουμε στην επιλογή του καταλληλότερου μοντέλου *ARIMA* για την περιγραφή της χρονοσειράς μας.

Όπως είδαμε και στο προηγούμενο κεφάλαιο, η επιλογή αυτή μπορεί να γίνει μέσω των διαγραμμάτων αυτοσυσχέτισης και μερικής αυτοσυσχέτισης. Ωστόσο, μέσω της εντολής *auto.arima*, η *R* υπολογίζει το καλύτερο μοντέλο *ARIMA* για τα δεδομένα μας, βάσει των κριτηρίων *AIC* και *BIC*. Συγκεκριμένα έχουμε:

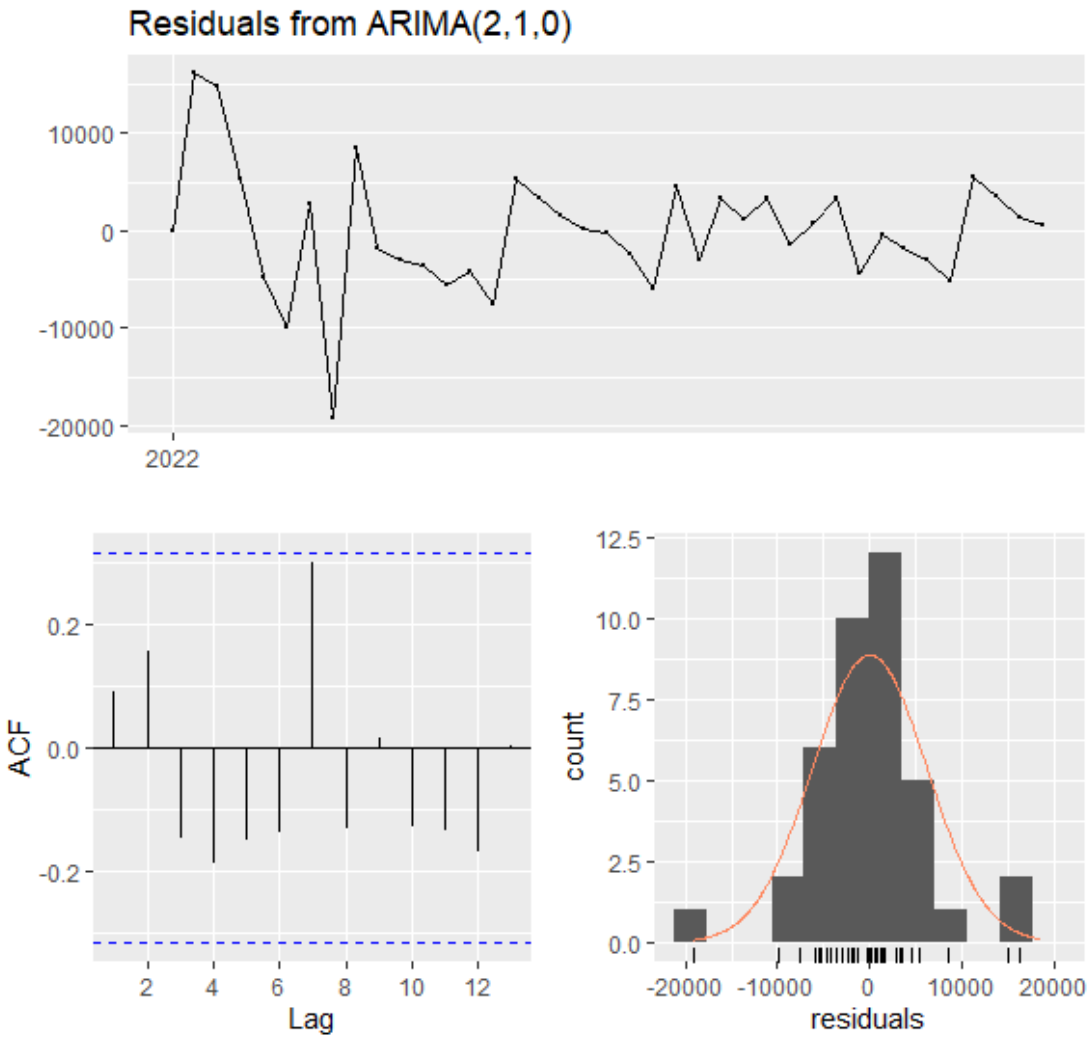
```
> model_arima<-auto.arima(training_set)
> model_arima
Series: training_set
ARIMA(2,1,0)

Coefficients:
          ar1      ar2
      -0.2377  -0.4689
s.e.   0.1715   0.1759

sigma^2 = 40631848:  log likelihood = -386.04
AIC=778.07   AICc=778.78   BIC=782.98
```

Με βάση τα παραπάνω, το καταλληλότερο μοντέλο για την περιγραφή των δεδομένων μας, είναι το *ARIMA*(2,1,0). Για το επιλεγμένο μοντέλο, θα πρέπει να γίνει έλεγχος για την ανεξαρτησία των υπολοίπων, δηλαδή δεν θα πρέπει τα υπόλοιπα να αυτοσυσχετίζονται, και πιο συγκεκριμένα να συμπεριφέρονται ως μια διαδικασία λευκού θορύβου. Το παραπάνω, επαληθεύεται μέσω του ελέγχου των *Ljung – Box*. Ο έλεγχος που πραγματοποιείται είναι ο ακόλουθος:

H_0: *Τα υπόλοιπα δεν αυτοσυσχετίζονται*

H_1: *Τα υπόλοιπα αυτοσυσχετίζονται*

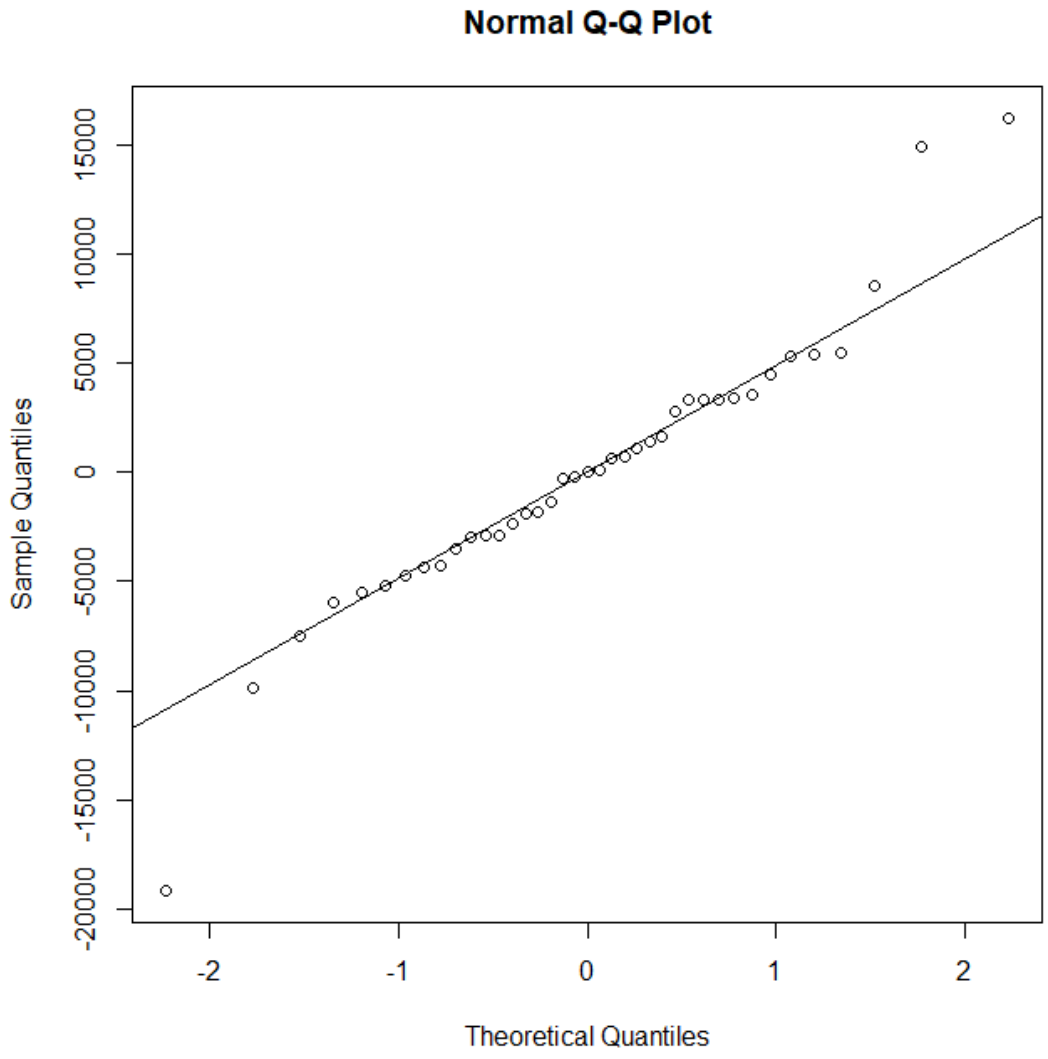
```
> Box.test(model_arima$residuals, type = "Ljung-Box")

        Box-Ljung test

data:  model_arima$residuals
X-squared = 0.33987, df = 1, p-value = 0.5599
```



Διάγραμμα 5.1.4.
Διαγράμματα συμπεριφοράς υπολοίπων του μοντέλου *ARIMA* (2,1,0)

Από το παραπάνω, διαπιστώνουμε ότι το *p – value* του ελέγχου είναι 0.5599, συνεπώς σε επίπεδο στατιστικής σημαντικότητας α = 5%, δεν απορρίπτουμε την μηδενική υπόθεση, δηλαδή τα υπόλοιπα συμπεριφέρονται ως μία διαδικασία λευκού θορύβου.

Επιπλέον θα πρέπει να γίνει έλεγχος για την κανονικότητα των υπολοίπων. Για τον σκοπό αυτό θα κατασκευάσουμε το διάγραμμα *Quantile – Quantile*.



Διάγραμμα 5.1.5.
Διάγραμμα *QQ – plot* των υπολοίπων του μοντέλου *ARIMA* (2,1,0)

Από το παραπάνω διάγραμμα δεν φαίνεται να απορρίπτεται η υπόθεση της κανονικότητας των καταλοίπων. Ωστόσο, θα χρειαστεί να προχωρήσουμε σε έλεγχο *Shapiro – Wilk*.

```
> shapiro.test(model_arima$residuals)

        Shapiro-Wilk normality test

data:  model_arima$residuals
W = 0.94535, p-value = 0.05733
```

Πράγματι, από τον έλεγχο *Shapiro – Wilk*, σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$, δεν απορρίπτεται η μηδενική υπόθεση, συνεπώς τα υπόλοιπα ακολουθούν την κανονική κατανομή.

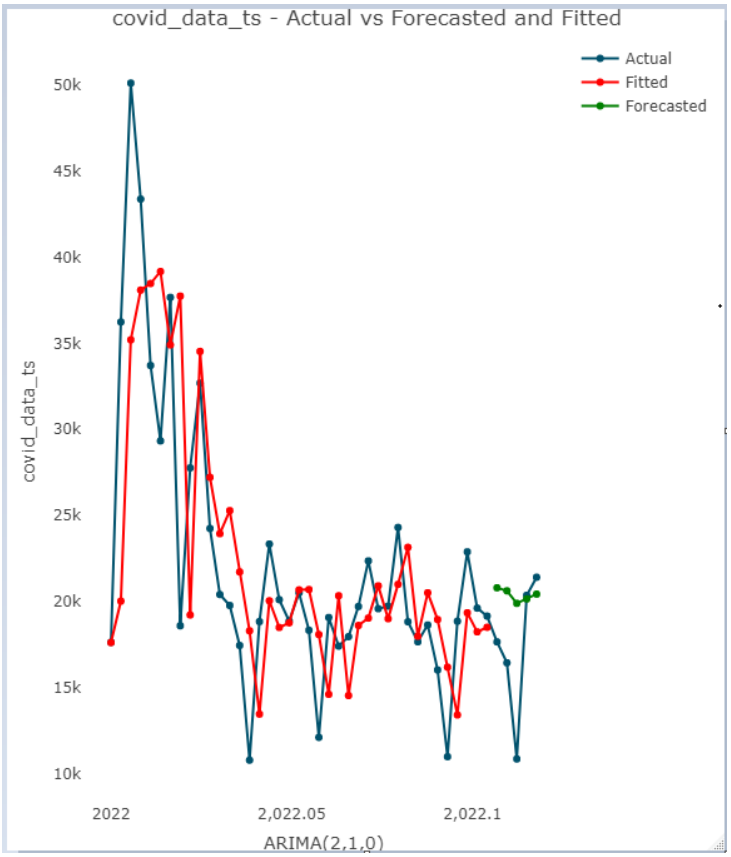
Έτσι, καθώς το επιλεγμένο μοντέλο *ARIMA* πληροί τις απαραίτητες προϋποθέσεις, πριν προχωρήσουμε στην διαδικασία προβλέψεων, θα χρησιμοποιήσουμε το training set για να προσπαθήσουμε να εκτιμήσουμε το test set, όπως το είχαμε ορίσει παραπάνω.

```
> covid_model_forecast<-forecast(model_arima,level = c(95),h = 5)
> covid_model_forecast
          Point Forecast     Lo 95    Hi 95
2022.1068       20798.21 8304.7875 33291.63
2022.1096       20624.90 4915.7644 36334.03
2022.1123       19895.07 3589.2439 36200.89
2022.1151       20149.85 2409.5190 37890.19
2022.1178       20431.52  636.3419 40226.71
```

Ημερομηνία	**Αναμενόμενα Κρούσματα**	**Επιβεβαιωμένα Κρούσματα**
11/02/2022	20798	17656
12/02/2022	20624	16442
13/02/2022	19895	10853
14/02/2022	20149	20361
15/02/2022	20431	21412

Πίνακας 5.1.1. Αναμενόμενα και Επιβεβαιωμένα κρούσματα

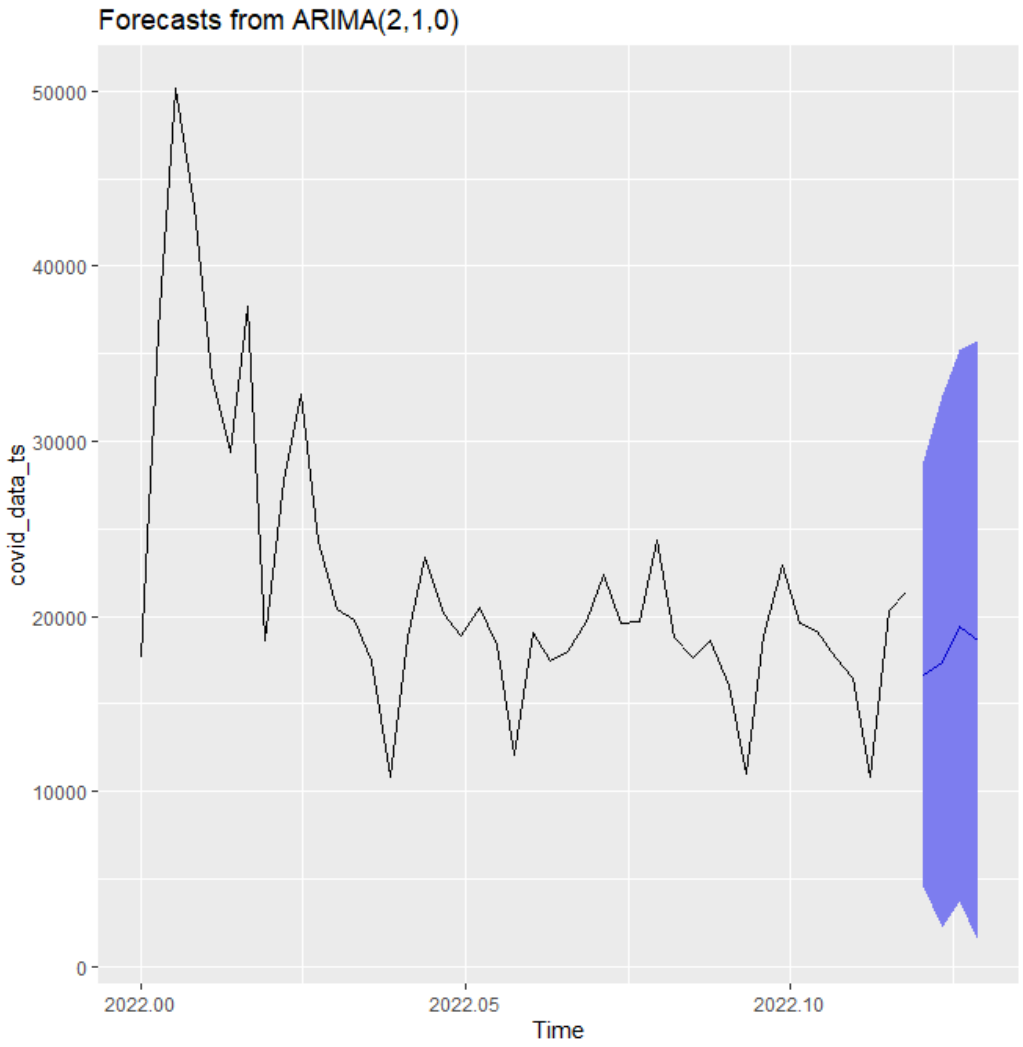
Το μοντέλο μας προσεγγίζει ικανοποιητικά τις πραγματικές τιμές των επιβεβαιωμένων κρουσμάτων, με εξαίρεση ωστόσο στις 13/02/2022 όπου τα αναμενόμενα κρούσματα είναι κατά 9000 παραπάνω από τα επιβεβαιωμένα. Ωστόσο κάτι τέτοιο μπορεί να οφείλεται στον χαμηλό αριθμό εργαστηριακών τεστ, αλλά και στην ημέρα, καθώς τις Κυριακές καταγράφεται ο χαμηλότερος αριθμός κρουσμάτων σε σύγκριση με τις υπόλοιπες ημέρες της εβδομάδας. Θα προχωρήσουμε σε πρόβλεψη 4 ημερών για την πορεία των κρουσμάτων κορονοϊού στην Ελλάδα. Καθώς οι ιοί είναι δυναμικά φαινόμενα, των οποίων ο ρυθμός μεταδοτικότητας αλλάζει διαρκώς, δεν συνίσταται η διεξαγωγή μακροπρόθεσμων προβλέψεων, καθώς τα εξαγόμενα συμπεράσματα ενδέχεται να μην είναι ασφαλή.



Διάγραμμα 5.1.6.
Συμπεριφορά Πραγματικών Δεδομένων έναντι Προβλεπόμενων και Προσαρμοσμένων

Πλέον, μέσω του μοντέλου *ARIMA* (2,1,0), μπορούμε να προχωρήσουμε στην δημιουργία βραχυπρόθεσμων προβλέψεων, και πιο συγκεκριμένα θα γίνει προσπάθεια πρόβλεψης των κρουσμάτων για χρονικό ορίζοντα 4 ημερών.

```
> model_final<-Arima(covid_data_ts,order = c(2,1,0))
> forecast_of_covid_cases<-forecast(model_final,level = c(95),h=4)
> forecast_of_covid_cases
          Point Forecast    Lo 95    Hi 95
2022.1205       16676.92 4521.816 28832.03
2022.1233       17383.11 2214.409 32551.80
2022.1260       19429.57 3705.439 35153.70
2022.1288       18578.98 1431.422 35726.53
```



Διάγραμμα 5.1.7.
Διάγραμμα προβλέψεων μοντέλου *ARIMA* (2,1,0)

Παρακάτω παρατίθεται πίνακας με τα κρούσματα τα οποία προέβλεψε το μοντέλο *ARIMA* (2,1,0) καθώς και τα επιβεβαιωμένα κρούσματα τα οποία ανακοινώθηκαν από τον ΕΟΔΥ για το χρονικό διάστημα από 16/02/2022 έως τις 19/02/2022.

Ημερομηνία	**Κρούσματα από *ARIMA* (2,1,0)**	**Επιβεβαιωμένα Κρούσματα**
16/02/2022	16677	19509
17/02/2022	17383	19504
18/02/2022	19430	18605
19/02/2022	18579	15305

Πίνακας 5.1.2.Κρούσματα από μοντέλο *ARIMA* (2,1,0) και Επιβεβαιωμένα κρούσματα

5.2 Εφαρμογή *p* – Διαγράμματος Ελέγχου σε δεδομένα κορονοϊού

Στην παρούσα εφαρμογή, χρησιμοποιήσαμε τους δείκτες θετικότητας της Ελλάδας, για την περίοδο από 01/10/2020 έως τις 28/10/2020. Θεωρούμε ένα νοσοκομειακό ίδρυμα, στο οποίο για την περίοδο που αναφέραμε, πραγματοποιούνται καθημερινά έλεγχοι στις πύλες εισόδου σε 80 νεοεισερχόμενους ασθενείς προκειμένου να ελεγχθεί αν αυτοί είναι θετικοί στον *Covid* – 19. Η επιλογή του αριθμού 80 έγινε για την καλύτερη προσαρμογή των δεδομένων μας στην διωνυμική κατανομή και δεν ανταποκρίνεται στο πραγματικό αριθμό εισαγωγών ασθενών στο νοσοκομείο. Οι δείκτες θετικότητας που καταγράφηκαν είναι οι ακόλουθοι:

Date	Positivite rate	Positive Cases for n = 80
1/10/2020	0,0381	3
2/10/2020	0,0474	4
3/10/2020	0,023	2
4/10/2020	0,0242	2
5/10/2020	0,0505	4
6/10/2020	0,033	3
7/10/2020	0,0345	3
8/10/2020	0,0388	3
9/10/2020	0,0369	3
10/10/2020	0,0236	2
11/10/2020	0,0306	2
12/10/2020	0,0442	4
13/10/2020	0,0209	2
14/10/2020	0,0217	2
15/10/2020	0,0228	2
16/10/2020	0,0264	2
17/10/2020	0,0256	2
18/10/2020	0,0294	2
19/10/2020	0,0605	5
20/10/2020	0,0344	3
21/10/2020	0,0434	3
22/10/2020	0,0434	3
23/10/2020	0,0411	3
24/10/2020	0,0466	4
25/10/2020	0,0476	4
26/10/2020	0,0597	5
27/10/2020	0,0623	5
28/10/2020	0,0769	6

Πίνακας 5.2.1. Δείκτες Θετικότητας και Προσομοιωμένα κρούσματα

Καθώς τον Οκτώβριο του 2020 η πανδημία στην Ελλάδα βρίσκονταν σε ελεγχόμενα πλαίσια, θα χρησιμοποιήσουμε τους παραπάνω δείκτες θετικότητας, για ανάλυση φάσης *I*, και πιο συγκεκριμένα μέσω του *p* – διαγράμματος ελέγχου, καθώς μας ενδιαφέρει η μελέτη του ποσοστού των θετικών διαγνώσεων.

Αρχικά, θα πρέπει να ελέγξουμε αν τα δεδομένα μας προέρχονται από την διωνυμική κατανομή. Αυτό θα πραγματοποιηθεί μέσω του ελέγχου καλής προσαρμογής, για τον οποίο πραγματοποιείται ο εξής έλεγχος υποθέσεων:

H_0: Τα δεδομένα προέρχονται από την Διωνυμική Κατανομή
H_1: Τα δεδομένα δεν προέρχονται από την Διωνυμική Κατανομή

Επειδή πραγματοποιούμε ανάλυση φάσης *I*, το ποσοστό θετικότητας δεν είναι εκ των προτέρων γνωστό. Συνεπώς θα χρειαστεί να εκτιμηθεί μέσω του δειγματικού ποσοστού:

$$\hat{p} = \frac{p_1+p_2+\ldots+p_m}{m} = \frac{X_1+X_2+\ldots+X_m}{mn} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{mn} = 0.038839$$

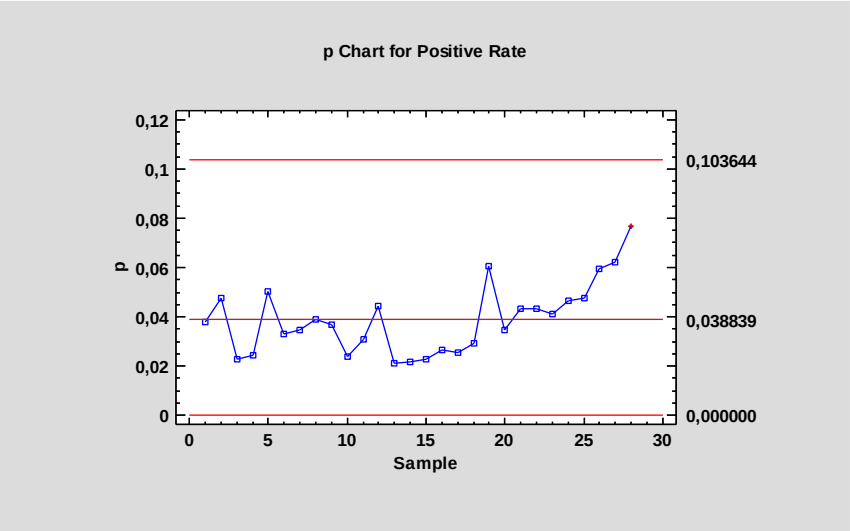
Μέσω *Excel*, υπολογίστηκαν τα αναμενόμενα θετικά τεστ υπό την *B*(*80, 0.038839*) και στην συνέχεια υπολογίστηκε το $\chi^2 = 37.22783$.

Για τον έλεγχο καλής προσαρμογής έχουμε ότι:

$$\chi^2 = 37.22783 < \chi^2_{27}\,(0.05) = 40.11327$$

Συνεπώς, σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$, δεν απορρίπτουμε την μηδενική υπόθεση, οπότε μπορούμε να προχωρήσουμε στην δημιουργία του *p* – διαγράμματος ελέγχου φάσης *I*.

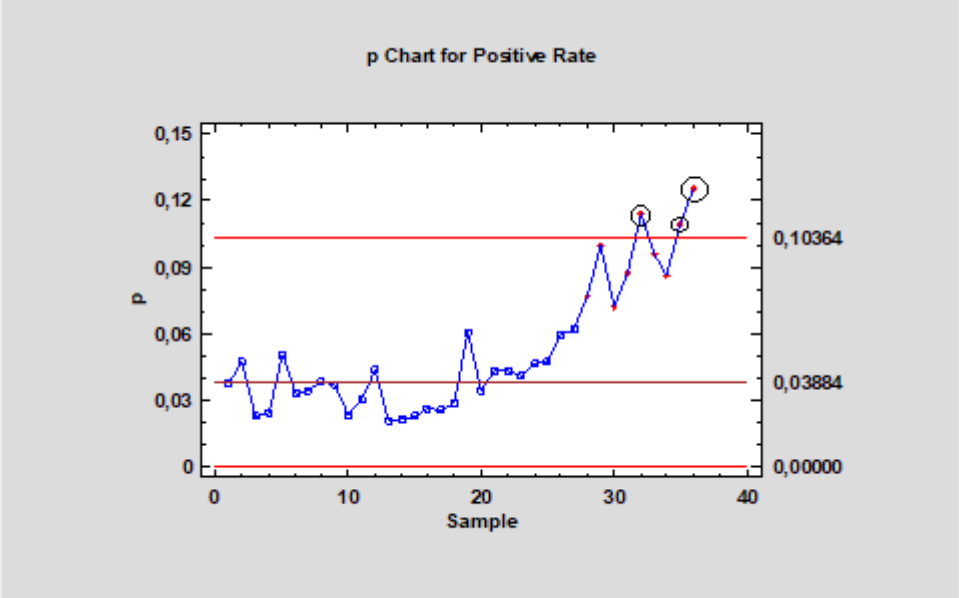
Για την δημιουργία του διαγράμματος θα χρησιμοποιηθεί το πρόγραμμα *Statgraphics*. Το διάγραμμα που λαμβάνουμε είναι το παρακάτω:



Διάγραμμα 5.2.1.
p – Διάγραμμα Ελέγχου Φάσης *I* για το ποσοστό των θετικών τεστ κορονοϊού

Από το παραπάνω διάγραμμα, παρατηρούμε ότι η διεργασία βρίσκεται εντός στατιστικού ελέγχου, καθώς κανένα σημείο δεν βρίσκεται εκτός των ορίων ελέγχου. Τα 3σ – όρια ελέγχου που μπορούν πλέον να χρησιμοποιηθούν για μελλοντική παρακολούθηση του δείκτη θετικότητας – ανάλυση φάσης *II* θα είναι από 0 (*LCL*) έως 0.103644. Η κεντρική γραμμή είναι 0.038839 (*CL*), δηλαδή για να είναι η διεργασία εντός στατιστικού ελέγχου επιτρέπεται να έχουμε κατά μέσο όρο 3.8% θετικά τεστ.

Στην συνέχεια, τα παραπάνω όρια ελέγχου χρησιμοποιήθηκαν για μελλοντική παρακολούθηση της διεργασίας – ανάλυση φάσης *II*. Συγκεκριμένα μελετήθηκαν τα ποσοστά θετικότητας για την περίοδο 29/10/2020 έως 05/11/2020.



Διάγραμμα 5.2.2.
p – Διάγραμμα Ελέγχου Φάσης *II* για το ποσοστό των θετικών τεστ κορονοϊού

Από το παραπάνω διάγραμμα, παρατηρούμε ότι η διεργασία δίνει ένδειξη ότι βρίσκεται εκτός στατιστικού ελέγχου. Συγκεκριμένα 3 από τα οχτώ καινούργια σημεία που προστέθηκαν βρίσκονται εκτός του άνω ορίου ελέγχου. Επιπλέον, από την 23η ημέρα της επιτήρησής μας, φαίνεται πως ο μέσος της διεργασίας έχει μετατοπιστεί σε υψηλότερα επίπεδα. Συνεπώς, οι υγειονομικές αρχές θα πρέπει να ερευνήσουν τις ειδικές αιτίες μεταβλητότητας που οδήγησαν στην αύξηση των κρουσμάτων (χαλάρωση υγειονομικών μέτρων, εμφάνιση νέου μεταδοτικότερου στελέχους της νόσου), καθώς επίσης θα πρέπει να τεθούν σε αυξημένη επαγρύπνηση και να λάβουν τα κατάλληλα μέτρα αναχαίτησης της αυξητικής πορείας της νόσου, καθώς είναι φανερό πως ο δείκτης θετικότητας των κρουσμάτων *Covid* – 19 έχει αρχίσει να αυξάνεται.

5.3 Εφαρμογή Μεθόδου Στατιστικής Σάρωσης σε δεδομένα κορονοϊού

Για την παρούσα εφαρμογή χρησιμοποιήθηκαν τα επιβεβαιωμένα κρούσματα *Covid* – 19 για την περίοδο 20/06/2022 έως 10/07/2022 που καταγράφηκαν για την Αττική, και πιο συγκεκριμένα για την Ανατολική Αττική, τον Βόρειο Τομέα Αθηνών, την Δυτική Αττική, τον Δυτικό Τομέα Αθηνών, τον Κεντρικό Τομέα Αθηνών, τον Νότιο Τομέα Αθηνών και τον Τομέα Πειραιά. Θέλουμε να βρεθεί η πιο πιθανή συστάδα, καθώς και η δευτερεύουσες συστάδες χρησιμοποιώντας καθαρά χωρική στατιστική συνάρτηση σάρωσης, λαμβάνοντας ως δεδομένο ότι οι καταγραμμένες περιπτώσεις

νόσου ακολουθούν την κατανομή *Bernoulli*. Για το σκοπό αυτό, θα χρησιμοποιηθεί το λογισμικό *SatScan*, το οποίο αναλύθηκε στο κεφάλαιο 2.

Το *SatScan* απαιτεί να δοθούν οι γεωγραφικές συντεταγμένες των υπό μελέτη περιοχών. Για τον σκοπό, χρησιμοποιήθηκαν οι γεωγραφικές συντεταγμένες των εδρών των περιοχών που θέλουμε να αναλύσουμε, και πιο συγκεκριμένα ο Ωρωπός για την Ανατολική Αττική, το Χαλάνδρι για τον Βόρειο Τομέα Αθηνών, η Ελευσίνα για την Δυτική Αττική, το Περιστέρι για τον Δυτικό Τομέα Αθηνών, η Αθήνα για τον Κεντρικό Τομέα Αθηνών, η Γλυφάδα για τον Νότιο Τομέα Αθηνών και ο Πειραιάς για Τομέα Πειραιά. Οι συντεταγμένες των παραπάνω περιοχών βρέθηκαν μέσω της εφαρμογής *Google Maps* της *Google*.

ΑΝΑΤΟΛΙΚΗ ΑΤΤΙΚΗ	38.304270	23.75362
ΒΟΡΕΙΟΣ ΤΟΜΕΑΣ ΑΘΗΝΩΝ	38.02000	23.81245
ΔΥΤΙΚΗ ΑΤΤΙΚΗ	38.04272	23.53642
ΔΥΤΙΚΟΣ ΤΟΜΕΑΣ ΑΘΗΝΩΝ	38.01229	23.6875
ΚΕΝΤΡΙΚΟΣ ΤΟΜΕΑΣ ΑΘΗΝΩΝ	37.98512	23.72731
ΝΟΤΙΟΣ ΤΟΜΕΑΣ ΑΘΗΝΩΝ	37.87789	23.75911
ΠΕΙΡΑΙΑΣ	37.9429	23.75362

Πίνακας 5.3.1. Γεωγραφικές Συντεταγμένες των υπό μελέτη περιοχών

Τέλος, επιλέχθηκε πως το μέγιστο μέγεθος του παραθύρου θα περιλαμβάνει το 10% του πληθυσμού σε κίνδυνο. Πραγματοποιώντας την χωρική σάρωση λαμβάνουμε τα κάτωθι αποτελέσματα:

```
Purely Spatial analysis
scanning for clusters with high rates
using the Bernoulli model.
```

```
SUMMARY OF DATA

Study period.....................: 2022/06/20 to 2022/07/10
Number of locations..............: 7
Total population.................: 4201148
Total number of cases............: 169558
Percent cases in area............: 4.0
```

```
CLUSTERS DETECTED

1.Location IDs included.: Boreios_Tomeas_Athinwn
  Overlap with clusters.: No Overlap
  Coordinates / radius..: (38.020000 N, 23.812450 E) / 0 km
  Gini Cluster.........: Yes
  Population...........: 621449
  Number of cases......: 28959
  Expected cases.......: 25081.63
  Observed / expected...: 1.15
  Relative risk........: 1.19
  Percent cases in area.: 4.7
  Log likelihood ratio..: 352.671090
  P-value..............: < 0.00000000000000001
2.Location IDs included.: Anatoliki_Attiki
  Overlap with clusters.: No Overlap
  Coordinates / radius..: (38.304270 N, 23.753620 E) / 0 km
  Gini Cluster.........: Yes
  Population...........: 424200
  Number of cases......: 20282
  Expected cases.......: 17120.68
  Observed / expected...: 1.18
  Relative risk........: 1.21
  Percent cases in area.: 4.8
  Log likelihood ratio..: 322.093226
  P-value..............: < 0.00000000000000001

3.Location IDs included.: Peiraias
  Overlap with clusters.: No Overlap
  Coordinates / radius..: (37.942900 N, 23.753620 E) / 0 km
  Gini Cluster.........: Yes
  Population...........: 469658
  Number of cases......: 20661
  Expected cases.......: 18955.36
  Observed / expected...: 1.09
  Relative risk........: 1.10
  Percent cases in area.: 4.4
  Log likelihood ratio..: 87.867892
  P-value..............: < 0.00000000000000001

4.Location IDs included.: Notios_Tomeas_Athinwn
  Overlap with clusters.: No Overlap
  Coordinates / radius..: (37.877890 N, 23.759110 E) / 0 km
  Gini Cluster.........: Yes
  Population...........: 552799
  Number of cases......: 22973
  Expected cases.......: 22310.92
  Observed / expected...: 1.03
  Relative risk........: 1.03
  Percent cases in area.: 4.2
  Log likelihood ratio..: 11.694419
  P-value..............: 0.000000084
```

Από την ανάλυση που προηγήθηκε, μελετήθηκαν 7 περιοχές με συνολικό πληθυσμό 4201148. Για την περίοδο από 20/06/2022 έως 10/07/2022 καταγράφηκαν στο σύνολο 169558 περιπτώσεις κρουσμάτων *Covid* – 19. Η πιο πιθανή συστάδα περιλαμβάνει τον Βόρειο Τομέα Αθηνών όπου βρέθηκαν 28959 κρούσματα κορονοϊού. Ο αναμενόμενος αριθμός περιπτώσεων με βάση την κατανομή *Bernoulli* είναι 25082. Ο λόγος *Observed*/*Expected* είναι ίσος με 1.16, δηλαδή σε αυτή την συστάδα υπάρχει 16% μεγαλύτερος κίνδυνος εμφάνισης θετικού κρούσματος *Covid* - 19 σε σχέση με την συνολική υπό μελέτη περιοχή. . Ο σχετικός κίνδυνος είναι ίσος με 1.19, ενώ η αύξηση του κινδύνου σε αυτή στην συστάδα είναι στατιστικά σημαντική σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$, καθώς το p – value του ελέγχου είναι ίσο με 0. Συνεπώς η διασπορά των υπό μελέτη γεγονότων στον χώρο δεν είναι τυχαία και οι υγειονομικές αρχές θα πρέπει να λάβουν τα απαραίτητα μέτρα προστασίας έναντι της *Covid* – 19. Αντίστοιχα, από την ανάλυση προέκυψαν άλλες 3 δευτερεύουσες συστάδες. . Συγκεκριμένα, η δεύτερη συστάδα αποτελείται από την Ανατολική Αττική, η τρίτη από τον τομέα Πειραιά ενώ η τέταρτη από τον Νότιο Τομέα Αθηνών.

Στόχος της παραπάνω μελέτης, είναι να βρεθεί αν η συχνότητα εμφάνισης κρουσμάτων *Covid* – 19, σχετίζεται με τις υπό μελέτη περιοχές και πιο συγκεκριμένα με τις περιφερειακές ενότητες του νομού Αττικής. Από την ανάλυση των αποτελεσμάτων, μεγαλύτερη συγκέντρωση κρουσμάτων σημειώθηκε στον Βόρειο Τομέα Αθηνών, ενώ ακολουθούν η Ανατολική Αττική, ο Πειραιάς, καθώς και ο Νότιος Τομέας Αθηνών. Συνεπώς, για τις περιφέρειες για τις οποίες βρέθηκαν στατιστικά σημαντικές συστάδες, θα πρέπει οι υγειονομικές αρχές να λάβουν τα απαραίτητα μέτρα για την αναχαίτηση του αυξανόμενου αριθμού κρουσμάτων. Η δυνατότητα που προσφέρει το πακέτο *SatScan* μέσω της χωρικής σάρωσης, να εντοπίζει ακριβώς τις περιοχές εκείνες που αποτελούν στατιστικά σημαντικές συστάδες, δίνει την ευκαιρία στους ειδικούς να ερευνήσουν περαιτέρω και πιο στοχευμένα τις περιοχές, εξετάζοντας τους λόγους που εμφανίζουν σημαντικό αριθμό κρουσμάτων και εν συνεχεία προχωρήσουν στις απαραίτητες ενέργειας με στόχο την αναχαίτηση μιας επικείμενης πανδημικής κατάστασης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Αγιακλόγλου. Φ. (2018). Σημειώσεις Μαθήματος *Χρονοσειρές*, ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιώς.

Αντζουλάκος Δ. (2019). Σημειώσεις Μαθήματος *Στατιστικός Έλεγχος Ποιότητας*, ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιώς.

Ασημακόπουλος Β. (2020). Σημειώσεις μαθήματος *Τεχνικές Προβλέψεων*, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών.

Κυριακίδης Μ. (2017). Σημειώσεις μαθήματος *Τεχνικές Ανάλυσης και Πρόβλεψης Τηλεπικοινωνιακών Αγορών*, Εθνικό Καποδιστριακό Πανεπιστήμιο Αθηνών.

Κούτρας Μ. Β. (2019). Σημειώσεις μαθήματος *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση - Ανάλυση σε συστάδες*, ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιώς.

Μπερσίμης Σ. (2019). Σημειώσεις Μαθήματος *Βιοστατιστική και Επιδημιολογία*, ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιά, Πανεπιστήμιο Πειραιώς.

Μπερσίμης Σ. (2020). Σημειώσεις μαθήματος *Βιοστατιστική και Επιδημιολογία*, Μοντέλα Επιτήρησης της Δημόσιας Υγείας. Στόχοι - Πλεονεκτήματα – Εφαρμογές, , ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιώς.

Μπούτσικας. Χ. (2019). Σημειώσεις Μαθήματος *Μέθοδοι Προσομοίωσης και Εφαρμογές*, ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιώς.

Ντζούφρας Ι. (2006). Σημειώσεις Μαθήματος *Βιοστατιστική ΙΙ*, Οικονομικό Πανεπιστήμιο Αθηνών.

Σαχλάς. Α. (2020). Σημειώσεις μαθήματος *Βιοστατιστική και Επιδημιολογία*, Στατιστικές Συναρτήσεις Σάρωσης στην Επιδημιολογία, ΠΜΣ στην *Εφαρμοσμένη Στατιστική*, Πανεπιστήμιο Πειραιώς.

Ταγάρας Ν.Γ. (2001). *Στατιστικός Έλεγχος Ποιότητας*, Εκδόσεις ΖΗΤΗ.

Αγγλική

Adeoti O. (2013). Application of Cusum Chart for Monitoring HIV/AIDS Patients in Nigeria, Federal University of Technology, Akure.

Alemi, F., Hankins R., Rice J. (1990). Predicting in-hospital survival of myocardial infarction. *Medical Care*, **28,** No 9.

Alemi F., Rom WA., Eisenstein E. (1996). Risk adjusted control charts for health care assessment, *Annals of Operations Research*, **67**, 45 – 60.

Arul E., Mark I C., Donald N., Leo Y. (2004). Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, *BMC Health Services Research*, **5**:36.

Benneyan J.C. (1998). Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part I Introduction and Basic Theory, *The Society for Healthcare Epidemiology of America*, **19**, 194 – 214.

Benneyan J.C., Lioyd R.C., Plsek P.E. (2003). Statistical Process Control as a tool for research and healthcare improvement, *Quality and Safety Health* Care, **12**, 458 – 464.

Bersimis S, Sachlas A, Koutras M.V. (2020). Health Monitoring Techniques Using Scan Statistics, Handbook of Scan Statistics, Springer Science+Business Media, LLC.

Burkom H. S. (2003). Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources, *Journal of Urban Health*: *Bulletin of the New York Academy of Medicine*, The New York Academy of Medicine, **80**, 2.

Correia, F., Nêveda, R., Oliveira, P. (2011). Chronic respiratory patient control by multivariate charts, *International journal of health care quality assurance*, **24**, 621– 643.

Cucala L. (2016). Scan Statistics for Detecting High-Variance Clusters, *Journal of Probability and statistics*, **2016**.

Desjardins, M.R., Hohl A., Delmelle E.M. (2020). Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters, *Applied Geography*, Elsevier, **118**.

Earnest, A., Chen, M.I., Ng, D. (2005). Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research*, **5**, 36.

Finison, J., Finison K, Bliersbach C.M. (1993). The Use of Control Charts to Improve Healthcare Quality, *Journal of Healthcare Quality*, **15**, 9 – 23.

Hanslik, T., Boelle P., Flahault A. (2001). The control chart :an epidemiological tool for public health monitoring, *Public Health*, **115**, 277-281.

Hendryx MS., Dyck DG., Srebnik D. (1999). Risk-adjusted outcome models for public mental health outpatient programs, *Health Services Research*, **34**(1), 171 – 195.

Imam N., Spelman T., Jognson S.A. (2019). Statistical Process Control Charts for Monitoring Staphylococcus aureus Bloodstream Infections in Australian Health Care Facilities, Walters Kluwer Health Inc, *Quality Management in Health Care*, **28**, 39 – 44.

Jebb A.T., Tay L., Wang W., Huang Q. (2015). Time series analysis for psychological research: examining and forecasting change, *Frontiers in Psychology*, **6**, Article 727.

Koetsier A., F de Keizer N., Jonge E., Cook D.A., Peek N. (2012). Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: A simulation study, *Critical Care Medicine*, **40**, 1799 – 1807.

Kulldorf, M. (1997). Communications in Statistics – Theory and Methods, A Spatial Scan Statistic, *Communications in Statistics – Theory and Methods*, Taylor & Francis, **26**, 1481 – 1496.

Kulldorf, M. (1999). Spatial Scan Statistics: Models, Calculations, and Applications, Glaz J, Balakrishnan N, *Scan Statistics and Applications*, Boston, MA: Birkhauser, 303 – 322.

Kulldorf, M. (2001). Prospective time periodic geographical disease surveillance using scan statistic, *Journal of the Royal Statistical Society*, **164**, 61 – 72.

Lai D. (2005). Monitoring the SARS Epidemic in China: A Time Series Analysis, *Journal of Data Science*, **3**, 279 – 293

Limaye Shreyas S., Mastrangelo Christina M., Zerr Danielle M. (2008). A Case Study in Monitoring Hospital-Associated Infections with Count Control Charts, *Quality Engineering*, **20**:4, 404 – 413.

Lucas, J. (1985). Counted data CUSUM's. *Technometrics*, JSTOR, **27**, 129 – 144.

Mahmood Y., Ishtiaq S., Khoo M., Khan H. (2020). Monitoring of three - pase variations in the mortality of Covid 19 pandemic using control charts: where does Pakistan stand?, *International Journal for Quality in Health Care*, **33**, 1 – 8.

Mbaye M. F., Sarr n., Ngom B. (2021). Construction of Control Charts for Monitoring Various Parameters Related to the Management of the COVID-19 Pandemic, *Journal of Biosciences and Medicines*, Scientific Research and Publishing, **9**, 9 – 19.

Mohammed M.A., Worthington P., Woodall. (2007). Plotting basic control charts: tutorial notes for healthcare practitioners, *Quality and Safety in Health Care*, **17**, 137 – 145.

Mohammed M.A, Laney D. (2006). Overdispresion in health care performance data: Laney's Approach, *Quality and Safety in Health Care*, **15**, 383 – 384.

Mohtashemi M., Szolovits P., Dunyak J., Mandi K.D. (2006). A susceptible – infected model of early detection of respiratory infection outbreaks on a background of influenza, J Theor Biol., **241**, 954 – 963.

Montgomery D., Jennings C., Kulahci M.. (2015). *Introduction to Time Series Analysis and Forecasting*, Wiley Series in Probability and Statistics, John Wiley, New York.

Montgomery, D.C. (1991). *Introduction to Statistical Quality Control*, 6th Edition. New York, John Wiley and Sons, Inc, New York.

Naus, J., Wallenstein, S. (2006). Temporal surveillance using scan statistics, *Statistics in Medicine*, **25**, 311 – 324.

Neill D. B. (2006), Detection of Spatial and Spatio-Temporal Clusters, School of Computer Science Carnegie Mellon University Pittsburgh.

Otunuga M.O. (2021). Time – dependent probability distribution for number of infection in a stochastic SIS model: case study COVID – 19, Department of Mathematics, Marshall University, Huntington.

Ramesh C. (2020). Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries, *Chaos, Solitons and Fractals*, **140**.

Reis B.Y., Pagano M., Mandl K.D.(2003). Using temporal context to improve biosurveillance, *Proceedings of the National Academy of Sciences*, **100**, 1961 – 1965.

Renato C. (2012). Disease management with ARIMA model in time series, *Einstein* (*Sao Paolo*), **11**, 128 – 131.

Robert Nau, Lecture notes on forecasting, Fuqua School of Business Duke University.

Rogerson, P. A. (1997). Surveillance systems for monitoring the development of spatial patterns, *Statistics in Medicine*, **16**, 2081 – 2093.

Sellick Jr. J.A. (1993). The Use of Statistical Process Control Charts in Hospital Epidemiology, *Infection Control and Hospital Epidemiology*, **14**, 649 – 656.

Sen D., Sen D. (2021). Use of a Modified SIRD Model to Analyze Covid – 19 Data, *Industrial & Engineering Chemistry Research*, **60**, 4251 – 4260.

Shtatland E.S., Shtatland T. (2008). Another Look at Low-Order Autoregressive Models in Early Detection of Epidemic Outbreaks and Explosive Behaviors in Economic and Financial Time Series, SAS Global Forum, *Statistics and Data Analysis*.

Shtatland E.S., Kleinman K.P., Cain E.M., (2008), Biosurveillance and Outbreak Detection Using the Arima and LOGISTIC Procedures, *Statistics and Data Analysis*, 197 – 31.

Singh B.P., Madhusudan J.V. (2020). Evaluation of EWMA Control Charts for Monitoring Spread of Transformed Observations of Covid 19 in India, *Asian Journal of Research in Infectious Diseases*, **5**, 25 – 36.

Stanley J., Edwin M. (2016). Modelling Epidemiological Data Using Box-Jenkins Procedure, *Open Journal of Statistics*, **6**, 295 – 302.

Tracy N.D., Young J.C., Mason R.L. (2018), Multivariate Control Charts for Individual Observations, *Journal of Quality Technology*, **24**:2, 88 – 95.

Toshiro T., Kunihiko T., Kazuaki K. (2011). A Space-Time Scan Statistic for Detecting Emerging Outbreaks, *International Biometric Society*, **67**, 106 – 115.

Tsui K. L., Chiu, W., Gierlich, P., Goldsman, D., Liu, X. and Maschek, T. (2008). A review of healthcare, public health, and syndromic surveillance, *Quality Engineering*, **20**(4), 435–450.

Turnbull, B. W., E. J. Iwano, W. S. Burnett, H. L. Howe, and L. C. Clark. (1990). Monitoring for clusters of disease: Application to leukemia in upstate New York. *American Journal of Epidemiology*, **132**(1), 136 – 143.

Wolfe H., Taylor A., Subramanyam R. (2020). Statistics in quality improvement: Measurement and statistical process control, *Pediatric Anesthesia*, **31**, 539 – 547.

Woodall W.H. (2006). The use of control charts in Health – Care and Public Health Surveillance, Journal of Quality Technology, **38**, 89 – 104.