

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**



**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ  
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ  
ΣΠΟΥΔΩΝ  
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

**ΕΦΑΡΜΟΓΗ ΤΩΝ ΤΕΧΝΙΚΩΝ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ  
ΕΛΕΓΧΟΥ ΔΙΕΡΓΑΣΙΑΣ ΣΤΗ ΒΙΟΕΠΙΤΗΡΗΣΗ**

**Αθανάσιος Β. Καβάζης**

*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,  
Ιανουάριος 2011*



Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίστηκε από τη ΓΣΕΣ του τμήματος 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:

- Κούτρας Μάρκος ( Επιβλέπων Καθηγητής)
- Αντζουλάκος Δημήτριος
- Στέγος Δημήτριος

Η έγκριση της Διπλωματικής Εργασίας από το τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνώμων του συγγραφέα.



**UNIVERSITY OF PIRAEUS**



**DEPARTMENT OF STATISTICS  
AND INSURANCE SCIENCE**

**POSTGRADUATE PROGRAM IN  
APPLIED STATISTICS**

**STATISTICAL PROCESS CONTROL METHODS IN  
BIOSURVEILLANCE**

By

**Athanasios V. Kavazis**

MSc Dissertation

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

Piraeus, Greece

January 2011



# ΠΑΝΕΓΧΕΙΜΟ ΠΑΤΕΡΑ

Στη μνήμη του πατέρα μου





## Ευχαριστίες

Θερμές ευχαριστίες θα ήθελα να απευθύνω στον Καθηγητή κ. Μάρκο Κούτρα για την καθοριστική συμβολή του στην εκπόνηση της παρούσας διπλωματικής εργασίας. Οι παρατηρήσεις, οι διορθώσεις και οι συμβουλές του συντέλεσαν αποφασιστικά στην ολοκλήρωση της εργασίας. Αισθάνομαι δε την υποχρέωση να εκφράσω την ευγνωμοσύνη μου στον Καθηγητή κ. Μάρκο Κούτρα γιατί χωρίς την υπομονή και την κατανόηση που επέδειξε στις δυσκολίες που παρουσιάστηκαν και χωρίς τη συνεχή παρότρυνσή του, η παρούσα προσπάθεια θα είχε μείνει ατελέσφορη.

Επιπλέον θα ήθελα να ευχαριστήσω όλους τους καθηγητές μου στο μεταπτυχιακό πρόγραμμα σπουδών και ιδιαίτερα τον Αναπληρωτή Καθηγητή κ. Δημήτριο Αντζουλάκο και τον Επίκουρο Καθηγητή κ. Δημήτριο Στέγγο, μέλη της τριμελούς επιτροπής, για την επίβλεψη της παρούσας εργασίας και τις χρήσιμες υποδείξεις τους.

Τέλος, θα ήταν σημαντική παράλειψη να μην ευχαριστήσω την Ελένη για τη διαρκή στήριξη και τις συμβουλές της, καθώς επίσης και την οικογένειά μου για την αμέριστη συμπαράστασή της καθ' όλη τη διάρκεια αυτής της προσπάθειας.



## ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια η ραγδαία αύξηση των κρουσμάτων μεταδοτικών ασθενειών, σε παγκόσμια κλίμακα, καθώς και η απειλή χρήσης βιολογικών παραγόντων σε τρομοκρατικές επιθέσεις, αναδεικνύουν την ανάγκη για έγκαιρη και έγκυρη ανίχνευση του ξεσπάσματος μιας επιδημίας. Η έγκαιρη ανίχνευση μιας επιδημίας προϋποθέτει, όχι μόνο τη συλλογή δεδομένων σε πραγματικό χρόνο, αλλά και την ανάλυση και ερμηνεία αυτών των δεδομένων σε πραγματικό χρόνο. Την τελευταία δεκαετία έχει γίνει σημαντική πρόοδος ως προς την ταχύτητα συλλογής των δεδομένων και την ποιότητά τους, ενώ για την ανάλυσή τους μπορούν να εφαρμοστούν αποτελεσματικά οι μέθοδοι του στατιστικού ελέγχου διεργασίας.

Στην παρούσα εργασία, περιγράφονται αναλυτικά διάφορες τεχνικές του στατιστικού ελέγχου διεργασίας που μπορούν να χρησιμοποιηθούν με σκοπό την ανίχνευση του ξεσπάσματος μιας επιδημίας. Επίσης, δίνονται παραδείγματα εφαρμογής των τεχνικών αυτών, σε πραγματικά ή προσομοιωμένα δεδομένα, που αφορούν τη δημόσια υγεία.



## **ABSTRACT**

The emergence of new infectious diseases at a global scale as well as the threat of bioterrorism acts have highlighted the need for early detection of epidemics. In order to detect disease outbreaks as soon as possible, both real-time data collection and real-time interpretation of them are required. In the last decade, important advances in setting up techniques for real-time data collection has been noted. In addition statistical process control methods have been applied efficiently in order to detect an epidemic in a timely maner.

In the present dissertation we first provide a short presentation of statical process control techniques, then we describe the application of them methods in disease outbreak detection. Illustrative examples are also presented using either real or simulated public health data.



# ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη	xi
Abstract	xiii
Περιεχόμενα	xv
Κατάλογος Πινάκων	xvii
Κατάλογος Σχημάτων	xix
Κατάλογων Συντομογραφιών	xxi
<b>ΚΕΦΑΛΑΙΟ 1: Εισαγωγή</b>	
1.1 Επιδημιολογία	1
1.2 Βιοεπιτήρηση	1
1.3 Ιστορικά στοιχεία	3
1.4 Συστήματα και οργανισμοί που διενεργούν βιοεπιτήρηση	5
1.5 Βιοτρομοκρατία	8
1.6 Στατιστικός Έλεγχος Διεργασίας	9
<b>ΚΕΦΑΛΑΙΟ 2: Διαγράμματα Ελέγχου</b>	
2.1 Εισαγωγή	11
2.2 Διαγράμματα Στατιστικού Ελέγχου Διεργασίας τύπου <i>Shewhart</i>	12
2.2.1 Βασικές έννοιες	12
2.2.2 Κατανομές Πιθανότητας – Επιλογή Κατάλληλου Διαγράμματος	20
2.2.3 Τα διαγράμματα ελέγχου $p$ και $np$	21
2.2.4 Τα διαγράμματα ελέγχου $c$ και $u$	28
2.2.5 Τα διαγράμματα ελέγχου $\bar{X}$ , $S$ και $R$	33
2.2.6 Το διάγραμμα ελέγχου μεμονωμένων παρατηρήσεων και το $MR$ διάγραμμα	38
2.3 Διαγράμματα ελέγχου με μνήμη	42
2.3.1 Διαγράμματα Ελέγχου Κινούμενου Μέσου	43
2.3.2 Διάγραμμα Ελέγχου Εκθετικά Σταθμισμένου Κινούμενου Μέσου	45
2.3.3 Το αθροιστικό διάγραμμα	48

2.4 Διαγράμματα Προσαρμοσμένα στον Κίνδυνο	53
--	----

### **ΚΕΦΑΛΑΙΟ 3: Χρήση και αξιολόγηση μεθόδων σάρωσης στην βιοεπιτήρηση**

3.1 Εισαγωγή	59
3.2 Προοδευτική, χρονολογική και γεωγραφική βιοεπιτήρηση με τη χρήση μεθόδων σάρωσης	64
3.3 Μόνο γεωγραφική βιοεπιτήρηση	67
3.4 Διαγραφή αναδυόμενων συστάδων	71
3.5 Περιοδική ως προς το χρόνο επιτήρηση	74
3.6 Η μέθοδος <i>P</i> -σάρωσης	79
3.6.1 Εισαγωγή	79
3.6.2 Αναδρομική Ανάλυση, συνεχής χρόνος	82
3.6.3 Παράδειγμα	82
3.6.4 Προοδευτική ανάλυση, συνεχής χρόνος	83
3.6.5 Προοδευτική ανάλυση, ομαδοποιημένα δεδομένα	84
3.6.6 Παράδειγμα	84
3.6.7 Προοδευτική ανάλυση, δίτιμα δεδομένα	85
3.6.8 Παράδειγμα	85

### **ΚΕΦΑΛΑΙΟ 4: Χρονολογικές Σειρές**

4.1 Εισαγωγή	89
4.2 Τα μοντέλα SIR και η εφαρμογή τους στην έγκαιρη αντίχνευση	90
4.3 Εφαρμογή του Αυτοπαλίνδρομου Μοντέλου στη Βιοεπιτήρηση	94
4.4 Εφαρμογή του Μοντέλου <i>ARMA</i> στην Βιοεπιτήρηση: Πλεονεκτήματα και Μειονεκτήματα	97
4.5 Συνδυασμός των μοντέλων αυτοπαλίνδρομου κινούμενου μέσου και των μεθόδων παλινδρόμησης	99

<b>Βιβλιογραφία</b>	<b>101</b>
---------------------	------------



## ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

2.1	Όρια ελέγχου για το $np$ –διάγραμμα	23
2.2	Όρια ελέγχου για το $p$ –διάγραμμα	23
2.3	Αριθμός ασθενών που μολύνθηκαν σε επέμβαση αφαίρεσης σπύλου	25
2.4	Αριθμός ασθενών που μολύνθηκαν σε επέμβαση αφαίρεσης σπύλου μετά την εφαρμογή της νέας μεθόδου	26
2.5	Αριθμός ασθενών που μολύνθηκαν σε επέμβαση αφαίρεσης σπύλου τις εβδομάδες 11 – 20 μετά την εφαρμογή της νέας μεθόδου	28
2.6	Όρια ελέγχου για το $c$ - διάγραμμα	30
2.7	Όρια ελέγχου για το $u$ - διάγραμμα	31
2.8	Πτώσεις ασθενών που νοσηλεύονται σε ένα νοσοκομείο	32
2.9	Όρια ελέγχου Φάσης I για το $\bar{X}$ - διάγραμμα	36
2.10	Όρια Ελέγχου για το $R$ –διάγραμμα	37
2.11	Όρια Ελέγχου για το $S$ –διάγραμμα	38
2.12	Όρια Ελέγχου για το $x$ –διάγραμμα	40
2.13	Όρια Ελέγχου για το $MR$ –διάγραμμα	40
2.14	Συστολική πίεση ασθενούς (σε $mmHg$ )	41
2.15	Όρια Ελέγχου για το $MA$ –διάγραμμα	44
2.16	Όρια Ελέγχου για το $EWMA$ διάγραμμα	46
2.17	$Poisson$ $CUSUM$ με παραμέτρους $h = 10$ , $k = 5$ . [Πηγή: <i>Lucas</i> , (1985)]	51
3.1	Γεωγραφική ανάλυση περιστατικών καρκίνου του θυρεοειδούς σε άντρες στο Νέο Μεξικό, με χρήση χωρικής συνάρτησης σάρωσης [Πηγή: <i>Kulldorf</i> , (2000)]	70
3.2	Γεωγραφική ανάλυση περιστατικών καρκίνου του θυρεοειδούς σε άντρες στο Νέο Μεξικό, με χρήση χωρικής – χρονολογικής συνάρτησης σάρωσης [Πηγή: <i>Kulldorf</i> , (2000)]	73
3.3	Ανίχνευση συστάδων καρκίνου θυρεοειδούς σε άντρες στο Νέο Μεξικό.[Πηγή: <i>Kulldorff</i> , (2000)]	75

# ТАНЕЦЫ И ТЕАТР

## ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Εικόνα 1.1	Ο χάρτης του <i>John Snow</i> με τους θανάτους από χολέρα	4
Εικόνα 1.2	Περιβάλλον χρήστη στο <i>RODS</i>	6
Διάγραμμα 2.1	Τυπικό Διάγραμμα Ελέγχου	14
Διάγραμμα 2.2	Διάγραμμα Ελέγχου με προειδοποιητικά όρια	16
Διάγραμμα 2.3	$p$ –διάγραμμα ελέγχου Φάσης I για το ποσοστό των μολύνσεων από καθετήρα	26
Διάγραμμα 2.4	$p$ –διάγραμμα ελέγχου Φάσης II για το ποσοστό των μολύνσεων από καθετήρα, 4 <sup>η</sup> εβδομάδα	27
Διάγραμμα 2.5	$p$ –διάγραμμα ελέγχου Φάσης II για το ποσοστό των μολύνσεων από καθετήρα, 14 <sup>η</sup> εβδομάδα μετά την εφαρμογή της νέας μεθόδου	28
Διάγραμμα 2.6	$u$ -διάγραμμα ελέγχου για τις πτώσεις ασθενών εντός ενός νοσοκομείου	32
Διάγραμμα 2.7	$x$ -διάγραμμα για τη συστολική πίεση ασθενούς	41
Διάγραμμα 2.8	<i>MR</i> - διάγραμμα για τη συστολική πίεση ασθενούς	42
Διάγραμμα 2.9	<i>MA</i> διάγραμμα με $k = 6$	44
Διάγραμμα 2.10	<i>MA</i> διάγραμμα για κρούσματα γονόρροιας [Πηγή: <i>Shmueli</i> , (2006)]	45
Διάγραμμα 2.11	<i>EWMA</i> διάγραμμα με $\lambda = 0.1$	47
Διάγραμμα 2.12	<i>EWMA</i> διάγραμμα με $\lambda = 0.2$	47
Διάγραμμα 2.13	Μη προσαρμοσμένο στον κίνδυνο διάγραμμα ελέγχου με μεταβλητά όρια πιθανότητας ( $\alpha = 0.1$ ) για παρακολούθηση του ποσοστού θνησιμότητας. [Πηγή: <i>Alemi et al.</i> , (1996)]	55
Διάγραμμα 2.14	Προσαρμοσμένο στον κίνδυνο διάγραμμα ελέγχου με μεταβλητά όρια πιθανότητας ( $\alpha = 0.1$ ) για παρακολούθηση του ποσοστού θνησιμότητας. Πηγή: <i>Alemi et al.</i> , (1996)	58
Εικόνα 3.1	Χάρτης της Πολιτείας του Νέου Μεξικού , Γεωγραφική συστάδα με την μεγαλύτερη πιθανότητα [Πηγή: <i>Kulldorf</i> , (2000)]	70
Διάγραμμα 3.1	Αθροιστικές Συχνότητες Εμφάνισης Καρκίνου του Θυρεοειδούς σε άντρες στο <i>Los Alamos</i>	75
Διάγραμμα 4.1	Ανίχνευση έξαρσης της μεταδοτικότητας αναπνευστικού συνδρόμου [Πηγή: <i>Mohtashemy et al.</i> (2006)]	94

# РАНЕЕЗНАМО ПЕРПАА

## ΚΑΤΑΛΟΓΟΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

<i>CL</i>	Κεντρική γραμμή
<i>UCL</i>	Άνω όριο ελέγχου
<i>LCL</i>	Κάτω όριο ελέγχου
<i>ARL<sup>0</sup></i>	Εντός ελέγχου μέσο μήκος ροής
<i>ARL<sup>1</sup></i>	Εκτός ελέγχου μέσο μήκος ροής
<i>MR</i>	Κινούμενο εύρος
<i>MA</i>	Κινούμενος μέσος
<i>EWMA</i>	Εκθετικά σταθμισμένος κινούμενος μέσος
<i>CUSUM</i>	Αθροιστικό διάγραμμα
<i>AR</i>	Αυτοπαλίνδρομο Μοντέλο
<i>ARMA</i>	Μοντέλο αυτοπαλίνδρομου κινούμενου μέσου
<i>ARIMA</i>	Μοντέλο αυτοπαλίνδρομου ολοκληρωμένου κινούμενου μέσου
<i>SIR</i>	Μοντέλο ευαισθησίας-μόλυνσης-ανάρρωσης



# ΚΕΦΑΛΑΙΟ 1

## Εισαγωγή

### 1.1 Επιδημιολογία

Επιδημιολογία είναι ο κλάδος της επιστήμης που μελετά την εξάπλωση μιας ασθένειας σε διάφορους πληθυσμούς (ανθρώπους, ζώα κ.λ.π.) και τους λόγους που επηρεάζουν αυτή την εξάπλωση. Είναι η δυναμική μελέτη των προσδιοριστικών παραγόντων, της εμφάνισης, και της διασποράς της υγείας και της νόσου σε ένα πληθυσμό. Περιλαμβάνει τον προσδιορισμό, την ανάλυση, και την ερμηνεία των ποσοστών εξάπλωσης μιας νόσου στον πληθυσμό. Η επιδημιολογία στοχεύει όχι μόνο στον εντοπισμό και την αξιολόγηση του προβλήματος, αλλά και στη λήψη μέτρων αντιμετώπισης του προβλήματος και την εκτίμηση της αποτελεσματικότητας των μέτρων αυτών. Πρόκειται για την ακρογωνιαία λίθο στην προστασία της δημόσιας υγείας.

Σε αντίθεση με την κλινική ιατρική, όπου δίδεται έμφαση στη θεραπεία του ατόμου, η επιδημιολογία ασχολείται με την εκτίμηση διαφόρων μορφών ασθενειών που αναπτύσσονται μέσα σε ομάδες ατόμων. Συνεπώς, τα συμπεράσματα που εξάγονται από τις επιδημιολογικές μελέτες είναι γενικού ενδιαφέροντος και δεν αναφέρονται σε συγκεκριμένα άτομα. Έτσι, οι επιδημιολόγοι χρησιμοποιούν ένα πλήθος από άλλες επιστήμες όπως η βιολογία, η κοινωνιολογία, η γεωγραφία και η στατιστική. Συγκεκριμένα η τελευταία βοηθάει στο σχεδιασμό μιας δειγματοληψίας, την ανάλυση των δεδομένων και την εξαγωγή, επιστημονικά τεκμηριωμένων, συμπερασμάτων.

Ένα πρόβλημα που προκύπτει από τη φύση της επιδημιολογίας είναι ότι δεν μπορεί να ελέγξει το ποιος θα εκτεθεί στον κίνδυνο.

### 1.2 Βιοεπιτήρηση

Μια διεργασία βιοεπιτήρησης είναι η συστηματική συλλογή και ανάλυση δεδομένων με σκοπό την ανίχνευση μιας ραγδαίας αύξησης του αριθμού των κρουσμάτων μιας νόσου και την αναγνώριση των συνθηκών που δημιουργούν προδιάθεση για τη νόσο.

Σκοπός της βιοεπιτήρησης είναι η καταγραφή των ασθενειών που υπάρχουν σε ένα πληθυσμό, η αναζήτηση των αιτιών που την προκαλούν, η κατανόηση του τρόπου με τον

οποίο αυτές εξαπλώνονται και η εκτίμηση της μελλοντικής εξάπλωσης. Έτσι, η βιοεπιτήρηση σχετίζεται έντονα με την επιδημιολογία και από αρκετούς θεωρείται ένα πεδίο της επιδημιολογίας. Όμως, η βιοεπιτήρηση διαφέρει από την επιδημιολογία καθώς επικεντρώνεται στη συνεχή παρακολούθηση, τη χρήση της πληροφορικής για την παροχή ποσοτικών αναφορών και στη δυνατότητα έγκαιρης παρέμβασης.

Τα συμπτώματα μιας νόσου είναι πιθανό να επιδεινωθούν δραματικά πολύ γρήγορα, όπως επίσης και τα κρούσματα μιας νόσου είναι πιθανό να εξαπλωθούν σε μια μεγάλη γεωγραφική περιοχή πολύ γρήγορα. Έτσι η επιδημία μιας νόσου μπορεί σε ελάχιστο χρονικό διάστημα να προκαλέσει μεγάλη ζημιά. Το περιθώριο αντίδρασης, για τον περιορισμό αυτής της ζημιάς, μπορεί να είναι λίγες μέρες.

Στις Η.Π.Α. και την Ε.Ε. αφιερώνονται δισεκατομμύρια δολάρια σε διάφορες μορφές βιοεπιτήρησης, με κυριότερες τον έλεγχο ενδονοσοκομειακών λοιμώξεων, την παρακολούθηση της δημόσιας υγείας, της ποιότητας του νερού και του αέρα. Ενδεικτικά, το 2005 ο προϋπολογισμός του Κέντρου Ελέγχου και Πρόληψης Ασθενειών (Centre for Disease Control and Prevention) στις Η.Π.Α. ήταν 8 δισεκατομμύρια δολάρια, τη στιγμή που η διάμεσος του προϋπολογισμού δαπανών υγείας, κάθε πολιτείας, ήταν 2.8 δισεκατομμύρια.

Οι οργανισμοί που ασχολούνται με την βιοεπιτήρηση συλλέγουν και αναλύουν δεδομένα ασταμάτητα. Με βάση τα αποτελέσματα αυτών των αναλύσεων, οι οργανισμοί αποφασίζουν αν πρέπει να ληφθούν επιπλέον μέτρα για την προστασία της δημόσιας υγείας. Όταν παρατηρηθεί μια, μη φυσιολογική, αύξηση του αριθμού των κρουσμάτων μιας ασθένειας, ή όταν αναφερθεί τουλάχιστον ένα κρούσμα μιας άκρως μεταδοτικής νόσου, τότε οι ερευνητές συλλέγουν επιπλέον πληροφορίες με στόχο την αναγνώριση των συνθηκών γύρω από τα περιστατικά. Αυτό οδηγεί σε περισσότερα ερωτήματα που πρέπει να απαντηθούν, δηλ. περισσότερα δεδομένα που πρέπει να συγκεντρωθούν και να αναλυθούν. Παράλληλα με αυτές τις διερευνητικές προσπάθειες, ο οργανισμός θα πρέπει να ξεκινήσει μια σειρά από ενέργειες, για τον περιορισμό της επιδημίας, όπως εμβολιασμούς, απομόνωση των ασθενών κλπ.

Στη σύγχρονη βιοεπιτήρηση, τα δεδομένα είναι, συνήθως, μετρήσεις που μπορεί να περιλαμβάνουν ενδείξεις αύξησης των κρουσμάτων μιας νόσου. Για παράδειγμα, τέτοια δεδομένα είναι ο αριθμός των επισκέψεων σε τμήματα επειγόντων περιστατικών νοσοκομείων, πωλήσεις φαρμάκων, απουσίες σε σχολεία, επισκέψεις σε ιατρεία και το πλήθος των αναζητήσεων στο διαδίκτυο με θέμα κάποια νόσο. Αυτά συλλέγονται από συγκεκριμένους οργανισμούς που ασχολούνται με τη δημόσια υγεία, αναλύονται και



αξιολογούνται. Το ενδιαφέρον δεν επικεντρώνεται μόνο σε επιδημίες που προκαλούνται από φυσιολογικά αίτια, π.χ. η γρίπη των πτηνών, αλλά και σε τρομοκρατικές ενέργειες, π.χ. η απελευθέρωση του αερίου Σαρίν (*Sarin*) στο μετρό του Τόκιο το 1995. Σε κάθε περίπτωση, σκοπός της βιοεπιτήρησης είναι να ειδοποιηθούν οι αρχές ώστε να λάβουν εγκαίρως τα κατάλληλα μέτρα.

Για να είναι αποτελεσματική η αντίδραση της πολιτείας θα πρέπει η διεργασία της βιοεπιτήρησης να ανιχνεύει γρήγορα το ξέσπασμα μιας επιδημίας. Επίσης, είναι αναγκαίο η βιοεπιτήρηση να μη περιορίζεται σε ένα αριθμό ασθενειών αλλά να μπορεί να ανιχνεύσει μια μη φυσιολογική αύξηση κρουσμάτων οποιασδήποτε ασθένειας. Τέλος, θα πρέπει η πιθανότητα εσφαλμένης ανίχνευσης μιας επιδημίας να είναι μικρή. Επομένως, η αποτελεσματικότητα της βιοεπιτήρησης μπορεί να επηρεαστεί από μια σειρά παραγόντων, όπως η εποχικότητα μιας νόσου, η έλλειψη στασιμότητας ή η αυτοσυσχέτιση στα δεδομένα. Άλλοι παράγοντες που μπορεί να επηρεάσουν αρνητικά την αποτελεσματικότητα της διεργασίας είναι η καθυστέρηση στη συλλογή των δεδομένων, το μη εξειδικευμένο προσωπικό κλπ.

### 1.3 Ιστορικά στοιχεία

Η μελέτη του Ιπποκράτη για τη σχέση ανάμεσα στην ασθένεια και τους περιβαλλοντικούς παράγοντες είναι η πρώτη γνωστή επιδημιολογική μελέτη. Ο Ιπποκράτης διέκρινε τα νοσήματα σε επιδημικά, ενδημικά και σποραδικά. Έτσι, πολλοί θεωρούν ότι ο «πατέρας της ιατρικής» είναι και «πατέρας της επιδημιολογίας». Πολύ αργότερα από τον Ιπποκράτη, κατά τον 14<sup>ο</sup> αιώνα, όταν η βουβωνική πανώλη εξαπλώθηκε στην Ιβηρική χερσόνησο, γράφτηκαν δύο μελέτες σχετικά με τον τρόπο μετάδοσης λοιμωδών νοσημάτων, από τους Μαυριτανούς *Ibn Khatima* και *Ibn al-Khatib*. Η πραγματική εξέλιξη της επιδημιολογίας ξεκίνησε μετά την ανάπτυξη της θεωρίας ότι πολλές ασθένειες προκαλούνται από μικρόβια από τον Λουί Παστέρ (*Louis Jean Pasteur*).

Όταν στο Λονδίνο, το 1854, ξέσπασε επιδημία χολέρας, ο γιατρός *John Snow* μπόρεσε να εντοπίσει την αιτία που την προκάλεσε. Αρχικά παρατήρησε ότι το ποσοστό θανάτων στη συνοικία *Soho* ήταν σημαντικά υψηλότερο απ' ό,τι στο υπόλοιπο Λονδίνο. Το κοινό γνώρισμα ανάμεσα στις συνοικίες αυτές ήταν ότι προμηθεύονταν νερό από την ίδια εταιρία την *Southwark Company*. Παράλληλα, σημειώνοντας σε ένα χάρτη τους θανάτους από χολέρα μπόρεσε να εντοπίσει τελικά ότι η πηγή της μόλυνσης ήταν μια αντλία νερού στην οδό

*Broad*. Στην Εικόνα 1.1 φαίνεται ένα τμήμα του χάρτη της περιοχής, που χρησιμοποίησε ο *John Snow*, στον οποίο με τελεία σημειώνονται οι θάνατοι από χολέρα και με X οι θέσεις όπου βρίσκονταν οι αντλίες νερού. Είναι φανερό η σχέση ανάμεσα στην αντλία της οδού *Broad* και την επιδημία χολέρας. Αυτή η έρευνα θεωρείται πλέον ως ένα κλασικό παράδειγμα της επιδημιολογίας. Φυσικά, ο *John Snow*, δεν περιορίστηκε στον εντοπισμό της αιτίας της μόλυνσης αλλά παρενέβη και σταμάτησε την επιδημία, καθαρίζοντας με χλωρίνη την αντλία και αντικαθιστώντας τη χειρολαβή.

Κατά το 19<sup>ο</sup> αιώνα καθώς και το πρώτο μισό του 20<sup>ου</sup> η επιδημιολογία ασχολήθηκε με την απόδειξη της ύπαρξης περιβαλλοντικών παραγόντων που μπορούν να προκαλέσουν ασθένειες. Σπουδαίο παράδειγμα αποτελεί η συσχέτιση του καπνίσματος με τον καρκίνο του πνεύμονα, από τους *Doll* και *Hill* το 1956.



Εικόνα 1.1 Ο χάρτης του *John Snow* με τους θανάτους από χολέρα

Η ανάπτυξη της επιστήμης της βιοπληροφορικής (*bioinformatics*) είχε σαν αποτέλεσμα την ευκολότερη και ταχύτερη πρόσβαση σε πληροφορίες που αφορούν τη δημόσια υγεία. Έτσι, εκτός από τη χρήση των πληροφοριών αυτών για το σχεδιασμό μιας μελέτης, κατέστη εφικτή η συστηματική καταγραφή τους με σκοπό την παρακολούθηση της δημόσιας υγείας. Μπορούμε λοιπόν να πούμε ότι η βιοεπιτήρηση είναι η εξέλιξη των επιδημιολογικών μεθόδων για τη συνεχή παρακολούθηση της δημόσιας υγείας.

Η βιοεπιτήρηση έχει εξελιχθεί ραγδαία την τελευταία δεκαετία κυρίως λόγω του φόβου που υπάρχει για δύο ενδεχόμενα. Πρώτον, η απειλή μιας τρομοκρατικής επίθεσης με βιολογικά όπλα και δεύτερον η απειλή μιας πανδημίας από φυσιολογικά αίτια. Συνεπώς, το ενδιαφέρον έχει στραφεί στην έγκαιρη ανίχνευση του ξεσπάσματος μιας ασθένειας.

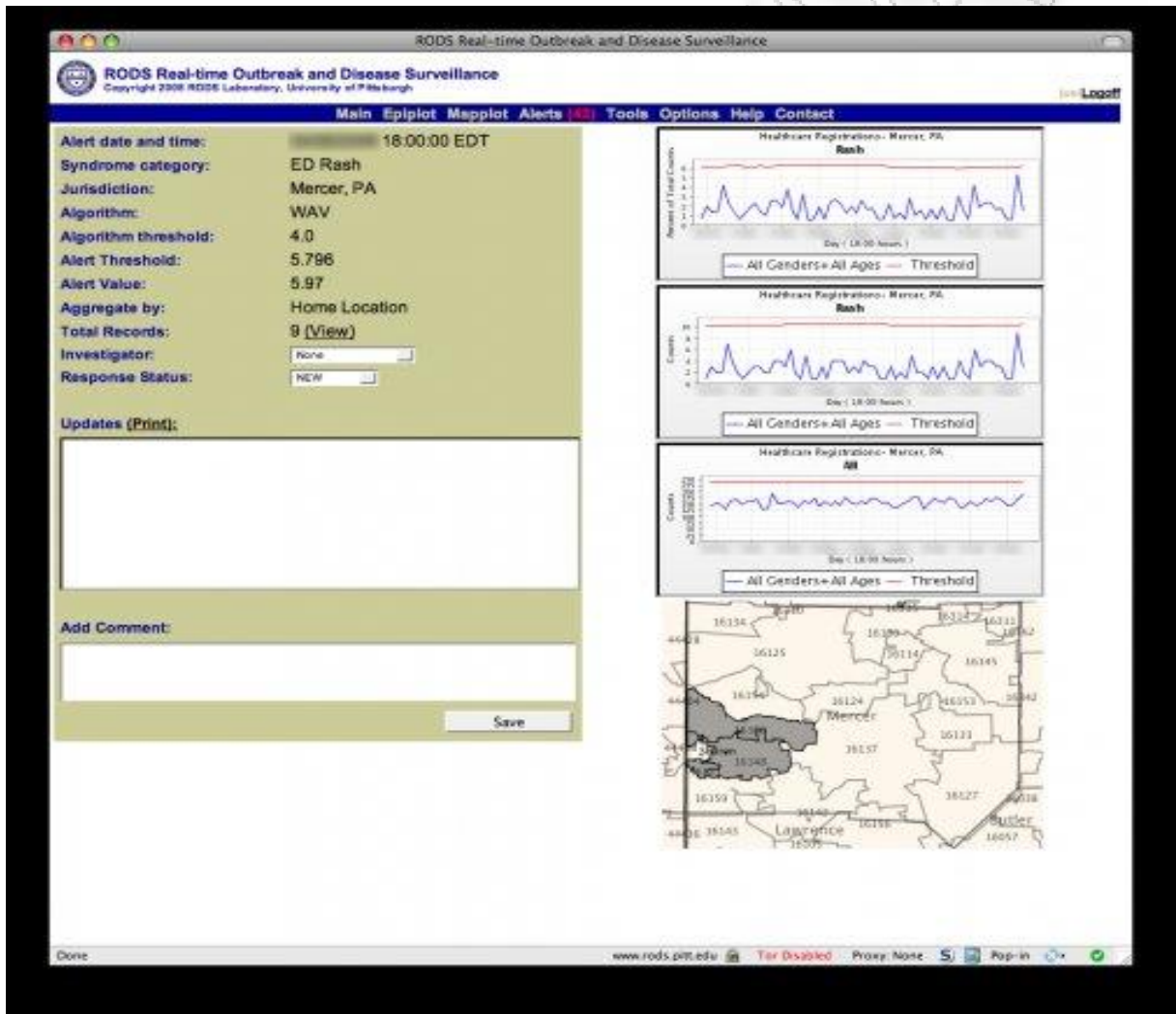
#### 1.4 Συστήματα και οργανισμοί που διενεργούν βιοεπιτήρηση

Οι οργανισμοί που ασχολούνται με την βιοεπιτήρηση είναι, κατά κύριο λόγο, κρατικοί ή διεθνείς. Ανάμεσα σε αυτούς, ξεχωρίζουν το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (*Centers for Disease Control and Prevention*) στις Η.Π.Α., ο Παγκόσμιος Οργανισμός Υγείας (*World Health Organisation*), το σύστημα *RODS* του Πανεπιστήμιο του Πίτσμπουργκ των Η.Π.Α, το Ευρωπαϊκό Κέντρο Πρόληψης και Ελέγχου Νοσημάτων (*Center for Disease Prevention and Control*) και στην Ελλάδα το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (ΚΕ.ΕΛ.Π.ΝΟ). Καθώς η λεπτομερής περιγραφή των παραπάνω οργανισμών ξεφεύγει από τους σκοπούς αυτής της διπλωματικής, δίνονται, εν συντομία, κάποια στοιχεία για τους τρεις τελευταίους.

##### ❖ RODS

Το 1999, στο Πανεπιστήμιο του Πίτσμπουργκ των Η.Π.Α. αναπτύχθηκε το πρώτο αυτοματοποιημένο σύστημα εντοπισμού ενός χτυπήματος βιο – τρομοκρατίας, το *RODS* (*Real-Time Outbreak Disease Surveillance*) . Το σύστημα αυτό είναι σχεδιασμένο να συλλέγει δεδομένα από διάφορες πηγές και να τα χρησιμοποιεί για την ανίχνευση, το συντομότερο δυνατό, ενός χτυπήματος με βιολογικούς παράγοντες. Το 2000 οι υπεύθυνοι του προγράμματος άρχισαν να το τροφοδοτούν με μη ιατρικά δεδομένα, όπως πωλήσεις καταναλωτικών προϊόντων, δείκτες ψυχολογίας κοινού κλπ. Σήμερα, χρησιμοποιούνται, μεταξύ άλλων, δεδομένα από πωλήσεις φαρμακευτικών προϊόντων σε 29.000 καταστήματα

λιανικής στην Αμερική. Το 2004, το *RODS* μετετράπη σε ελεύθερο λογισμικό (*open source software*) και διανέμεται δωρεάν υπό την προϋπόθεση οποιαδήποτε βελτίωση γίνει στο πρόγραμμα να δημοσιοποιείται. Σήμερα χρησιμοποιείται από εκατοντάδες υπηρεσίες που ασχολούνται με τη δημόσια υγεία τόσο στις Η.Π.Α. όσο και διεθνώς. Στην Εικόνα 1.2 φαίνεται το περιβάλλον χρήστη (*user interface*) του λογισμικού όταν ζητηθούν οι λεπτομέρειες ενός συναγερμού.



Εικόνα 1.2 Περιβάλλον χρήστη στο RODS

❖ Ευρωπαϊκό Κέντρο Ελέγχου και Πρόληψης Νοσημάτων

Το Ευρωπαϊκό Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (*European Centre for Disease Prevention and Control*), έχει σχεδιάσει μία βάση δεδομένων με την ονομασία *Threat Tracking Tool (TTT)* στην οποία καταχωρούνται και παρακολουθούνται απειλές που το

Κέντρο έχει εντοπίσει μέσα από διάφορες πηγές πληροφόρησης. Σκοπός του ΤΤΤ είναι ο εντοπισμός επιβεβαιωμένων κρουσμάτων με γνωστές ή πιθανές επιπτώσεις στη δημόσια υγεία.

Οι βασικές αποστολές της βιοεπιτήρησης που διεξάγεται από το Κέντρο είναι:

- Ο εντοπισμός και η αξιολόγηση νέων απειλών κατά της δημόσιας υγείας, από μεταδοτικές νόσους
- Η καθιέρωση, σε συνεργασία με τα Κράτη – Μέλη, διαδικασιών για την συστηματική έρευνα, συλλογή, αντιπαραβολή και ανάλυση πληροφοριών και δεδομένων με σκοπό τον προσδιορισμό νέων απειλών για την υγεία που ενδέχεται να έχουν ψυχολογικές και σωματικές επιπτώσεις και οι οποίες θα μπορούσαν να πλήξουν την Κοινότητα.

Επιπλέον κάποιοι από τους σκοπούς που επιτυγχάνονται μέσω των εργασιών του Ευρωπαϊκού Κέντρου είναι:

- Η έρευνα, η συλλογή, η αντιπαραβολή, η αξιολόγηση και η διάδοση των σχετικών επιστημονικών και τεχνικών δεδομένων
- Ο συντονισμός και η διασφάλιση της λειτουργίας των διαφόρων δικτύων επιτήρησης.
- Η ανάπτυξη ολοκληρωμένων συστημάτων συλλογής δεδομένων που καλύπτουν όλο το φάσμα των μεταδοτικών νοσημάτων και η διατήρηση βάσεων δεδομένων.
- Η διεξαγωγή, εφαρμοσμένων, επιστημονικών μελετών για την καλύτερη ανάπτυξη και προετοιμασία του οργανισμού
- Η στενή συνεργασία με οργανισμούς που ασχολούνται με τη συλλογή δεδομένων
- Η παρακολούθηση της τάσης εξάπλωσης των ασθενειών στην Ευρώπη, με σκοπό τη χάραξη ενός τρόπου δράσης για την προστασία της δημόσιας υγείας
- Η ενδυνάμωση των εθνικών συστημάτων βιοεπιτήρησης.

Τέλος στην Ευρωπαϊκή Ένωση έχει καθιερωθεί η χρήση του Συστήματος Έγκαιρης Προειδοποίησης και Αντίδρασης (*EWRs*). Το *EWRs* είναι ένα ηλεκτρονικό σύστημα πληροφοριών εμπιστευτικού χαρακτήρα μέσω του οποίου τα κράτη μέλη μπορούν να στέλνουν προειδοποιήσεις για περιστατικά που έχουν σημαντικές επιπτώσεις στην ΕΕ, να ανταλλάσσουν πληροφορίες και να συντονίζουν την αντίδρασή τους. Το σύστημα έχει ήδη χρησιμοποιηθεί με επιτυχία στο παρελθόν για την αντιμετώπιση της επιδημίας του σοβαρού οξέος αναπνευστικού συνδρόμου (*SARS*), της γρίπης *A(H1N1)* και άλλων μεταδοτικών νοσημάτων.

#### ❖ ΚΕ.ΕΛ.Π.ΝΟ

Στην Ελλάδα λειτουργεί από το 1992 το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (ΚΕ.ΕΛ.Π.ΝΟ). Σύμφωνα με το Νόμο, 2071/92 ένας από τους κύριους σκοπούς του ΚΕ.ΕΛ.Π.ΝΟ είναι η «*επιδημιολογική επιτήρηση των λοιμωδών νοσημάτων μέσω των συστημάτων επιτήρησης και συγκεκριμένα: του συστήματος υποχρεωτικής δήλωσης, του συστήματος παρατηρητών νοσηρότητας στην πρωτοβάθμια φροντίδα υγείας (sentinel), του εργαστηριακού συστήματος επιτήρησης και την παρέμβαση στους χώρους υγείας*».

### 1.5 Βιοτρομοκρατία

Σύμφωνα με το αμερικανικό Κέντρο Ελέγχου και Πρόληψης Ασθενειών, μια βιοτρομοκρατική επίθεση είναι «*η εκούσια απελευθέρωση ιών, βακτηρίων και άλλων μικροβίων (παράγοντες), με σκοπό την πρόκληση ασθένειας ή θανάτου σε ανθρώπους, ζώα ή φυτά. Αυτοί οι παράγοντες βρίσκονται στη φύση, είναι όμως δυνατό να αλλοιωθούν για να αυξηθεί η ανθεκτικότητά τους στα υπάρχοντα φάρμακα ή για να αυξηθεί η μεταδοτικότητά τους. Οι βιολογικοί παράγοντες μεταδίδονται μέσω του αέρα, του νερού ή των τροφίμων, ... , Κάποιοι βιολογικοί παράγοντες, όπως ο ιός της ευλογιάς, μεταδίδονται από άνθρωπο σε άνθρωπο και άλλοι, όπως ο βάκιλος του άνθρακα, όχι.*»

Στοιχεία που παρουσιάστηκαν από το ΝΑΤΟ το 1969, αποκάλυψαν ότι για μια επιχείρηση που σκοπό θα έχει να προκληθούν απώλειες σε άμαχο πληθυσμό ίσες με το 50% αυτού, το κόστος ανά τετραγωνικό χιλιόμετρο είναι 2.000 δολάρια όταν χρησιμοποιούνται συμβατικά όπλα, 800 δολάρια όταν χρησιμοποιούνται πυρηνικά όπλα, 600 δολάρια όταν χρησιμοποιούνται χημικά όπλα και μόλις 1 δολάριο όταν χρησιμοποιούνται βιολογικά όπλα. Αξίζει να αναφέρουμε ότι η πληθυσμιακή πυκνότητα στη Νέα Υόρκη είναι 10,630 κάτοικοι ανά τετραγωνικό χιλιόμετρο (7<sup>ος</sup> /2009), στο Παρίσι ήταν 20.807 κάτοικοι ανά τετραγωνικό χιλιόμετρο (1<sup>ος</sup> /2007) και στη Νεάπολη Θεσσαλονίκης είναι 27.252 κάτοικοι ανά τετραγωνικό χιλιόμετρο (απογραφή 2001). Εύκολα λοιπόν γίνεται αντιληπτό το μέγεθος της καταστροφής που μπορεί να προκληθεί από τη χρήση βιολογικών όπλων ή παραγόντων.

Ένας επιπλέον λόγος, που τρομοκράτες μπορεί να απελευθερώσουν βιολογικούς παράγοντες, είναι ότι, οι τελευταίοι, είναι δύσκολο να ανιχνευτούν και χρειάζονται αρκετές ώρες ή και μέρες μέχρι να εκδηλωθούν τα πρώτα συμπτώματα.

Η πρώτη αναφορά για βιολογικά όπλα βρίσκεται στην αρχαία Ρώμη, όταν οι Ρωμαίοι χρησιμοποιούσαν καταπέλτες για να πετούν περιττώματα στους εχθρούς, αργότερα τον 14<sup>ο</sup>

αιώνα οι Τάταροι πετούσαν με καταπέλτες πτώματα θυμάτων της πανούκλας μέσα στην πολιορκούμενη πόλη της Κάφφα (Κριμαία). Οι Ινδιάνοι της Αμερικής μολύνθηκαν σκόπιμα από τους Άγγλους, με τον ιό της ευλογιάς, που τους μεταδόθηκε από μολυσμένες κουβέρτες. Στη διάρκεια του Α΄ Παγκόσμιου Πολέμου υπολογίζεται πως το 20% των στρατιωτών στο Δυτικό Μέτωπο σκοτώθηκε από ένα συγγενή των βιολογικών όπλων: τα χημικά αέρια.

Κατά τη δεκαετία του 1930 οι Ιάπωνες μελέτησαν τις επιδράσεις της βουβωνικής πανώλης, του άνθρακα, της βοτουλιάσης (τοξίνη), της βρουκέλωσης, της χολέρας, της ευλογιάς και του τύφου, πάνω σε Κινέζους αιχμαλώτους.

Οι βιολογικοί παράγοντες χωρίζονται σε δύο κατηγορίες. Στην πρώτη ανήκουν αυτοί που αποτελούν κίνδυνο για την εθνική ασφάλεια, που μπορούν να μεταδοθούν εύκολα, που έχουν μεγάλη πιθανότητα να προκαλέσουν θάνατο, που μπορεί να προκαλέσουν πανικό και που απαιτούν υψηλά επίπεδα προετοιμασίας για να αντιμετωπιστούν. Στην κατηγορία αυτή ανήκουν, μεταξύ άλλων, ο βάκιλος του άνθρακα, η τουλαραιμία, η ευλογία, η βουβωνική πανώλη και ο ιογενής αιμορραγικός πυρετός. Στη δεύτερη κατηγορία ανήκουν παράγοντες που εντοπίζονται σχετικά εύκολα και έχουν χαμηλά ποσοστά θνησιμότητας, όπως η βρουκέλωση, η σαλμονέλα, η ψιττάκωση, ο πυρετός Q, ο τύφος και η ιογενής εγκεφαλίτιδα.

## 1.6 Στατιστικός Έλεγχος Διεργασίας

Με τον όρο Στατιστικός Έλεγχος Διεργασίας (*Statistical Process Control*, *SPC*) εννοούμε την εφαρμογή στατιστικών μεθόδων με σκοπό την παρακολούθηση και τον έλεγχο μίας διεργασίας ώστε να εξασφαλιστεί ότι η διεργασία παράγει προϊόντα που ανταποκρίνονται στις προδιαγραφές. Παρόλο που οι πιο συνηθισμένες εφαρμογές του στατιστικού ελέγχου διεργασίας βρίσκονται στη βιομηχανική παραγωγή, οι τεχνικές του μπορούν να χρησιμοποιηθούν σε οποιαδήποτε διεργασία κατά την οποία λαμβάνονται μετρήσιμα δεδομένα, άρα και σε μία διεργασία βιοεπιτήρησης.

Ο Στατιστικός Έλεγχος Διεργασίας μαζί με το Σχεδιασμό Πειραμάτων (*Design of Experiments*) και τη Δειγματοληψία Αποδοχής (*Acceptance Sampling*) αποτελούν τα τρία βασικά εργαλεία του Στατιστικού Ελέγχου Ποιότητας (*Statistical Quality Control*, *SQC*). Τα συστήματα βιοεπιτήρησης εστιάζονται σε θέματα Στατιστικού Ελέγχου Διεργασιών καθώς ο Σχεδιασμός Πειραμάτων στοχεύει στη βέλτιστη σχεδίαση μιας διεργασίας, συνεπώς προϋποθέτει ότι μπορούμε να ελέγξουμε ή να καθορίσουμε τις συνθήκες υπό τις οποίες θα διεξαχθεί η έρευνα κάτι που, όπως αναφέρθηκε στην Παράγραφο 1.1, δεν συμβαίνει στην

επιδημιολογία. Αφετέρου η Δειγματοληψία Αποδοχής στοχεύει στη λήψη αποφάσεων σχετικά με την αποδοχή ή απόρριψη μιας παρτίδας παραγόμενου προϊόντος, έννοιες που απουσιάζουν από την επιδημιολογία.

Ένα πολύ σημαντικό γνώρισμα του Στατιστικού Ελέγχου Διεργασίας είναι ότι εισάγει για πρώτη φορά την έννοια της πρόληψης στον Έλεγχο Ποιότητας [βλ. Ταγαράς, (2001)]. Αυτό είναι και το ζητούμενο στη βιοεπιτήρηση, η πρόληψη έναντι μιας επιδημίας, δηλαδή η έγκαιρη ανίχνευση ενός επικείμενου ξεσπάσματος μιας ασθένειας και η λήψη μέτρων προκειμένου να αντιμετωπιστεί το συντομότερο δυνατό.

Ο Στατιστικός Έλεγχος Διεργασίας αναπτύχθηκε αρχικά από τον *Walter A. Shewhart*, στις αρχές της δεκαετίας του 1920. Ο *Shewhart* συνειδητοποίησε ότι η διασπορά των τιμών ενός χαρακτηριστικού ποιότητας είναι αναπόφευκτη κατά την παραγωγή και ότι οφείλεται είτε σε τυχαία (άρα μη ελεγχόμενα) αίτια, ή σε συγκεκριμένα, μη τυχαία, αίτια τα οποία μπορούν να εντοπιστούν και να διορθωθούν .

Τα βασικά εργαλεία του Στατιστικού Ελέγχου Διεργασιών είναι το ιστόγραμμα, το φύλλο ελέγχου, το διάγραμμα *Pareto*, το διάγραμμα αιτίας – αποτελέσματος, το διάγραμμα διασποράς, το διάγραμμα συγκέντρωσης ελαττωμάτων και το διάγραμμα ελέγχου. Από αυτά, στη βιοεπιτήρηση χρησιμοποιείται το διάγραμμα ελέγχου, το οποίο περιγράφεται στο 2<sup>ο</sup> Κεφάλαιο της εργασίας, ενώ ως φύλλο ελέγχου μπορούμε να θεωρήσουμε το ιστορικό νοσηλείας ενός ασθενούς, μια έκθεση περιστατικών που συμπληρώνεται σε τακτά διαστήματα από το διευθυντή του τμήματος επειγόντων περιστατικών ενός νοσοκομείου κλπ.



## ΚΕΦΑΛΑΙΟ 2

### Διαγράμματα Ελέγχου

#### 2.1 Εισαγωγή

Όπως είδαμε και στο προηγούμενο κεφάλαιο, όλες οι διεργασίες, που αφορούν την υγεία, παρουσιάζουν κάποια τυχαία μεταβλητότητα. Για παράδειγμα, παρόλο που το ποσοστό των μολύνσεων που συνέβησαν εντός του χειρουργείου είναι περίπου σταθερό από μήνα σε μήνα, ο ακριβής αριθμός των μολύνσεων δεν είναι ο ίδιος για κάθε μήνα, αλλά μεταβάλλεται θετικά ή αρνητικά κινούμενος γύρω από μία μακροχρόνια μέση τιμή. Οι αιτίες που προκαλούν τις διακυμάνσεις σε μία διεργασία μπορούν να χωριστούν σε δύο κατηγορίες, τις φυσικές (*natural*) και τις μη φυσικές ή ειδικές (*unnatural*). Ως φυσική αιτία μεταβλητότητας μιας διεργασίας εννοούμε ένα φυσιολογικό χαρακτηριστικό της διεργασίας που προκαλεί διακύμανση στις τιμές και που δεν μπορεί να εξηγηθεί γιατί εξαρτάται από τυχαίους παράγοντες. Για παράδειγμα, μία μικρή αύξηση στα κρούσματα του κοινού κρυολογήματος μπορεί να οφείλεται σε μία απότομη μεταβολή του καιρού ή σε άλλους τυχαίους παράγοντες. Αντίθετα, παρατηρήσεις που έχουν πολύ μικρή πιθανότητα να συμβούν με βάση τη φυσιολογική συμπεριφορά της διεργασίας υποθέτουμε ότι οφείλονται σε ακραία γεγονότα και αποκλίσεις από τη διεργασία. Τέτοια περιστατικά αποτελούν ένδειξη ότι έχει συμβεί κάποια αλλαγή στη διεργασία ή στο περιβάλλον της και είναι πραγματοποιήσιμες μη συστημικές, μη φυσικής διακύμανσης. Με άλλα λόγια οι ειδικές αιτίες διακύμανσης δεν είναι μέρος της φυσιολογικής εξέλιξης της διεργασίας.

Γενικά, ο όρος στατιστικός έλεγχος αναφέρεται στη σταθερότητα και την προβλεψιμότητα της διεργασίας με την πάροδο του χρόνου. Αν μία διεργασία παραμένει απόλυτα σταθερή και προβλέψιμη τότε παρουσιάζει μόνο φυσική μεταβλητότητα καθώς τα στοιχεία που την προκαλούν παραμένουν αμετάβλητα. Μία τέτοια διεργασία θα λέμε ότι βρίσκεται σε εντός στατιστικού ελέγχου κατάσταση ή απλά ότι είναι εντός ελέγχου. Από την άλλη, αν η συμπεριφορά της υποκείμενης διεργασίας αλλάξει από τα συνήθη πλαίσια, θα παρουσιάζει ειδική μεταβλητότητα. Σε αυτή την περίπτωση θα λέμε ότι η διεργασία βρίσκεται σε εκτός στατιστικού ελέγχου κατάσταση ή απλά ότι είναι εκτός ελέγχου. Αξίζει να σημειωθεί ότι τόσο στην πρώτη όσο και στην δεύτερη περίπτωση θα πρέπει να λαμβάνονται μέτρα για την βελτίωση της διεργασίας. Ένα από τα βασικά χαρακτηριστικά του στατιστικού ελέγχου είναι

η ταυτοποίηση και η απομάκρυνση των ειδικών αιτίων μεταβλητότητας ή, αν αυτό δεν είναι δυνατό, η ελαχιστοποίηση της επίδρασής τους.

Για παράδειγμα, ας θεωρήσουμε ότι σε ένα νοσοκομείο προσλαμβάνεται καινούριο προσωπικό που εφαρμόζει μια διαφορετική κλινική μεθοδολογία (πρωτόκολλο). Αυτό μπορεί να αποτελέσει πηγή μη φυσιολογικής απόδοσης της διεργασίας. Αν αυτή η νέα μεθοδολογία προκαλέσει μία αλλαγή (αύξηση ή μείωση) στο ποσοστό των μολύνσεων εντός χειρουργείου τότε θα πρέπει το νοσοκομείο να υιοθετήσει μια τυποποιημένη μέθοδο που θα εφαρμόζεται υποχρεωτικά. Αν μετά την πρόσληψη του νέου προσωπικού τα ποσοστά αυξήθηκαν τότε το νοσοκομείο θα πρέπει να απαγορέψει την εφαρμογή της νέας τεχνικής. Αν όμως η τεχνογνωσία, που έφεραν μαζί τους οι νεοπροσληφθέντες, οδηγήσει σε μείωση του ποσοστού των μολύνσεων τότε η διοίκηση του νοσοκομείου θα πρέπει να εκπαιδεύσει και το προϋπάρχον προσωπικό στη νέα μεθοδολογία.

Το παραπάνω αποτελεί ένα απλουστευτικό παράδειγμα για τις έννοιες που αναπτύξαμε. Σε πραγματικές συνθήκες, χωρίς τη χρήση στατιστικών μεθόδων, είναι δύσκολο να καθοριστεί διαισθητικά αν η υπάρχουσα μεταβλητότητα είναι φυσική ή ειδική και άρα είναι δύσκολο να αποφασιστεί το είδος των ενεργειών που θα εκτελεστούν. Τέλος όταν, με την χρήση των διαγραμμάτων ελέγχου, που θα αναπτύξουμε παρακάτω, ανιχνευτεί ένα μη φυσιολογικό συμβάν ίσως χρειαστεί να εφαρμοστούν πιο προηγμένες επιδημιολογικές μέθοδοι για να βρεθεί η αιτία που το προκάλεσε.

## **2.2 Διαγράμματα Στατιστικού Ελέγχου Διεργασίας τύπου *Shewhart***

### **2.2.1 Βασικές έννοιες**

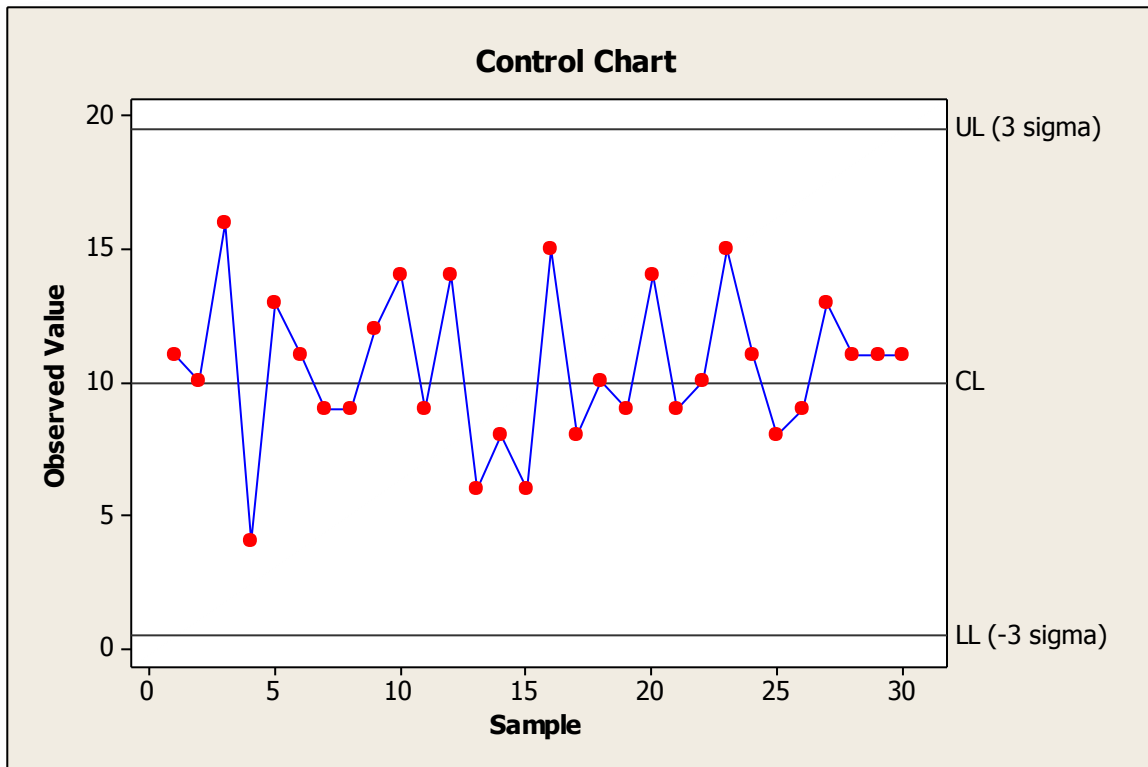
Τα διαγράμματα ελέγχου εφαρμόζονται ώστε να είναι εφικτή η διάκριση ανάμεσα σε δεδομένα που προέρχονται από μία εντός ελέγχου διεργασία και σε αυτά που προέρχονται από μία διεργασία που παρουσιάζει κάποια μη φυσιολογική μεταβλητότητα, δίνοντας ένδειξη ότι υπάρχει τουλάχιστον μία ειδική αιτία μεταβλητότητας. Στο εξής όταν ένα διάγραμμα (ή κάποια από τις μεθόδους που θα δούμε σε επόμενα κεφάλαια) παρέχει ένδειξη ότι υπάρχει μία ειδική αιτία μεταβλητότητας θα λέμε, για λόγους συντομίας και απλότητας, ότι έχουμε ένδειξη αλλαγής ή ότι το διάγραμμα δίνει συναγερμό, ή ότι έχουμε ένδειξη ότι η διεργασία είναι εκτός ελέγχου ή απλά ένδειξη εκτός ελέγχου διεργασίας. Οι όροι αυτοί θα χρησιμοποιούνται ως ισοδύναμοι.

Η θεωρία και η εφαρμογή των διαγραμμάτων ελέγχου, στην βιοεπιτήρηση, έχει αναπτυχθεί λεπτομερώς σε πολλές εργασίες. Ανάμεσα τους ξεχωρίζουν αυτές των *Dunkan* (1986), *Montgomery* (1991), *Finison et al.* (1993), *Bennayan* (1995) και *Woodall* (2006).

Τα διαγράμματα ελέγχου αναπτύχθηκαν στα εργαστήρια *Bell* από τον *Walter Shewhart* το 1924 και είναι ένα από τα πρωταρχικά εργαλεία στην βελτίωση της ποιότητας και τον Στατιστικό Έλεγχο Διεργασίας. Παρόλο που βασίζονται στην στατιστική, έχουν σχεδιαστεί ώστε να μπορούν να εφαρμοστούν και να ερμηνευτούν και από μη στατιστικούς.

Ένα σύνολο παρατηρήσεων, ή αλλιώς, στην ορολογία του Στατιστικού Ελέγχου Διεργασίας, μια υποομάδα (*subgroup*), για το υπό μελέτη χαρακτηριστικό, λαμβάνεται περιοδικά από την διεργασία και οι τιμές του ποσοτικού μεγέθους σημειώνονται στο διάγραμμα ελέγχου. Το πλήθος των παρατηρήσεων σε κάθε υποομάδα καλείται μέγεθος της υποομάδας και συνήθως συμβολίζεται με  $n$ . Αξίζει να σημειωθεί ότι, σε αρκετές περιπτώσεις, για οικονομικούς λόγους, λαμβάνεται και αναλύεται μόνο ένα μέρος από τις παρατηρήσεις για κάθε υποομάδα. Όταν κατασκευάζουμε ένα διάγραμμα ελέγχου χρησιμοποιούμε και ιστορικά δεδομένα και κάθε τιμή που λαμβάνεται σημειώνεται αμέσως στο διάγραμμα. Στη συνέχεια με το που λαμβάνουμε κάθε παρατήρηση την σημειώνουμε στο διάγραμμα.

Ένα διάγραμμα ελέγχου περιλαμβάνει τρεις ευθείες παράλληλες στον οριζόντιο άξονα, την κεντρική γραμμή (*center line, CL*), το άνω όριο ελέγχου (*upper control limit, UCL*) και το κάτω όριο ελέγχου (*lower control limit, LCL*). Οι ευθείες αυτές χρησιμοποιούνται για να προσδιοριστεί η κεντρική τάση και η φυσική μεταβλητότητα της διεργασίας και, συνεπώς, για να ανιχνευτεί μια στατιστικά σημαντική αλλαγή στην διεργασία. Συνήθως, η κεντρική γραμμή σχεδιάζεται στον αριθμητικό μέσο ή την αναμενόμενη τιμή της τυχαίας μεταβλητής που παρακολουθούμε ώστε περίπου τα μισά σημεία να σχεδιάζονται στην περιοχή κάτω από αυτή και τα υπόλοιπα στην άλλη περιοχή. Σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν μεγάλη λοξότητα (*skewness*) χρησιμοποιείται εναλλακτικά η διάμεσος [βλ. *Bannayan & Kaminsky* (1994)]. Τα όρια του ελέγχου σχεδιάζονται σε απόσταση από την κεντρική γραμμή ίση με την τυπική απόκλιση πολλαπλασιασμένη επί ένα συντελεστή  $L$ , συνήθως χρησιμοποιείται η τιμή 3 ( $3\sigma$  (*3-sigma*)). Στο Διάγραμμα 2.1 δίνεται ένα διάγραμμα ελέγχου όπου έχουν σημειωθεί οι τιμές από μία μετρήσιμη ποσότητα ενός χαρακτηριστικού που ελήφθησαν για κάθε ένα από 30 δείγματα ή σε 30 χρονικές στιγμές. Η κατασκευή αυτού του διαγράμματος, όπως και όλων των διαγραμμάτων στατιστικού ελέγχου διεργασίας, έγινε με τα βοήθεια του προγράμματος *MINITAB*.



**Διάγραμμα 2.1** Τυπικό Διάγραμμα Ελέγχου

Παρατηρώντας τη συμπεριφορά ενός ποσοστού μολύνσεων ή γενικά του ποσοστού εμφάνισης μιας νόσου μπορούμε να αποφασίσουμε αν η διαδικασία είναι εντός ή εκτός ελέγχου. Όταν τα δεδομένα συλλέγονται και σημειώνονται στο διάγραμμα συχνά και σε λίγο χρόνο από την στιγμή που παρατηρούνται, μπορούμε να αποφασίσουμε εγκαίρως για τις ενέργειες που θα εκτελεστούν και είτε να διορθώσουμε την διαδικασία ή να καθιερώσουμε τις βελτιώσεις που έγιναν. Γενικά, μια διεργασία θεωρείται, με μεγάλη πιθανότητα, ότι είναι εντός ελέγχου αν όλες οι παρατηρήσεις των υποσυνόλων κατανέμονται τυχαία στο διάγραμμα για ένα ικανό διάστημα χρόνου. Η πιο κλασσική περίπτωση στην οποία έχουμε ένδειξη για εκτός ελέγχου διεργασία, είναι όταν τουλάχιστον ένα σημείο βρεθεί εκτός των ορίων ελέγχου. Στην περίπτωση αυτή θα πρέπει να γίνεται έλεγχος για να καθοριστεί αν υπάρχει ειδική αιτία μεταβλητότητας και αν υπάρχει να απομακρύνεται. Φυσικά, και ειδικά στην περίπτωση που η υποκείμενη διεργασία αφορά ένα χαρακτηριστικό της δημόσιας υγείας, μία αλλαγή δεν είναι κατ' ανάγκη ανεπιθύμητη. Τότε, εφόσον ανιχνευτεί μια αλλαγή και βρεθεί μία ειδική αιτία μεταβλητότητας θα πρέπει να εργαστούμε προς την καθιέρωση της. Για παράδειγμα έστω ότι η βιοεπιτήρηση εστιάζει στην παιδική παχυσαρκία και έστω επίσης ότι σε κάποια σχολεία τα ποσοστά παχύσαρκων μαθητών είναι χαμηλότερα από το κάτω όριο. Τα σημεία κάτω από το

κάτω όριο μας δείχνουν ότι η διεργασία είναι πιθανώς εκτός ελέγχου. Αν μετά τον έλεγχο διαπιστωθεί ότι υπάρχει μια ειδική αιτία μεταβλητότητας, πχ ο διευθυντής του σχολείου απαγόρευσε την πώληση αναψυκτικών και λιπαρών τροφών από το κυλικείο τότε όχι μόνο δεν θα πρέπει να απομακρύνουμε αυτή την αιτία (την λιγότερο ανθυγιεινή διατροφή ) αλλά να την καθιερώσουμε και στα άλλα σχολεία.

Επίσης δεν θα πρέπει να υπάρχουν ενδείξεις μη τυχαίας συμπεριφοράς ανάμεσα στα όρια του ελέγχου, όπως τάση, περιοδικότητα, μετατοπίσεις πάνω ή κάτω από την κεντρική γραμμή, ή άλλες μορφές συμπεριφοράς που έχει μικρή πιθανότητα να εμφανίζεται. Με άλλα λόγια τα όρια ελέγχου δεν αποτελούν το μοναδικό κριτήριο για να αποφασίσουμε αν η διαδικασία είναι εντός ή εκτός ελέγχου. Δυστυχώς, όταν η ανίχνευση τέτοιων συμπεριφορών γίνεται χωρίς αυστηρούς κανόνες, για παράδειγμα μέσω ενός διαγράμματος, τότε στηρίζεται στην υποκειμενική κρίση του αναλυτή.

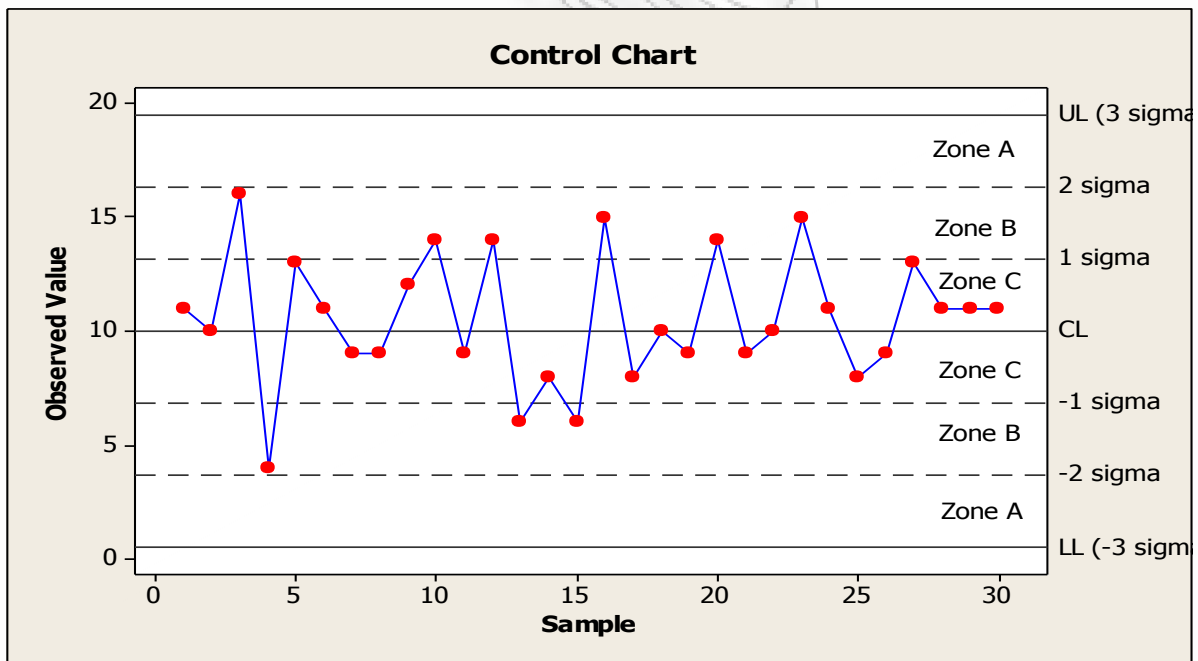
Στο παρόν κεφάλαιο θα ασχοληθούμε με στάσιμα δεδομένα, δηλ. δεδομένα που κινούνται γύρω από μία μέση τιμή με σταθερό ή προβλέψιμο τρόπο, τα οποία είναι ασυσχέτιστα δηλ. η τιμή μίας παρατήρησης δεν επηρεάζεται από κάποια άλλη.

Φυσικά, μεγάλο ενδιαφέρον δεν έχει μόνο να ανιχνεύσουμε πότε έχει ήδη συμβεί μια σημαντική αλλαγή αλλά και να την προβλέψουμε. Επίσης σημαντικό είναι να μπορούμε να ελέγξουμε την ύπαρξη μικρών μετατοπίσεων του μέσου. Για τους λόγους αυτούς χρησιμοποιούνται και τα προειδοποιητικά όρια (*warning limits*) τα οποία σχεδιάζονται εντός της περιοχής που ορίζουν τα όρια ελέγχου. Για παράδειγμα όταν χρησιμοποιούμε τα 3σ όρια ελέγχου, τα προειδοποιητικά όρια τίθενται συνήθως σε απόσταση  $1\sigma$  και  $2\sigma$  άνω και κάτω της κεντρικής γραμμής. Με βάση τα προειδοποιητικά όρια, η κάθε μία από τις δύο περιοχές που ορίζονται ανάμεσα στην κεντρική γραμμή και τα όρια ελέγχου, χωρίζεται σε τρεις Ζώνες. Τα σημεία που βρίσκονται άνω (ή κάτω) της κεντρικής γραμμής και σε απόσταση μικρότερη από  $1\sigma$  θα λέμε ότι βρίσκονται στη Ζώνη C. Ομοίως, τα σημεία που βρίσκονται σε απόσταση μεταξύ  $1\sigma$  και  $2\sigma$  θα λέμε ότι βρίσκονται στη Ζώνη B. Τέλος τα υπόλοιπα σημεία, έως το όριο ελέγχου θα λέμε ότι βρίσκονται στη Ζώνη A. Ένα παράδειγμα τέτοιου διαγράμματος δίνεται στο Διάγραμμα 2.2 .

Η χρήση προειδοποιητικών ορίων αυξάνει την ευαισθησία του διαγράμματος. Προκειμένου τα προειδοποιητικά όρια να είναι αποτελεσματικά στον εντοπισμό ύπαρξης ενός μοτίβου (*pattern*) έχουν θεσπιστεί κάποιοι κανόνες. Όταν κάποιος από αυτούς τους κανόνες δεν πληρείται τότε έχουμε ένδειξη εκτός ελέγχου διεργασίας. Δυστυχώς η αύξηση της

ευαισθησίας οδηγεί σε αύξηση της πιθανότητας εσφαλμένου συναγεμού άρα συνιστάται η προσεκτική, συνετή χρήση των κανόνων. Παρακάτω δίνονται οι πιο σημαντικοί κανόνες:

- Ένα ή περισσότερα σημεία εκτός των ορίων ελέγχου
- Δύο από τρία συνεχόμενα σημεία στην Ζώνη A (σε μία από τις δύο).
- Τέσσερα από πέντε συνεχόμενα σημεία στη Ζώνη C (σε μία από τις δύο).
- Οκτώ συνεχόμενα σημεία σε μία από τις δύο περιοχές που ορίζει η CL.
- Έξι συνεχόμενα σημεία με αύξουσα ή φθίνουσα διάταξη.
- Δεκαπέντε συνεχόμενα σημεία στις δύο Ζώνες C.
- Δεκατέσσερα συνεχόμενα σημεία που εναλλάσσονται της CL.
- Οκτώ συνεχόμενα σημεία έξω από τις δύο Ζώνες C.
- Οποιαδήποτε μη τυχαία ακολουθία σημείων.
- Τουλάχιστον ένα σημείο κοντά στα όρια ελέγχου.



Διάγραμμα 2.2 Διάγραμμα Ελέγχου με προειδοποιητικά όρια

Στη βιβλιογραφία υπάρχουν δύο φάσεις για τον έλεγχο μιας διεργασίας με τα διαγράμματα ελέγχου. Η Φάση I (*Phase I*) ή αναδρομική ανάλυση (*retrospective analysis*) κατά την οποία με βάση τα δεδομένα που συλλέχθηκαν στο παρελθόν αποφασίζουμε αν η διεργασία είναι εντός ή εκτός ελέγχου. Με τα δεδομένα αυτά μπορούμε να σχεδιάσουμε την κεντρική γραμμή και τα όρια ελέγχου για μελλοντική παρακολούθηση. Η Φάση II (*Phase II*) ή παρακολούθηση

διεργασίας (*process monitoring*) κατά την οποία ελέγχουμε αν η διεργασία είναι εντός στατιστικού ελέγχου με βάση τα όρια που προέκυψαν από τη Φάση I ή σύμφωνα με τα όρια που προκύπτουν αν οι παράμετροι της κατανομής του χαρακτηριστικού είναι εκ των προτέρων γνωστές.

Ένα σημαντικό κομμάτι της παρακολούθησης είναι η αξιολόγηση της μεθόδου που επιλέχτηκε. Πριν δούμε τον τρόπο αξιολόγησης θα εισάγουμε μερικές ακόμη έννοιες. Ας θεωρήσουμε ότι παρακολουθούμε σε πραγματικό χρόνο τη στοχαστική διαδικασία  $X = \{X(t); t = 1, 2, \dots\}$  με στόχο την έγκαιρη και έγκυρη ανίχνευση κάποιας σημαντικής αλλαγής στην διαδικασία, η οποία συνέβη στο χρόνο  $\tau$ . Σε κάθε χρονική στιγμή  $s$  η διεργασία μπορεί να βρίσκεται εντός ελέγχου  $D(s)$  ή εκτός ελέγχου  $C(s)$ . Έστω  $X_s = \{X(t); t \leq s\}$  οι τιμές που λάβαμε από το υπό παρακολούθηση σύστημα, και  $A(s)$  ένα σύνολο τέτοιο ώστε αν  $X_s \in A(s)$ , τότε να υπάρχει ένδειξη ότι η διαδικασία είναι εκτός ελέγχου. Εναλλακτικά θεωρούμε την συνάρτηση  $p(X_s)$ , ένα άνω όριο ελέγχου  $g(s)$  και ένα κάτω όριο ελέγχου  $h(s)$ , οπότε ο χρόνος  $t_A$  ενός συναγερμού δίνεται από την σχέση:

$$t_A = \min\{s; \{p(X_s) > g(s)\} \cup \{p(X_s) < h(s)\}\}$$

Οι αλλαγές, στην υπό παρακολούθηση διαδικασία, που προσπαθούμε να ανιχνεύσουμε ποικίλουν ανάλογα με τη διαδικασία. Συχνά αυτό που μας ενδιαφέρει είναι η αλλαγή στην κατανομή της μεταβλητής  $X$ . Για παράδειγμα μια αλλαγή στο υπό παρακολούθηση χαρακτηριστικό μπορεί να επιφέρει αλλαγή του μέσου, της διακύμανσης ή και των δύο. Συνήθως όμως το ενδιαφέρον στρέφεται στις αλλαγές του μέσου καθώς η μεταβλητή μεταφέρεται από το επίπεδο που βρίσκεται σε ένα άλλο επίπεδο. Τέλος, αλλαγές μπορεί να συμβαίνουν γραμμικά, ή εκθετικά.

Για την αξιολόγηση μιας μεθόδου στατιστικού ελέγχου χρησιμοποιούνται διάφορα μεγέθη, είτε η διεργασία βρίσκεται εντός ελέγχου είτε εκτός. Όταν η διεργασία είναι εντός ελέγχου οι συναγερμοί είναι εσφαλμένοι. Έτσι, δύο μέτρα αξιολόγησης της διεργασίας είναι το εντός ελέγχου μέσο μήκος ροής (*average in-control run length*):

$$ARL^0 = E[t_A | \tau = \infty]$$

και η πιθανότητα εσφαλμένου συναγερμού:

$$P(t_A < \tau) = \sum_{t=1}^{\infty} P(\tau = t)P(t_A < \tau | \tau = t)$$

Για τον υπολογισμό της παραπάνω πιθανότητας πρέπει να είναι γνωστή η κατανομή της  $\tau$ . Συνήθως υποθέτουμε ότι ακολουθεί την γεωμετρική κατανομή.

Κατά την εφαρμογή μιας μεθόδου στατιστικού ελέγχου διεργασίας σκοπός μας είναι να μειώσουμε τον αριθμό των εσφαλμένων συναγεργμών και να μειώσουμε τον χρόνο που μεσολαβεί από την στιγμή που η διεργασία μεταφέρεται εκτός ελέγχου μέχρι τον πρώτο συναγεργμό. Δυστυχώς όμως αυτές οι δύο συνθήκες δεν μπορούν να συμβαίνουν παράλληλα. Το ίδιο συμβαίνει και στον έλεγχο στατιστικών υποθέσεων, όπου όταν μειώνουμε το σφάλμα τύπου I, αυξάνεται το σφάλμα τύπου II. Έτσι όταν μειώνεται η πιθανότητα να λάβουμε λανθασμένο συναγεργμό, αυξάνεται ο χρόνος που απαιτείται μέχρι να εντοπιστεί μια αλλαγή στην διαδικασία. Το πρόβλημα αυτό το αντιμετωπίζουμε με αντίστοιχο τρόπο. Συγκρίνουμε δηλαδή τις διάφορες μεθόδους ελέγχου διεργασίας για μια δεδομένη τιμή του εντός ελέγχου μέσου μήκους ροής ( $ARL^0$ ).

Για την αξιολόγηση μιας μεθόδου ελέγχου, δεδομένου ότι έχει συμβεί μια αλλαγή στην διεργασία, χρησιμοποιούμε το εκτός ελέγχου μέσο μήκος ροής.

$$ARL^1 = E[t_A | \tau = 1]$$

Θεωρούμε δηλαδή ότι η πρώτη αλλαγή συνέβη την στιγμή που ξεκίνησε η διαδικασία. Αυτό είναι πολύ χρήσιμο όταν πρόκειται για κάποια βιομηχανική γραμμή παραγωγής καθώς αναμένεται να παρουσιαστούν διάφορα προβλήματα κατά την έναρξη της παραγωγικής διαδικασίας. Αντιθέτως σε προβλήματα βιοεπιτήρησης κάτι τέτοιο δεν αποτελεί σωστή προσέγγιση. Σε τέτοιες περιπτώσεις επικεντρωνόμαστε σε άλλα μέτρα αξιολόγησης τα οποία λαμβάνουν υπ' όψιν την πιθανότητα να συμβούν αλλαγές μετά την έναρξη της παρακολούθησης, καθώς η δυνατότητα να ανιχνευτεί μια αλλαγή εξαρτάται από την στιγμή που αυτή επήλθε. Ένα τέτοιο μέτρο αξιολόγησης είναι η έκφραση της υπό συνθήκη αναμενόμενης καθυστέρησης (*conditional expected delay, CED*) ως συνάρτηση του χρόνου που συνέβη η αλλαγή. Δηλαδή πρόκειται για τον μέσο χρόνο καθυστέρησης ενός πραγματικού συναγεργμού όταν η αλλαγή έχει συμβεί την χρονική στιγμή  $t$ :

$$CED(t) = E[t_A - \tau | t_A \geq \tau = t]$$



Υποθέτοντας μία κατανομή για το  $\tau$ , μπορούμε να υπολογίσουμε την αναμενόμενη καθυστέρηση, δηλαδή τον σταθμισμένο μέσο χρόνο που μεσολάβησε από τη στιγμή που συνέβη η αλλαγή μέχρι τη στιγμή που έχουμε ένδειξη ότι η διαδικασία βρίσκεται εκτός ελέγχου:

$$ED_t = \sum_{t=1}^{\infty} P(\tau = t)P(t_A \geq t)CED(t).$$

Σε αρκετές εφαρμογές είναι αποδεκτό μόνο ένα πολύ μικρό χρονικό διάστημα  $d$  μέχρι την εκδήλωση συναγερμού. Για παράδειγμα στο ξέσπασμα μιας μεταδοτικής νόσου θα πρέπει να γίνουν το συντομότερο δυνατό άμεσες ενέργειες ώστε να αποφευχθεί το ενδεχόμενο επιδημίας. Σε τέτοιες περιπτώσεις χρησιμοποιούμε την πιθανότητα σωστής ανίχνευσης (*probability of successful detection*) :

$$PSD(d, t) = P(t_A - \tau \leq d | t_A \geq \tau = t)$$

Πρόκειται για την πιθανότητα η αλλαγή να γίνει αντιληπτή με καθυστέρηση όχι μεγαλύτερη από  $d$ .

Όταν ενεργοποιηθεί ένας συναγερμός θα πρέπει να γίνουν διάφορες προληπτικές ενέργειες. Για να καθοριστεί ποιες ενέργειες θα πραγματοποιηθούν θα πρέπει να είναι γνωστό πόση εμπιστοσύνη μπορούμε να δείξουμε σε ένα τέτοιο συναγερμό. Για κάθε μέθοδο βιοεπιτήρησης η κατανομή ως συνάρτηση του χρόνου των εσφαλμένων συναγερμών είναι διαφορετική. Συνεπώς ο λόγος των εσφαλμένων συναγερμών προς τους ορθούς σε κάποια χρονική στιγμή θα διαφέρει από μέθοδο σε μέθοδο και αντίστοιχα θα διαφέρει η εμπιστοσύνη που έχουμε σε ένα συναγερμό. Για την επιλογή των ενεργειών που θα γίνουν, μετά από ένα συναγερμό, χρησιμοποιούμε την προβλεπτική αξία (*predictive value*) του συναγερμού σαν συνάρτηση του χρόνου του συναγερμού. Είναι δηλ. η πιθανότητα ότι η διεργασία ήταν εκτός ελέγχου όταν ενεργοποιήθηκε ο συναγερμός:

$$PV(t) = P\{C(t) | t_A = t\}$$

Όταν η παραπάνω πιθανότητα είναι γνωστή μειώνεται ο χρόνος που απαιτείται μέχρι να εκτελεστούν οι ενέργειες. Γενικά, είναι επιθυμητό η predictive value να έχει σταθερή τιμή.

Με τον τρόπο αυτό κάθε φορά που ενεργοποιείται συναγερμός θα πρέπει να εκτελούνται οι ίδιες ενέργειες.

Όταν η παρακολούθηση γίνεται σε πραγματικό χρόνο τέτοια μέτρα αξιολόγησης είναι απαραίτητα. Στις οδηγίες των κέντρων ελέγχου και πρόληψης ασθενειών (*Centers for Disease control and Prevention*) τα χρονικά όρια, παρόλο που δεν καθορίζονται με σαφήνεια, αναφέρονται ως ένας σημαντικός παράγοντας για την αξιολόγηση της μεθόδου βιοεπιτήρησης. Τα χρονικά όρια που θα τεθούν θα πρέπει να είναι τέτοια ώστε να εξασφαλίζεται ικανή ποσότητα πληροφορίας για τον έλεγχο μιας ασθένειας. Αυτό αφορά τόσο τον χρόνο που θα περάσει από την στιγμή της εκδήλωση μέχρι την στιγμή του εντοπισμού του προβλήματος όσο και τον χρόνο που απαιτείται για την ταυτοποίηση της ασθένειας. Παράλληλα, στις οδηγίες, ως μέθοδοι αξιολόγησης συνιστούνται η ευαισθησία και η θετική προβλεπτική αξία. Αυτές αφορούν την σωστή κατηγοριοποίηση – ταξινόμηση (*classification*) των δεδομένων που συγκεντρώθηκαν από το σύστημα βιοεπιτήρησης. Τέτοια μέτρα αφορούν την ποιότητα των δεδομένων και δεν εκφράζουν την αποτελεσματικότητα ενός συστήματος στην ανίχνευση σημείων εκτός ελέγχου. Συνεπώς δεν μπορούν να χρησιμοποιηθούν ως υποκατάστατα των μεθόδων αξιολόγησης που αναπτύχθηκαν παραπάνω.

Συχνά τα δεδομένα που συλλέγονται για την δημόσια υγεία αφορούν το πλήθος των περιστατικών μιας ασθένειας. Παρόλα αυτά οι συνεχείς μεταβλητές είναι πιο συνηθισμένες.

### 2.2.2 Κατανομές Πιθανότητας – Επιλογή Κατάλληλου Διαγράμματος

Τα διαγράμματα στατιστικού ελέγχου διεργασίας βασίζονται στην θεωρία πιθανοτήτων. Υπάρχουν διάφορες μορφές των διαγραμμάτων αυτών όπως τα  $np, p, c, u, \bar{X}$  και  $S$  και επιλέγονται ανάλογα με το είδος των δεδομένων, τον τρόπο που εκδηλώνονται ή τον τρόπο που συλλέγονται κλπ. Τα δεδομένα μπορεί να είναι τιμές κάποιου χαρακτηριστικού ή συμπτώματος της νόσου, πλήθος των ασθενών, το πλήθος των κλήσεων στο κέντρο άμεσης βοήθειας, οι απουσίες των μαθητών, ποσοστά ασθενών στον πληθυσμό ή ρυθμοί εκδήλωσης της νόσου κλπ. Έτσι, όπως υπάρχουν διαφορετικοί έλεγχοι για το μέσο, την διακύμανση, τις διαφορές ή τα ποσοστά υπάρχουν και διαφορετικά διαγράμματα ελέγχου που προκύπτουν θεωρητικά σύμφωνα με την κατανομή που ακολουθούν τα εμπειρικά δεδομένα από την βιοεπιτήρηση.

Επομένως, η επιλογή του κατάλληλου διαγράμματος ελέγχου στηρίζεται στην αναγνώριση του είδους των δεδομένων της υποκείμενης διεργασίας. Στις πιο συνήθεις περιπτώσεις η

απόφαση για την κατανομή πιθανότητας που ακολουθούν τα δεδομένα γίνεται ακόμη πιο εύκολη καθώς για διάφορες περιπτώσεις εφαρμόζονται πολλά διαγράμματα. Για παράδειγμα:

- Όταν λαμβάνουμε διακριτά δεδομένα από την διωνυμική κατανομή τότε επιλέγουμε είτε το  $np$  ή το  $p$  διάγραμμα ελέγχου.
- Όταν τα δεδομένα προέρχονται από την κατανομή *Poisson* επιλέγουμε το  $c$  ή το  $u$  διάγραμμα και
- Όταν τα δεδομένα προέρχονται από την κανονική κατανομή πρέπει να χρησιμοποιήσουμε είτε το  $\bar{X}$  και το  $S$  διάγραμμα παράλληλα.

Αυτές οι τρεις κατανομές, διωνυμική, *Poisson* και κανονική, είναι οι πιο συχνές με αποτέλεσμα τα παραπάνω διαγράμματα να είναι κατάλληλα για τις περισσότερες, αλλά όχι για όλες, τις περιπτώσεις. Για παράδειγμα, σε πολλές περιπτώσεις, συνεχή δεδομένα μπορούμε να υποθέσουμε ότι προέρχονται από την κανονική κατανομή. Αυτά μπορεί να είναι: η αναμονή των ασθενών, η διάρκεια μίας επέμβασης, ο χρόνος λήψης αντιβιοτικών που χορηγούνται πριν από μία επέμβαση, διάφορες μετρήσεις συνεχών μεταβλητών που εκφράζουν βιολογικά μεγέθη, κλπ.

Επίσης, σε πολλές περιπτώσεις, διακριτά συμβάντα θεωρούμε ότι ακολουθούν την κατανομή *Poisson*. Για παράδειγμα το πλήθος των πτώσεων σε ένα μήνα ενός ασθενούς με αδυναμία στήριξης, αφίξεις στο τμήμα επειγόντων περιστατικών ενός νοσοκομείου σε μία εβδομάδα, τηλεφωνικές κλήσεις στο κέντρο άμεσης βοήθειας σε μία εβδομάδα ή ημέρα, κρούσματα μιας μεταδοτικής νόσου σε εβδομαδιαία βάση, κλπ.

Μία επίσης συνηθισμένη περίπτωση για διακριτά δεδομένα είναι αυτά να ακολουθούν την διωνυμική κατανομή. Ενδεικτικά αναφέρουμε τις γεννήσεις με φυσιολογικό τοκετό ή με καισαρική τομή, η μόλυνση ή μη ενός ασθενούς στο χειρουργείο, η επανεμφάνιση της νόσου μετά την θεραπεία ενός ασθενούς, αν μετά από μια συγκεκριμένη επέμβαση ο ασθενής επιβίωσε, αν το αποτέλεσμα από δύο ή περισσότερες διαγνωστικές εξετάσεις συμφωνούν ή όχι.

### 2.2.3 Τα διαγράμματα ελέγχου $p$ και $np$

Τα διαγράμματα  $np$  και  $p$  χρησιμοποιούνται για να ελέγξουμε το πλήθος των ασθενών και το ποσοστό των ασθενών, αντίστοιχα (ή το πλήθος των επισκέψεων στα επείγοντα και το αντίστοιχο ποσοστό, κλπ.) . Όταν θεωρούμε ταυτόχρονα περισσότερες από μία δοκιμές *Bernoulli* ίδιου τύπου, για παράδειγμα τις χειρουργικές επεμβάσεις που έγιναν με την ίδια

μέθοδο για την ίδια πάθηση σε ένα μήνα, που θεωρούμε ότι είναι ανεξάρτητες μεταξύ τους και σε κάθε δοκιμή υπάρχει πιθανότητα  $p$  (ισόνομες) να έρθει το ίδιο αποτέλεσμα, για παράδειγμα ο ασθενής επιβιώνει, το πλήθος από αυτές τις επιτυχίες ή το ποσοστό επιτυχιών στο σύνολο είναι διωνυμικές τυχαίες μεταβλητές. Αξίζει να σημειωθεί ότι παρόλο που πολλά από τα παρατηρούμενα χαρακτηριστικά δεν είναι διωνυμικά μπορούμε να τα τροποποιήσουμε ώστε να είναι. Για παράδειγμα ο χρόνος θεραπείας είναι μια συνεχής τυχαία μεταβλητή αλλά το αν ο χρόνος θεραπείας ξεπερνάει τις 72 ώρες είναι μια δοκιμή *Bernoulli*. Συνεπώς τα διαγράμματα  $p$  και  $np$  μπορούν να εφαρμοστούν σε ένα μεγάλο εύρος περιπτώσεων.

Έστω ότι το ποσοστό εμφάνισης μίας νόσου στον πληθυσμό είναι γνωστό και ίσο με  $p$ . Επιλέγουμε από τον πληθυσμό με τυχαίο τρόπο  $m$  ανεξάρτητα δείγματα  $n$  ατόμων το καθένα. Θεωρούμε την τυχαία μεταβλητή  $X_{ij}$ ,  $i = 1, 2, \dots, m$  και  $j = 1, 2, \dots, n$  η οποία παίρνει τις τιμή 1 αν το  $j$ -οστό άτομο στο  $i$ -οστό δείγμα είναι ασθενής και την τιμή 0 αν δεν είναι. Η  $X_{ij}$  θα ακολουθεί την κατανομή Bernoulli με πιθανότητα  $p$ . Το πλήθος των ασθενών στο  $i$ -οστό δείγμα δίνεται από την μεταβλητή:  $X_i = X_{i1} + \dots + X_{in}$  που ως άθροισμα από  $n$  το πλήθος μεταβλητές Bernoulli θα ακολουθεί διωνυμική κατανομή δηλ.  $X_i \sim B(n, p)$ . Συνεπώς, για κάθε  $i$ , η μέση τιμή και η διακύμανση της  $X_i$  θα είναι αντίστοιχα:

$$\mu_{X_i} = n_i p$$

και

$$\sigma_{X_i}^2 = n_i p(1 - p)$$

Αν το ποσοστό των ασθενών στον πληθυσμό δεν είναι γνωστό θα πρέπει να εκτιμηθεί. Από τα  $m$  δείγματα μεγέθους  $n$  που έχουμε λάβει η αμερόληπτη εκτιμήτρια για το ποσοστό είναι:

$$\hat{p} = \bar{p} = \frac{p_1 + p_2 + \dots + p_m}{m} = \frac{\sum_{i=1}^m \sum_{j=1}^n X_{ij}}{nm}$$

Αφού  $\sum_{i=1}^m \sum_{j=1}^n X_{ij} \sim B(mn, p)$  θα είναι και  $E(\hat{p}) = p$ , άρα είναι πράγματι αμερόληπτη εκτιμήτρια.

Όπως περιγράψαμε στην παράγραφο 2.2.1. οι γενικές εξισώσεις για την κεντρική γραμμή και τα  $L\sigma$  όρια ελέγχου της τυχαίας μεταβλητής  $X_i$  είναι:

$$UCL = \mu_{X_i} + L\sigma_{X_i}$$

$$CL = \mu_{X_i}$$

$$LCL = \mu_{X_i} - L\sigma_{X_i}$$

Έτσι αν αντικαταστήσουμε τις τιμές των  $\mu_{X_i}, \sigma_{X_i}$  στις παραπάνω εξισώσεις και για  $L = 3$ , προκύπτουν τα  $3\sigma$  όρια ελέγχου Φάσης I και Φάσης II για την παρακολούθηση του αριθμού των ασθενών, Πίνακας 2.1 :

Όρια Ελέγχου $np$ –διάγραμμα			
<b>ΦΑΣΗ I</b>	$LCL = n\hat{p} - 3\sqrt{n\hat{p}(1 - \hat{p})}$	$CL = n\hat{p}$	$UCL = n\hat{p} + 3\sqrt{n\hat{p}(1 - \hat{p})}$
<b>ΦΑΣΗ II</b>	$LCL = np - 3\sqrt{np(1 - p)}$	$CL = np$	$UCL = np + 3\sqrt{np(1 - p)}$

**Πίνακας 2.1** Όρια ελέγχου για το  $np$  –διάγραμμα.

Το ποσοστό των ασθενών στο  $i$  δείγμα,  $p_i = X_i/n$ , θα έχει μέση τιμή και διακύμανση  $p$  και  $p(1 - p)/n$  αντίστοιχα. Όπως είδαμε μια αμερόληπτη εκτιμήτρια για το ποσοστό  $p$  είναι το  $\hat{p}$ . Και η εκτιμήτρια για την διακύμανση θα είναι το πηλίκο  $\hat{p}(1 - \hat{p})/n$ .

Συνεπώς τα όρια ελέγχου Φάσης I και Φάσης II για την παρακολούθηση του ποσοστού  $p$  των ασθενών είναι:

Όρια Ελέγχου $p$ –διάγραμμα			
<b>ΦΑΣΗ I</b>	$LCL = \hat{p} - 3\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$	$CL = \hat{p}$	$UCL = \hat{p} + 3\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$
<b>ΦΑΣΗ II</b>	$LCL = p - 3\sqrt{\frac{p(1 - p)}{n}}$	$CL = p$	$UCL = p + 3\sqrt{\frac{p(1 - p)}{n}}$

**Πίνακας 2.2** Όρια ελέγχου για το  $p$  –διάγραμμα

Τα όρια ελέγχου που δόθηκαν στον Πίνακα 2.2 αφορούν την περίπτωση που το μέγεθος των δειγμάτων είναι σταθερό.

Πράγματι αν έχουμε  $m$  ανεξάρτητα δείγματα μεγέθους  $n_i$  το καθένα υπολογίζουμε:

$$\hat{p} = \bar{p} = \frac{X_1 + X_2 + \dots + X_m}{n_1 + n_2 + \dots + n_m} = \frac{\sum_{i=1}^m \sum_{j=1}^n X_{ij}}{\sum_{i=1}^m n_i}$$

Και πάλι θα είναι  $E(\hat{p}) = p$  αφού  $\sum_{i=1}^m \sum_{j=1}^n X_{ij} \sim B(n_1 + n_2 + \dots + n_m, p)$

Αν τα δείγματα έχουν διαφορετικό μέγεθος τότε χρησιμοποιούμε προσεγγιστικά όρια. Θα θεωρούμε ότι όλα τα δείγματα έχουν κοινό μέγεθος ίσο με τη δειγματική μέση τιμή τους και αντικαθιστούμε στους παραπάνω τύπους το  $n$  με το  $\bar{n} = (n_1 + n_2 + \dots + n_m)/m$ . Η προσέγγιση αυτή έρχεται από τον βιομηχανικό έλεγχο διεργασίας. Έχει προταθεί από την Ford Motor Company υπό τον περιορισμό ότι  $\bar{n} \in [0.75n_i, 1.25n_i]$  για κάθε  $i$ . Όταν χρησιμοποιούμε αυτά τα προσεγγιστικά όρια απαιτείται μεγάλη προσοχή όταν υπάρχουν παρατηρήσεις σε μικρή απόσταση από τα όρια του ελέγχου.

Για παράδειγμα ας υποθέσουμε ότι σε ένα νοσοκομείο πρόκειται να εφαρμοστεί μία νέα χειρουργική μέθοδος αφαίρεσης σπύλων σε ασθενείς. Ο Διευθυντής του νοσοκομείου επιθυμεί να διαπιστώσει αν η νέα μέθοδος θα αυξήσει ή θα ελαττώσει το ποσοστό των μολύνσεων από την αφαίρεση σπύλων. Από τα αρχεία που τηρούνται στο νοσοκομείο, επιλέγει για κάθε μία από τις προηγούμενες  $m = 30$  εβδομάδες ένα τυχαίο δείγμα  $n = 50$  ασθενών που έκαναν αφαίρεση σπύλου και καταγράφει τον αριθμό των ασθενών που μολύνθηκαν. Στον Πίνακα 2.3 καταγράφονται τα αποτελέσματα της δειγματοληψίας. Ως  $X_i$  και  $p_i$  ορίζεται ο αριθμός των ασθενών που μολύνθηκαν και το ποσοστό των μολύνσεων αντίστοιχα την  $i$ -οστή εβδομάδα.

Αξίζει να σημειωθεί ότι οι νέες δοκιμές δεν θα είναι ίδιου τύπου με τις προηγούμενες 30 αλλά στο παρόν παράδειγμα απλώς χρησιμοποιούμε τα όρια ελέγχου που καθορίστηκαν ως επιθυμητά για να ελέγξουμε τα αποτελέσματα από τις επόμενες δοκιμές. Επίσης θεωρούμε ότι ο χρόνος που διαρκεί η επέμβαση (χρόνος έκθεσης στον κίνδυνο) είναι σταθερός για όλους τους ασθενείς και με τις δύο μεθόδους.

Εβδ.	$n$	$X_i$	$p_i$	Εβδ.	$n$	$X_i$	$p_i$	Εβδ.	$n$	$X_i$	$p_i$
1	50	3	0,06	11	50	3	0,06	21	50	3	0,06
2	50	2	0,04	12	50	4	0,08	22	50	5	0,1
3	50	4	0,08	13	50	3	0,06	23	50	3	0,06
4	50	3	0,06	14	50	3	0,06	24	50	1	0,02
5	50	3	0,06	15	50	5	0,1	25	50	2	0,04
6	50	5	0,1	16	50	5	0,1	26	50	5	0,1

7	50	2	0,04	17	50	3	0,06	27	50	2	0,04
8	50	3	0,06	18	50	5	0,1	28	50	0	0
9	50	4	0,08	19	50	7	0,14	29	50	7	0,14
10	50	3	0,06	20	50	7	0,14	30	50	4	0,08

**Πίνακας 2.3** Αριθμός ασθενών που μολύνθηκαν σε επέμβαση αφαίρεσης σπίλου.

Ο διευθυντής αρχικά ελέγχει αν τα δεδομένα προέρχονται από την διωνυμική κατανομή. Ο δειγματικός μέσος για το ποσοστό είναι:

$$\hat{p} = \frac{\sum_{i=1}^{30} X_i}{nm} = 0.0726$$

Και για τον  $\chi^2$  έλεγχο καλής προσαρμογής στην  $B(n, \hat{p})$  είναι:

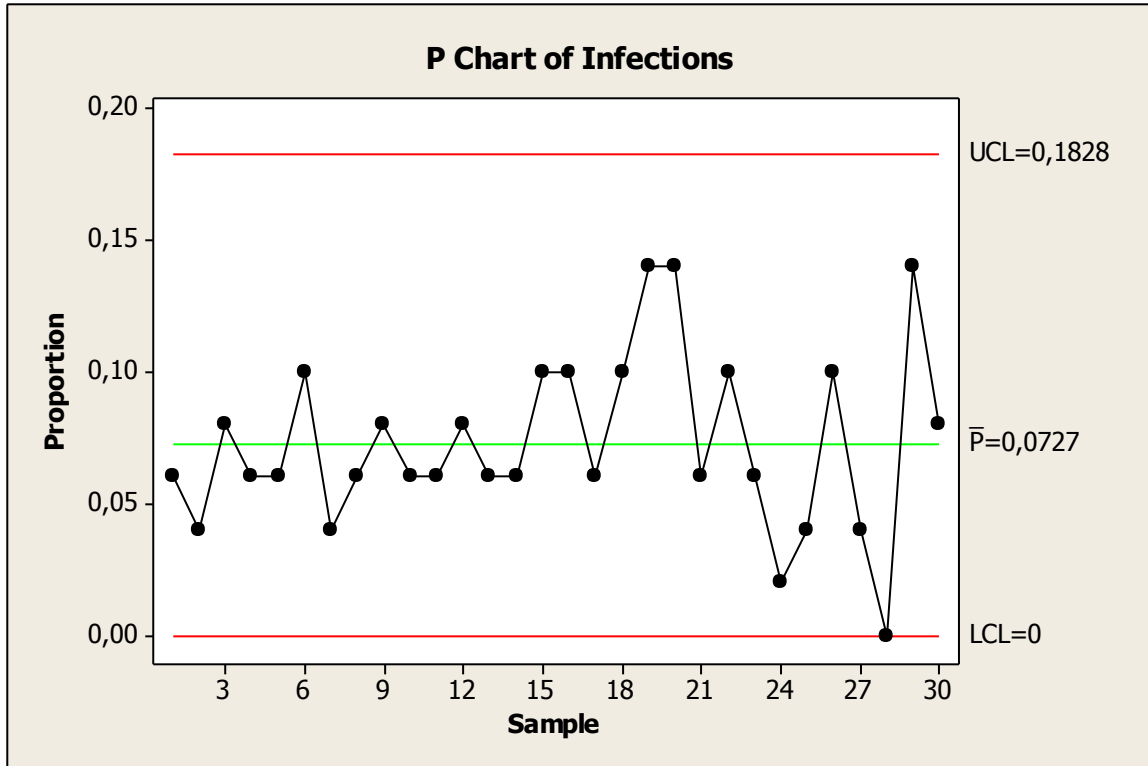
$$\chi^2 = 11.6 < 66.39 = \chi_{29}^2(0.05)$$

( $p - value = 0$ )

Επομένως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση της καλής προσαρμογής. Στην συνέχεια υπολογίζει (Φάση I) :

$$CL = 0.0727, UCL = 0.1828, LCL = -0.037.$$

Επειδή δεν είναι δυνατό το ποσοστό των ασθενών που εμφάνισαν μόλυνση να είναι αρνητικός ( $p_i \geq 0$ ) αντικαθιστά την τιμή του  $LCL$  που υπολόγισε με  $LCL = 0$ . και κατασκευάζει το  $p$ -διάγραμμα ελέγχου, Διάγραμμα 2.3 . Πράγματι κανένα σημείο δεν βρίσκεται πάνω από το άνω όριο του ελέγχου ενώ μόνο ένα σημείο βρίσκεται στο κάτω όριο. Το αναμενόμενο πλήθος δειγμάτων με κανένα ασθενή που μολύνθηκε σε επέμβαση αφαίρεσης σπίλου είναι 0.69. Ο Διευθυντής κρίνει ότι το μέσο ποσοστό μολύνσεων και τα όρια του ελέγχου είναι αποδεκτά και τα χρησιμοποιεί για έλεγχο Φάσης II μετά την εφαρμογή της νέας μεθόδου.



**Διάγραμμα 2.3**  $p$ -διάγραμμα ελέγχου Φάσης I για το ποσοστό των μολύνσεων από καθετήρα.

Τα δεδομένα από τις τέσσερις πρώτες εβδομάδες εφαρμογής της νέας μεθόδου φαίνονται στον Πίνακα 2.4 .

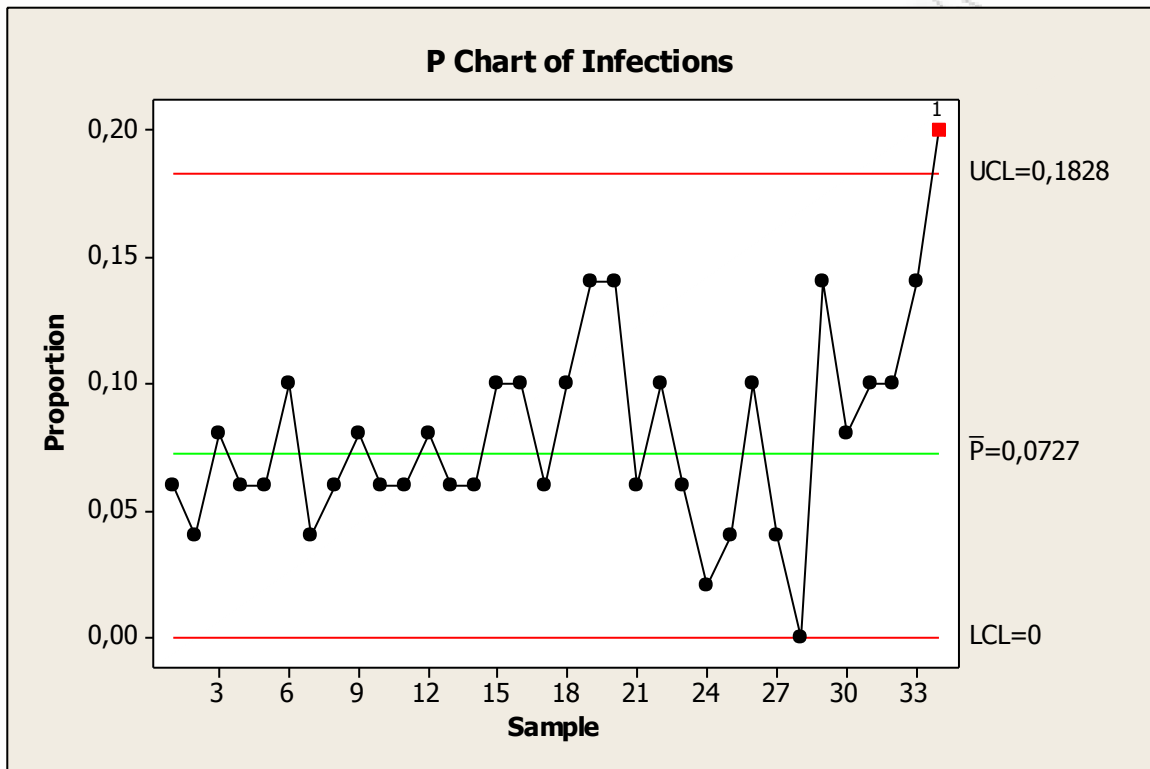
Εβδομάδα	$n$	$X_i$	$p_i$
31	50	5	0.1
32	50	5	0.10
33	50	7	0.14
34	50	10	0.20

**Πίνακας 2.4** Αριθμός ασθενών που μολύνθηκαν σε επέμβαση αφαίρεσης σπίλου μετά την εφαρμογή της νέας μεθόδου.

Μετά από 4 εβδομάδες που εφαρμόζεται η νέα μέθοδος προκύπτει το πρώτο σημείο πάνω από το άνω όριο ελέγχου, Διάγραμμα 2.4. Το δειγματικό ποσοστό των ασθενών που



μολύνθηκαν είναι 0.2 . Υπό την μηδενική υπόθεση η πιθανότητα να συμβεί αυτό είναι μόλις 0.002 .



**Διάγραμμα 2.4**  $p$  –διάγραμμα ελέγχου Φάσης II για το ποσοστό των μολύνσεων από καθετήρα, 4<sup>η</sup> εβδομάδα.

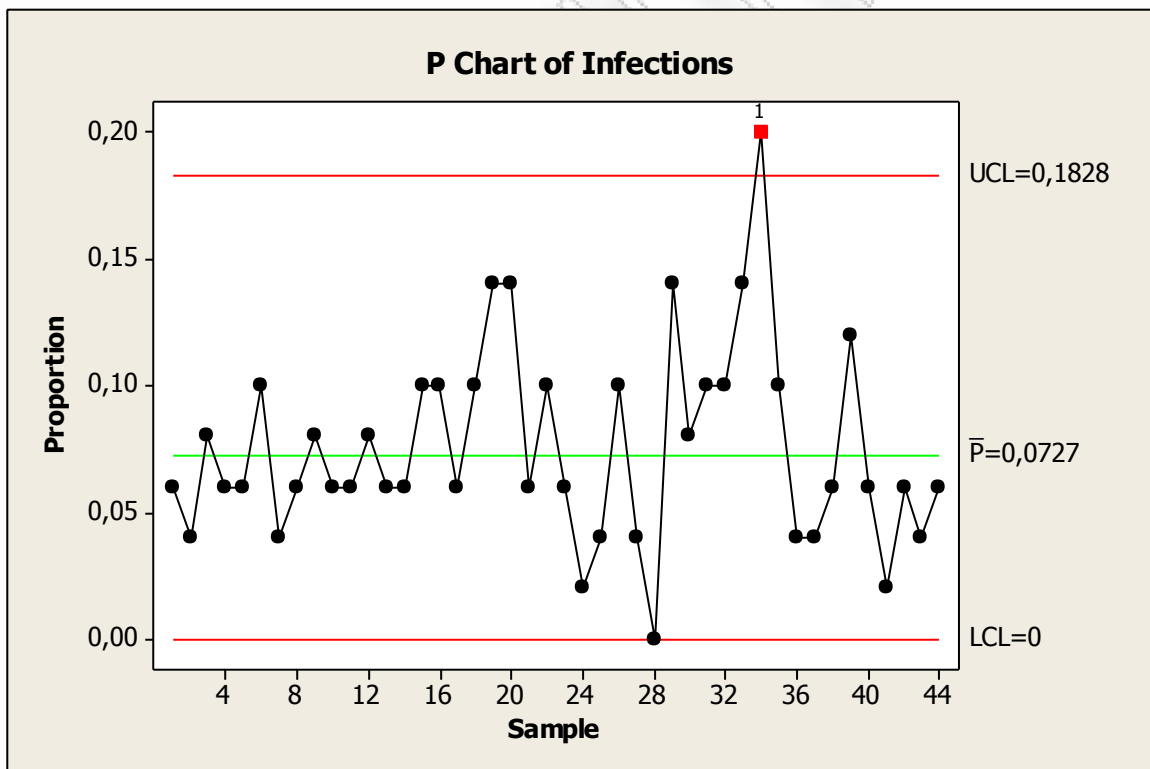
Ύστερα από έλεγχο που διενεργήθηκε προέκυψε ότι το προσωπικό δεν ήταν σωστά εκπαιδευμένο στην νέα μέθοδο και επανεκπαιδεύτηκε. Δέκα εβδομάδες μετά από αυτό, τα δεδομένα από τις μολύνσεις φαίνονται στον Πίνακα 2.5 και το διάγραμμα ελέγχου είναι το Διάγραμμα 2.5 .

Εβδομάδα	$n$	$X_i$	$p_i$
35	50	5	0,10
36	50	2	0,04
37	50	2	0,04
38	50	3	0,06
39	50	6	0,12
40	50	3	0,06

41	50	1	0,02
42	50	3	0,06
43	50	2	0,04
44	50	3	0,06

**Πίνακας 2.5** Αριθμός ασθενών που μολύνθηκαν σε επέμβαση αφαίρεσης σπίλου τις εβδομάδες 11 – 20 μετά την εφαρμογή της νέας μεθόδου.

Από το p-διάγραμμα βλέπουμε ότι η διεργασία επανήλθε εντός των ορίων ελέγχου. Επιπλέον υπάρχει και μία ασθενής ένδειξη ότι ο μέσος μετατοπίστηκε σε χαμηλότερα επίπεδα καθώς οκτώ από τα δέκα καινούρια σημεία βρίσκονται κάτω από την κεντρική γραμμή.



**Διάγραμμα 2.5** p-διάγραμμα ελέγχου Φάσης II για το ποσοστό των μολύνσεων από καθετήρα, 14<sup>η</sup> εβδομάδα μετά την εφαρμογή της νέας μεθόδου.

### 2.2.4 Τα διαγράμματα ελέγχου c και u

Στην προηγούμενη παράγραφο είδαμε την ανάπτυξη των διαγραμμάτων ελέγχου για δεδομένα που μπορούν να αναλυθούν με την διωνυμική κατανομή. Όταν εξετάζουμε το

πλήθος από περιστατικά σε μία μονάδα χρόνου ή χώρου, που έχουν μικρή αλλά σταθερή πιθανότητα να συμβούν και δεν είναι σαφές αν τα δεδομένα προκύπτουν από δοκιμές *Bernoulli*, μπορούμε να υποθέσουμε την κατανομή *Poisson*. Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε το  $c$  ή το  $u$  διάγραμμα ελέγχου για το πλήθος των περιστατικών ή το μέσο αριθμό περιστατικών αντίστοιχα. Η διαφορά ανάμεσα στα δύο διαγράμματα είναι ότι στο  $c$  διάγραμμα θεωρούμε ότι τα περιστατικά καταγράφονται σε ίσα διαστήματα, με άλλα λόγια σταθερό μέγεθος δείγματος. Αντίθετα στο  $u$  διάγραμμα μπορούμε να έχουμε διαστήματα που διαφέρουν μεταξύ τους.

Στο παράδειγμα με τις μολύνσεις από επέμβαση για αφαίρεση σπύλου ο χρόνος που διαρκεί η επέμβαση είναι ίδιος για όλους τους ασθενείς. Αντίθετα αν εξετάζαμε τον αριθμό των ασθενών που μολύνθηκαν κατά την παραμονή τους στο νοσοκομείο τότε ο χρόνος νοσηλείας για κάθε ασθενή θα ήταν διαφορετικός. Έτσι θα μπορούσαμε να θεωρήσουμε τον εβδομαδιαίο μέσο αριθμό των ασθενών που μολύνθηκαν ανά 1000 ώρες θεραπείας. Επιπλέον υποθέσαμε ότι οι ασθενείς μπορεί να εμφανίσουν ένα είδος μόλυνσης. Εναλλακτικά μπορεί να εμφανιστούν δύο ή περισσότεροι τύποι μολύνσεων στον ίδιο ασθενή.

Η κατανομή *Poisson* και η διωνυμική κατανομή είναι δύο διαφορετικές κατανομές που προσαρμόζονται σε διαφορετικές περιπτώσεις. Στην περίπτωση της παρακολούθησης της δημόσιας υγείας μπορεί η εφαρμογή τους να οδηγήσει σε παρόμοια συμπεράσματα. Η επιλογή ανάμεσα στις δύο εξαρτάται τόσο από τα δεδομένα, όσο και από τον τρόπο που συλλέχτηκαν αλλά και από τον τρόπο που θέλουμε να τα παρουσιάσουμε. Φυσικά αλλάζοντας την υφή των δεδομένων, μειώνεται η ισχύς του ελέγχου. Τέλος είναι γνωστό ότι η διωνυμική κατανομή, όταν το μέγεθος του δείγματος τείνει στο άπειρο και η πιθανότητα  $p$  είναι πολύ μικρή, προσεγγίζεται από την κατανομή *Poisson* με παράμετρο  $\lambda = np$ .

Έστω, λοιπόν, η τυχαία μεταβλητή  $X$  που εκφράζει τον αριθμό διαφορετικών μολύνσεων που εμφανίζει ένας ασθενής σε συγκεκριμένο χρονικό διάστημα, υπό συγκεκριμένες συνθήκες. Η συνάρτηση πιθανότητας, η μέση τιμή και η διακύμανση της τ.μ.  $X$ , υπό την υπόθεση ότι ακολουθεί την κατανομή *Poisson* με παράμετρο  $c$ , είναι:

$$P(X = x) = e^{-c} \frac{c^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$\mu_x = \sigma_x^2 = c.$$

Το πρόβλημα που εγείρεται πάλι είναι ότι πολλές φορές η παράμετρος  $c$  της κατανομής δεν είναι γνωστή πριν από την έναρξη της βιοεπιτήρησης. Σε αυτή την περίπτωση θα πρέπει να

την εκτιμήσουμε. Έστω ότι λαμβάνουμε τυχαίο δείγμα από  $m$  ανεξάρτητα άτομα και έστω η τυχαία μεταβλητή  $X_i$  που εκφράζει τον αριθμό των μολύνσεων (ή ασθενειών, ή πτώσεων κλπ) για το  $i$  άτομο, τότε η δειγματική μέση τιμή για το  $c$  είναι:

$$\hat{c} = \bar{c} = \frac{X_1 + X_2 + \dots + X_m}{m}$$

Είναι  $E(\bar{c}) = c$ , δηλαδή πρόκειται για αμερόληπτη εκτιμήτρια.

Έτσι τα όρια ελέγχου (3σ όρια) για το διάγραμμα  $c$  δίνονται στον ακόλουθο πίνακα:

<b>3σ Όρια Ελέγχου <math>c</math> –διάγραμμα</b>			
<b>ΦΑΣΗ Ι</b>	$LCL = \hat{c} - 3\sqrt{\hat{c}}$	$CL = \hat{c}$	$UCL = \hat{c} + 3\sqrt{\hat{c}}$
<b>ΦΑΣΗ ΙΙ</b>	$LCL = c - 3\sqrt{c}$	$CL = c$	$UCL = c + 3\sqrt{c}$

**Πίνακας 2.6** Όρια ελέγχου για το  $c$ - διάγραμμα

Έστω τώρα ότι λαμβάνουμε παρατηρήσεις από  $m$  δείγματα μεγέθους  $n$  το καθένα. Για να υπολογίσουμε τα όρια του  $u$  διαγράμματος θα θεωρήσουμε την τυχαία μεταβλητή  $X_{ij}$  από την κατανομή *Poisson* με παράμετρο  $c$  που εκφράζει τον αριθμό των μολύνσεων (ή παθήσεων κλπ) για  $j$ -οστό το άτομο στο  $i$ -οστό δείγμα. Τότε ο συνολικός αριθμός μολύνσεων στο  $i$  δείγμα θα δίνεται από την τυχαία μεταβλητή  $X_i = \sum_{j=1}^n X_{ij} \sim P(nc)$ . Θεωρούμε επίσης την τυχαία μεταβλητή  $U_i = \frac{X_i}{n}$  η οποία εκφράζει το μέσο αριθμό μολύνσεων ανά άτομο στο δείγμα  $i$ . Τότε η μέση τιμή και η διακύμανση της  $U_i$  θα είναι:

$$\mu_{U_i} = \frac{1}{n} nc = c$$

και

$$\sigma_{U_i}^2 = \frac{1}{n^2} nc = \frac{1}{n} c$$

Στην περίπτωση που το  $c$  δεν είναι γνωστό εκτιμάται από το δειγματικό μέσο για το  $U_i$ . Πράγματι:

$$\bar{U} = \frac{U_1 + \dots + U_m}{m} = \frac{\sum_{i=1}^m X_i}{mn}$$

Αφού  $X_i = \sum_{j=1}^n X_{ij} \sim P(nc)$  τότε  $\sum_{i=1}^m X_i \sim P(mnc)$  άρα θα είναι  $E(\bar{U}) = c$ .

Έτσι τα όρια ελέγχου (3σ όρια) για το διάγραμμα  $u$  δίνονται στον ακόλουθο πίνακα:

3σ Όρια Ελέγχου $u$ – διάγραμμα			
ΦΑΣΗ Ι	$LCL = \bar{U} - 3\sqrt{\frac{\bar{U}}{n}}$	$CL = \bar{U}$	$UCL = \bar{U} + 3\sqrt{\frac{\bar{U}}{n}}$
ΦΑΣΗ ΙΙ	$LCL = c - 3\sqrt{\frac{c}{n}}$	$CL = c$	$UCL = c + 3\sqrt{\frac{c}{n}}$

Πίνακας 2.7 Όρια ελέγχου για το  $u$ - διάγραμμα

Τα παραπάνω όρια ελέγχου αφορούν την περίπτωση σταθερού μεγέθους δείγματος. Αν κάθε δείγμα έχει διαφορετικό μέγεθος  $n_i$  τότε, όπως και στην περίπτωση του  $p$  διαγράμματος, χρησιμοποιούμε προσεγγιστικά όρια αντικαθιστώντας στα όρια Φάσης Ι το  $n$  με το  $\bar{n}$ . Όπου

$$\bar{n} = \frac{(n_1 + n_2 + \dots + n_m)}{m}$$

Προκειμένου να γίνει αντιληπτή η εφαρμογή των διαγραμμάτων ελέγχου αυτής της παραγράφου θα χρησιμοποιήσουμε τα δεδομένα του Πίνακα 2.8 [βλ. Mohammed et al, (2008)]. Αυτά αναφέρονται σε πτώσεις ασθενών που νοσηλεύονται σε ένα συγκεκριμένο νοσοκομείο την περίοδο Οκτώβριος 2004 – Οκτώβριος 2005 (13 μήνες). Στη δεύτερη στήλη δίνεται το πλήθος των πτώσεων ανά μήνα και στην Τρίτη στήλη, αθροιστικά για όλους τους ασθενείς, ο χρόνος νοσηλείας σε ημέρες.

Μήνας	Πτώσεις Ασθενών	Ημέρες Νοσηλείας	Πτώσεις ανα ημέρα νοσηλείας	LCL	UCL
Οκτ-04	1	1048	0,0009542	0	0,009361
Νοε-04	4	896	0,0044643	0	0,009822
Δεκ-04	3	918	0,003268	0	0,009748
Ιαν-05	4	995	0,0040201	0	0,00951

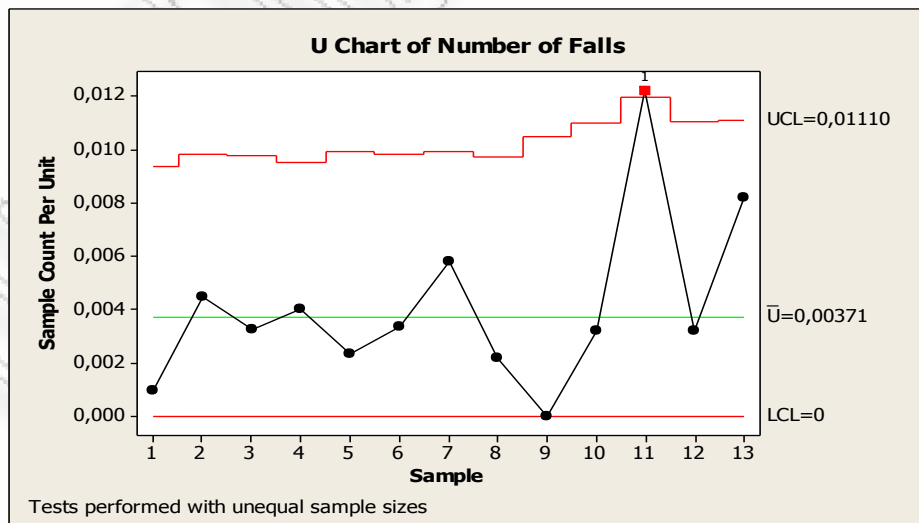
Φεβ-05	2	866	0,0023095	0	0,009927
Μαρ-05	3	896	0,0033482	0	0,009822
Απρ-05	5	864	0,005787	0	0,009934
Μαϊ-05	2	930	0,0021505	0	0,009709
Ιουν-05	0	732	0	0	0,010471
Ιουλ-05	2	630	0,0031746	0	0,010998
Αυγ-05	6	492	0,0121951	0	0,011956
Σεπ-05	2	622	0,0032154	0	0,011045
Οκτ-05	5	612	0,0081699	0	0,011104

**Πίνακας 2.8** Πτώσεις ασθενών που νοσηλεύονται σε ένα νοσοκομείο

Θα κατασκευάσουμε το  $u$  - διάγραμμα για μεταβλητό μέγεθος δείγματος. Η κεντρική γραμμή του διαγράμματος είναι:

$$\bar{U} = \frac{U_1 + \dots + U_m}{m} = \frac{\sum_{i=1}^m X_i}{mn} = \frac{39}{10501} = 0.00371$$

Δηλ. ο μέσος των πτώσεων ανά ημέρα νοσηλείας είναι 0.00371. Το άνω όριο ελέγχου για κάθε μήνα δίνεται στην τελευταία στήλη του Πίνακα 2.8 , ενώ το κάτω όριο του διαγράμματος έχει οριστεί ίσο με το μηδέν καθώς η επίλυση της εξίσωσης, που δίνει το  $LCL$ , δίνει αρνητικές τιμές κάτι που δεν έχει νόημα όταν μετράμε το πλήθος πτώσεων.



**Διάγραμμα 2.6**  $u$ -διάγραμμα ελέγχου για τις πτώσεις ασθενών εντός ενός νοσοκομείου

Παρατηρούμε, στο Διάγραμμα 2.6, ότι η 11<sup>η</sup> παρατήρηση είναι πάνω από το άνω όριο ελέγχου. Δηλαδή, τον Αύγουστο του 2005 υπάρχει ένδειξη ύπαρξης ειδικής αιτίας μεταβλητότητας. Αξίζει να σημειώσουμε ότι έχουμε υποθέσει ότι κάθε πτώση ασθενούς συμβαίνει ανεξάρτητα από τις υπόλοιπες και ότι κάθε χρονική στιγμή στην ίδια γεωγραφική θέση συμβαίνει το πολύ μία πτώση.

### 2.2.5 Τα διαγράμματα ελέγχου $\bar{X}$ , $S$ και $R$ .

Σε πολλές περιπτώσεις η μεταβλητή που παρατηρούμε σε μία διαδικασία βιοεπιτήρησης είναι συνεχής τυχαία μεταβλητή. Σαν παράδειγμα μπορούμε να αναφέρουμε τις μετρήσεις σακχάρου στο αίμα, την αρτηριακή πίεση κλπ. Σε αυτές τις περιπτώσεις, συχνά θεωρούμε ότι τα δεδομένα προέρχονται από την κανονική κατανομή. Τότε χρησιμοποιούμε τα διαγράμματα  $\bar{X}$  και  $S$  για την παρακολούθηση της μέσης τιμής και της μεταβλητότητας. Η κανονική κατανομή περιγράφεται από δύο παραμέτρους, την μέση τιμή  $\mu$  και την διασπορά  $\sigma^2$ . Μία στατιστικά σημαντική αλλαγή στην μία από τις δύο παραμέτρους δεν συνεπάγεται υποχρεωτικά αλλαγή και στην δεύτερη. Όμως για να θεωρήσουμε ότι η υποκείμενη διεργασία είναι εκτός ελέγχου αρκεί μία από τις δύο να έχει αλλάξει. Για αυτό, τα δύο διαγράμματα ελέγχου,  $\bar{X}$  και  $S$ , θα πρέπει να χρησιμοποιούνται ταυτόχρονα.

Ας υποθέσουμε ότι η υπό μελέτη μεταβλητή της υποκείμενης διεργασίας είναι η τυχαία μεταβλητή  $X$  που ακολουθεί κανονική κατανομή με παραμέτρους  $\mu$  και  $\sigma^2$ . Έστω, επίσης, ότι λαμβάνουμε  $m$  ανεξάρτητα τυχαία δείγματα μεγέθους  $n$  το καθένα. Έτσι για το  $j$ -οστό άτομο στο δείγμα  $i$  η τιμή του χαρακτηριστικού είναι  $X_{ij} \sim N(\mu, \sigma^2)$ . Η δειγματική μέση τιμή για το χαρακτηριστικό στο  $i$ -οστό δείγμα θα είναι:

$$\bar{X}_i = \frac{\sum_{j=1}^n X_{ij}}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Ως γνωστόν ισχύει  $E(\bar{X}) = \mu$  άρα είναι αμερόληπτη εκτιμήτρια. Ένα διάστημα συντελεστού εμπιστοσύνης  $1 - \alpha$  για την εκτιμήτρια είναι:

$$\left( \mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

όπου  $z \sim N(0,1)$ .

Έτσι τα όρια και η κεντρική γραμμή του ελέγχου μπορούμε να θεωρήσουμε ότι είναι:

$$UCL = \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$CL = \mu$$

$$LCL = \mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Έστω τώρα ότι ζητάμε να κατασκευάσουμε όρια ελέγχου  $3\sigma$  Φάσης II. Αυτά θα είναι:

$$UCL = \mu + 3 \frac{\sigma}{\sqrt{n}}$$

$$CL = \mu$$

$$LCL = \mu - 3 \frac{\sigma}{\sqrt{n}}$$

Επιλύοντας τα δύο παραπάνω συστήματα, ως προς  $\alpha$ , παίρνουμε την τιμή 0.0027. Με άλλα λόγια, στην περίπτωση της κανονικής κατανομής, όταν χρησιμοποιούμε  $3\sigma$  όρια ελέγχου για δείγματα, η πιθανότητα να βρεθεί ο δειγματικός μέσος εντός των ορίων ελέγχου είναι 99.73%. Έτσι, επειδή η πιθανότητα ένα οποιοδήποτε σημείο να βρεθεί εκτός των ορίων είναι πολύ μικρή, ίση με 0.0027, θεωρούμε ότι υπάρχει ένδειξη εκτός ελέγχου διεργασίας.

Για τον υπολογισμό των παραπάνω ορίων θεωρήσαμε ότι ο μέσος και η διασπορά της κανονικής κατανομής είναι γνωστά και ειδικά για την διασπορά ότι είναι σταθερή. Όπως και στα διαγράμματα για διωνυμικά δεδομένα ή για δεδομένα από την *Poisson* θα εργαστούμε για την κατασκευή του διαγράμματος στην περίπτωση που ο μέσος και η διασπορά είναι άγνωστα. Η διαφορά εδώ είναι ότι τώρα έχουμε δύο παραμέτρους να εκτιμήσουμε.

Η εκτίμηση για το μέσο είναι η δειγματική μέση τιμή των ποσοτήτων  $\bar{X}_i$ . Είναι δηλαδή ο μέσος των μέσων. Γράφουμε:

$$\hat{\mu} = \bar{\bar{X}} = \frac{1}{m} \sum_{i=1}^m \bar{X}_i = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n X_{ij}$$

Η διασπορά στο παρατηρούμενο δείγμα  $i$ ,  $i = 1, 2, \dots, m$  είναι τυχαία μεταβλητή που δίνεται από την σχέση:

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



Επίσης το εύρος για το δείγμα  $i$  είναι:

$$R_i = X_i^{max} - X_i^{min}$$

όπου  $X_i^{max} = \max \{x_{ij}, j = 1, 2, \dots, n\}$  και  $X_i^{min} = \min \{x_{ij}, j = 1, 2, \dots, n\}$

Η εκτίμηση για τη διασπορά μπορεί να γίνει είτε με βάση την μέση τιμή των δειγματικών διασπορών ή με βάση την μέση τιμή των ευρών των δειγμάτων. Δηλαδή επιλέγουμε μία από τις ακόλουθες εκτιμήτριες:

Την 
$$\hat{\sigma} = \bar{R}/d_2, \text{ όπου } \bar{R} = \frac{1}{m} \sum_{i=1}^m R_i$$

ή την 
$$\hat{\sigma} = \bar{S}/c_4, \text{ όπου } \bar{S} = \frac{1}{m} \sum_{i=1}^m S_i$$

ή την 
$$\hat{\sigma} = \sqrt{\bar{S}^2}, \text{ όπου } \bar{S}^2 = \frac{1}{m} \sum_{i=1}^m S_i^2$$

Οι σταθερές  $c_4$  και  $d_2$  εξαρτώνται μόνο από το μέγεθος του δείγματος και προκύπτουν από τις παρακάτω εξισώσεις:

$$c_4 = \left(\frac{2}{n-1}\right)^2 \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

και

$$d_2 = n(n-1) \int_0^{+\infty} \int_{-\infty}^{+\infty} u \varphi(v) [\Phi(v+u) - \Phi(v)]^{n-2} \varphi(v+u) dv du, \quad u = \frac{r}{\sigma}, \quad r > 0$$

Μάλιστα επειδή  $\mu_{S_i} = E(S_i) = \sigma c_4$  προκύπτει ότι  $E(\bar{S}) = \sigma c_4$  η εκτιμήτρια  $\hat{\sigma} = \bar{S}/c_4$  είναι αμερόληπτη.

Συνοψίζοντας, για κάθε μία από τις δυνατές επιλογές τα όρια του ελέγχου (Φάση I) είναι αυτά που δίνονται στον Πίνακα 2.9.

3σ Όρια Ελέγχου Φάσης I			
$\bar{X}$ – διάγραμμα			
$\hat{\sigma} = \bar{R}/d_2$	$LCL = \bar{\bar{X}} - \frac{3}{d_2\sqrt{n}}\bar{R}$	$CL = \bar{\bar{X}}$	$UCL = \bar{\bar{X}} + \frac{3}{d_2\sqrt{n}}\bar{R}$
$\hat{\sigma} = \bar{S}/c_4$	$LCL = \bar{\bar{X}} - \frac{3}{c_4\sqrt{n}}\bar{S}$	$CL = \bar{\bar{X}}$	$UCL = \bar{\bar{X}} + \frac{3}{c_4\sqrt{n}}\bar{S}$
$\hat{\sigma} = \sqrt{\bar{S}^2}$	$LCL = \bar{\bar{X}} - \frac{3}{\sqrt{n}}\sqrt{\bar{S}^2}$	$CL = \bar{\bar{X}}$	$UCL = \bar{\bar{X}} + \frac{3}{\sqrt{n}}\sqrt{\bar{S}^2}$

Πίνακας 2.9 Όρια ελέγχου Φάσης I για το  $\bar{X}$ - διάγραμμα

Όπως είδαμε και στην αρχή της παραγράφου, όταν παρακολουθούμε μια τυχαία μεταβλητή από την κανονική κατανομή, μπορεί να συμβεί μία στατιστικά σημαντική αλλαγή στην διασπορά, χωρίς απαραίτητα να έχει μετατοπιστεί ο μέσος. Και σε αυτή την περίπτωση η διεργασία είναι εκτός στατιστικού ελέγχου. Για τον έλεγχο της διασποράς των τιμών του χαρακτηριστικού χρησιμοποιούνται δύο είδη διαγραμμάτων. Το  $R$  –διάγραμμα ελέγχου του εύρους και το  $S$  – διάγραμμα ελέγχου της τυπικής απόκλισης.

Όταν τα δείγματα προέρχονται από κανονικό πληθυσμό η μέση τιμή και η τυπική απόκλιση του εύρους είναι αντίστοιχα  $\mu_R = d_2\sigma$  και  $\sigma_R = d_3\sigma$ , όπου  $\sigma$  είναι η τυπική απόκλιση του πληθυσμού και  $d_3$  μία σταθερά που εξαρτάται, όπως και η  $d_2$ , μόνο από το μέγεθος δείγματος. Αν η τυπική απόκλιση του πληθυσμού είναι γνωστή τότε τα όρια ελέγχου Φάσης II είναι αυτά που φαίνονται στον Πίνακα 2.9.

Αν η πληθυσμιακή τυπική απόκλιση είναι άγνωστη θα πρέπει να εκτιμηθεί από ιστορικά δεδομένα και να υπολογιστούν τα όρια Φάσης I. Αφού η εκτιμήτρια της πληθυσμιακής τυπικής απόκλισης είναι  $\hat{\sigma} = \bar{R}/d_2$ , η εκτιμήτρια για την τυπική απόκλιση του εύρους είναι:

$$\hat{\sigma}_R = d_3(\bar{R}/d_2)$$

Έτσι υπολογίζουμε τα όρια Φάσης I για το R-διάγραμμα:

$$UCL = \bar{R} + 3d_3(\bar{R}/d_2)$$

$$CL = \bar{R}$$

$$LCL = \bar{R} - 3d_3(\bar{R}/d_2)$$

Και αν θέσουμε  $D_3 = 1 - 3\frac{d_3}{d_2}$  και  $D_4 = 1 + 3\frac{d_3}{d_2}$  τα  $3\sigma$  όρια ελέγχου για το  $R$ -διάγραμμα είναι:

3σ Όρια Ελέγχου $R$ -διάγραμμα			
<b>ΦΑΣΗ Ι</b>	$LCL = D_3\bar{R}$	$CL = \bar{R}$	$UCL = D_4\bar{R}$
<b>ΦΑΣΗ ΙΙ</b>	$LCL = d_2\sigma - 3d_3\sigma$	$CL = d_2\sigma$	$UCL = d_2\sigma + 3d_3\sigma$

**Πίνακας 2.10** Όρια Ελέγχου για το  $R$ -διάγραμμα

Το  $R$ -διάγραμμα βασίζεται στο εύρος του δείγματος άρα αντλεί πληροφορία μόνο από δύο παρατηρήσεις, την μικρότερη και τη μεγαλύτερη. Αντίθετα η τυπική απόκλιση περιλαμβάνει πληροφορία από όλες τις παρατηρήσεις του δείγματος. Από την άλλη το  $R$ -διάγραμμα είναι πολύ πιο εύκολο να κατασκευαστεί, για τον ίδιο ακριβώς λόγο.

Είδαμε προηγούμενα ότι μια αμερόληπτη εκτιμήτρια για τη μέση τιμή της τυχαίας μεταβλητής  $S$  είναι η ποσότητα  $\sigma c_4$ . Επίσης η τυπική απόκλιση της  $S$  είναι:

$$\sigma_S = \sigma \sqrt{1 - c_4^2}$$

Έτσι η κεντρική γραμμή και τα  $3\sigma$  όρια Φάσης ΙΙ για το  $S$ -διάγραμμα θα είναι:

$$UCL = \sigma c_4 + 3\sigma \sqrt{1 - c_4^2}$$

$$CL = \sigma c_4$$

$$LCL = \sigma c_4 - 3\sigma \sqrt{1 - c_4^2}$$

Όταν η πληθυσμιακή τυπική απόκλιση δεν είναι γνωστή την εκτιμούμε από την μέση τιμή των δειγματικών τυπικών αποκλίσεων. Συνεπώς η τυπική απόκλιση της τ.μ.  $S$  θα δίνεται από την σχέση:

$$\hat{\sigma}_S = \frac{\bar{s}}{c_4} \sqrt{1 - c_4^2}$$

Συνεπώς η κεντρική γραμμή και τα  $3\sigma$  όρια Φάσης I για το  $S$ -διάγραμμα θα είναι:

$$UCL = \bar{s} + 3 \frac{\bar{s}}{c_4} \sqrt{1 - c_4^2}$$

$$CL = \bar{s}$$

$$LCL = \bar{s} - 3 \frac{\bar{s}}{c_4} \sqrt{1 - c_4^2}$$

Θέτουμε:

$$B_3 = 1 - \frac{3}{c_4} \sqrt{1 - c_4^2}, B_4 = 1 + \frac{3}{c_4} \sqrt{1 - c_4^2}, B_5 = c_4 - 3\sqrt{1 - c_4^2}, B_6 = c_4 + 3\sqrt{1 - c_4^2}$$

Τα αποτελέσματα της αντικατάστασης των παραπάνω ποσοτήτων στις σχέσεις που δίνουν τα όρια για το  $S$  –διάγραμμα δίνονται στον Πίνακα 2.11.

<b>3σ Όρια Ελέγχου <math>S</math> –διάγραμμα</b>			
<b>ΦΑΣΗ I</b>	$LCL = B_3\bar{s}$	$CL = \bar{s}$	$UCL = B_4\bar{s}$
<b>ΦΑΣΗ II</b>	$LCL = B_5\sigma$	$CL = c_4\sigma$	$UCL = B_6\sigma$

**Πίνακας 2.11** Όρια Ελέγχου για το  $S$  –διάγραμμα

### 2.2.6 Το διάγραμμα ελέγχου μεμονωμένων παρατηρήσεων και το $MR$ – διάγραμμα

Υπάρχουν διαδικασίες βιοεπιτήρησης, για παράδειγμα μέτρηση δεικτών στο αίμα, το pH κλπ, όπου το κάθε δείγμα που λαμβάνεται αποτελείται από μία μόνο μέτρηση. Σε αυτή την περίπτωση δεν είναι δυνατός ο υπολογισμός της φυσικής μεταβλητότητας. Αυτό το πρόβλημα

μπορεί να αντιμετωπιστεί με τη χρήση του κινούμενου εύρους (*moving range, MR*) μεταξύ δύο διαδοχικών μετρήσεων. Αυτό δίνεται από την σχέση:

$$MR_j = |x_j - x_{j-1}|$$

Όπου  $x_j$ ,  $j = 1, 2, \dots, m$  είναι η τιμή του υπό μελέτη χαρακτηριστικού στην  $j$ -οστή μέτρηση, με  $X \sim N(\mu, \sigma^2)$ . Το μέσο κινούμενο εύρος δίνεται από την σχέση:

$$\overline{MR} = \frac{1}{m-1} \sum_{j=2}^m MR_j .$$

Ο παρονομαστής στην παραπάνω εξίσωση προκύπτει από το πλήθος των τιμών που λαμβάνουμε για το κινούμενο εύρος που προφανώς είναι κατά μία λιγότερες από τις διαθέσιμες παρατηρήσεις. Μία εκτίμηση για την τυπική απόκλιση είναι:

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{\overline{MR}}{1.128}$$

αφού για  $n = 2$  είναι  $d_2 = 1.128$ .

Αν ο μέσος των τιμών του υπό μελέτη χαρακτηριστικού είναι γνωστός, δηλ. για τη Φάση II, η κεντρική γραμμή και τα  $3\sigma$  όρια ελέγχου για το διάγραμμα ελέγχου μεμονωμένων παρατηρήσεων ( $x$  - διάγραμμα) θα δίνονται από τις εξισώσεις στον Πίνακα 2.12.

Αν όμως ο μέσος δεν είναι γνωστός εκτιμάται από το δειγματικό μέσο των  $m$  παρατηρήσεων:

$$\hat{\mu} = \bar{x} = \frac{1}{m} \sum_{j=1}^m x_j$$

Επομένως, τα  $3\sigma$  όρια του ελέγχου και η κεντρική γραμμή, για την Φάση I, στην περίπτωση των μεμονωμένων παρατηρήσεων ( $x$  - διάγραμμα ή  $I$  - διάγραμμα) είναι αυτά που δίνονται στον Πίνακα 2.12.

Το  $x$  - διάγραμμα δεν θα πρέπει να συγχέεται με το  $\bar{x}$  - διάγραμμα . Η διαφορά τους είναι στο μέγεθος δείγματος που συνεπάγεται διαφορά στην ισχύ. Το  $x$  - διάγραμμα έχει σαφώς χαμηλότερη ισχύ. Επιπρόσθετα δεν μπορούμε να εφαρμόσουμε το κεντρικό οριακό θεώρημα με αποτέλεσμα το  $x$  - διάγραμμα να είναι ευαίσθητο στις αποκλίσεις από την κανονικότητα.

Όπως και στο  $\bar{x}$ -διάγραμμα έτσι και εδώ δεν αρκεί μόνο η εφαρμογή του  $x$ -διαγράμματος. Είναι σκόπιμο να κατασκευάζουμε και το  $MR$ -διάγραμμα. Βέβαια παρόλο που το εύρος αποτελεί έκφραση της διασποράς, ένα σημείο εκτός ελέγχου στο  $MR$ -διάγραμμα δεν μεταφράζεται υποχρεωτικά σε ένδειξη αλλαγής της διασποράς αλλά μπορεί να δηλώνει μια μεταβολή στο μέσο της διεργασίας. Η κεντρική γραμμή και τα 3σ όρια για το διάγραμμα ελέγχου του εύρους είναι:

$$LCL = D_3 \overline{MR}, \quad CL = \overline{MR}, \quad UCL = D_4 \overline{MR}.$$

Αν θέσουμε τις τιμές των σταθερών  $D_3$  και  $D_4$  για  $n = 2$ , που είναι  $D_4 = 3.267$  και  $D_3 = 0$ , προκύπτουν τα όρια του Πίνακα 2.12.

<b>3σ Όρια Ελέγχου <math>x</math>-διάγραμμα</b>			
<b>ΦΑΣΗ Ι</b>	$LCL = \bar{x} - 2.66 \overline{MR}$	$CL = \bar{x}$	$UCL = \bar{x} + 2.66 \overline{MR}$
<b>ΦΑΣΗ ΙΙ</b>	$LCL = \mu - 2.66 \overline{MR}$	$CL = \mu$	$UCL = \mu + 2.66 \overline{MR}$

**Πίνακας 2.12** Όρια Ελέγχου για το  $x$ -διάγραμμα

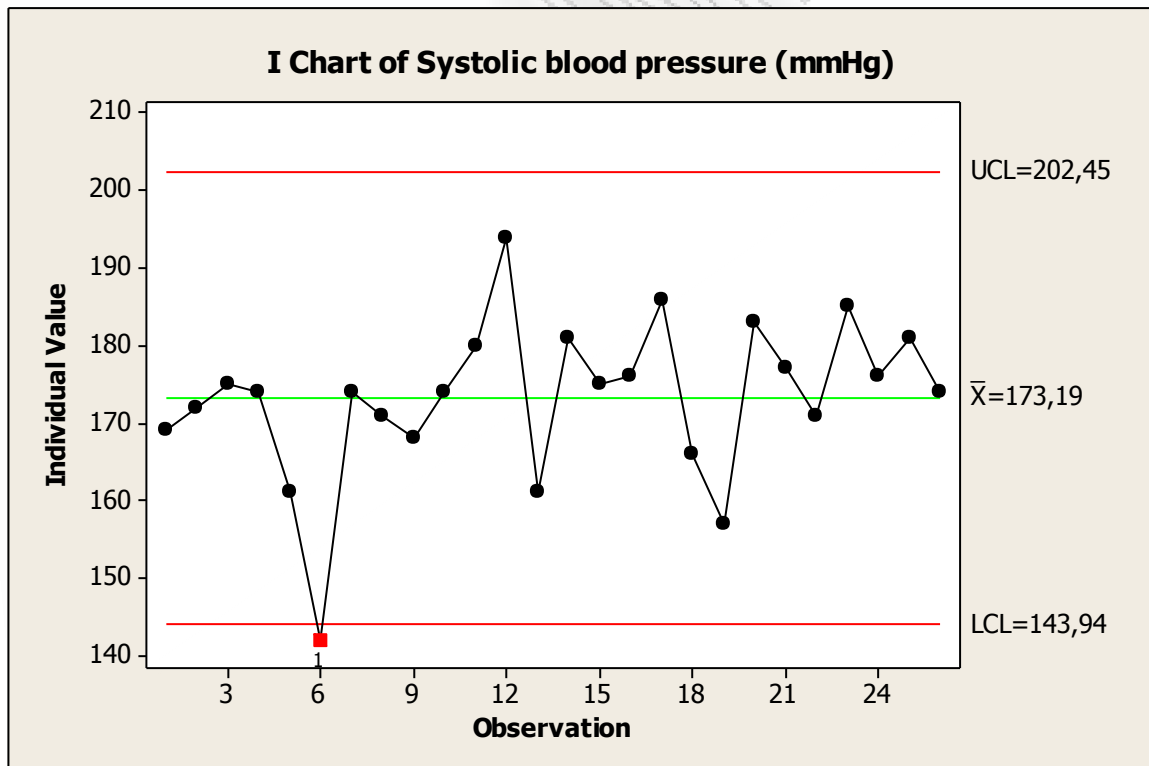
<b>3σ Όρια Ελέγχου <math>MR</math>-διάγραμμα</b>			
<b>ΦΑΣΗ Ι</b>	$LCL = 0$	$CL = \overline{MR}$	$UCL = 3.267 \overline{MR}$

**Πίνακας 2.13** Όρια Ελέγχου για το  $MR$ -διάγραμμα

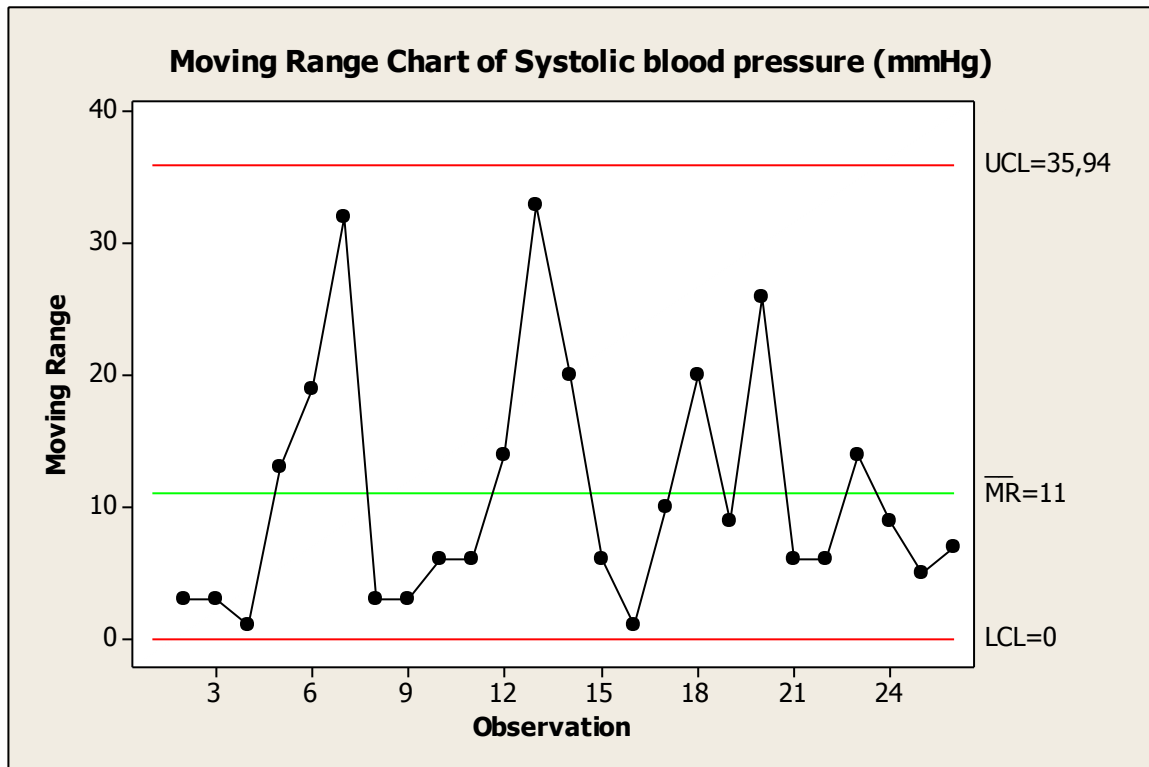
Θεωρούμε τα δεδομένα του Πίνακα 2.14 που αφορούν τη συστολική πίεση ενός ασθενούς κάθε πρωί για 26 συνεχόμενες ημέρες [βλ. Mohamed et al. (2008)]. Στα Διαγράμματα 2.7 και 2.8 δίνονται το  $x$ -διάγραμμα και το  $MR$ -διάγραμμα αντίστοιχα. Παρατηρούμε ότι στην έκτη μέτρηση το  $x$ -διάγραμμα δίνει ένδειξη εκτός ελέγχου διεργασίας, κάτι που δεν συμβαίνει στο  $MR$ -διάγραμμα.

Ημέρα	Πίεση σε mmHg	MR	Ημέρα	Πίεση σε mmHg	MR
1	169	-	14	181	20
2	172	3	15	175	6
3	175	3	16	176	1
4	174	1	17	186	10
5	161	13	18	166	20
6	142	19	19	157	9
7	174	32	20	183	26
8	171	3	21	177	6
9	168	3	22	171	6
10	174	6	23	185	14
11	180	6	24	176	9
12	194	14	25	181	5
13	161	33	26	174	7

Πίνακας 2.14 Συστολική πίεση ασθενούς (σε mmHg)



Διάγραμμα 2.7 x-διάγραμμα για τη συστολική πίεση ασθενούς



Διάγραμμα 2.8 MR- διάγραμμα για τη συστολική πίεση ασθενούς

### 2.3 Διαγράμματα ελέγχου με μνήμη

Τα διαγράμματα ελέγχου τύπου *Shewhart* που είδαμε στην προηγούμενη ενότητα χρησιμοποιούν πληροφορίες μόνο από το πιο πρόσφατο δείγμα. Συνεπώς αφήνουν αναξιοποίητη την πληροφορία που μας έχουν δώσει προηγούμενα δείγματα. Για αυτό και αποκαλούνται διαγράμματα ελέγχου χωρίς μνήμη. Σε αυτή την ενότητα θα ασχοληθούμε με διαγράμματα ελέγχου με μνήμη και θα δούμε την ανάπτυξή τους μόνο για τη Φάση II. Όπως και στα διαγράμματα τύπου *Shewhart*, έτσι και εδώ, από μία διαδικασία βιοεπιτήρησης λαμβάνουμε τυχαία δείγματα, ανεξάρτητα μεταξύ τους, και υπολογίζουμε την τιμή μιας στατιστικής συνάρτησης (μέσος, τυπική απόκλιση, ποσοστό κλπ. ) για το κάθε δείγμα. Έστω, λοιπόν, ότι παρακολουθούμε ένα χαρακτηριστικό που εκφράζεται από την τυχαία μεταβλητή  $X_{ij}$  και για το  $i$  δείγμα υπολογίζουμε την τιμή της στατιστικής συνάρτησης  $g(X_i)$ . Η διαφορά έγκειται στο ότι τώρα δεν σημειώνουμε στο διάγραμμα την τιμή της συνάρτησης αλλά εργαζόμαστε ως εξής:



Το σημείο που θα θέσουμε στο διάγραμμα έχει τεταγμένη  $Y_t$ ,  $t = 1, 2, \dots$  που δίνεται από τις σχέσεις:

$$Y_t = a_t + \sum_{i=1}^t b_i^t g(X_i)$$

και

$$Y_0 = a_0$$

Αξιίζει να σημειωθεί ότι για  $a_t = 0$ ,  $b_1^t = b_2^t = \dots = b_{t-1}^t = 0$  και  $b_t^t = 1$  τότε έχουμε διάγραμμα τύπου Shewhart.

### 2.3.1 Διαγράμματα Ελέγχου Κινούμενου Μέσου

Το διάγραμμα ελέγχου κινούμενου μέσου (*Moving Average control chart*) ή απλώς *MA* διάγραμμα είναι πιο αποτελεσματικό από τα κλασσικά διαγράμματα στον εντοπισμό μικρών αλλαγών του μέσου. Η στατιστική συνάρτηση που χρησιμοποιούμε είναι η :

$$Y_t = \frac{1}{k} \sum_{i=t-k+1}^t g(X_i), \quad t \geq 1$$

Συνήθως θέτουμε  $g(X_i) = \bar{X}_i$ .

Σε αυτή την περίπτωση είναι:

$$E(Y_t) = \mu_0 \text{ και } \sigma_{Y_t}^2 = \frac{\sigma^2}{nk}$$

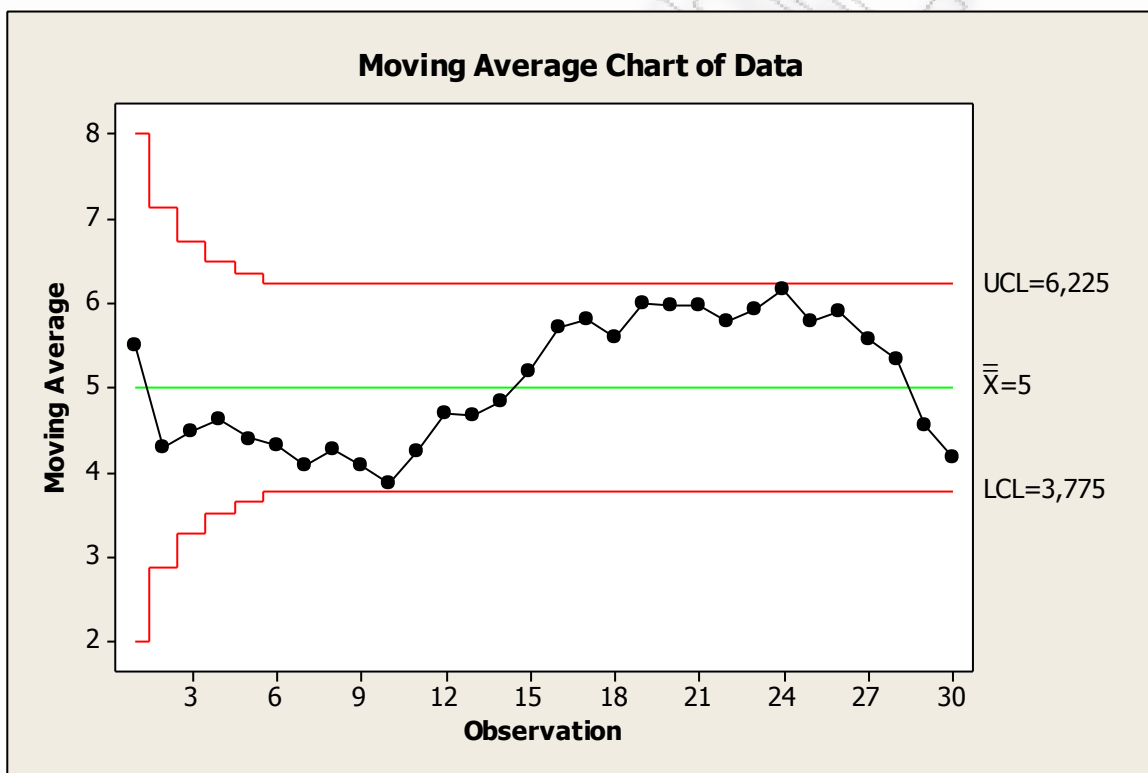
όπου  $\mu_0$  και  $\sigma^2$  είναι η εντός ελέγχου μέση τιμή και διακύμανση και  $n \geq 1$  το μέγεθος του δείγματος. Για  $n=1$  θέτουμε  $g(X_i) = X_i$ .

Έτσι η κεντρική γραμμή και τα όρια ελέγχου  $L$  τυπικών αποκλίσεων είναι αυτά που δίνονται στον Πίνακα 2.15.

Lσ Όρια Ελέγχου MA –διάγραμμα		
$LCL = \mu_0 - L \frac{\sigma}{\sqrt{nk}}$	$CL = \mu_0$	$UCL = \mu_0 + L \frac{\sigma}{\sqrt{nk}}$

Πίνακας 2.15 Όρια Ελέγχου για το MA –διάγραμμα.

Στο ακόλουθο διάγραμμα δίνεται ένα MA διάγραμμα με 3σ όρια ελέγχου για μεμονωμένες παρατηρήσεις από την  $N(5,1)$  για  $k = 6$ .

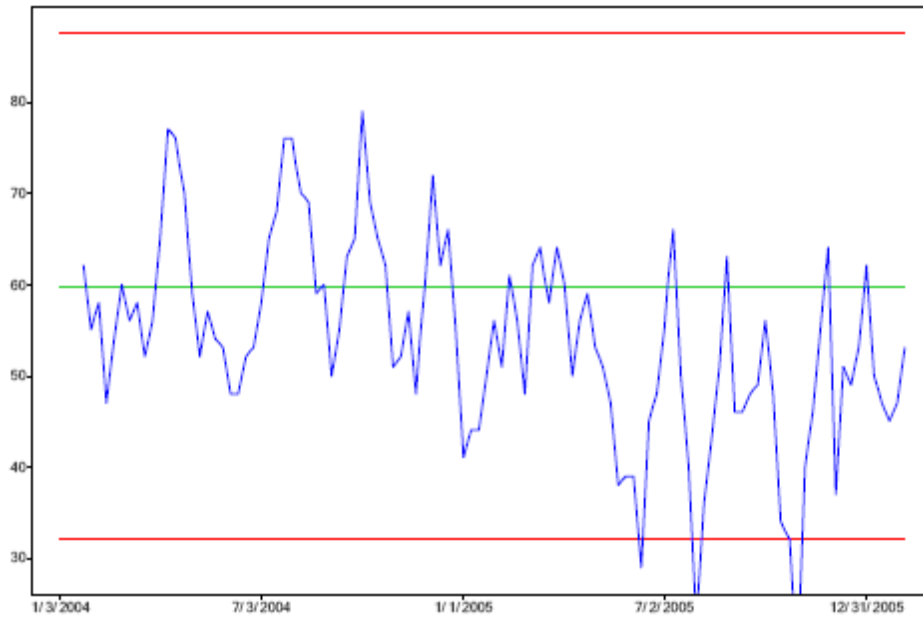


Διάγραμμα 2.9 MA διάγραμμα με  $k = 6$

Αξίζει να σημειωθεί ότι το MA διάγραμμα έχει περιορισμένη μνήμη αφού αντλεί πληροφορία από τις τελευταίες  $k$  παρατηρήσεις ή δείγματα και ομοιόμορφη μνήμη καθώς σε κάθε παρατήρηση δίνεται το ίδιο βάρος  $1/k$ .

Στο Διάγραμμα 2.10, δίνεται ένα MA διάγραμμα για κρούσματα γονόρροιας στη Μασαχουσέτη, την περίοδο 1/1/2004 – 31/12/2005, με  $k = 4$  εβδομάδες.

## Gonorrhoea



**Διάγραμμα 2.10** MA διάγραμμα για κρούσματα γονόρροιας [Πηγή: Shmueli (2006)]

### 2.3.2 Διάγραμμα Ελέγχου Εκθετικά Σταθμισμένου Κινούμενου Μέσου

Το διάγραμμα ελέγχου εκθετικά κινούμενου μέσου (*Exponential Weighted Moving Average control chart*) ή απλώς *EWMA* διάγραμμα είναι πιο αποτελεσματικό από το *MA* διάγραμμα. Η στατιστική συνάρτηση που χρησιμοποιούμε είναι η :

$$Z_t = (1 - \lambda)Z_{t-1} + \lambda g(X_t)$$

Για  $g(X_t) = X_t$  είναι:

$$Z_t = (1 - \lambda)^t Z_0 + \lambda \sum_{i=1}^t (1 - \lambda)^{t-i} X_i$$

με

$$\mu_{Z_t} = \mu_0$$

και

$$\sigma_{Z_t}^2 = \sigma \sqrt{\frac{\lambda}{2-\lambda} [1 - (1 - \lambda)^{2t}]}$$

Επομένως η κεντρική γραμμή και τα όρια του EWMA διαγράμματος ελέγχου είναι:

Lσ Όρια Ελέγχου EWMA –διάγραμμα		
$LCL = \mu_0 - L\sigma \sqrt{\frac{\lambda}{2-\lambda} [1 - (1-\lambda)^{2t}]}$	$CL = \mu_0$	$UCL = \mu_0 + L\sigma \sqrt{\frac{\lambda}{2-\lambda} [1 - (1-\lambda)^{2t}]}$

**Πίνακας 2.16** Όρια Ελέγχου για το EWMA διάγραμμα

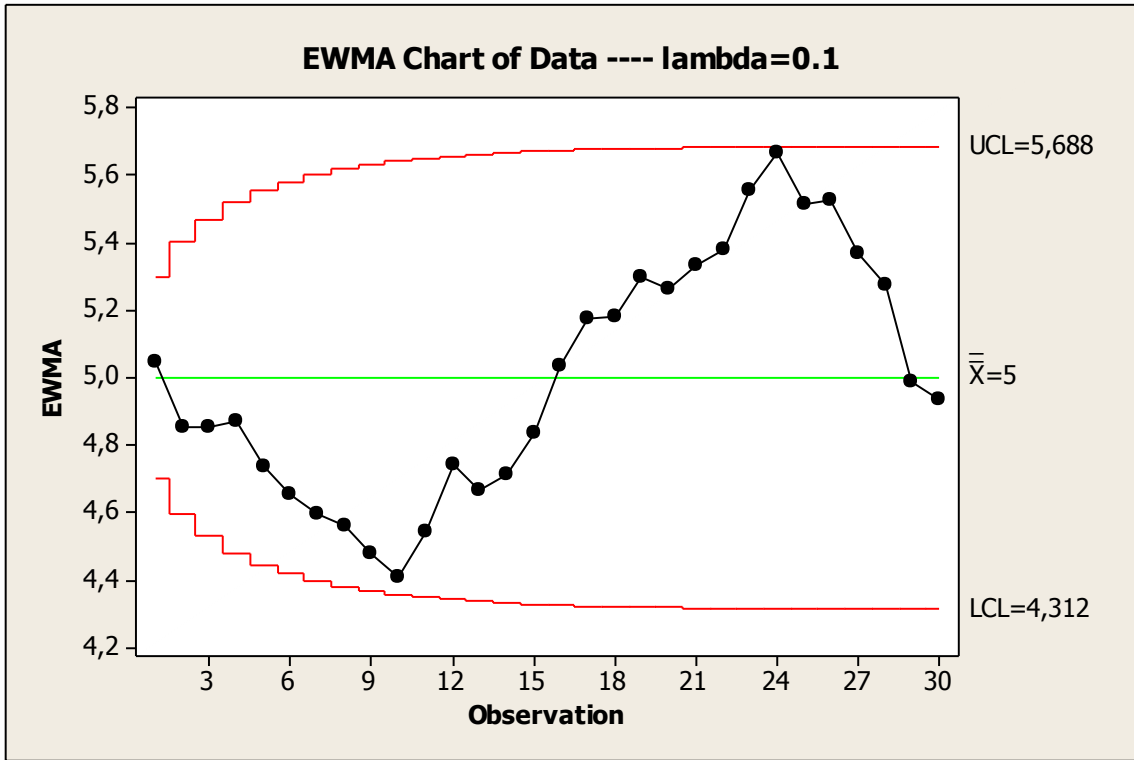
Όπως και στο διάγραμμα κινούμενου μέσου τα όρια του EWMA είναι μεταβλητά και με την πάροδο του χρόνου η ποσότητα  $(1-\lambda)^{2t}$  τείνει στο μηδέν. Τότε τα όρια δίνονται από τις σχέσεις:

$$UCL = \mu_0 + L\sigma \sqrt{\frac{\lambda}{2-\lambda}}$$

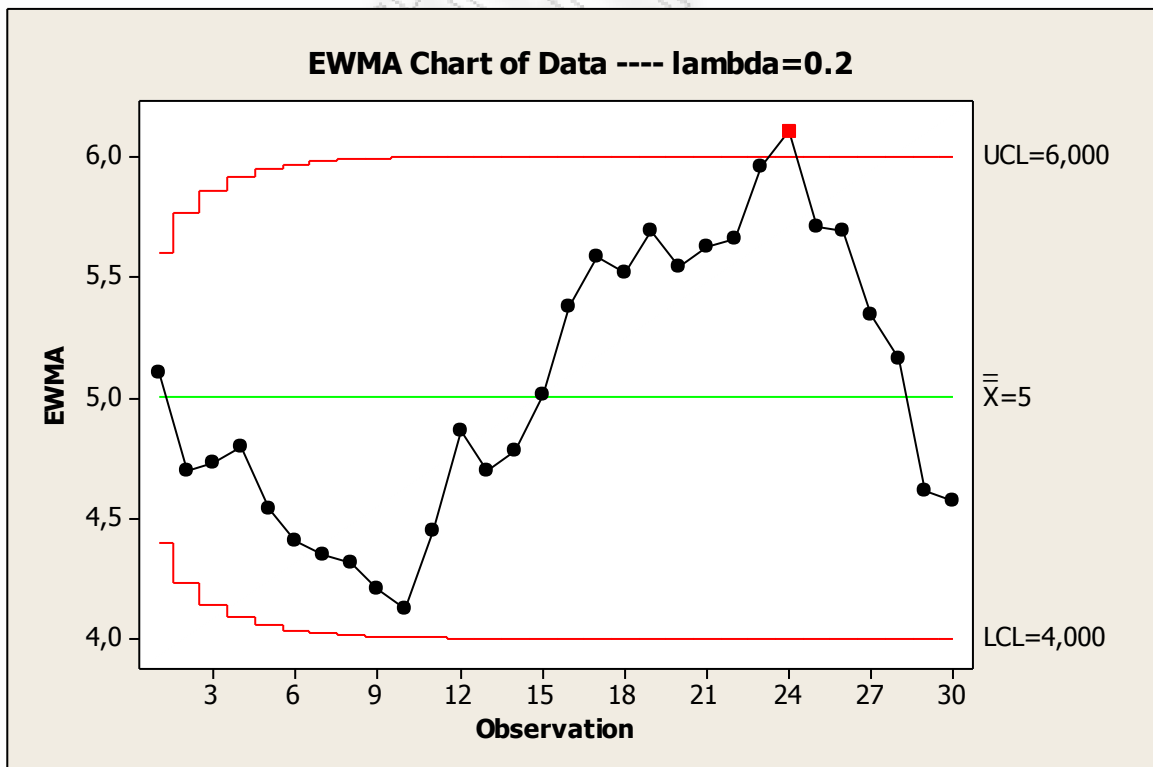
και

$$LCL = \mu_0 - L\sigma \sqrt{\frac{\lambda}{2-\lambda}}$$

Στα παρακάτω διαγράμματα 2.11 και 2.12 έχουν αποτυπωθεί, για τα ίδια δεδομένα από την  $N(5,1)$  της παραγράφου 2.3.1, EWMA διαγράμματα ελέγχου με  $\lambda = 0.1$  και  $\lambda = 0.2$  αντίστοιχα. Παρατηρούμε ότι στο δεύτερο διάγραμμα το 24<sup>ο</sup> σημείο βρίσκεται πάνω από το άνω όριο ελέγχου. Συνεπώς, αν αυτά τα δεδομένα, αφορούσαν μια πραγματική διεργασία βιοεπιτήρησης, θα έπρεπε να γίνει έλεγχος ώστε να διερευνηθεί τυχόν ύπαρξη ειδικής αιτίας μεταβλητότητας.



Διάγραμμα 2.11 EWMA διάγραμμα με  $\lambda = 0.1$



Διάγραμμα 2.12 EWMA διάγραμμα με  $\lambda = 0.2$

Τα *EWMA* διαγράμματα ελέγχου είναι πιο αποτελεσματικά από τα διαγράμματα τύπου *Shewhart* για μικρές μετατοπίσεις στο μέσο της διεργασίας, αλλά σε μεγάλες μετατοπίσεις τα *Shewhart* είναι καλύτερα. Το αντίστροφο ισχύει για τα *EWMA* και τα *CUSUM* που θα δούμε στην επόμενη παράγραφο. Επιπλέον είναι πιο αποτελεσματικά από τα *MA* διαγράμματα. Τέλος από τον τρόπο υπολογισμού των ορίων ελέγχου είναι προφανές ότι εκμεταλλεύονται τις πληροφορίες από όλα τα προηγούμενα δείγματα και το καθένα δείγμα σταθμίζεται διαφορετικά. Λέμε λοιπόν ότι τα *EWMA* διαγράμματα έχουν μη περιορισμένη και μη ομοιόμορφη μνήμη.

### 2.3.3 Το αθροιστικό διάγραμμα

Προκειμένου να εφαρμόσουμε μια μέθοδο βιοεπιτήρησης πρέπει να κάνουμε κάποιες υποθέσεις για την διαδικασία ανάλογα με τη φύση του προβλήματος και με τα δεδομένα που συλλέχθηκαν. Η διαδικασία Poisson χρησιμοποιείται συχνά για το πλήθος των περιστατικών της ασθένειας, ενώ, αν αυτή δεν προσαρμόζεται καλά στα δεδομένα, χρησιμοποιούνται πιο σύνθετες διαδικασίες με ανεξάρτητη μεταβλητή το χρόνο.

Υποθέτοντας την διαδικασία Poisson για το πλήθος των περιστατικών μιας ασθένειας ένα αυξημένο ποσοστό συμβάντος αντιστοιχεί σε αυξημένη ένταση (*intensity*) της διαδικασίας. Η πιθανότητα ανίχνευσης τέτοιας έντασης εξαρτάται από τον τρόπο που εξετάζουμε τη διαδικασία και από την μέθοδο βιοεπιτήρησης που χρησιμοποιούμε.

Σε κάποιες περιπτώσεις εστιάζουμε στα διαστήματα ανάμεσα στα συμβάντα. Αυτά μετρώνται είτε με το χρόνο ως συνεχή μεταβλητή είτε με μια διακριτή κλίμακα χρόνου που μετράει το πλήθος των μη σημαντικών συμβάντων μεταξύ σημαντικών συμβάντων. Και στις δύο περιπτώσεις δεν υπάρχει απώλεια πληροφορίας για τη διαδικασία. Αύξηση στην ένταση της διαδικασίας θα ανιχνεύεται από την μείωση του χρονικού διαστήματος ανάμεσα σε σημαντικά περιστατικά ή αντίστοιχα από τη μείωση του πλήθους των μη σημαντικών περιστατικών ανάμεσα σε σημαντικά.

Όταν υποθέτουμε ότι οι χρόνοι ανάμεσα σε σημαντικά συμβάντα είναι συνεχείς και ακολουθούν την εκθετική κατανομή χρησιμοποιούμε μεθόδους όπως το αθροιστικό διάγραμμα (*Cumulative Sum chart, CUSUM*) και τα διαγράμματα εκθετικά σταθμισμένου κινητού μέσου (*Exponentially Weighted Moving Average, EWMA*). Για την μέθοδο *CUSUM* η στατιστική συνάρτηση συναγερού βασίζεται στο *cumulative sum* των διαφορών ανάμεσα στις παρατηρούμενες και στις αναμενόμενες τιμές. Έστω  $L(s, t)$  ο υπό συνθήκη λόγος πιθανοφάνειας την χρονική στιγμή  $s$  για την περίπτωση που  $\tau = t$ . Όταν ένα από τους λόγους

αυτούς θα είναι σημαντικά μεγάλος τότε έχουμε ένδειξη ότι η διαδικασία είναι εκτός ελέγχου. Έτσι, για μία σταθερά  $K$ , ο χρόνος που θα έχουμε τον πρώτο συναγερμό θα δίνεται από την σχέση:

$$t_A = \min [s: \max_t \{L(s, t) > K\}] \quad (2.9)$$

Η μέθοδος CUSUM είναι βέλτιστο κριτήριο που επιλέγει το ελάχιστο από μια σειρά από μέγιστες τιμές (*minimax optimum*). Δηλαδή ελαχιστοποιεί την μέγιστη τιμή της υπό συνθήκη αναμενόμενης καθυστέρησης όταν το αποτέλεσμα των παρατηρήσεων, πριν συμβεί η αλλαγή, είναι το χειρότερο δυνατό [βλ. Moustakides (1986)].

Αν η καταγραφή του χρόνου που συμβαίνει ένα περιστατικό γίνεται στρωματοποιημένα, ήτοι για προκαθορισμένα χρονικά διαστήματα, τότε θα υπάρχει απώλεια πληροφορίας και η μέθοδος παρακολούθησης που εφαρμόστηκε θα υστερεί σε ταχύτητα εντοπισμού της αλλαγής. Έτσι, η στρωματοποίηση του χρόνου θα πρέπει να αποφεύγεται, εκτός από τις περιπτώσεις που αυτό επιβάλλεται, από το υπό παρακολούθηση σύστημα. Σε περιπτώσεις που υπάρχει τέτοια αδυναμία η μέθοδος που χρησιμοποιείται είναι η μέθοδος CUSUM για Poisson δεδομένα (*Poisson CUSUM method*). Η μέθοδος αυτή συγκρίνει τις παρατηρούμενες συχνότητες για κάθε χρονική περίοδο με τις αναμενόμενες και χρησιμοποιεί το *cumulated sum* των αποκλίσεων (*deviations*) για να σχηματιστεί η συνάρτηση συναγερμού. Επίσης, στην περίπτωση που ο χρόνος είναι στρωματοποιημένος, για την συνάρτηση συναγερμού, μπορεί να χρησιμοποιηθεί και το μέγιστο των λόγων πιθανοφάνειας.

Τα διαγράμματα ελέγχου τύπου CUSUM δίνουν συνήθως στενότερα όρια ελέγχου σε σχέση με τα κλασσικά διαγράμματα, όπως τα διαγράμματα *Shewhart*. Λόγω του αυστηρότερου ελέγχου που γίνεται με τη μέθοδο CUSUM, θα πρέπει να δοθεί περισσότερη έμφαση στο να παραμένει η διεργασία γύρω από μία μέση τιμή και να μην κινείται κοντά στα όρια. Επιπλέον μία παρατήρηση εκτός των ορίων ελέγχου της διεργασίας θα αποτελεί ένδειξη ότι έχει ήδη συμβεί μια αλλαγή αντί να υποδηλώνει ότι πρέπει να ληφθούν μέτρα για να μην συμβεί η αλλαγή. Καθώς η μέθοδος CUSUM δίνει εγκαίρως ενδείξεις ότι η διεργασία είναι εκτός ελέγχου, είναι συνεπής με τις οδηγίες των Οργανισμών Υγείας που απαιτούν ταχύτητα στον εντοπισμό μιας αλλαγής. Ένα άλλο πλεονέκτημα είναι ότι η μέθοδος αυτή μπορεί να δηλώσει περιοχές που απαιτούν προσοχή.

Η μέθοδος CUSUM, χρησιμοποιείται περισσότερο για τον έλεγχο των μεταβλητών παρά για τον έλεγχο των συχνοτήτων.

Η μέθοδος ελέγχου CUSUM υπολογίζει την διαφορά ανάμεσα σε μια παρατηρούμενη τιμή  $Y_i$  και μια τιμή αναφοράς  $k$ . Αν η διαφορά είναι μεγαλύτερη ή ίση από μια τιμή ελέγχου  $h$

τότε έχουμε ένδειξη ότι η διαδικασία είναι εκτός ελέγχου. Για την ανίχνευση αύξησης του πλήθους των περιστατικών, με την μέθοδο *Poisson CUSUM*, η συνάρτηση ελέγχου είναι:

$$S_i = \max(0, Y_i - k + S_{i-1}).$$

Ένα κανονικό *CUSUM* διάγραμμα θα έχει αρχική τιμή  $S_0 = 0$ , ενώ μπορεί να λάβει και θετική αρχική τιμή. Συνιστάται η αρχική τιμή να είναι ίση προς το μισό της τιμής ελέγχου  $S_0 = h/2$ . Με αυτό τον τρόπο αν η διεργασία είναι εκτός ελέγχου από τη στιγμή που ξεκίνησε, το *CUSUM* θα δώσει ταχύτερα ένδειξη εκτός ελέγχου διεργασίας. Για μια διαδικασία που βρίσκεται στο επιθυμητό επίπεδο η τιμή  $S$  σύντομα θα μηδενιστεί επομένως έχει μικρή επίδραση, [βλ. *Lucas (1985)*]. Η επιλογή της αρχικής τιμής θα πρέπει να γίνεται αφού ληφθούν υπόψη τα σφάλματα  $\alpha$  και  $\beta$ . Ως σφάλμα  $\alpha$  ορίζουμε την πιθανότητα να δεχτούμε ότι η διεργασία είναι εκτός ελέγχου ενώ στην πραγματικότητα είναι εντός ελέγχου. Ως σφάλμα  $\beta$  ορίζουμε την πιθανότητα να δεχτούμε ότι η διεργασία είναι εντός ελέγχου ενώ στην πραγματικότητα δεν είναι.

Ένα ανθεκτικό *CUSUM* διάγραμμα (*robust CUSUM*) μπορεί να κατασκευαστεί αν υιοθετήσουμε τον κανόνα: δύο στην σειρά (*two-in-a-row rule*). Στην περίπτωση αυτή απαιτείται ένα όριο έκτροπων παρατηρήσεων (*outlier limit*). Μία μοναδική παρατήρηση εκτός του ορίου δεν λαμβάνεται υπ' όψιν. Αντιθέτως δύο συνεχόμενες παρατηρήσεις εκτός του ορίου αυτού αποτελούν ένδειξη ότι υπάρχει έξαρση της υπό μελέτη νόσου.

Τα διαγράμματα *CUSUM* αξιολογούνται υπολογίζοντας το εντός ελέγχου μέσο μήκος ροής  $ARL^0$ . Όπως ειπώθηκε και στην προηγούμενη παράγραφο, το  $ARL^0$  είναι το μέσο μέγεθος των δειγμάτων που ελήφθησαν ανάμεσα σε δύο συναγερμούς. Θα πρέπει να λαμβάνει μεγάλες τιμές όταν η εξάπλωση της νόσου στο δείγμα είναι στο επιθυμητό επίπεδο και μικρές τιμές όταν επέλθει μία αλλαγή. Η κατανομή του  $ARL^0$  προσεγγίζεται από την γεωμετρική, αφού το  $ARL^0$  χαρακτηρίζει την κατανομή του μήκους ροών [ βλ. *Brook and Evans (1972)*]. Για αρχική τιμή  $S_0 = 0$  η κατανομή του  $ARL^0$  είναι σχεδόν γεωμετρική, με την διαφορά ότι η πιθανότητα να έχουμε εξαιρετικά μικρό μήκος ροής είναι μικρή. Όταν χρησιμοποιούμε θετική αρχική τιμή η κατανομή είναι σχεδόν γεωμετρική με την διαφορά ότι αυξάνεται η πιθανότητα να έχουμε μικρό μήκος ροής [βλ. *Lucas and Crosier (1982a)*].

Στον παρακάτω πίνακα δίνεται η εφαρμογή ενός *CUSUM* για δεδομένα Poisson με παραμέτρους  $k = 5$  και  $h = 10$  χωρίς αρχική τιμή και για αρχική τιμή  $S_0 = 5$ . Όπως προκύπτει η μέθοδος προσαρμόζεται καλύτερα αν το επιθυμητό επίπεδο για την διεργασία



είναι το 4 και αν επιθυμούσαμε να ανιχνεύεται γρήγορα μια αύξηση του πλήθους των περιστατικών στο 7 και παραπάνω. Σε αυτή την περίπτωση το  $ARL^0$  είναι περίπου 422 ή 397 και το  $ARL^1$  περίπου 5.59 ή 3.35 χωρίς αρχική τιμή και για αρχική τιμή ίση με 5 αντίστοιχα. Οι πρώτες 10 παρατηρήσεις προέρχονται από μία διαδικασία που βρίσκεται στα επιθυμητά επίπεδα και οι υπόλοιπες 5 από μια διαδικασία με υψηλότερο μέσο. Η διαδικασία ξεκινάει αρχικά ως εντός ελέγχου οπότε η αρχική τιμή έχει μικρή επίδραση και τελικά μηδενίζει. Το πλεονέκτημα της αρχικής τιμής φαίνεται αν χρησιμοποιήσουμε μόνο τις τελευταίες 5 παρατηρήσεις. Αυτό σημαίνει ότι η διαδικασία ξεκινάει ευρισκόμενη εκτός ελέγχου. Σε αυτή την περίπτωση το CUSUM με αρχική τιμή δίνει ένδειξη από την πρώτη παρατήρηση. Αντιθέτως χωρίς αρχική τιμή η ένδειξη λαμβάνεται στην 4<sup>η</sup> παρατήρηση.

<i>Poisson CUSUM</i> ( $h = 10, k = 5$ )				
			$S_i$	
$i$	$Y_i$	$Y_i - k$	$S_0 = 0$	$S_0 = 5$
0	-	-	0	5
1	3	-2	0	3
2	7	2	2	4
3	2	-3	0	2
4	0	-5	0	0
5	2	-3	0	0
6	8	3	3	3
7	4	-1	2	2
8	0	-5	0	0
9	2	-3	0	0
10	3	-2	0	0
11	10	5	5	5
12	8	3	8	8
13	4	-1	7	7
14	9	4	11	11
15	11	6	17	17

**Πίνακας 2.17** *Poisson CUSUM* με παραμέτρους  $h = 10, k = 5$ . [Πηγή: Lucas (1985)]

Ο ρυθμός εκδήλωσης μιας νόσου μπορεί είτε να αυξάνεται είτε να μειώνεται. Όταν επιθυμούμε να ανιχνεύουμε με την ίδια μέθοδο την αύξηση ή την μείωση μιλάμε για δίπλευρο έλεγχο. Αυτό επιτυγχάνεται με την ταυτόχρονη εφαρμογή δύο μονόπλευρων τεχνικών [βλ. *Lucas (1985)*]. Από τον σχεδιασμό προκύπτει πώς θα επιλέξουμε την τιμή αναφοράς  $k$  και την τιμή του διαστήματος απόφασης  $h$ . Η τιμή της παραμέτρου  $k$  καθορίζεται από τον αποδεκτό ρυθμό περιστατικών. Αφού υπολογιστεί η  $k$  χρησιμοποιούμε πίνακες για να επιλέξουμε την παράμετρο  $h$ . Η τελευταία υπολογίζεται από το επιθυμητό εντός ελέγχου μέσο μήκος ροής δηλαδή από την αποδεκτή συχνότητα εσφαλμένων συναγερμών. Συχνά αποφεύγουμε να θεωρούμε ότι η διεργασία βρίσκεται εκτός ελέγχου στο ξεκίνημα της παρακολούθησης. Η ανθεκτικότητα των παραμέτρων υπολογίζεται μετά την ολοκλήρωση του σχεδιασμού.

Για την *Poisson CUSUM*, η τιμή της  $k$  επιλέγεται ανάμεσα στην αποδεκτή τιμή του μέσου της διεργασίας  $\mu_a$  και στην μέση τιμή του επιπέδου των περιστατικών  $\mu_d$  που θέλουμε να ανιχνεύουμε γρήγορα. Οι  $\mu_a$  και  $\mu_d$  είναι οι μέσες τιμές των περιστατικών ανά κλάση της δειγματοληψίας. Παρόλο που η επιθυμητή τιμή (ή τιμή στόχος) για την  $\mu_a$  είναι συχνά το μηδέν, δεν χρησιμοποιείται στις διάφορες μεθόδους σχεδιασμού. Αν το  $\mu_a$  είναι μηδέν τότε το *CUSUM* σχεδιάζεται για  $h = 1$  και  $k = 0$ . Όταν το πλήθος των περιστατικών είναι χαμηλό και μία περεταίρω μείωση του υποδηλώνει βελτίωση της διαδικασίας τότε κάθε περιστατικό θα πρέπει να εξετάζεται προσεκτικά. Οι συνθήκες γύρω από κάθε περιστατικό θα πρέπει να ελέγχονται ώστε να εντοπίζονται οι αιτίες που το προκάλεσαν και να απομακρύνονται. Στην πράξη το  $\mu_a$  επιλέγεται κοντά στον μέσο που παρατηρούμε εκείνη την στιγμή. Η τιμή αναφοράς  $k_p$  της *Poisson CUSUM* πρέπει να επιλέγεται κοντά στην τιμή:

$$k_p = (\mu_d - \mu_a) / (\ln(\mu_d) - \ln(\mu_a))$$

Η τιμή αναφοράς που υπολογίζεται από την παραπάνω σχέση είναι ίδια με την τιμή αναφοράς ενός ακολουθιακού ελέγχου λόγου πιθανοφάνειας (*Sequential Probability Ratio Test, SPRT*) ελέγχου της μηδενικής υπόθεσης ότι ο μέσος είναι  $\mu_a$  έναντι της εναλλακτικής ότι ο μέσος είναι  $\mu_d$ . Όταν το  $k_p \geq 1$  τότε στρογγυλοποιούμε στο πλησιέστερο ακέραιο αριθμό. Έτσι όλοι οι υπολογισμοί για την μέθοδο *CUSUM* γίνονται με ακέραιους αριθμούς.

Με την μέθοδο *CUSUM*, όταν ερευνούμε τον χρόνο που μεσολαβεί ανάμεσα σε δύο διαδοχικά περιστατικά, ο μέσος χρόνος ανάμεσα στα περιστατικά είναι αντιστρόφως ανάλογος του πλήθους των περιστατικών σε κάθε κλάση της δειγματοληψίας. Συνεπώς τα  $\mu_a$

και  $\mu_d$  αφορούν το ρυθμό των περιστατικών στο time-between-events CUSUM. Για παράδειγμα, αν  $\mu = 2$ , δηλαδή εκδηλώνονται κατά μέσο όρο 2 περιστατικά στην μονάδα του χρόνου, ο μέσος χρόνος ανάμεσα σε διαδοχικά περιστατικά είναι  $1/\mu = 0.5$ . Η τιμή αναφοράς  $k_b$  στην περίπτωση αυτή είναι:

$$k_b = (\ln(\mu_d) - \ln(\mu_a)) / (\mu_d - \mu_a).$$

Αν ο χρόνος μετράται στην ίδια κλίμακα στο *time-between-events CUSUM* και στο *Poisson CUSUM* τότε η τιμή αναφοράς  $k_b$  του πρώτου είναι ο αντίστροφος της τιμής αναφοράς  $k_p$  του δεύτερου. Αξίζει να σημειωθεί ότι η τιμή αναφοράς και στις δύο περιπτώσεις είναι κοντά στο επίπεδο του μέσου και με την μικρότερη διακύμανση.

## 2.4 Διαγράμματα Προσαρμοσμένα στον Κίνδυνο

Τα διαγράμματα στατιστικού ελέγχου διεργασίας που αναπτύξαμε μέχρι τώρα δεν λαμβάνουν υπ' όψιν τη δριμύτητα (*severity*) ή βαρύτητα ή επιθετικότητα μίας ασθένειας. Σε αντίθεση με τον βιομηχανικό έλεγχο ποιότητας μια διαδικασία βιοεπιτήρησης εστιάζεται συχνά σε συμπτώματα ασθενειών σε ανθρώπους ή ζώα που μπορεί να παρουσιάζουν σημαντικές διακυμάνσεις. Με άλλα λόγια η εντός ελέγχου πιθανότητα εμφάνισης της νόσου (ή θανάτου κλπ) μπορεί να διαφέρει από ασθενή σε ασθενή. Επιπλέον στα αρχεία των νοσοκομείων καταγράφεται μια πληθώρα χαρακτηριστικών ανά ασθενή. Είναι λοιπόν χρήσιμο να μπορούν τα διαγράμματα ελέγχου να εμπεριέχουν με κάποιο τρόπο την δριμύτητα της νόσου προκειμένου να ελέγχουν την ποιότητα της δημόσιας υγείας.

Ο όρος δριμύτητα μιας νόσου είναι δύσκολο να προσδιοριστεί με σαφή και αποκλειστικό τρόπο. Ένας από τους πολλούς τρόπους που μπορούμε να αποδώσουμε ποιοτικά και ποσοτικά την έννοια της δριμύτητας είναι η θνησιμότητα. Αυτή ίσως είναι και το απλούστερο γνώρισμα της δριμύτητας. Δυστυχώς ακόμη και για την θνησιμότητα υπάρχουν διάφορες μέθοδοι μοντελοποίησης [βλ. *Hardon et al.*, (1993)]. Γενικά, με βάση τη θνησιμότητα μπορούμε να χωρίσουμε σε ομάδες τους ασθενείς με σοβαρές παθήσεις, όπως έμφραγμα του μυοκαρδίου, που συχνά επιφέρουν θάνατο. Σε αυτή την εργασία θα θεωρήσουμε την θνησιμότητα ως κατηγορική μεταβλητή με δύο επίπεδα, ο ασθενής επιβιώνει και ο ασθενής δεν επιβιώνει και ως μέτρο της θνησιμότητας θα χρησιμοποιήσουμε το ποσοστό θνησιμότητας.

Ως ποσοστό θνησιμότητας θεωρούμε τον λόγο του αριθμού των ασθενών που αποβίωσαν προς το συνολικό αριθμό ασθενών σε ένα προκαθορισμένο χρονικό διάστημα. Προφανώς τα δείγματα που λαμβάνουμε δεν έχουν σταθερό μέγεθος καθώς περιλαμβάνουν το σύνολο των ασθενών στο συγκεκριμένο χρονικό διάστημα. Παρόλο που το μεταβλητό μέγεθος δείγματος δεν είναι προτιμητέο, στην παρούσα θεώρηση είναι αναπόφευκτο.

Έστω λοιπόν ότι σε  $T_i, i = 1, 2, \dots, m$  χρονικές περιόδους λαμβάνουμε  $m$  ανεξάρτητα δείγματα, μεγέθους  $n_i$  το καθένα. Ο αριθμός των θανάτων στο δείγμα  $i$  είναι  $d_i$ . Το ποσοστό θνησιμότητας στο δείγμα  $i$  (παρατηρούμενο δειγματικό ποσοστό) δίνεται από την σχέση:

$$p_i = \frac{d_i}{n_i}.$$

Το ποσοστό θνησιμότητας για το σύνολο των παρατηρήσεων  $P$  εκτιμάται από το δειγματικό μέσο (εκτίμηση πληθυσμιακού ποσοστού):

$$\hat{p} = \bar{p} = \frac{\sum_{i=1}^m p_i}{\sum_{i=1}^m n_i}.$$

Στην Παράγραφο 2.2.3. είδαμε μία προσέγγιση για την περίπτωση που το μέγεθος δείγματος είναι μεταβλητό.

Σε αυτή την παράγραφο θα χρησιμοποιήσουμε μία διαφορετική προσέγγιση. Μία εκτιμήτρια της δειγματικής τυπικής απόκλισης είναι η:

$$S_i = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n_i}}.$$

Συνεπώς θα χρησιμοποιήσουμε μεταβλητά όρια πιθανότητας όπως φαίνονται στις παρακάτω εξισώσεις [βλ. *Alemi et al.* (1996)]:

$$UCL_i = \bar{p} + t_{\alpha/2} S_i$$

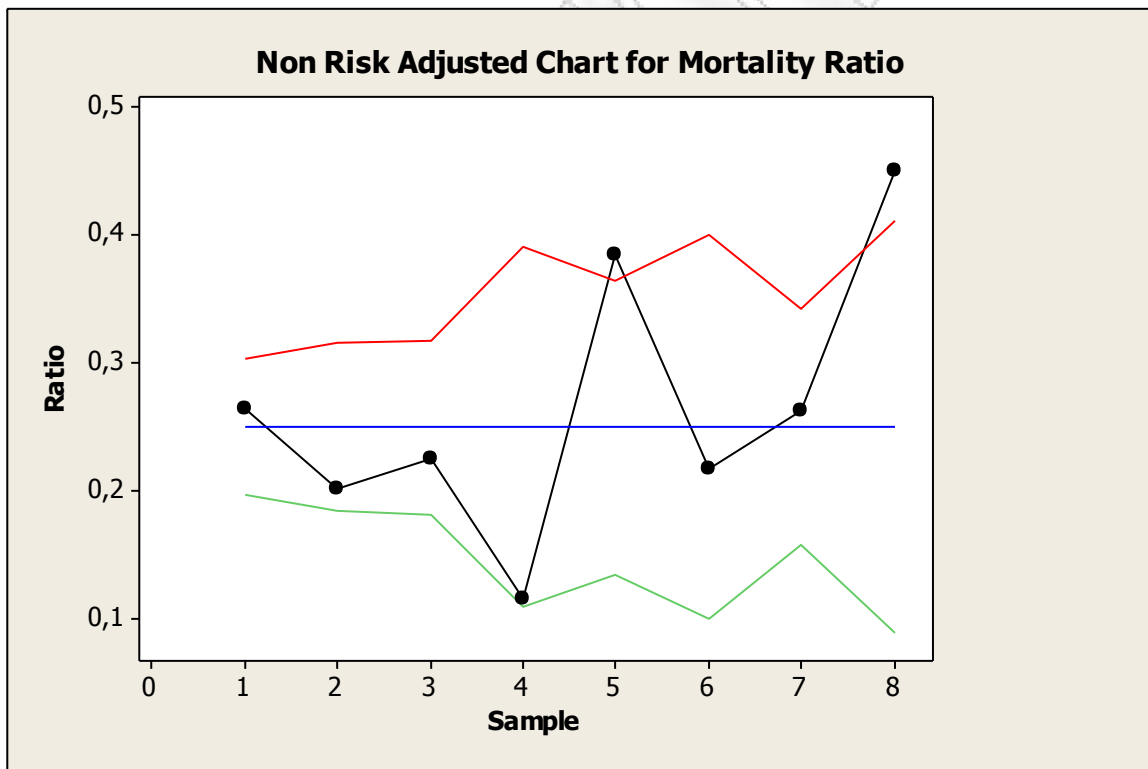
$$LCL_i = \bar{p} - t_{\alpha/2} S_i$$

Επίσης η κεντρική γραμμή είναι  $CL = \bar{p}$ , σταθερή για κάθε  $i$ .

Η ερμηνεία των αποτελεσμάτων όταν έχει γίνει η προσέγγιση της μη συμμετρικής διωνυμικής κατανομής από τη συμμετρική κατανομή θα πρέπει να γίνεται με προσοχή, ειδικότερα αν υπάρχουν μικρές πιθανότητες ή μικρά δείγματα.

Από τον τρόπο υπολογισμού της δειγματικής τυπικής απόκλισης και κατ' επέκταση των ορίων του ελέγχου είναι προφανές ότι όσο αυξάνει το μέγεθος του δείγματος, θα μικραίνει η τυπική απόκλιση και άρα θα στενεύουν τα όρια ελέγχου.

Στο Διάγραμμα 2.13 δίνεται ένα μη προσαρμοσμένο στον κίνδυνο διάγραμμα ελέγχου με μεταβλητά όρια πιθανότητας για παρακολούθηση του ποσοστού θνησιμότητας. [βλ. *Alemi et al.* (1996)]. Τα όρια του ελέγχου έχουν υπολογιστεί με συντελεστή εμπιστοσύνης 90%. Διαπιστώνουμε ότι το παρατηρούμενο ποσοστό θνησιμότητας στο πέμπτο και το όγδοο δείγμα είναι πάνω από το άνω όριο του ελέγχου. Ο αναμενόμενος αριθμός δειγμάτων εκτός των ορίων ελέγχου είναι 0.8 (8 δείγματα,  $\alpha = 0.1$ ). Συνεπώς υπάρχει ένδειξη για εκτός ελέγχου διεργασία και θα πρέπει να αποφασιστεί αν υπάρχει ειδική ή φυσική αιτία μεταβλητότητας.



**Διάγραμμα 2.13** Μη προσαρμοσμένο στον κίνδυνο διάγραμμα ελέγχου με μεταβλητά όρια πιθανότητας ( $\alpha = 0.1$ ) για παρακολούθηση του ποσοστού θνησιμότητας. [Πηγή: *Alemi et al.* (1996)]

Το παραπάνω διάγραμμα κατασκευάστηκε λαμβάνοντας υπ' όψη το παρατηρούμενο ποσοστό θνησιμότητας στο σύνολο των ατόμων. Για να κατασκευάσουμε ένα διάγραμμα προσαρμοσμένο στον κίνδυνο θα θεωρήσουμε όχι το παρατηρούμενο αλλά το αναμενόμενο ποσοστό θνησιμότητας. Αυτό μπορεί να γίνει αναθέτοντας σε κάθε ασθενή μεταβλητές

(scores) που εκφράζουν τη δριμύτητα της νόσου και άλλες επεξηγηματικές μεταβλητές όπως το φύλο, η ηλικία, το βάρος κλπ. Αυτό δεν γίνεται αυθαίρετα αλλά ανατρέχοντας σε σχετική βιβλιογραφία, για παράδειγμα κατηγοριοποιούμε τους ασθενείς με βάση τους δείκτες *PMC* (*Patient Management Categories*) [βλ. *Young* (1985)], *CSI* (*Computerized Severity Index*) [βλ. *Backofen et al.* (1987)], *APACHE II* [βλ. *Knaus et al.* (1985)], κλπ. Στη συνέχεια εφαρμόζουμε μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή το αν ο ασθενής επιβιώνει ή αποβιώνει και ανεξάρτητες μεταβλητές τα παραπάνω scores των ασθενών. Έτσι, με βάση το μοντέλο παλινδρόμησης, εκτιμούμε αν ο κάθε ασθενής επιβιώνει ή αποβιώνει.

Έχοντας εκτιμήσει τον αναμενόμενο αριθμό θανάτων σε κάθε δείγμα (ή χρονική περίοδο) θα υπολογίσουμε το αναμενόμενο ποσοστό θνησιμότητας (και όχι το παρατηρούμενο όπως κάναμε μέχρι τώρα). Έστω η μεταβλητή  $Y_{ij}$  που παίρνει την τιμή 1 αν ο ασθενής  $j$  στο δείγμα  $i$  αποβιώνει και την τιμή 0 αν επιβιώνει. Από το μοντέλο παλινδρόμησης εκτιμούμε τα  $\hat{Y}_{ij}$ . Έστω επίσης  $\hat{P}_{ij}$  η πιθανότητα, υπό το μοντέλο παλινδρόμησης, να αποβιώσει ο  $j$  ασθενής στο  $i$  δείγμα. Το αναμενόμενο ποσοστό θνησιμότητας στο  $i$ -οστό δείγμα θα είναι:

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

Για μεγάλα δείγματα το ποσοστό θνησιμότητας ακολουθεί ασυμπτωτικά κανονική κατανομή. [βλ. *Fisz*, (1963)]

Η εκτίμηση για την τυπική απόκλιση στο δείγμα  $i$  δίνεται από την σχέση:

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

Τα όρια του ελέγχου θα είναι και πάλι μεταβλητά και δίνονται από τις σχέσεις:

$$UCL_i = \hat{P}_i + t_{\alpha/2} \hat{S}_i$$

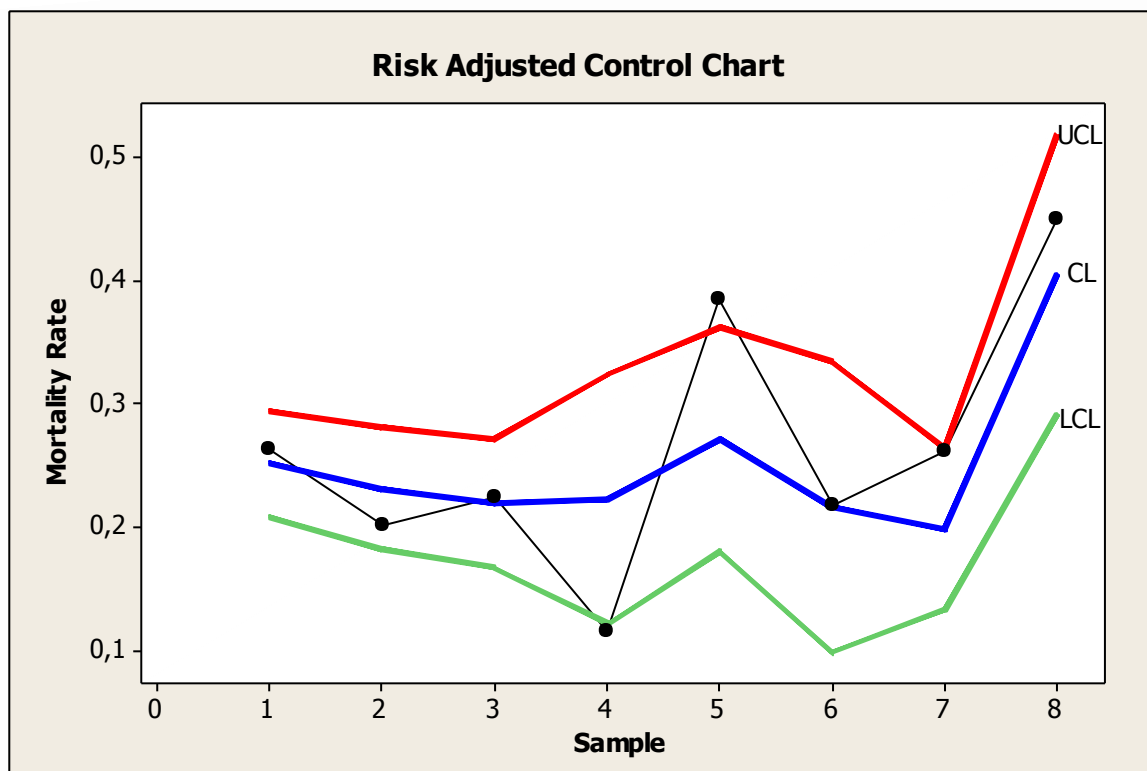
$$LCL_i = \hat{P}_i - t_{\alpha/2} \hat{S}_i$$

Από τις παραπάνω εξισώσεις βλέπουμε ότι τα όρια του ελέγχου έχουν υπολογιστεί με βάση τα αναμενόμενα ποσοστά θνησιμότητας. Άρα μπορούμε να μιλάμε για αναμενόμενα όρια ελέγχου. Μία ακόμα σημαντική διαφοροποίηση είναι η απουσία ευθείας κεντρικής

γραμμής που τώρα έχει αντικατασταθεί από ευθύγραμμα τμήματα που ενώνουν τις αναμενόμενες πιθανότητες. Αφού υπολογίσουμε τα όρια και την κεντρική γραμμή σημειώνουμε στο διάγραμμα τα παρατηρούμενα δειγματικά ποσοστά θνησιμότητας.

Στο Διάγραμμα 2.14 δίνεται ένα παράδειγμα. Βλέπουμε πάλι ότι υπάρχουν 2 σημεία εκτός των ορίων ελέγχου. Το ποσοστό θνησιμότητας στο πέμπτο δείγμα εξακολουθεί να είναι πάνω από το άνω όριο ελέγχου. Αυτή τη φορά όμως το τέταρτο δείγμα είναι κάτω από το κάτω όριο. Αυτό σημαίνει ότι η διεργασία εκείνο το χρονικό διάστημα ήταν σε καλύτερο και από το αναμενόμενο επίπεδο. Θα πρέπει να γίνει έλεγχος ώστε αν ανακαλυφθεί ότι κάποιο ειδικό αίτιο λειτούργησε να καθιερωθεί. Αντίθετα με το προηγούμενο διάγραμμα, το ποσοστό θνησιμότητας στο όγδοο δείγμα δεν είναι πια εκτός ελέγχου. Ενδεχομένως αυτό να οφείλεται στο ότι κάποιος ή κάποιοι ασθενείς να εμφάνιζαν την νόσο με μεγάλη δριμύτητα. Αν είχαμε αρκεστεί στο μη προσαρμοσμένο διάγραμμα θα είχε εφαρμοστεί σχέδιο αναζήτησης ειδικών αιτιών μεταβλητότητας χωρίς λόγο.

Θα πρέπει να σημειωθεί και πάλι ότι ο έλεγχος έγινε υπό την υπόθεση της κανονικότητας. Για να ευσταθεί η υπόθεση αυτή, θα πρέπει τα δείγματα να έχουν μέγεθος τουλάχιστον τριάντα παρατηρήσεις. Στο παράδειγμα αυτό, έχουμε επιλέξει επίπεδο σημαντικότητας έχουμε 10%, εναλλακτικά μπορεί να χρησιμοποιηθεί επίπεδο 5% ή 1% ανάλογα με την περίπτωση. Επίσης ως παράδειγμα χρησιμοποιήθηκε το  $p$ -διάγραμμα. Ομοίως τα διαγράμματα  $\bar{x}$  και  $c$  μπορούν να προσαρμοστούν ώστε να συμπεριλαμβάνουν την δριμύτητα της υποκείμενης νόσου. Η επιλογή του μέτρο της δριμύτητας εξαρτάται από το είδος της νόσου. Έτσι αντί της θνησιμότητας μπορεί να επιλεγεί ο χρόνος θεραπείας. Σε μία τέτοια περίπτωση υπολογίζουμε το ποσοστό των ασθενών που ανάρρωσαν σε μία προκαθορισμένη χρονική περίοδο.



**Διάγραμμα 2.14** Προσαρμοσμένο στον κίνδυνο διάγραμμα ελέγχου με μεταβλητά όρια πιθανότητας ( $\alpha = 0.1$ ) για παρακολούθηση του ποσοστού θνησιμότητας.  
Πηγή: Alemi et al. (1996)



## ΚΕΦΑΛΑΙΟ 3

### Χρήση και αξιολόγηση μεθόδων σάρωσης στην βιοεπιτήρηση

#### 3.1 Εισαγωγή

Οι υπηρεσίες που ασχολούνται με την δημόσια υγεία και οι επιδημιολόγοι συχνά επικεντρώνουν το ενδιαφέρον τους σε μεγάλες χρονολογικές συστάδες με μεγάλα ποσοστά εμφάνισης της νόσου. Με τον όρο συστάδες, εδώ, θα εννοούμε σύνολα από χρονικές περιόδους ή όπως θα δούμε παρακάτω γεωγραφικές περιοχές ή και τα δύο. Η ερώτηση που ακολουθεί είναι αν η συστάδα εμφανίστηκε λόγω κάποιων τυχαίων αιτιών ή λόγω της ύπαρξης κάποιου παράγοντα κινδύνου. Στο εξής θα λέμε ότι μια συστάδα ανιχνεύεται όταν η ανάλυση των δεδομένων παρέχει ένδειξη ότι σε ένα σύνολο από χρονικές περιόδους ή από γεωγραφικές περιοχές έχει αυξηθεί ο κίνδυνος εμφάνισης της νόσου. Αυτός είναι και ο σκοπός της ανάλυσης που ακολουθεί. Να αναδειχθούν οι περιοχές ή οι χρονικές περίοδοι στις οποίες υπάρχει μια στατιστικά σημαντική αύξηση του ποσοστού εμφάνισης της νόσου που θα αποτελείται, όχι υποχρεωτικά, από άλλες μικρότερες στις οποίες έχει ήδη γίνει ανάλυση των δεδομένων και δεν κρίθηκε σημαντική η αύξηση του ποσοστού σε κάθε μία χωριστά. Για να ερευνηθούν τα αίτια που οδήγησαν σε αυτή την αύξηση θα πρέπει να λάβουμε υπ' όψη τους πολλαπλούς ελέγχους που γίνονται όταν σαρώνουμε μια μεγαλύτερη χρονική περίοδο. Οι μέθοδοι σάρωσης (*Scan Methods* ή *Scan Statistics*) είναι τα εργαλεία για να απαντήσουμε σε τέτοια ερωτήματα.

Οι μέθοδοι σάρωσης έχουν χρησιμοποιηθεί για τον αναδρομικό έλεγχο ύπαρξης συστάδων για διάφορες μορφές καρκίνου, το σύνδρομο *Down*, τις αυτοκτονίες, το *SIDS*, τον ιό *HIV* και σε άλλες εφαρμογές. Μετά την 11<sup>η</sup> Σεπτεμβρίου 2001 και τις επιστολές με άνθρακα που ακολούθησαν το ενδιαφέρον στράφηκε στην προοδευτική ανάλυση. Όταν υπάρχουν πολλαπλές επιθέσεις μέσα σε ένα μικρό χρονικό διάστημα οι στατιστικές μέθοδοι σάρωσης είναι ιδιαίτερα χρήσιμες για να λαμβάνουμε εγκαίρως προειδοποιήσεις.

Οι μέθοδοι σάρωσης έχουν αναπτυχθεί και εφαρμόζονται για τρεις περιπτώσεις χρονολογικής ανάλυσης. Στην πρώτη περίπτωση, καταγράφονται οι χρόνοι των περιστατικών και η συνεχής στατιστική σάρωσης είναι ο μέγιστος αριθμός περιστατικών σε ένα παράθυρο μήκους  $w$  που σαρώνει το χρονικό διάστημα  $(0, T)$ . Στην δεύτερη περίπτωση, ο χρόνος διαμερίζεται σε ισομήκη  $T$  διαστήματα, σε ώρες ή μέρες, και τα δεδομένα αφορούν το πλήθος

των περιστατικών σε κάθε διάστημα. Στην τρίτη περίπτωση θεωρούμε μια ακολουθία  $T$  το πλήθος δοκιμών και σε κάθε ακολουθία σημειώνεται αν συνέβη ένα περιστατικό ή όχι. Η στατιστική σάρωσης για την δίτιμη περίπτωση είναι ο μέγιστος αριθμός περιστατικών για οποιεσδήποτε  $w$  συνεχόμενες δοκιμές. Φυσικά μπορεί να γίνει και ένας συνδυασμός των δύο πρώτων περιπτώσεων όπου καταγράφεται ο μέγιστος αριθμός περιστατικών σε  $w$  συνεχόμενα διαστήματα.

Πολλές φορές, πριν από την έναρξη της επιτήρησης, η εκδήλωση μιας νόσου παρουσιάζει μια μέτρια εποχικότητα, τάση ή γενικά υπάρχουν ενδείξεις ότι ακολουθεί κάποιο μοτίβο ή απλά μια λιγότερο τυχαία συμπεριφορά. Στην βιβλιογραφία υπάρχουν δύο είδη προσέγγισης στο πρόβλημα. Πρώτον μία τροποποίηση της μεθόδου σάρωσης με μεγέθυνση του παραθύρου σάρωσης ή, δεύτερον, μπορεί να εφαρμοστεί ένας έλεγχος γενικευμένου λόγου πιθανοφάνειας (*Generalized Likelihood Ratio Test, GLRT*). Αυτός ο έλεγχος βασίζεται στο μέγιστο μίας συνάρτησης των παρατηρούμενων και των αναμενόμενων συχνοτήτων των περιστατικών μέσα σε ένα παράθυρο της μεθόδου σάρωσης προκαθορισμένου πλάτους. Το μειονέκτημα είναι ότι τα  $p$  – values υπολογίζονται με προσομοίωση. Μια ακόμη λύση, στο ίδιο πνεύμα, για να αντιμετωπιστεί η προ-ύπαρξη μη σταθερών ποσοστών εμφάνισης της νόσου, είναι να χρησιμοποιήσουμε ένα παράθυρο με μεταβλητό μέγεθος που πάντα θα περιέχει το  $w/T$  του αναμενόμενου πλήθους παρατηρήσεων. Αυτό αποτελεί μια τροποποίηση της αναδρομικής συνεχούς μεθόδου σάρωσης [βλ. *Weinstock, (1981)*]. Όπως επισημάνθηκε από τους *Naus* και *Wallenstein* το 2005, η μέθοδος αυτή υστερεί γιατί, επιλέγοντας μεταβλητό μέγεθος παραθύρου, αποκλείει περιοδικές επιδράσεις όπως την μέρα της εβδομάδας (*day-of-week effect, w = 7*), ή την εποχικότητα (*seasonal effect, w = 365*) και λαμβάνει υπ' όψη πρόσφατες περιόδους που οδηγούν σε μια εκ των προτέρων (*a priori*) επιλογή του  $w$ .

Οι τελευταίοι πρότειναν μια προσέγγιση και για τις τρεις περιπτώσεις χρονολογικής ανάλυσης που είδαμε παραπάνω, όταν υπάρχουν σταθερά ποσοστά εμφάνισης της ασθένειας πριν από την έναρξη της βιοεπιτήρησης (μέθοδος  $P$  – σάρωσης (*P – scan approach*)). Η μέθοδος αυτή έχει το πλεονέκτημα ότι χρησιμοποιεί σταθερό μέγεθος παραθύρου και τα  $p$ -values υπολογίζονται χωρίς προσομοίωση.

Για την συνεχή συνάρτηση σάρωσης το πιο απλό αναδρομικό μοντέλο υπό την μηδενική υπόθεση της τυχειότητας, είναι ότι ένας γνωστός αριθμός από  $N$  περιπτώσεις κατανέμονται ομοιόμορφα στο  $(0, T)$ . Ο έλεγχος που απορρίπτει την μηδενική υπόθεση της ομοιόμορφης

κατανομής για μεγάλες τιμές σε ένα παράθυρο προκαθορισμένου μεγέθους έναντι της εναλλακτικής υπόθεσης ότι η πυκνότητα είναι μεγαλύτερη κατά ένα σταθερό ποσοστό σε ένα υποδιάστημα μήκους  $w$ , είναι ένας έλεγχος γενικευμένου λόγου πιθανοφάνειας [βλ. *Naus*, (1966)].

Στην περίπτωση που το προ-υπάρχον ποσοστό εμφάνισης δεν είναι σταθερό η απλή μέθοδος σάρωσης και το GLRT χρησιμοποιούν διαφορετικές στατιστικές συναρτήσεις σάρωσης. Σε επόμενη παράγραφο δίνεται μία προσέγγιση για την περίπτωση της προοδευτικής ανάλυσης που προτάθηκε από τον *Kulldorff* το 2001.

Όπως είδαμε οι μέθοδοι σάρωσης (*Scan Methods*) μπορούν να εφαρμοστούν τόσο στην περίπτωση που οι παρατηρήσεις είναι χρόνοι (*temporal case*), δηλαδή το μόνο που είναι γνωστό είναι ο χρόνος στον οποίο εκδηλώθηκε το κάθε συμβάν, όσο και στην περίπτωση που παρατηρούμε ζεύγη θέσης και χρόνων (*spatiotemporal case*), δηλαδή και η γεωγραφική θέση και η χρονική στιγμή που εκδηλώθηκε το συμβάν είναι γνωστά. Ειδικά για την δεύτερη περίπτωση προτιμάται η χρήση μεθόδων scan από τις πιο κλασσικές μεθόδους των διαγραμμάτων ελέγχου, όπως τα *CUSUM* [βλ. *Burkom*, (2006)]. Οι μέθοδοι scan εφαρμόστηκαν για πρώτη φορά στην παρακολούθηση χρόνιων παθήσεων, κυρίως του καρκίνου και έχουν πρόσφατα υιοθετηθεί στην παρακολούθηση λοιμωδών νόσων. Οι περισσότερες εργασίες πάνω στις μεθόδους αυτές αφορούν την αναδρομική (Φάσης I) ανάλυση ιστορικών δεδομένων.

Στην περίπτωση που η μόνη μεταβλητή είναι ο χρόνος, υπάρχουν δύο γενικοί τύποι προοδευτικών μεθόδων ανίχνευσης. Ο απλούστερος τρόπος αφορά την μέτρηση του πλήθους των περιστατικών, σε μία μόνο γεωγραφική περιοχή, για μία πρόσφατη χρονική περίοδο, δεδομένης διάρκειας. Όταν το πλήθος των περιστατικών ξεπεράσει ένα προκαθορισμένο όριο, θα έχουμε ένδειξη ότι υπάρχει αύξηση στο ρυθμό εκδήλωσης συμβάντων σε αυτή την περιοχή. Μία πιο σύνθετη μέθοδος θα ήταν να θεωρήσουμε ότι η διάρκεια της περιόδου παρακολούθησης εξαρτάται από διάφορες παραμέτρους [βλ. *Kulldorf* (2001)].

Σε πολλές περιπτώσεις τα διαγράμματα τύπου *CUSUM* έχουν καλύτερη απόδοση από τις μεθόδους σάρωσης των *Naus* και *Wallenstein* (2006), υπό την υπόθεση της αύξησης στον ρυθμό εκδήλωσης της υπό παρακολούθηση ασθένειας [βλ. *Joner et al.* (2008)]. Για το λόγο αυτό είναι χρήσιμο να γίνεται σύγκριση της απόδοσης των δύο μεθόδων. Τα διαγράμματα τύπου *CUSUM* έχουν την πολλή καλή ιδιότητα να ανιχνεύουν την ύπαρξη έντονης μετατόπισης (*sustained shift*) στην παράμετρο της κατανομής πιθανότητας. Από την άλλη,

δεν μπορεί να αγνοηθεί η απλότητα της βασικής μεθόδου σάρωσης, ειδικά για την περίπτωση όπου η μόνη μεταβλητή είναι ο χρόνος.

Στην παραπάνω περίπτωση, οι μετρήσεις που λαμβάνουμε μπορεί να είναι είτε οι ακριβείς χρονικές στιγμές που εκδηλώθηκε ένα περιστατικό της νόσου, σε μία συγκεκριμένη γεωγραφική περιοχή, είτε το πλήθος των περιστατικών που σημειώθηκαν σε τακτά χρονικά διαστήματα. Υπό την υπόθεση ότι τα δεδομένα προέρχονται από την διαδικασία *Poisson*, κάτι που γίνεται συχνά, οι χρόνοι ανάμεσα στα συμβάντα κατανέμονται εκθετικά και το πλήθος των παρατηρήσεων (*aggregated counts*) θα ακολουθεί την κατανομή *Poisson*. Μια άλλη μέθοδος θα ήταν να παρατηρούμε τα αποτελέσματα διαδοχικών (*successive*) ανεξάρτητων δοκιμών *Bernoulli* που μπορούν να αναχθούν σε δυωνυμικές παρατηρήσεις σε τακτά χρονικά διαστήματα.

Και οι τέσσερις περιπτώσεις παραπάνω βασίζονται στις κλασικές υποθέσεις των μεθόδων σάρωσης με μεταβλητή το χρόνο, όμως, στην βιοεπιτήρηση, σε αντίθεση με τον βιομηχανικό έλεγχο ποιότητας, τα δεδομένα είναι πολύ πιο συχνά μη-στάσιμα (*non stationary*) με εποχικότητα και άλλες επιδράσεις (*effects*), [βλ. *Shmueli* (2007)]. Μία μέθοδος scan με μεταβλητή μόνο το χρόνο, η οποία λαμβάνει υπ' όψη την ημέρα της εβδομάδας και άλλες επιδράσεις εποχικότητας προτάθηκε από τους *Wallenstein & Naus* (2004). Αν παρακολουθούμε ποσοστά θνησιμότητας, με βάση δεδομένα από νοσοκομεία, τότε μπορεί να παρατηρήσουμε διαφοροποιήσεις και στην περίπτωση που η διεργασία είναι εντός ελέγχου. Αυτό οφείλεται, συχνά, στους διάφορους παράγοντες κινδύνου των ασθενών και τότε χρειάζεται να χρησιμοποιήσουμε μία μέθοδο προσαρμοσμένη στον κίνδυνο (*risk adjusted*). Τέτοιες μέθοδοι έχουν προταθεί από τους *Grigg* και *Farewell* (2004).

Σε κάθε περίπτωση, που η μόνη μεταβλητή είναι ο χρόνος, υποθέτουμε την ύπαρξη ενός ρυθμού εκδήλωσης περιστατικών της νόσου. Σκοπός της παρακολούθησης είναι να ανιχνεύσουμε μια αύξηση αυτού του ρυθμού το συντομότερο δυνατό από την στιγμή που θα εμφανιστεί. Ο τρόπος που θα εκτιμήσουμε το βασικό ρυθμό εκδήλωσης αποτελεί μια σημαντική περιοχή έρευνας [βλ. *Jensen et al.* (2006)]. Γενικά, αυτή η εκτίμηση είναι πολύ πιο δύσκολο να γίνει στην βιοεπιτήρηση απ' ό,τι στον βιομηχανικό έλεγχο.

Στην προοδευτική ανάλυση με μεθόδους scan, έχουμε ένδειξη αύξησης του ρυθμού εκδήλωσης όταν το πλήθος των περιστατικών στο πιο πρόσφατο χρονικό διάστημα, προκαθορισμένου μήκους, είναι ασυνήθιστα μεγάλο. Αυτό αποτελεί προσέγγιση της μεθόδου κινητού παραθύρου (*moving window*). Για παράδειγμα, στην περίπτωση που χρησιμοποιούμε τις δοκιμές *Bernoulli*, όπως εξηγήθηκε παραπάνω, έχουμε ένδειξη ότι αυξήθηκε ο ρυθμός

εκδήλωσης της νόσου όταν  $m$  ή περισσότερες «αποτυχίες» έχουν συμβεί στις  $n$  τελευταίες δοκιμές. Παρόλο που μπορεί να θεωρηθεί τεχνική διαγραμμάτων στατιστικού ελέγχου, συχνά χρησιμοποιείται ένα διάγραμμα της στατιστικής συνάρτησης με το χρόνο με ένα άνω όριο ελέγχου. Σε αντίθεση με τα παραδοσιακά διαγράμματα ελέγχου δεν υπάρχουν συγκεκριμένες τεχνικές για τον σχεδιασμό μιας μεθόδου σάρωσης. Μία μέθοδος scan με  $m = 2$  είναι ισοδύναμη με το μονοδιάστατο διάγραμμα τύπου *CUSUM* με όριο ελέγχου  $H = 1$  και τιμή αναφοράς  $K = 1/n$ . Για μια ακολουθία τυχαίων μεταβλητών  $X_1, X_2, X_3$  που παρατηρούνται με την πάροδο του χρόνου, η στατιστική ελέγχου για το μονοδιάστατο διάγραμμα τύπου *CUSUM* είναι:

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

όπου  $i = 1, 2, 3, \dots$ , το  $X_i$  είναι 1 αν η  $i$ -οστή παρατήρηση είναι αποτυχία και 0 αλλιώς,  $K$  είναι μία μη αρνητική σταθερά και  $S_0 = 0$ . Όταν το  $S_i$  είναι ίσο ή μεγαλύτερο από την μη αρνητική τιμή  $H$  τότε έχουμε ένδειξη εκτός ελέγχου διεργασίας.

Οι βασικές μέθοδοι scan που χρησιμοποιούνται στην βιοεπιτήρηση βασίζονται στο ίδια λογική κινητού παράθυρου όπως και το διάγραμμα κινητού μέσου στο βιομηχανικό έλεγχο. Για παράδειγμα, για την παρακολούθηση του μέσου μίας διεργασίας με την μέθοδο κινητού μέσου, η στατιστική ελέγχου είναι ο δειγματικός μέσος των παρατηρήσεων στα τελευταία  $w$  δείγματα, όπου το  $w$  είναι προκαθορισμένο. Αξίζει να σημειωθεί ότι υπάρχουν λίγες εργασίες πάνω στα διαγράμματα κινητού μέσου.

Στο βιομηχανικό έλεγχο ποιότητας, τα διαγράμματα τύπου *CUSUM* και τα διαγράμματα εκθετικά σταθμισμένου κινητού μέσου (*exponentially weighted moving average, EWMA*) χρησιμοποιούνται πολύ πιο συχνά από τις μεθόδους scan κινητού μέσου στην προοδευτική ανάλυση. Αυτό συμβαίνει για τρεις λόγους. Πρώτον, στο διάγραμμα *EWMA*, τα βάρη στα δεδομένα από προηγούμενα δείγματα μειώνονται με την πάροδο του χρόνου. Αυτό φαίνεται καλύτερο από το να δίνονται ίσα βάρη στα τελευταία  $w$  δείγματα. Δεύτερον, η ακολουθία των μεταβλητών στην περίπτωση του κινητού μέσου δεν ικανοποιεί την ιδιότητα Markov καθώς τα τελευταία  $w - 1$  δείγματα επηρεάζουν την μελλοντικές τιμές της ακολουθίας. Τέλος η προσαρμογή της μεθόδου κινητού μέσου δεν είναι το ίδιο καλή όπως με άλλες μεθόδους [βλ. Roberts (1966)]. Αντίθετα με την μέθοδο *CUSUM*, για την μέθοδο κινητού μέσου δεν υπάρχει ιδιότητα βελτιστοποίησης.

### 3.2 Προοδευτική, χρονολογική και γεωγραφική βιοεπιτήρηση με τη χρήση μεθόδων σάρωσης.

Ένας συνηθισμένος τρόπος γεωγραφικής παρακολούθησης μιας ασθένειας είναι να επιλέξουμε την ασθένεια, την γεωγραφική περιοχή και μια προκαθορισμένη χρονική περίοδο για την χαρτογράφηση των ποσοστών της νόσου. Επιπλέον των βασικών χαρτών με παρατηρούμενα ποσοστά από γεωγραφικές περιοχές έχουν αναπτυχθεί διάφορες στατιστικές τεχνικές τόσο για την χαρτογράφηση όσο και για να ελέγχουμε αν μια αύξηση ή μείωση των ποσοστών αυτών οφείλεται σε φυσικά αίτια (τυχαία μεταβολή) ή όχι. Στην πρώτη περίπτωση χρησιμοποιούνται διαστημικές τεχνικές εξομάλυνσης (*spatial smoothing techniques*) ενώ στην δεύτερη εφαρμόζονται έλεγχοι τυχειότητας.

Πολλές χώρες διατηρούν ιατρικά αρχεία με πιο συνηθισμένα αυτά για τον καρκίνο, τις γεννήσεις παιδιών με συγγενή νοσήματα, ενώ πολλά εθνικά ή τοπικά κέντρα συλλέγουν συνεχώς πληροφορίες για περιστατικά λοιμωδών ασθενειών. Τέτοια περιστατικά καταχωρούνται σε ημερήσια, μηνιαία ή ετήσια κλίμακα ανάλογα με την φύση της ασθένειας και την χωρητικότητα του συστήματος. Η ανίχνευση γεωγραφικών συστάδων, όπου αυξάνονται σημαντικά τα ποσοστά μιας νόσου εξ' αιτίας κάποιου παράγοντα κινδύνου, με ταχύτητα, έχει μεγάλη σημασία για την δημόσια υγεία. Μπορεί να χρησιμοποιηθεί και για την ανίχνευση άγνωστων μέχρι τώρα ασθενειών, αγνώστων παραγόντων κινδύνου για γνωστές ασθένειες ή για την ύπαρξη γνωστών παραγόντων κινδύνου που μέχρι τώρα αγνοούσαμε ότι υπάρχουν σε μία συγκεκριμένη περιοχή.

Για μια καλά καθορισμένη γεωγραφική περιοχή, υπάρχουν διάφορες χρονολογικές στατιστικές τεχνικές για την παρακολούθηση και την ανίχνευση μιας ξαφνικής αύξησης του κινδύνου εκδήλωσης μιας νόσου. Αν συμβεί ένα ξέσπασμα σε μια μικρή γεωγραφική περιοχή, η ξαφνική και σημαντική αύξηση των ποσοστών μπορεί να μην γίνει αντιληπτή αν τα δεδομένα λαμβάνονται για την ευρύτερη περιοχή. Ένας τρόπος για να ξεπεραστεί αυτό το πρόβλημα είναι να χωριστεί μια μεγάλη γεωγραφική περιφέρεια σε μικρότερες περιοχές και να παρακολουθούνται αυτές οι περιοχές ξεχωριστά. Απ' την άλλη έχοντας πολλές περιοχές απαιτείται η εφαρμογή ενός πολλαπλού ελέγχου και θα υπάρχουν πολλοί εσφαλμένοι συναγερμοί αν δεν χρησιμοποιηθεί το κατάλληλο επίπεδο σημαντικότητας. Κυρίως όμως, αν ένα ξέσπασμα της ασθένειας συμβεί κοντά στο όριο που χωρίζει δύο περιοχές ή σε ένα ακόμη

μικρότερο κομμάτι μιας περιοχής, μπορεί να το αγνοήσουμε απλώς και μόνο επειδή δεν «προσαρμόστηκε» σωστά στα προκαθορισμένα γεωγραφικά όρια.

Η μονοδιάστατη μέθοδος σάρωσης χρησιμοποιείται συχνά στην περίπτωση των χρονολογικών δεδομένων (χρονολογική βιοεπιτήρηση) ενώ μια γεωγραφική μέθοδος scan χρησιμοποιείται για την γεωγραφική βιοεπιτήρηση. Ο *Kulldorff* (2001) πρότεινε μία μέθοδο για την περιοδική γεωγραφική βιοεπιτήρηση, βασισμένη σε μία χωρική – χρονολογική μέθοδο σάρωσης (*space-time scan statistic*). Σ' αυτή τη μέθοδο, τα συμπεράσματα που προκύπτουν, προσαρμόζονται στον πολλαπλό έλεγχο που προκύπτει λόγω των πολλών πιθανών θέσεων και μεγεθών των συστάδων της ασθένειας. Η μέθοδος ανιχνεύει μόνο τις συστάδες στις οποίες υπάρχουν ακόμη άτομα σε κίνδυνο, στην πιο πρόσφατη χρονική περίοδο. Μπορεί να ανιχνεύσει συστάδες ανεξάρτητα από την ηλικία τους, δηλαδή ανεξάρτητα από το χρόνο που πέρασε από τότε που αρχίσαμε να παρακολουθούμε την συστάδα. Έτσι μπορεί να ανιχνεύσει συστάδες που υπάρχουν πολύ καιρό στο σύστημα και ο κίνδυνος είναι μέτριος αλλά και συστάδες που προέκυψαν πρόσφατα και στις οποίες ο κίνδυνος είναι αυξημένος. Η προσαρμογή της μεθόδου είναι το ίδιο καλή ανεξάρτητα από το μήκος του χρονικού διαστήματος που έχουμε προκαθορίσει. Τέλος το σύστημα παρακολούθησης που προκύπτει από αυτή την μέθοδο, προσαρμόζεται στην ανάλυση που έχει ήδη πραγματοποιηθεί καθώς περιλαμβάνει πολλαπλούς ελέγχους για τις περιοδικά επαναλαμβανόμενες μετρήσεις.

Η παραπάνω μέθοδος δεν θα πρέπει να συγχέεται με τις παραδοσιακές μεθόδους παρακολούθησης μιας ασθένειας με ανάλυση κατά συστάδες χώρου-χρόνου. Οι τελευταίες είναι αναδρομικές μελέτες σχεδιασμένες για να ελέγχουν αν τα περιστατικά της νόσου είναι τυχαία κατανομημένα στο χώρο και το χρόνο, για μια προκαθορισμένη γεωγραφική περιοχή και σε ένα προκαθορισμένο χρονικό διάστημα ενώ η μέθοδος του *Kulldorff* είναι προοδευτική με περιοδικά επαναλαμβανόμενες αναλύσεις. [βλ. *Rogerson*, (1997)].

Προκειμένου να δούμε την εφαρμογή των μεθόδων στην πράξη θα χρησιμοποιήσουμε, ως παράδειγμα, δεδομένα που αφορούν καρκίνο θυρεοειδούς σε άντρες από την πολιτεία του Νέου Μεξικού, στις Η.Π.Α.. Η ηλικία έχει τυποποιηθεί και τα p-values βασίζονται στον *Monte Carlo* έλεγχο υποθέσεων [βλ. *Dwass*,(1957)], χρησιμοποιώντας 999 επαναλήψεις των δεδομένων υπό την μηδενική υπόθεση ότι κατανέμονται τυχαία στο χώρο. Οι αναλύσεις έχουν γίνει με την χρήση του λογισμικού *SaTScan* του Εθνικού Ινστιτούτου για τον Καρκίνο των Ηνωμένων Πολιτειών. Το λογισμικό αυτό διανέμεται δωρεάν.

Τα δεδομένα συλλέχτηκαν στο Νέο Μεξικό από το 1973 έως το 1992. Η συλλογή έγινε από το Ογκολογικό Αρχείο του Νέου Μεξικού (*New Mexico Tumor Registry*) για λογαριασμό του προγράμματος «Επιτήρηση, Επιδημιολογία και Τελικά Αποτελέσματα», (*Surveillance, epidemiology and end results program*) του Εθνικού Ινστιτούτου για τον Καρκίνο (*National Cancer Institute*). Τα δεδομένα προέρχονται από τις 32 κομητείες που υπήρχαν στην πολιτεία το 1973. Συνολικά αναφέρθηκαν 333 περιπτώσεις. Το ποσοστό εμφάνισης ήταν 2,4 περιστατικά ανά 100.000 άνδρες κάθε χρονιά για κάθε κομητεία. Οι εκτιμήσεις έγιναν λαμβάνοντας υπ' όψιν τα επίσημα στατιστικά στοιχεία, που λαμβάνονται με απογραφή κάθε δεκαετία, για τις γεννήσεις, τους θανάτους και την μετανάστευση. Ο συνολικός πληθυσμός αρρένων αυξήθηκε από 546.000 το 1973 σε 779.000 το 1992.

Ο καρκίνος του θυρεοειδούς είναι σπάνια ασθένεια. Το ετήσιο ποσοστό εμφάνισης της νόσου στις ΗΠΑ εκτιμάται ότι είναι 5.2 περιπτώσεις ανά 100.000 άτομα. Το ποσοστό αυτό είναι μικρότερο στους άντρες απ' ότι στις γυναίκες, 2,9 ανά 100.000 άντρες έναντι του 7,4 ανά 100.000 που εκτιμάται για τις γυναίκες. Η πιθανότητα να διαγνωστεί κάποιος με καρκίνο θυρεοειδούς κατά την διάρκεια της ζωής του είναι 0,26% για τους άνδρες και 0,72% για τις γυναίκες [βλ. *Ries et al.*, (1999)]. Η ίδια μελέτη έγινε σε διάφορες περιοχές του πλανήτη, όπου τηρούνται τέτοια αρχεία, και προέκυψε ότι τα ποσοστά με βάση την ηλικία, ήταν υψηλότερα στην Χαβάη και την Ισλανδία και χαμηλότερα στην Βομβάη, την Αγγλία και την Ουαλία [βλ. *Parkin et al.*, (1992)]. Γνωστοί παράγοντες κινδύνου είναι η έκθεση σε ιονίζουσα ραδιενεργή ακτινοβολία κατά την διάρκεια της παιδικής ηλικίας [βλ. *Duffy & Fitzgerald*, (1950) και *Ron et al.*, (1989)], η ακτινοθεραπεία για παθήσεις της κεφαλής και του αυχένα [βλ. *Shore et al.*, (1985)], η ραδιενέργεια από πυρηνικές εκρήξεις [βλ. *Robbins and Adams*, 1989, *Prentice et al.*, 1982] και ο χειρισμός μηχανημάτων ακτινών X [βλ. *Carstensen et al.*, (1990)].

Ο σχετικός ρυθμός επιβίωσης για τον καρκίνο του θυρεοειδούς είναι 95%, όπου θεωρούμε ότι ένας ασθενής επιβιώνει από την νόσο αν μετά την πάροδο 5 ετών από την αρχική εκδήλωση δεν έχουν παρουσιαστεί υποτροπές. Συνεπώς ο καρκίνος του θυρεοειδούς είναι η λιγότερο θανατηφόρα μορφή καρκίνου. Τα ετήσια ποσοστά θνησιμότητας στις ΗΠΑ είναι περίπου ίδια για άντρες και γυναίκες, με 0,3 θανάτους ανά 100.000 άτομα κατά την περίοδο 1992-1996. Για την περίοδο αυτή η αντρική θνησιμότητα στο Νέο Μεξικό ήταν η 21<sup>η</sup> ανάμεσα στις 50 πολιτείες των ΗΠΑ με 0,3 θανάτους ανά 100.000 [βλ. *Ries et al.*, (1999)].



### 3.3 Μόνο γεωγραφική βιοεπιτήρηση

Η μέθοδος σάρωσης στο χώρο ή μέθοδος χωρικής σάρωσης (*spatial scan statistic*) αναπτύχθηκε για να ανιχνεύει την ύπαρξη γεωγραφικών συστάδων και να προσδιορίζει την θέση τους στο χώρο [βλ. *Kulldorff*, (1997)]. Με τον όρο χώρος εδώ θα εννοούμε τη δισδιάστατη περιοχή που απεικονίζεται σε ένα γεωγραφικό χάρτη. Μπορούμε να υποθέσουμε ότι ο αριθμός των συμβάντων, για παράδειγμα ο αριθμός των διαγνώσεων καρκίνου του μαστού τελικού σταδίου ή ο αριθμός των θανάτων, ακολουθούν την κατανομή *Poisson* ή την κατανομή *Bernoulli*. Ένα παράδειγμα για το οποίο χρησιμοποιούμε το μοντέλο *Poisson* είναι αν μελετάμε την γεωγραφική κατανομή περιστατικών καρκίνου του μαστού σε μία πενταετία και ο παρονομαστής στο ποσοστό εμφάνισης αντιστοιχεί στα χρόνια που έζησε ένα άτομο μέσα σε στην πενταετία αυτή, πάντα με επίπεδο αναφοράς τον πληθυσμό σε κίνδυνο. Η κατανομή *Bernoulli* πρέπει να εφαρμόζεται όταν μελετάμε την γεωγραφική κατανομή των περιστατικών καρκίνου του μαστού τελικού σταδίου και ο παρονομαστής αφορά όλα τα περιστατικά καρκίνου του μαστού. Ανάλογα με τα δεδομένα που έχουμε στην διάθεση μας η χωρική σάρωση μπορεί να εφαρμοστεί είτε για συγκεντρωτικά δεδομένα σε περιοχές με ευαισθησία, είτε για περιπτώσεις όπου μία περιοχή αντιστοιχεί σε ένα άτομο σε κίνδυνο, δηλ. η γεωγραφική περιοχή είναι η ακριβής θέση του ατόμου.

Με την μέθοδο αυτή θεωρούμε ένα κυκλικό παράθυρο στο χάρτη και το κέντρο του κύκλου να κινείται ώστε σε διαφορετικά σημεία το παράθυρο να περιλαμβάνει διαφορετικούς συνδυασμούς γειτονικών περιοχών με ευαισθησία. Αν στο παράθυρο περιέχεται το κέντρο μιας περιοχής, τότε θεωρούμε ότι όλη η περιοχή περιλαμβάνεται στο παράθυρο. Για κάθε κέντρο του κύκλου, η ακτίνα του παραθύρου λαμβάνει συνεχείς τιμές από το 0 μέχρι κάποιο προκαθορισμένο μέγιστο, έτσι ώστε να μην περιλαμβάνει ποτέ περισσότερο από το μισό του συνολικού πληθυσμού σε κίνδυνο. Με αυτόν τον τρόπο το παράθυρο έχει ευελιξία τόσο ως προς την θέση που βρίσκεται όσο και ως προς το μέγεθος του χώρου που περιλαμβάνει. Η μέθοδος δημιουργεί ένα πολύ μεγάλο αριθμό διακεκριμένων κυκλικών παραθύρων, το καθένα με διαφορετικό συνδυασμό γειτονικών περιοχών, ενώ σε κάθε παράθυρο υπάρχει πιθανότητα να περιέχεται μια ομάδα περιστατικών.

Υπό την υπόθεση ότι ο συνολικός αριθμός περιστατικών που παρατηρήθηκαν είναι  $N$ , η στατιστική συνάρτηση  $S$ , για την μέθοδο, ορίζεται ως ο λόγος μέγιστης πιθανοφάνειας όλων των πιθανών κύκλων  $Z$ .

$$S = \frac{\max_Z \{L(Z)\}}{L_0} = \max_Z \left\{ \frac{L(Z)}{L_0} \right\} \quad (3.2)$$

όπου  $L_0$  είναι η συνάρτηση πιθανοφάνειας υπό την μηδενική υπόθεση και  $L(Z)$  είναι η μέγιστη πιθανοφάνεια του κύκλου  $Z$  που εκφράζει πόσο πιθανό είναι το ποσοστό των περιστατικών να είναι διαφορετικό εντός και εκτός του κύκλου.

Έστω  $n_Z$  ο αριθμός των περιπτώσεων στον κύκλο  $Z$ . Για το μοντέλο Poisson, έστω ότι  $\mu(Z)$  είναι ο αναμενόμενος αριθμός περιστατικών υπό την μηδενική υπόθεση, έτσι ώστε  $\mu(A) = N$ , όπου  $A$  η συνολική περιοχή που παρακολουθούμε. Αποδεικνύεται (Kulldorff, 1997) ότι :

$$\frac{L(Z)}{L_0} = \begin{cases} \left\{ \frac{n_Z}{\mu(Z)} \right\}^{n_Z} \left\{ \frac{N-n_Z}{N-\mu(Z)} \right\}^{N-n_Z}, & \text{αν } n_Z > \mu(Z) \\ 1 & , \text{αλλού} \end{cases} \quad (3.3)$$

Καθώς αυτός ο λόγος πιθανοφάνειας μεγιστοποιείται για όλους τους κύκλους μπορεί να προσδιορίσει τον κύκλο που έχει την μεγαλύτερη πιθανότητα να αποτελεί μια ομάδα. Το  $p$ -value προκύπτει από τον Monte Carlo έλεγχο υποθέσεων.

Η μέθοδος χωρικής σύρωσης έχει τις ακόλουθες ιδιότητες που την κάνουν ιδιαίτερα χρήσιμο εργαλείο της βιοεπιτήρησης:

- Προσαρμόζεται το ίδιο για την πυκνότητα ενός ανομοιογενούς πληθυσμού και για οποιοδήποτε αριθμό συμμεταβλητών.
- Μηδενίζεται η μεροληψία που θα προέκυπτε από την προεπιλογή των συστάδων, καθώς η ανίχνευση για συστάδες γίνεται χωρίς να έχει καθοριστεί το μέγεθος και η θέση τους.
- Ο έλεγχος λόγων πιθανοφάνειας λαμβάνει υπ' όψιν τους πολλαπλούς ελέγχους και δίνει μία μόνο  $p$  - value για τον έλεγχο της μηδενικής υπόθεσης.
- Αν απορριφθεί η μηδενική υπόθεση, μπορούμε να υπολογίσουμε περίπου την θέση της συστάδας που προκάλεσε την απόρριψη.

Με την μέθοδο αυτή, εκτός από την πιο πιθανή συστάδα, μπορούμε να ανιχνεύσουμε και δευτερεύουσες συστάδες και να τις ταξινομήσουμε με βάση το λόγο πιθανοφάνειας. Βέβαια, θα υπάρχουν πάντα ομάδες περιοχών που θα πλησιάζουν την πιο πιθανή συστάδα καθώς

προσθαφαιρώντας ένα μικρό αριθμό περιοχών δεν αλλάζει σημαντικά η πιθανοφάνεια. Αυτό σημαίνει ότι παρόλο που μπορούμε να προσδιορίσουμε την ευρύτερη περιοχή που βρίσκεται η συστάδα, δεν μπορούμε να καθορίσουμε με ακρίβεια τα όρια της. Για αυτό το λόγο θα αναφερόμαστε πάντα στην «κατά προσέγγιση θέση» της συστάδας.

Η *spatial scan statistic* έχει χρησιμοποιηθεί για διάφορες ασθένειες όπως η λευχαιμία (Hjalmarsson et al, 1996), ο καρκίνος του μαστού τελικού σταδίου (Sheehan et al, 2000) και το σάρκωμα μαλακών μορίων (Viel et al, 2000).

Για το παράδειγμα του καρκίνου του θυρεοειδούς σε άντρες στο Νέο Μεξικό, υποθέτουμε ότι έχουμε στην διάθεση μας μόνο τα δεδομένα τις περιόδου 1973 – 1978. Σε αυτή την περίοδο, 99 άντρες διαγνώστηκαν με καρκίνο του θυρεοειδούς, σε όλη την πολιτεία. Η πιο πιθανή συστάδα αποτελείται από τις περιοχές *Bernalillo* και *Valencia* που βρίσκονται στο κέντρο της πολιτείας και περιλαμβάνουν το *Albuquerque*, την μεγαλύτερη πόλη του Νέου Μεξικό. Σ' αυτήν την περιοχή παρατηρήθηκαν 46 περιπτώσεις ενώ ο αναμενόμενος αριθμός περιπτώσεων ήταν 34,6, με  $p = 0,441$ , Πίνακας 3.1. Καταλήγουμε λοιπόν στο συμπέρασμα ότι δεν υπάρχει γεωγραφική διαφοροποίηση του καρκίνου του θυρεοειδούς στο Νέο Μεξικό την χρονική περίοδο που μελετάμε.

Στην περίπτωση του καρκίνου τα νέα δεδομένα συλλέγονται και καταγράφονται σε ετήσια βάση. Έτσι μπορούμε να επαναλαμβάνουμε την διαδικασία κάθε χρόνο, λαμβάνοντας πλέον υπ' όψη τα νέα δεδομένα που προστέθηκαν την τελευταία χρονιά. Τα αποτελέσματα μιας τέτοιας ανάλυσης κάθε χρόνο δίνονται στον Πίνακα 3.1. Η αναμενόμενη συχνότητα δίνεται με στρογγυλοποίηση στον πλησιέστερο ακέραιο αριθμό, καθώς εκφράζει τον αναμενόμενο αριθμό ανδρών που αναμένεται να νοσήσουν. Ως σχετικός κίνδυνος ορίζεται, εδώ, ο λόγος της συχνότητας προς την αναμενόμενη συχνότητα.

Χρονική Περίοδος	Συστάδες με την Μεγαλύτερη Πιθανότητα	Διάρκεια Συστάδας	Συχνότητα	Αναμενόμενη Συχνότητα	Σχετικός Κίνδυνος	p-value
1973-1978	Bernalillo, Valencia	1973-1978	46	35	1,3	0,44
1973-1979	Bernalillo, Valencia	1973-1979	52	40	1,3	0,48
1973-1980	Bernalillo, Valencia	1973-1980	58	45	1,3	0,39
1973-1981	Bernalillo, Valencia	1973-1981	68	51	1,3	0,1
1973-1982	Bernalillo, Valencia	1973-1982	79	58	1,4	0,04
1973-1983	Bernalillo, Valencia	1973-1983	84	62	1,4	0,04
1973-1984	North Central Counties	1973-1984	113	90	1,3	0,05
1973-1985	North Central Counties - San Miguel	1973-1985	115	95	1,2	0,18

1973-1986	North Central Counties + Colfax, Harding	1973-1986	129	108	1,2	0,16
1973-1987	North Central Counties + Colfax, Harding	1973-1987	142	117	1,2	0,05
1973-1988	North Central Counties - San Miguel	1973-1988	143	115	1,2	0,02
1973-1989	North Central Counties + Colfax, Harding	1973-1989	165	134	1,2	0,01
1973-1990	North Central Counties + Torrance	1973-1990	174	144	1,2	0,02
1973-1991	North Central Counties + Colfax, Harding	1973-1991	186	155	1,2	0,02
1973-1992	North Central Counties + Torrance	1973-1992	199	164	1,2	0,01

**Πίνακας 3.1** Γεωγραφική ανάλυση περιστατικών καρκίνου του θυρεοειδούς σε άντρες στο Νέο Μεξικό, με χρήση χωρικής συνάρτησης σάρωσης [Πηγή: Kulldorf, (2000)].



**Εικόνα 3.1** Χάρτης της Πολιτείας του Νέου Μεξικού, Γεωγραφική συστάδα με την μεγαλύτερη πιθανότητα [Πηγή: Kulldorf, (2000)]

Όταν παρακολουθούμε το σύστημα επαναλαμβανόμενα χρησιμοποιώντας αποκλειστικά την γεωγραφική ανάλυση αναδύονται δύο προβλήματα. Πρώτον, η ισχύς του ελέγχου, για τον γρήγορο εντοπισμό αναδυόμενων συστάδων, είναι μικρή. Η καθαρά γεωγραφική ανάλυση για

μια μεγάλη χρονική περίοδο, περιλαμβάνει διακυμάνσεις του ποσοστού εμφάνισης της ασθένειας που οφείλονται σε τυχαίους παράγοντες. Έτσι αν ο πραγματικός κίνδυνος είναι αυξημένος τα τελευταία μόνο χρόνια, η δύναμη της συστάδας εξασθενεί καθώς συνυπολογίζονται και μικρότερα ποσοστά κινδύνου από παλαιότερα έτη. Δεύτερον, παρόλο που τα δεδομένα προσαρμόζονται για πολλαπλούς ελέγχους με βάση τις θέσεις και τα μεγέθη των συστάδων, δεν προσαρμόζονται για πολλαπλούς ελέγχους με βάση τις επαναλαμβανόμενες μετρήσεις κάθε χρόνο. Για παράδειγμα, από τις 15 αναλύσεις που παρουσιάζονται στον Πίνακα 3.1 κάποιες είναι στατιστικά σημαντικές ενώ άλλες όχι, γεγονός που δυσκολεύει την ανάλυση.

### 3.4 Ανίχνευση αναδυόμενων συστάδων

Όπως αναφέρθηκε, αν εφαρμόσουμε μια αποκλειστικά γεωγραφική ανάλυση των δεδομένων για μία εκτεταμένη χρονική περιοχή, η ισχύς για την ανίχνευση προσφάτως αναδυόμενων συστάδων, είναι χαμηλή. Ένας τρόπος για την επίλυση αυτού του προβλήματος θα ήταν να εφαρμόσουμε την μέθοδο μόνο για τα τελευταία χρόνια. Ο κατάλληλος αριθμός ετών που θα πρέπει να συμπεριλάβουμε είναι θεωρητικά άγνωστος, καθώς δεν υπάρχει σχετική βιβλιογραφία ακόμη. Αν συμπεριλάβουμε λιγότερα χρόνια από όσο πρέπει, τότε η ισχύς θα είναι τόσο χαμηλή ώστε η διαδικασία δεν θα είναι ικανή να ανιχνεύσει μια μικρότερη του μετρίου αύξηση του κινδύνου, αν αυτή υπάρχει σε ένα σεβαστό χρονικό διάστημα. Απ' την άλλη, αν συμπεριλάβουμε περισσότερα χρόνια, η ισχύς θα είναι πάλι χαμηλή και δεν θα μπορούμε να ανιχνεύσουμε μια συστάδα όπου παρουσιάστηκε αύξηση στον κίνδυνο, αν αυτή η αύξηση συνέβη πολύ πρόσφατα. Η λύση στο πρόβλημα αυτό είναι να εφαρμόσουμε την μέθοδο στον χώρο και στον χρόνο ταυτόχρονα.

Αντί του κυκλικού παραθύρου που χρησιμοποιήσαμε στην γεωγραφική ανάλυση, δισδιάστατη θεώρηση, θα χρησιμοποιήσουμε τώρα ένα κυλινδρικό παράθυρο θέσης και χρόνου, τρισδιάστατη θεώρηση. Η βάση του κυλίνδρου εκφράζει την επιφάνεια στο χάρτη και λειτουργεί όπως στην περίπτωση της χωρικής συνάρτησης σάρωσης (*spatial scan statistic*), ενώ το ύψος του κυλίνδρου εκφράζει τον χώρο. Ο κύλινδρος είναι ευέλικτος ως προς την θέση και το μέγεθος της βάσης αλλά και ως προς την χρονική στιγμή έναρξης της παρακολούθησης. Το μέγεθος και η θέση της βάσης είναι ανεξάρτητα από το ύψος του κυλίνδρου. Αυτό σημαίνει ότι για κάθε πιθανή θέση και μέγεθος της βάσης εξετάζουμε κάθε

πιθανή ημερομηνία έναρξης της παρακολούθησης και αντίστροφα. Λαμβάνουμε υπ' όψη μόνο εκείνους τους κυλίνδρους που φτάνουν έως την λήξη της χρονικής περιόδου παρακολούθησης. Συνεπώς, λαμβάνουμε υπ' όψη μόνο εκείνες τις συστάδες που είναι «ζωντανές», που υπάρχουν, που είναι σε κίνδυνο κατά την λήξη της παρακολούθησης και αγνοούμε εκείνες που υπήρξαν ιστορικά αλλά πλέον δεν αποτελούν πρόβλημα στην δημόσια υγεία. Έστω ότι το διάστημα  $[Y_1, Y_2]$  είναι το χρονικό διάστημα για το οποίο έχουμε δεδομένα και έστω  $s$  και  $t$  οι ημερομηνίες έναρξης και λήξης του κυλίνδρου αντίστοιχα. Θεωρούμε τότε όλους τους κυλίνδρους για τους οποίους ισχύει:

$$Y_1 \leq s \leq t = Y_2.$$

Ο έλεγχος λόγου πιθανοφάνειας κατασκευάζεται με τον ίδιο τρόπο όπως και στην περίπτωση της γεωγραφικής ανάλυσης με την χρήση των εξισώσεων (3.2) και (3.3). Για τυχαία δειγματοληψία, οι γεωγραφικές συστάδες και τα χρονικά διαστήματα περιλαμβάνουν τυχαίο αριθμό περιπτώσεων, ανεξάρτητα τα μεν από τα δε. Οι αλγόριθμοι για την εύρεση της πιθανοφάνειας σε κάθε παράθυρο εφαρμόζονται τώρα όχι σε δύο αλλά σε τρεις διαστάσεις κάτι που τους κάνει ευαίσθητους στους υπολογισμούς.

Στο παράδειγμα με τον καρκίνο του θυρεοειδούς σε άντρες στην πολιτεία του Νέου Μεξικού, μελετάμε πάλι την περίοδο 1973-1978. Προκύπτει ότι η πιο πιθανή συστάδα αποτελείται από τις κομητείες Bernalillo, McKinley, Valencia, San Juan, Carton, Sandoval, Los Alamos και Rio Arriba στο χρονικό διάστημα 1975-1978. Δηλαδή από τα έξι χρόνια που συνολικά διήρκεσε η επιτήρηση μόνο τα τελευταία τέσσερα περιλαμβάνονται στην συστάδα. Ενώ το αναμενόμενο πλήθος συμβάντων ήταν 36 παρατηρήθηκαν τελικά 48 συμβάντα με  $p - value = 0,6$ . Συνεπώς, το 1978, δεν μπορούμε να υποθέσουμε ότι υπήρχαν γεωγραφικές περιοχές στο Νέο Μεξικό με αυξημένο κίνδυνο εμφάνισης καρκίνου του θυρεοειδούς στους άντρες.

Όπως έγινε και προηγούμενα με την αποκλειστικά γεωγραφική ανάλυση του προβλήματος μπορούμε να επαναλαμβάνουμε την διαδικασία κάθε χρόνο μετά το 1978 προσθέτοντας κάθε φορά τα δεδομένα της χρονιάς που προηγήθηκε. Τα αποτελέσματα δίνονται στον Πίνακα 3.2. Παρατηρούμε ότι το 1985 προκύπτει ξαφνικά μία, μικρή γεωγραφικά, συστάδα στην κομητεία Lincoln με τρεις διαγνώσεις καρκίνου θυρεοειδούς και σχετικό κίνδυνο ίσο προς 13,8. Παρόλο που το εύρημα αυτό δεν είναι στατιστικά σημαντικό,  $p - value = 0,23$ ,

υπάρχει ένδειξη ότι η περιοχή αυτή είναι πιθανό να αποτελέσει συστάδα την επόμενη χρονιά. Το 1986, σύμφωνα με τα αρχεία που τηρούνται στην πολιτεία του Νέου Μεξικού, δεν υπήρξε κανένα νέο περιστατικό στην περιοχή άρα δεν αποτέλεσε συστάδα και η απότομη αύξηση του προηγούμενου έτους μάλλον οφειλόταν σε τυχαίους παράγοντες.

Το 1991 αναδύεται μια στατιστικά σημαντική συστάδα, που αποτελείται από την κομητεία Los Alamos όπου την περίοδο 1989 – 1981 είχαμε 7 νέα περιστατικά ενώ αναμενόταν μόλις 0,9 ,  $p - value = 0,02$  . Η ανάλυση τα επόμενα χρόνια έδειξε ότι υπήρχε μία σημαντική συστάδα σε αυτήν την περιοχή για την περίοδο 1989 – 1982, καθώς συνέβησαν ακόμη δύο περιστατικά, συνολικά 9, ενώ το αναμενόμενο πλήθος ήταν 1,2 ,  $p - value = 0,002$  . Αξίζει να σημειωθεί ότι παρόλο που τα p-value που δίνονται στον πίνακα έχουν υπολογιστεί για πολλά πιθανά μεγέθη και πιθανές θέσεις των συστάδων καθώς και για πολλά πιθανά μήκη χρονικών περιόδων , αποτελούν το p-value ενός και μόνο ελέγχου και δεν αφορούν τους πολλαπλούς, περιοδικούς ελέγχους που πραγματοποιήθηκαν για τα χρόνια αυτά.

Χρονική Περίοδος	Συστάδες με την Μεγαλύτερη Πιθανότητα	Διάρκεια Συστάδας	Συχνότητα	Αναμενόμενη Συχνότητα	Σχετικός Κίνδυνος	p-value
1973-1978	Bernalillo + 7 πολιτείες δυτικά	1975-1978	48	36	1,4	0,6
1973-1979	Los Alamos, Rio Arriba	1975-1979	9	3,3	2,7	0,58
1973-1980	Los Alamos, Rio Arriba	1975-1980	10	3,8	2,6	0,54
1973-1981	North Central Counties - San Miguel	1975-1981	72	53	1,4	0,19
1973-1982	North Central Counties - San Miguel	1975-1982	85	62	1,4	0,08
1973-1983	Bernalillo, Valencia	1975-1983	84	62	1,4	0,13
1973-1984	North Central Counties	1973-1984	113	90	1,3	0,14
1973-1985	Lincoln	1985	3	0,2	13,8	0,23
1973-1986	North Central Counties + Colfax, Harding	1973-1986	129	108	1,2	0,49
1973-1987	North Central Counties + Colfax, Harding	1973-1987	142	117	1,2	0,21
1973-1988	North Central Counties - San Miguel	1973-1988	143	115	1,2	0,08
1973-1989	North Central Counties + Colfax, Harding	1973-1989	165	134	1,2	0,06
1973-1990	North Central Counties, Rio Arriba, Santa Fe, Taos	1979-1990	41	22	1,8	0,06
1973-1991	Los Alamos	1989-1991	7	0,9	7,6	0,02
1973-1992	Los Alamos	1989-1992	9	1,2	7,6	0,002

**Πίνακας 3.2** Γεωγραφική ανάλυση περιστατικών καρκίνου του θυρεοειδούς σε άντρες στο Νέο Μεξικό, με χρήση χωρικής – χρονολογικής συνάρτησης σάρωσης [Πηγή: Kulldorf, (2000)].

### 3.5 Περιοδική ως προς τον χρόνο επιτήρηση.

Συνήθως οι νέες περιπτώσεις εισάγονται στο σύστημα καταγραφής μόλις συμβούν και για τους σκοπούς της βιοεπιτήρησης είναι σημαντικό να αναλύουμε τα δεδομένα σε περιοδικά διαστήματα. Επομένως θα πρέπει να προσαρμόζουμε τα δεδομένα για πολλαπλούς ελέγχους που γίνονται κατά τη διάρκεια της επιτήρησης. Η προσαρμογή πρέπει να γίνεται όχι μόνο για τις πολλές πιθανές θέσεις και τα πολλά μεγέθη των συστάδων στην επιφάνεια του χάρτη και στο χρόνο, αλλά και για τις αναλύσεις που προηγήθηκαν. Αυτό επιτυγχάνεται μέσω μιας μικρής τροποποίησης στη χωρική – χρονολογική μέθοδο σάρωσης.

Το ενδιαφέρον επικεντρώνεται και πάλι στις «ζωντανές» συστάδες. Η πιθανοφάνεια για τα πραγματικά δεδομένα ορίζεται, όπως και προηγουμένως, ως ο μέγιστος κύλινδρος από αυτούς που εξακολουθούν να υπάρχουν τη στιγμή που λήγει η παρακολούθηση. Για παράδειγμα θεωρούμε τους κυλίνδρους για τους οποίους  $Y_1 \leq s \leq t = Y_2$ . Για δεδομένα που προκύπτουν από τυχαία δειγματοληψία η πιθανοφάνεια ορίζεται ως ο μέγιστος από όλους τους κυλίνδρους που χρησιμοποιήθηκαν σε προηγούμενες αναλύσεις και από τους κυλίνδρους που υπάρχουν στην λήξη. Αυτό συμβαίνει για να προσαρμοστούν τα δεδομένα στις πολλαπλές αναλύσεις. Για παράδειγμα θεωρούμε όλους τους κυλίνδρους για τους οποίους  $Y_1 \leq s \leq t \leq Y_2$  και  $t \geq Y_m$ , όπου  $Y_m$  είναι η χρονιά στην οποία ξεκίνησε η επιτήρηση. Σε οποιαδήποτε χρονική στιγμή, μια παρατηρηθείσα συστάδα είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας  $\alpha$  αν η πιθανότητα να έχει ανιχνευτεί μία συστάδα με μεγαλύτερη πιθανοφάνεια στην τρέχουσα ή σε προηγούμενες αναλύσεις είναι το πολύ  $\alpha$ . Οι συστάδες με την μεγαλύτερη πιθανότητα να επιλεγούν θα είναι και πάλι οι ίδιες με αυτές στην προηγούμενη μέθοδο, μόνο που τα  $p$  – values θα είναι μεγαλύτερα λόγω της προσαρμογής στις περιοδικές αναλύσεις.

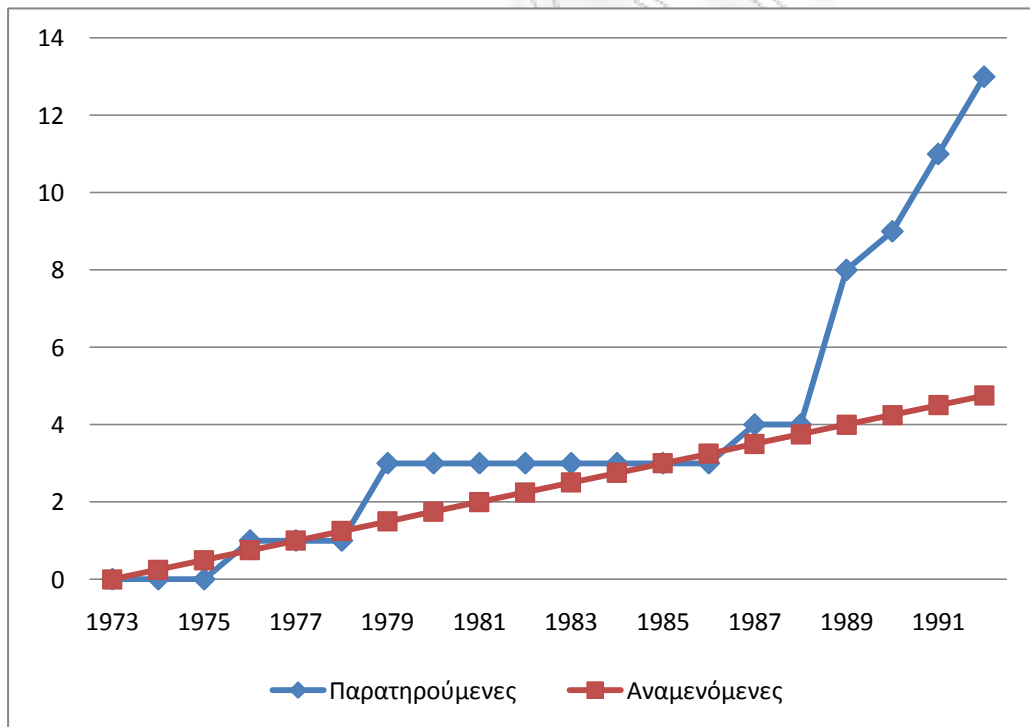
Τα αποτελέσματα από αυτού του είδους την ανάλυση για την περίπτωση του παραδείγματος με τον θυρεοειδή καρκίνο σε άντρες στο Νέο Μεξικό δίνονται στον Πίνακα 3.3 . Παρατηρούμε ότι για τα δεδομένα της περιόδου 1973-1991, με την νέα μέθοδο, η συστάδα που αποτελείται από το Los Alamos δεν είναι πια στατιστικά σημαντική,  $p$  – value = 0.13 . Αντίθετα, αν εισάγουμε στην ανάλυση τα δεδομένα και του επόμενου έτους, δηλαδή αν η ανάλυση γίνει για την περίοδο 1973-1992, η συστάδα αυτή είναι στατιστικά σημαντική με  $p$  – value = 0.016. Στον παρακάτω πίνακα δεν δίνονται και τα αποτελέσματα



για τις ετήσιες αναλύσεις έως και το 1991 καθώς είναι ίδια με αυτά του πίνακα 4.2 με εξαίρεση τα p-values που είναι μεγαλύτερα από πριν και άρα μη σημαντικά.

Χρονική Περίοδος	Συστάδες με την Μεγαλύτερη Πιθανότητα	Διάρκεια Συστάδας	Συχνότητα	Αναμενόμενη Συχνότητα	Σχετικός Κίνδυνος	p-value
1973-1991	Los Alamos	1989-1991	7	0,9	7,6	0,13
1973-1992	Los Alamos	1989-1992	9	1,2	7,6	0,016

**Πίνακας 3.3** Ανίχνευση συστάδων καρκίνου θυρεοειδούς σε άντρες στο Νέο Μεξικό. [Πηγή: Kulldorff (2000)]



**Διάγραμμα 3.1** Αθροιστικές Συχνότητες Εμφάνισης Καρκίνου του Θυρεοειδούς σε άντρες στο Los Alamos

Στο Διάγραμμα 3.1, δίνονται γραφικά οι παρατηρηθείσες και οι αναμενόμενες αθροιστικές συχνότητες εμφάνισης καρκίνου του θυρεοειδούς σε άντρες στην κομητεία του Los Alamos την περίοδο 1973 – 1992. Παρόλο που το 1989 παρατηρήθηκαν 4 νέα κρούσματα της νόσου η συστάδα δεν θεωρείται στατιστικά σημαντική. Όμως το 1992 ήταν πια εμφανές ότι η συστάδα που αποτελείται από την κομητεία Los Alamos είναι στατιστικά σημαντική.

Υπάρχει δηλαδή μία αύξηση του κινδύνου που δεν μπορεί να οφείλεται σε τυχαία αίτια αλλά σε ειδικούς παράγοντες μεταβλητότητας.

Για ένα σύστημα βιοεπιτήρησης που υπάρχει για πολύ μεγάλο χρονικό διάστημα η προσαρμογή της συμπερασματολογίας σε όλες τις προηγούμενες αναλύσεις, αποτελεί μία συντηρητική προσέγγιση. Ένας καλύτερος τρόπος θα ήταν να καθορίσουμε μια μέγιστη τιμή για την πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης  $\alpha$ , για κάποιο ορισμένο χρονικό διάστημα, για παράδειγμα 20 έτη για μία χρόνια πάθηση. Αυτό επιτυγχάνεται εύκολα, αν θεωρήσουμε μόνο τους κυλίνδρους των τελευταίων 20 ετών και υπολογίσουμε την πιθανοφάνεια της συστάδας με την μεγαλύτερη πιθανότητα επιλογής στα δεδομένα της τυχαίας δειγματοληψίας. Για παράδειγμα αν θεωρήσουμε τους κυλίνδρους για τους οποίους  $Y_1 \leq s \leq t$  και  $Y_2 - 20 < t \leq Y_2$ .

Όπως και στην περίπτωση των χρόνιων παθήσεων έτσι και με τα μεταδοτικά νοσήματα το ενδιαφέρον επικεντρώνεται στην ανίχνευση της αύξησης του κινδύνου τοπικά, το συντομότερο δυνατό. Ο προσδιορισμός της θέσης και του μεγέθους της γεωγραφικής συστάδας και του χρονικού διαστήματος που επήλθε η μεταβολή δεν πρέπει να γίνεται με ανάλυση εκ των προτέρων (a priori). Ειδικά για την περίπτωση των μεταδοτικών νοσημάτων οι συστάδες που εμφανίζουν αυξημένο κίνδυνο για μεγάλη χρονική διάρκεια δεν αξιολογούνται σημαντικά καθώς αυτές πιθανότατα αφορούν ομάδες υψηλού κινδύνου του πληθυσμού και δεν αποτελούν κίνδυνο για την δημόσια υγεία συνολικά. Παρόλο που μπορούμε να μελετήσουμε και τις δύο περιπτώσεις, ο σκοπός της βιοεπιτήρησης συνήθως αφορά τον συνολικό πληθυσμό και όχι ομάδες υψηλού κινδύνου. Για τέτοιου είδους διαδικασίες θα πρέπει να θεωρήσουμε μόνο εκείνους τους κυλίνδρους που έχουν μια προκαθορισμένη μέγιστη χρονική διάρκεια. Ανάλογα με το είδος της μεταδοτικής νόσου αυτή η διάρκεια μπορεί να είναι από λίγες μέρες έως μερικά χρόνια.

Ύστερα από τον εντοπισμό των συστάδων στις οποίες υπάρχουν ενδείξεις ότι αυξήθηκε η πιθανότητα εμφάνισης της νόσου θα πρέπει να αποφασίσουμε αν θα διερευνήσουμε περαιτέρω για την ύπαρξη ειδικών επιδράσεων που αυξάνουν τον κίνδυνο. Η απόφαση αυτή μπορεί να ληφθεί με κριτήριο τα  $p$  – values. Στην βιοεπιτήρηση μια τέτοια αυστηρή προσέγγιση δεν συνιστάται. Το p-value θα πρέπει να λειτουργεί ως ένδειξη ότι σε μία συγκεκριμένη περιοχή, συστάδα, υπάρχει αύξηση του κινδύνου. Η διερεύνηση για την ύπαρξη τυχαίων παραγόντων ή ειδικών αιτιών μεταβλητότητας θα πρέπει να γίνεται με βάση το πόσο ισχυρή είναι η ένδειξη αυτή. Όταν υπάρχουν ισχυρές ενδείξεις θα πρέπει να γίνεται μια λεπτομερής επιδημιολογική έρευνα, χρησιμοποιώντας το πλήρες ιστορικό γύρω από κάθε

κρούσμα ξεχωριστά. Πολύ σημαντικό όμως είναι να γνωρίζουμε ότι δεν οφείλονται όλες οι περιπτώσεις σε γεωγραφικούς παράγοντες αλλά κάποιες από αυτές θα προέκυπταν ανεξάρτητα από την αύξηση του κινδύνου στην περιοχή.

Η προοδευτική μέθοδος βιοεπιτήρησης με χρήση των τεχνικών scan δεν μπορεί να εντοπίσει συστάδες που ξεκίνησαν και έληξαν πριν την έναρξη της διαδικασίας. Παρόλο που πλέον δεν αποτελούν κίνδυνο για την δημόσια υγεία παρουσιάζουν σημαντικό ενδιαφέρον για τον προσδιορισμό των αιτιών που οδηγούν σε έξαρση του κινδύνου. Για μία τέτοια ανάλυση θα πρέπει να χρησιμοποιήσουμε την χωρική – χρονολογική συνάρτηση σάρωσης (*space-time scan statistic*) για ποικίλα μεγέθη των κυλίνδρων (Kulldorff et al. 1998a), η άλλες αναδρομικές μεθόδους βιοεπιτήρησης.

Εκτός από την παρούσα μέθοδο έχουν προταθεί και άλλες προοδευτικές μέθοδοι περιοδικής γεωγραφικής βιοεπιτήρησης, όπως του Rogerson (1997,2001) και Jarpe (1999). Αυτές επικεντρώνονται στον εντοπισμό συστάδων στις οποίες υπάρχει απότομη έξαρση του κινδύνου σε παγκόσμιο επίπεδο μέσω της παρακολούθησης μιας συγκεκριμένης γεωγραφικής περιοχής. Αντίθετα η μέθοδος που αναπτύξαμε παραπάνω έχει σκοπό τον σαφή γεωγραφικό προσδιορισμό της θέσης της συστάδας και την αξιολόγηση της σημαντικότητας των αποτελεσμάτων. Συνεπώς αυτές οι δύο μέθοδοι μπορούν να λειτουργήσουν παράλληλα συμπληρώνοντας η μία την άλλη και να εφαρμοστούν σε διαφόρων τύπων έρευνες.

Μόλις ανιχνεύσουμε μία συστάδα όπου υπάρχει σημαντική αύξηση στα ποσοστά εμφάνισης της νόσου η πρώτη ενέργεια που θα γίνει πρέπει να είναι ο έλεγχος για λάθη στα δεδομένα. Αυτά τα λάθη μπορούν να βρίσκονται στον αριθμητή ή στον παρονομαστή του ποσοστού ή στις γεωγραφικές συντεταγμένες. Στην έρευνα για την εμφάνιση καρκίνου του θυρεοειδούς σε άντρες στην πολιτεία του Νέου Μεξικού εντοπίστηκε μία στατιστικά σημαντική συστάδα το 1992 στην κομητεία Los Alamos, ενώ υπήρχαν ενδείξεις ήδη από το 1989. Σε αυτή την περίπτωση δεν εντοπίστηκαν λάθη στα δεδομένα. Η γραμματεία υγείας της πολιτείας εξέτασε προσεκτικά όλες τις περιπτώσεις από το 1970 και μετά. Η κατανομή του σταδίου της νόσου βρέθηκε ότι ήταν ίδια με την υπόλοιπη πολιτεία κάτι που οδήγησε στο συμπέρασμα ότι ο λόγος πίσω από την ξαφνική ανίχνευση της συστάδας ήταν οι καλύτερες διαγνωστικές μέθοδοι που εφαρμόζονταν στην περιοχή. Σε τρεις από τις 17 περιπτώσεις προέκυψε ότι οι ασθενείς είχαν υποβληθεί σε θεραπεία με ακτινοθεραπεία για παθήσεις του εγκεφάλου ή του αυχένα τις δεκαετίες του 1940 και 1950, κάτι που αποτελεί γνωστό παράγοντα του κινδύνου. Δύο από αυτές τις περιπτώσεις διαγνώστηκαν μετά το 1988. Τουλάχιστον οκτώ από τους 17 ασθενείς παρακολουθούνταν ήδη για έκθεση σε ραδιενέργεια

λόγω της εργασίας τους στο Εθνικό Ερευνητικό Κέντρο του Los Alamos (Los Alamos National Laboratory). Συγκεκριμένα 2 από αυτούς είχαν εκτεθεί σημαντικά σε ιονίζουσα ραδιενεργή ακτινοβολία στο πεδίο δοκιμών πυρηνικών όπλων, επίσης γνωστός παράγοντας αύξησης του κινδύνου. Και οι δύο διαγνώστηκαν μετά το 1988. Επομένως καταλήγουμε με ασφάλεια στο συμπέρασμα ότι η αύξηση του κινδύνου οφειλόταν σε γνωστούς παράγοντες που υπήρχαν ήδη στην περιοχή και που αποτελούν χαρακτηριστικό μόνο της συγκεκριμένης περιοχής σε ολόκληρη την πολιτεία.

Στην προοδευτική μέθοδο της βιοεπιτήρησης, η εξαγωγή συμπερασμάτων πρέπει να γίνεται λαμβάνοντας υπ' όψη το χρόνο επώασης της νόσου. Δηλαδή, στο προηγούμενο παράδειγμα πολλοί ασθενείς είχαν εκτεθεί στους παράγοντες κινδύνου πολλά χρόνια πριν εκδηλώσουν την νόσο. Ειδικά στην περίπτωση που οι παράγοντες υπάρχουν σε τοπικό επίπεδο, ο μεγάλος χρόνος επώασης της νόσου μειώνει την ισχύ της μεθόδου καθώς, μέσα σε αυτά τα χρόνια, άτομα που έχουν εκτεθεί στον παράγοντα μπορεί να έχουν μετακομίσει από την περιοχή ενώ άλλα άτομα από περιοχές που δεν υπάρχουν οι παράγοντες έχουν εισχωρήσει στο σύστημα. Επιπρόσθετα, στην μελέτη ασθενειών με μεγάλο χρόνο επώασης, όταν ανακαλύπτουμε μια στατιστικά σημαντική συστάδα θα πρέπει να ερευνούμε τους παράγοντες κινδύνου τόσο στο παρόν όσο και στο παρελθόν. Δεν θα πρέπει να αγνοούμε ότι ακόμα και αν ένας παράγοντας έχει ήδη απομακρυνθεί είναι πολύ πιθανό να συνεχίσουν να εμφανίζονται νέα κρούσματα. Πόσο μάλλον στην περίπτωση που ο παράγοντας ταυτοποιηθεί και απομακρυνθεί αποτέλεσμα της τρέχουσας έρευνας.

Η συνήθης επιλογή για το μέγιστο μέγεθος του κυκλικού παραθύρου ή της βάσης του κυλίνδρου είναι το ήμισυ του πληθυσμού σε κίνδυνο. Αυτό επιτρέπει την ανίχνευση και μικρών και μεγάλων συστάδων. Οι κύκλοι που περιλαμβάνουν περισσότερο από το μισό πληθυσμό σε κίνδυνο αφήνουν εκτός συστάδες με χαμηλότερο κίνδυνο από τον αναμενόμενο. Παρόλα αυτά η space-time scan statistic μπορεί να ανιχνεύσει και περιοχές με εξαιρετικά χαμηλά ποσοστά [βλ. *Kulldorff (1998b)*].

Τέλος η μέθοδος που αναπτύχθηκε παραπάνω αποτελεί ένα σημαντικό εργαλείο στην συστηματική περιοδική γεωγραφική βιοεπιτήρηση. Μπορεί να ανιχνεύσει αναδυόμενες συστάδες και να προσαρμοστεί σε πολλαπλούς ελέγχους που γίνονται στην πάροδο του χρόνου. Η μέθοδος εφαρμόζεται για διάφορα επίπεδα γεωγραφικής ή χρονολογικής συγκέντρωσης και για διάφορες μορφές ασθενειών. Επιπλέον, παρόλο που είναι ευαίσθητη σε υπολογισμούς, η space-time scan statistic δεν είναι ιδιαίτερα σύνθετη.

### 3.6 Η μέθοδος P-σάρωσης

#### 3.6.1 Εισαγωγή

Σε αυτή την παράγραφο θα αναπτύξουμε την μέθοδο P-scan για παράθυρο σάρωσης προκαθορισμένου μεγέθους. Η μέθοδος χρησιμοποιεί μια προσεγγιστική τεχνική στην περίπτωση των σταθερών ποσοστών για να εξάγει το  $p$  –  $value$  για την περίπτωση των μη σταθερών ποσοστών εμφάνισης που προ-υπάρχουν. Σε αντίθεση με το *GLRT*, η μέθοδος P-scan δεν χρησιμοποιεί προσομοίωση για τον υπολογισμό των κρίσιμων σημείων.

Για την περίπτωση που ο χρόνος είναι συνεχής μεταβλητή, καταγράφονται οι χρόνοι στους οποίους εκδηλώνονται τα περιστατικά της νόσου και σε κάθε χρονική στιγμή  $t$  ( $w \leq t \leq T$ ) γνωρίζουμε το πλήθος των περιστατικών που παρατηρήθηκαν  $Y_t(w)$  και το αναμενόμενο πλήθος περιστατικών  $E_t(w)$ , στο υποδιάστημα  $(t - w, t)$ .

Για τις περιπτώσεις των στρωματοποιημένων και των δίτιμων δεδομένων ο χρόνος διαμερίζεται σε  $T$  διαστήματα και τα δεδομένα που καταγράφουμε είναι το πλήθος των περιστατικών σε κάθε διάστημα. Στην πρώτη περίπτωση το πλήθος αυτό μπορεί να είναι οποιοσδήποτε φυσικός αριθμός ενώ στην δεύτερη περίπτωση παίρνει τις τιμές 0 ή 1, ανάλογα με το αν υπήρξε ή όχι κρούσμα στο διάστημα αυτό.

Έστω η στατιστική συνάρτηση σάρωσης:

$$S_w = \max_t \{Y_t(w)\}, \quad w \leq t \leq T.$$

Έστω επίσης για την περίπτωση προ-ύπαρξης σταθερού ποσοστού:

$$E_t(w) = \lambda, \text{ για κάθε } t$$

όπου  $\lambda$  μία σταθερά. Η πιθανότητα  $P[S_w \geq k | \lambda]$  είναι η p-value για ένα παρατηρούμενο μέγεθος συστάδας  $S_w = k$ . Αν συμβολίσουμε αυτήν την p-value ως  $P(k; \lambda, w, T)$  τότε για την περίπτωση του μη σταθερού ποσοστού εμφάνισης για κάθε παρατηρούμενη  $Y_t(w)$  υπολογίζουμε το  $P(Y_t(w); E_t(w), w, T)$  και συντεταγμένα γράφουμε  $P(Y_t; E_t, w, T)$ . Η στατιστική συνάρτηση P-σάρωσης (PSS) είναι:

$$PSS = \min_t \{P(Y_t; E_t, w, T)\}.$$

Η μηδενική υπόθεση της ύπαρξης μη σταθερού ποσοστού εμφάνισης της νόσου πριν την έναρξη της επιτήρησης με την μέθοδο P-σάρωσης απορρίπτεται αν  $PSS < \alpha$ . Επομένως, αν βρεθεί τουλάχιστον μία τιμή του  $t$  τέτοια ώστε:

$$P(Y_t; E_t, w, T) \leq \alpha$$

απορρίπτουμε τη μηδενική υπόθεση.

Εφαρμόζοντας την μέθοδο P-σάρωσης μπορούμε να υπολογίζουμε τις αναμενόμενες τιμές με διάφορους τρόπους, ανάλογα με τον τρόπο που ελήφθησαν τα δεδομένα για το προ-υπάρχον ποσοστό. Αν τα ποσοστά αυτά είναι διαθέσιμα σε κλίμακα μικρότερη του  $w$ , για παράδειγμα αν οι αναμενόμενες τιμές δίνονται σε ημερήσια βάση και  $w = 7$ , τότε δεν υπάρχει πρόβλημα με τον υπολογισμό του  $E_t(w)$ . Αν τα ποσοστά είναι διαθέσιμα στην ίδια κλίμακα με το  $w$  τότε το  $E_t(w)$  δίνεται από το μέσο ποσοστό σε αυτή την θέση του παραθύρου. Για παράδειγμα αν τα ποσοστά είναι διαθέσιμα σε επίπεδο ημερολογιακής εβδομάδας και  $w = 7$ , τότε το  $E_t(w)$  προκύπτει ως το μέσο ποσοστό για τις 2 ημερολογιακές εβδομάδες που εσωκλείει το κινητό παράθυρο.

Ο έλεγχος P-σάρωσης, όπως και η παραδοσιακή μέθοδος σάρωσης και το *GLRT*, βασίζεται στην συνάρτηση μιας μεταβλητής που παίρνει ακέραιες τιμές, την  $Y_t(w)$ . Τα κρίσιμα σημεία της στατιστικής συνάρτησης μπορούν να επιλεγούν κατά τέτοιο τρόπο ώστε το επίπεδο σημαντικότητας να είναι μικρότερο του  $\alpha$ , τόσο για την μηδενική υπόθεση των σταθερών όσο και για την μηδενική υπόθεση των μη σταθερών ποσοστών της νόσου που υπάρχουν πριν από την έναρξη της βιοεπιτήρησης. Αυτή η διαδικασία είναι συνηθισμένη σε πολλούς ελέγχους για διακριτές τυχαίες μεταβλητές, όπου η τυχαιοποίηση μπορεί να χρησιμοποιηθεί για να πετύχουμε ένα ποσοστό εσφαλμένων συναγερμών ίσο με  $\alpha$ .

Θεωρούμε την συνάρτηση:

$$f = \frac{\alpha - P(Y_t + 1; E_t, w, T)}{P(Y_t; E_t, w, T) - P(Y_t + 1; E_t, w, T)}$$

Και έστω  $h$  ένας τυχαίος αριθμός μεταξύ του 0 και του 1. Ο τυχαιοποιημένος έλεγχος P-σάρωσης δίνει συναγερομό αν σε οποιαδήποτε χρονική στιγμή  $t$  ισχύει:

$$P(Y_t; E_t, w, T) \leq \alpha \text{ ή } P(Y_t + 1; E_t, w, T) \leq \alpha \leq P(Y_t; E_t, w, T) \text{ ή } f \geq h.$$

Οι τυχαιοποιημένες διαδικασίες είναι χρήσιμες για θεωρητικές συγκρίσεις και επιτρέπουν να επιτύχουμε το επιθυμητό ποσοστό εσφαλμένων συναγερομών. Δυστυχώς όμως αν ο ίδιος έλεγχος στα ίδια δεδομένα εφαρμοστεί από δύο διαφορετικούς αναλυτές μπορεί να δώσει διαφορετικά αποτελέσματα.

Μία εναλλακτική διαδικασία, για την περίπτωση των μη σταθερών ποσοστών, λιγότερο συντηρητική από τους μη-τυχαιοποιημένους ελέγχους βασίζεται στα διάμεσα  $p$ -values (*mid-p-values*). Σ' αυτήν την περίπτωση απορρίπτουμε την μηδενική υπόθεση αν σε οποιαδήποτε χρονική στιγμή  $t$ :

$$f \geq 0.5$$

ή ισοδύναμα αν

$$\frac{1}{2} [P(Y_t; E_t, w, T) + P(Y_t + 1; E_t, w, T)] < \alpha.$$

Ως *mid-p-value* ορίζουμε το  $\min_t \left\{ \frac{1}{2} [P(Y_t; E_t, w, T) + P(Y_t + 1; E_t, w, T)] \right\}$ .

Όταν κάνουμε την ανάλυση μπορεί εύκολα να θεωρήσουμε αυτά τα  $p$ -values που εξαρτώνται από τα δεδομένα ως «παρατηρούμενα επίπεδα σημαντικότητας», [βλ. Hubbard & Byarri, (2003)]. Ένα καλό παράδειγμα που μια τέτοια παρερμηνεία είναι ακατάλληλη είναι η μέθοδος P-σάρωσης για την περίπτωση των μη σταθερών ποσοστών. Συγκεκριμένα, στην εφαρμογή προοδευτικών μεθόδων βιοεπιτήρησης, είναι καλύτερα να θεωρούμε την PSS και την  $P(Y_t; E_t, w, T)$  σαν στατιστικές συναρτήσεις ενός ελέγχου με προκαθορισμένο επίπεδο σημαντικότητας.

Ένα τοπικό κέντρο ελέγχου νοσημάτων μπορεί να λαμβάνει αναφορές για εκατοντάδες διαφορετικούς ιούς ή ασθένειες γενικότερα, να παρακολουθεί την κάθε μία ξεχωριστά και να έχει τα χρήματα να κάνει ένα συγκεκριμένο αριθμό ελέγχων μετά από συναγερομούς σε μία λογιστική περίοδο. Προκαθορίζοντας ένα μικρό επίπεδο σημαντικότητας,  $\alpha$ , για κάθε ένα από τους  $N$  ιούς, αναμένεται να ερευνήσουν περίπου  $N \times \alpha$  εσφαλμένους συναγερομούς. Ομοίως,

το Υπουργείο Υγείας μπορεί να εκτιμήσει το συνολικό αριθμό από έρευνες που θα γίνουν μετά από συναγερούς, ανά περιοχή όπου υπάρχει κέντρο ελέγχου νοσημάτων.

### 3.6.2 Αναδρομική ανάλυση , συνεχής χρόνος

Στην αναδρομική περίπτωση οι χρόνοι από  $N$  περιστατικά είναι ανεξάρτητες τυχαίες μεταβλητές, ομοιόμορφα κατανεμημένες στο  $(0, T)$ . Για αυτή την περίπτωση η βιβλιογραφία γύρω από τη θεωρία και τις εφαρμογές της μεθόδου σάρωσης με παράθυρο μεγέθους  $w$ , είναι ιδιαίτερα ανεπτυγμένη. Στην παρούσα παράγραφο θα δούμε την εφαρμογή της μεθόδου P-σάρωσης.

Η μέθοδος P-σάρωσης απορρίπτει την μηδενική υπόθεση αν  $w < t \leq T$ , σε οποιοδήποτε  $t$ . Είδαμε στην προηγούμενη παράγραφο ότι η εξίσωση υπολογισμού της  $p$ -value ισχύει για οποιαδήποτε χρονική στιγμή  $t$ . Έτσι αν θέλουμε να ελέγξουμε αν ισχύει για μια συγκεκριμένη στιγμή  $t^*$ , πρέπει να υπολογίσουμε το  $P(Y_{t^*}; E_{t^*}, w, T)$ . Για την περίπτωση του σταθερού ποσοστού το αναμενόμενο πλήθος των συμβάντων στο  $(t^* - w, t^*)$  είναι  $Nw/T$ . Για την περίπτωση του μη σταθερού ποσοστού χρησιμοποιούμε πληροφορίες γύρω από τη διακύμανση του πληθυσμού σε κίνδυνο ή τις επιδράσεις της εποχικότητας και αυξάνουμε ή μειώνουμε το λόγο  $Nw/T$  ώστε το πλήθος των περιστατικών στο  $(t^* - w, t^*)$  να είναι  $E_{t^*}$ . Στην αναδρομική περίπτωση, τώρα, η πιθανότητα  $P(Y_{t^*}; E_{t^*}, w, T)$  υπολογίζεται χρησιμοποιώντας την προσέγγιση των Wallenstein – Neff [βλ. Wallenstein & Neff (1987)] :

$$P_R(k; \lambda, w) \approx \left[ (k - \lambda) \left( \frac{T}{w} \right) + 1 \right] b \left( k; \frac{\lambda T}{w}, \frac{w}{T} \right) + 2 \sum_{i=k+1}^N b \left( i; \frac{\lambda T}{w}, \frac{w}{T} \right) \quad (3.1)$$

όπου  $b(k; M, p) = \binom{M}{k} p^k (1 - p)^{M-k}$ .

### 3.6.3 Παράδειγμα

Μία μέθοδος σάρωσης με παράθυρο 30 ημερών εφαρμόστηκε για να ελεγχθεί αν υπάρχει κάποια ασυνήθιστη συστάδα, φλεγμονώδους πάθησης του εντέρου (*Inflammatory Bowel Disease, IBD*), σε επαρχίες της Σουηδίας. Χρησιμοποιήθηκε για την σάρωση η μέθοδος μεγέθυνσης χρόνου του Weinstock (*Weinstock's method of stretching time*) [βλ. Weinstock, (1981)]. Σε μία συγκεκριμένη επαρχία καταγράφηκαν  $N = 194$  περιπτώσεις μέσα σε 13 χρόνια ( $T = 4748$  ημέρες). Μια συστάδα με 8 περιπτώσεις ανιχνεύτηκε σε μία περίοδο 30 ημερών. Για να ελέγξουμε την σημαντικότητα αυτής της συστάδας πρέπει πρώτα να



γνωρίζουμε το προ-υπάρχον ποσοστό εμφάνισης της νόσου για την περίοδο των 30 ημερών που εμφανίστηκε η συστάδα.

Για παράδειγμα, αν ο αριθμός των περιστατικών την περίοδο των 30 ημερών της συστάδας ήταν 4% παραπάνω από τον μέσο αριθμό γεννήσεων μιας οποιασδήποτε περιόδου 30 ημερών μέσα στα 13 χρόνια που διήρκεσε η μελέτη, δηλ. αν  $d = 1,04$ , θα θεωρήσουμε ότι ο αναμενόμενος αριθμός των περιστατικών της νόσου είναι 1,04 φορές μεγαλύτερος από το μέσο πλήθος περιστατικών ανά 30 ημέρες:

$$E_t(30) = d(Nw/T) = 1,04 \left( \frac{194 \times 30}{4748} \right) = 1275$$

Δηλαδή ο αναμενόμενος αριθμός των περιστατικών της νόσου για την συστάδα είναι 1275 περιστατικά.

Είναι από την εξίσωση 3.1:

$$P_R(8; 1275, 30, 4748) = 0,047$$

Άρα σε επίπεδο σημαντικότητας  $\alpha = 0.05$  δεν μπορούμε να αποδεχτούμε την μηδενική υπόθεση με βάση τα στοιχεία αυτής της συστάδας.

Για μία αύξηση 10 % πάνω από το μέσο, είναι  $P_R(8; 1.348, 30, 4748) = 0.0672$ ,  $P_R(9; 1.348, 30, 4748) = 0.0112$  και το *mid - p - value* είναι  $0.039 < 0.05$  οπότε πάλι δεν μπορούμε να αποδεχτούμε την μηδενική υπόθεση. Εάν πάλι αυτή η αύξηση ήταν 15% τότε οι αντίστοιχες τιμές είναι 0.091 και 0.016 με *mid - p - value* είναι  $0.054 > 0.05$  και δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση.

### 3.6.4 Προοδευτική ανάλυση, συνεχής χρόνος.

Στην προοδευτική ανάλυση, όταν ο χρόνος θεωρείται συνεχής μεταβλητή, τα δεδομένα προέρχονται από μία μη ομογενή διαδικασία Poisson. Η στατιστική συνάρτηση σάρωσης απορρίπτει την μηδενική υπόθεση αν για οποιοδήποτε  $t, t \in (w, T]$  ισχύει  $P(Y_t; E_t, w, T) \leq \alpha$  και δίνεται προσεγγιστικά [βλ. Alm (1983)] από την σχέση:

$$P_{PC}(Y_t; E_t, w, T) \approx 1 - F(Y_t - 1; E_t) \exp \left\{ - \frac{(Y_t - E_t)E_t}{Y_t} \times \frac{T - w}{w} p[Y_t - 1; E_t] \right\}$$

όπου  $p(r; \lambda) = e^{-\lambda} \lambda^r / r!$  και  $F(k; \lambda) = \sum_{r \leq k} p(r; \lambda)$ .

### 3.6.5 Προοδευτική Ανάλυση, ομαδοποιημένα δεδομένα

Στην περίπτωση των ομαδοποιημένων δεδομένων, σαρώνουμε τα τελευταία  $w$  χρονικά διαστήματα και το πλήθος των περιστατικών ανά χρονικό διάστημα ακολουθεί την κατανομή Poisson. Όταν το ποσοστό εμφάνισης της νόσου πριν την έναρξη της βιοεπιτήρησης είναι σταθερό και κάνουμε προοδευτική ανάλυση ο αριθμός των περιστατικών σε  $T$  χρονικά διαστήματα είναι  $T$  ανεξάρτητες ισόνομες τυχαίες μεταβλητές από την κατανομή Poisson με μέση τιμή  $\lambda/w$ . Η συνάρτηση ελέγχου στην περίπτωση αυτή δίνεται προσεγγιστικά [βλ. Naus & Wallenstein, (2006)] από τον τύπο:

$$P_{PG}(k; \lambda, w, T) \approx 1 - F(k-1; \lambda) \times \left( \frac{\sum_{j=0}^{k-1} F^2\left(k-j-1, \frac{\lambda}{w}\right) p\left[j, \frac{(w-1)\lambda}{w}\right]}{F(k-1; \lambda)} \right)^{T-w}$$

όπου  $p(r; \lambda) = e^{-\lambda} \lambda^r / r!$  και  $F(k; \lambda) = \sum_{r \leq k} p(r; \lambda)$ .

Η μηδενική υπόθεση απορρίπτεται όταν  $P_{PG}(k; \lambda, w, T) \leq \alpha$ .

### 3.6.6 Παράδειγμα

Στο Εθνικό Σύστημα Παρακολούθησης Σημαντικών Ασθενειών των Η.Π.Α. (*National Notifiable Disease Surveillance System*) καταγράφονται περιστατικά που προκαλούνται από μικροβιακούς παράγοντες που είναι πιθανό να σχετίζονται με την βιο-τρομοκρατία. Οι πιο συχνοί παράγοντες είναι η βρουκέλλωση ή αλλιώς μελιτοκοκκίαση (*brucellosis*) και η tularemia [βλ. Chang et al. (2003)]. Πολλές περιπτώσεις της μελιτοκοκκίασης στον άνθρωπο σχετίζονται με την κατανάλωση μολυσμένων μη παστεριωμένων γαλακτοκομικών προϊόντων και με την εργασία στην επεξεργασία κρέατος. Θα εφαρμόσουμε την μέθοδο προοδευτικής παρακολούθησης της brucellosis για το έτος 2002. Η εκτίμηση του εβδομαδιαίου μέσου ποσοστού εμφάνισης έγινε με δεδομένα από τα προηγούμενα πέντε χρόνια και συμπεριλήφθη η επίδραση της εποχικότητας. Κάθε εβδομάδα το Κέντρο Ελέγχου Νοσημάτων δίνει το πλήθος των περιπτώσεων που καταγράφηκαν τις 4 προηγούμενες εβδομάδες για τα επιλεγμένα μικρόβια. Για το λόγο αυτό επιλέγουμε παράθυρο 4 εβδομάδων που ξεκινά κάθε εβδομάδα. Η διαδικασία βιοεπιτήρησης ξεκίνησε την 1<sup>η</sup> Ιανουαρίου 2002 και διαρκεί 52 εβδομάδες. Επιθυμούμε να επιτύχουμε ποσοστό εσφαλμένων συναγεργμών μικρότερο ή ίσο

του 0.05 ανά μικρόβιο. Παρατηρούμε την 22<sup>η</sup> εβδομάδα ένα παράθυρο τεσσάρων εβδομάδων με 21 περιστατικά ενώ αναμέναμε μόλις 8 περιστατικά με βάση τα προ-υπάρχοντα ποσοστά της πενταετίας. Εφαρμόζοντας τις σχέσεις των προηγούμενων παραγράφων έχουμε:

$$P_{PG}(21; 8, 4, 52) \approx 1 - 0.999906(0.999828/0.999906)^{48} = 0.038$$

και

$$P_{PC}(21; 8, 4, 52) \approx 0.0095$$

### 3.6.7 Προοδευτική ανάλυση, δίτιμα δεδομένα

Στην απλή περίπτωση, υπό την υπόθεση σταθερού ποσοστού εμφάνισης, υπάρχουν  $T$  ανεξάρτητες δίτιμες δοκιμές με σταθερή πιθανότητα επιτυχίας σε κάθε μία. Εδώ, ως επιτυχία θεωρούμε την εμφάνιση ενός περιστατικού της νόσου και έστω  $p$  η πιθανότητα αυτή. Η πιθανότητα να υπάρχει μια υπο-ακολουθία από  $w$  περιστατικά με τουλάχιστον  $k$  επιτυχίες δίνεται προσεγγιστικά από την σχέση [βλ. *Wallenstein et al.* (2003)] :

$$P_{PB}(k|w, w, T) \approx 1 - \left[ C(D/C)^{\frac{T}{w} - 2} \right]$$

$$C = 2 \sum_{i < k} b(i|w, p) - 1 - (k - 1 - wp)b(k|w, p)$$

$$D = 2 \sum_{i < k} b(i|w, p) - 1 - (2k - 1 - 2wp)b(k|w, p)$$

### 3.6.8 Παράδειγμα

Μετά από εκπαίδευση ένας παιδίατρος καρδιοχειρουργός έκανε εγχειρήσεις σε νεογνά με σκοπό την απόφραξη αρτηριών. [βλ. *deLaval et al.* (1994)]. Στις πρώτες 52 επεμβάσεις συνέβη μόνο ένας θάνατος ποσοστό χαμηλότερο από τους 3.5 θανάτους που αναμένονταν με βάση μελέτες «control» που είχαν γίνει ήδη. Στις επόμενες 16 επεμβάσεις απεβίωσαν 7 νεογνά και συγκεκριμένα στις επεμβάσεις 53, 55, 59, 63, 64, 67 και 68. Μετά την 55<sup>η</sup> και την 64<sup>η</sup> επέμβαση ο γιατρός επισκέφτηκε ένα άλλο νοσοκομείο για να δει τις διαδικασίες που τηρούνται εκεί και μετά την 68<sup>η</sup> επέμβαση επανεκπαιδεύτηκε. Μετά την επανεκπαίδευσή του συνέβη ένας μόνο θάνατος σε 36 επεμβάσεις που ακολούθησαν. Ο χειρουργός και οι συνεργάτες του εφάρμοσαν διαδικασία ελέγχου τύπου *CUSUM* στα δεδομένα και

παρατήρησαν ότι αν ο μηχανισμός συνεχούς παρακολούθησης είχε τεθεί σε λειτουργία τότε ανεπιθύμητες τάσεις και η ανάγκη αλλαγής στο πρωτόκολλο θα είχαν ανιχνευτεί νωρίτερα.

Ο ιατρός αρχικά εκτίμησε με βάση τις 50 πρώτες επεμβάσεις ότι το εντός ελέγχου αναμενόμενο ποσοστό θνησιμότητας είναι 2% δηλ.  $E_t = w(0.02)$ . Ένα σχέδιο παρακολούθησης με μεθόδους σάρωσης επιλέχτηκε για να προειδοποιεί για απότομες αυξήσεις στο ποσοστό θνησιμότητας. Ως απότομη αύξηση εννοούμε μια συστάδα τριών θανάτων σε ένα ασυνήθιστα μικρό δείγμα διαδοχικών επεμβάσεων. Για παράδειγμα η μέθοδος μπορεί να σχεδιαστεί ώστε, για αναμενόμενο ποσοστό θνησιμότητας 0.02, η πιθανότητα εσφαλμένου συναγερμού να είναι 0.05 για τις επόμενες  $T = 100$  επεμβάσεις. Στο συγκεκριμένο παράδειγμα 100 επεμβάσεις γίνονται κάθε 5 χρόνια. Παρατηρούμε ότι για  $w = 15$  προκύπτει  $P(3|0.3,15,100) \approx 0.046$  επομένως επιλέγουμε αυτή την τιμή για το  $w$ . Αυτό σημαίνει ότι η πιθανότητα το εντός ελέγχου μήκος ροής μέχρι τον πρώτο συναγερμό να είναι μεγαλύτερο του 100 είναι 0.954.

Ομοίως, μπορούμε να υπολογίσουμε την κατανομή για το εκτός ελέγχου μήκος ροής. Αν το εκτός ελέγχου ποσοστό θνησιμότητας είναι 0.12 η πιθανότητα να έχουμε ένα συναγερμό στις επόμενες 100 επεμβάσεις είναι  $P(3|1.8, 15, 100) \approx 0.96$ . Αν το ποσοστό αυτό αυξηθεί στο 20% τότε η πιθανότητα συναγερμού στις πρώτες 15 επεμβάσεις είναι 0.6 και η πιθανότητα συναγερμού στις πρώτες 30 επεμβάσεις είναι 0.89.

Αν εφαρμόσουμε αυτόν τον έλεγχο μετά το θάνατο του ασθενούς στην 59<sup>η</sup> επέμβαση, θα υπάρχουν  $k = 3$  θάνατοι στις τελευταίες 15 επεμβάσεις ενώ ο αναμενόμενος αριθμός θανάτων είναι μόλις 0.3 και θα έχουμε ένδειξη ότι η διαδικασία είναι εκτός ελέγχου. Φυσικά αυτή η ανάλυση υποθέτει ότι η πιθανότητα επιβίωσης ανά ασθενή ήταν σε φυσιολογικά επίπεδα στις 15 τελευταίες επεμβάσεις. Με άλλα λόγια υποθέτουμε ότι κανένας από τους 15 τελευταίους ασθενείς δεν παρουσίαζε αυξημένη έκθεση σε κάποιο παράγοντα κινδύνου όπως π.χ. ηλικία, βάρος, ανατομία της στεφανιαίας αρτηρίας κλπ. Τη λύση σε αυτό το πρόβλημα δίνει η μέθοδος P-σάρωσης καθώς υπολογίζει τον κίνδυνο στην αξιολόγηση των συστάδων. Για κάθε ασθενή εκτιμούμε την πιθανότητα θνησιμότητας με βάση προηγούμενες μελέτες (*control studies*). Το αναμενόμενο ποσοστό θνησιμότητας για το παράθυρο των 15 επεμβάσεων προκύπτει από το άθροισμα των πιθανοτήτων θνησιμότητας για τους 15 ασθενείς. Έστω ότι για τις επεμβάσεις 45 – 59 η μέση αναμενόμενη πιθανότητα θνησιμότητας είναι 0.04 ανά ασθενή άρα ο αναμενόμενος αριθμός θανάτων θα είναι 0.6. Ο αριθμός των θανάτων ήταν 3. Με την μέθοδο P-σάρωσης υπολογίζουμε την πιθανότητα  $P(3|0.6,15,100) \approx 0.25$  και η  $mid - p - value = 0.15 > 0.05$  επομένως σ' αυτό το σημείο

δεν θα είχαμε ένδειξη εκτός ελέγχου διεργασίας. Στην 63<sup>η</sup> επέμβαση, θα είχαμε 4 θανάτους στο παράθυρο των 15 επεμβάσεων και η πιθανότητα  $P(4|0.6,15,100) \approx 0.045$  θα έδινε ένδειξη ότι η διαδικασία είναι εκτός ελέγχου.

Η μέθοδος P-σάρωσης που προτάθηκε από τους *Naus* και *Wallenstein* έχει σχεδιαστεί ώστε να είναι ευαίσθητη σε απότομες αυξήσεις και επικεντρώνεται σε ασυνήθιστες συστάδες. Μπορεί να συνδυαστεί με τα διαγράμματα τύπου *CUSUM* ώστε να ανιχνεύονται βαθμιαίες αυξήσεις του κινδύνου σε μεγάλες χρονικές περιόδους. Αυτό είναι σύμφωνο με την πρόταση να θεωρούμε ότι μία διαδικασία είναι εκτός ελέγχου όταν υπάρχει μία ακραία παρατήρηση (εκτός των ορίων 3-σ ) ή αν υπάρχουν  $k$  από  $n$  διαδοχικά σημεία έξω από τα όρια 2-σ [βλ. *Page*, (1954)]

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΡΔΑΛΙΑ

## ΚΕΦΑΛΑΙΟ 4

### Χρονολογικές Σειρές

#### 4.1 Εισαγωγή

Όπως έχουμε δει και παραπάνω, στη σύγχρονη εποχή, η έγκαιρη ανίχνευση μίας τρομοκρατικής ενέργειας με βιολογικά όπλα ή η πρόβλεψη μιας επιδημίας έχει τεράστια σημασία. Ασθένειες όπως ο ιός *SARS*, ο *H1N1*, ο ιός του Δυτικού Νείλου, η γρίπη κλπ, εξαπλώνονται με εκθετικό ρυθμό στον πληθυσμό. Είναι επομένως πρόκληση να κατασκευάσουμε ισχυρά μοντέλα για την πρόωρη ανίχνευση τέτοιων εκθετικά αυξανόμενων διαδικασιών. Όπως αναπτύξαμε στο προηγούμενο κεφάλαιο, οι χωρικές-χρονολογικές μέθοδοι είναι πολλές φορές πιο ισχυρές στην ανίχνευση μιας αλλαγής από τις καθαρά χρονολογικές. Δυστυχώς όμως τα χωρικά δεδομένα είναι συχνά λιγότερα από αυτά που χρειαζόμαστε. Επιπλέον πολλοί οργανισμοί που ασχολούνται με την δημόσια υγεία καταγράφουν τα δεδομένα με την μορφή χρονολογικών σειρών, για παράδειγμα τις επισκέψεις στα επείγοντα περιστατικά των νοσοκομείων, τις κλήσεις σε τηλεφωνικές γραμμές ιατρικής πληροφόρησης, τις πωλήσεις φαρμάκων, τις απουσίες σε σχολεία κλπ. Αυτές οι πηγές παρέχουν την έγκαιρη προειδοποίηση για το ξέσπασμα μιας επιδημίας. Γι αυτό οι καθαρά χρονολογικές μέθοδοι βιοεπιτήρησης είναι ακόμα χρήσιμες. Ένα από τα πιο συχνά χρησιμοποιούμενα μοντέλα είναι η διαδικασία αυτοπαλίνδρομου κινητού μέσου (*Autoregressive Moving Average, ARMA*), ενώ συχνά εφαρμόζονται και μοντέλα αυτοπαλίνδρομου ολοκληρωμένου κινητού μέσου (*Autoregressive Integrated Moving Average, ARIMA*). Η βιβλιογραφία γύρω από τα μοντέλα *ARMA* στην βιοεπιτήρηση είναι μεγάλη π.χ. *Reis & Mandl* (2003), *Reis et al.* (2003), *Earnest et al.* (2005) και *Lai* (2005). Παρά την δημοτικότητα τους τα *AR* και *ARMA* μοντέλα μειονεκτούν γιατί χρησιμοποιούνται ως εμπειρικά μοντέλα χωρίς να δίνεται κάποια ιδιαίτερη φυσική ερμηνεία στους συντελεστές τους.

Ένας ακόμη λόγος για την εφαρμογή των μοντέλων χρονολογικών σειρών στην βιοεπιτήρηση είναι η παρουσία αυτοσυσχέτισης. Οι κλασσικές μέθοδοι Στατιστικού Ελέγχου Διεργασίας υποθέτουν ότι οι παρατηρήσεις αποτελούν ανεξάρτητες πραγματοποιήσεις της υπό μελέτη διεργασίας. Η απόδοση των διαγραμμάτων ελέγχου σε συσχετισμένα δεδομένα που προέρχονται από *ARMA* ή *ARIMA* μοντέλα έχει γίνει αντικείμενο μελέτης σε πολλές

έρευνες. Στάσιμες διαδικασίες με συσχέτιση ανάμεσα στις παρατηρήσεις που μειώνεται γρήγορα μπορούν να παρακολουθούνται από διαγράμματα Shewhart καθώς και διαγράμματα τύπου *CUSUM* ή *EWMA* με κατάλληλες τροποποιήσεις. Ενδεικτικά αναφέρουμε εργασίες των *Montgomery & Mastrangelo* (1991) και των *Wardell, Moskowitz & Plante* (1992).

Πολλές χρονολογικές σειρές έχουν συσχέτιση μεταξύ παρατηρήσεων που απέχουν πολύ μεταξύ τους η οποία μηδενίζεται πολύ αργά. Τέτοιες διαδικασίες καλούνται μακροχρόνια εξαρτημένες (*long range dependent*) ή επίμονες (*persistent*). Αυτές οι διαδικασίες μπορεί να είναι τελικώς στάσιμες ακόμη και όταν παρουσιάζουν προφανώς μη στάσιμη συμπεριφορά σε μικρές περιόδους. Συνεπώς αποτελούν γέφυρα ανάμεσα στις *ARMA* και *ARIMA*. Οι διαδικασίες αυτοπαλίνδρομου κλασματικά ενσωματωμένου κινητού μέσου (*Autoregressive fractionally integrated moving average, ARFIMA*) με παράμετρο  $d$  μπορούν να περιγράψουν αυτή τη μακροχρόνια συμπεριφορά. Μια διαδικασία *ARFIMA* είναι στάσιμη και αντιστρέψιμη για τιμές τις παραμέτρου μεταξύ του  $-0.5$  και του  $0.5$  [βλ. *Beran* (1994)].

Στην βιβλιογραφία υπάρχουν πολλές αναφορές για στάσιμες συσχετισμένες διαδικασίες με μικρή μνήμη τύπου *ARMA* [βλ. *Montgomery & Mastrangelo* (1991), *Wardell et al.* (1992), (1994)] και για διαγράμματα ελέγχου σε μη στάσιμες *ARIMA* διαδικασίες [βλ. *Vander Wiel* (1996)]. Λόγω της χαμηλής απόδοσης των διαγραμμάτων σφαλμάτων τύπου *EWMA* έχει προταθεί ένα νέο είδος διαγραμμάτων. Τα διαγράμματα υπερβολικά σταθμισμένου κινητού μέσου (*Hyperbolically Weighted Moving Average, HWMA*) έχουν σχεδιαστεί για δεδομένα από διαδικασίες με μακροχρόνια εξάρτηση. Η απόδοση των τελευταίων στην ανίχνευση αλλαγών σε διαδικασίες με μεγάλη μνήμη είναι καλύτερη από την απόδοση των *EWMA*. Επιπλέον η ανίχνευση αλλαγών σε διαδικασίες με μη μακροχρόνια εξάρτηση είναι καλύτερη στην περίπτωση των *HWMA* αντί των *ARMA*.

## 4.2 Τα μοντέλα SIR και η εφαρμογή τους στην έγκαιρη ανίχνευση

Είτε μιλάμε για μία τρομοκρατική επίθεση με βιολογικούς, χημικούς παράγοντες είτε για μια επιδημία, το νόσημα που μελετάμε έχει την ιδιότητα να μεταδίδεται με ταχύτητα από άνθρωπο σε άνθρωπο. Οι ασθενείς ή οι άνθρωποι που πρόκειται να νοσήσουν χαρακτηρίζονται από ποικίλους παράγοντες, όπως η ηλικία, η κοινωνική θέση, το εισόδημα, η οικογενειακή κατάσταση, το επάγγελμα, η σωματική ή ψυχική υγεία, ο τόπος κατοικίας κλπ. που συχνά διαφέρουν από ασθενή σε ασθενή. Προκειμένου να μοντελοποιήσουμε την πρόοδο μιας επιδημίας σε ένα μεγάλο πληθυσμό πρέπει να ελαχιστοποιήσουμε αυτές τις



διαφορετικότητες σε μερικά χαρακτηριστικά γνωρίσματα σχετικά με την υπό μελέτη νόσο. Οι παιδικές ασθένειες, για παράδειγμα, έχουν το γνώρισμα να προσφέρουν ανοσία σε όποιον έχει ασθενήσει και αναρρώσει, επομένως έχει νόημα να χωρίζουμε τον πληθυσμό σε αυτούς που είναι ευαίσθητοι (*susceptible*) στην ασθένεια, σ' αυτούς που είναι ασθενείς με άλλα λόγια που έχουν μολυνθεί (*infected*) και σε αυτούς που έχουν αναρρώσει (*recovered*) και έχουν ανοσία. Αυτοί οι υποπληθυσμοί είναι, φυσικά, διαμερίσεις (*compartments*) του πληθυσμού. Έτσι μιλάμε για μοντέλο ευαισθησίας-μόλυνσης-ανάρρωσης (*recovered – infected – susceptible, SIR*). Αυτό το απλό μοντέλο μπορεί να εφαρμοστεί σε πολλές μεταδοτικές ασθένειες όπως η ιλαρά, η παρωτίτιδα και η ερυθρά, ασθένειες που δεν αφορούν μόνο παιδιά ή το περιβάλλον τους αλλά και που σύμφωνα με εκθέσεις μπορεί να προκληθούν από κάποιο τρομοκρατικό χτύπημα.

Ένα *SIR* μοντέλο περιγράφεται από το παρακάτω μη γραμμικό, πρώτης τάξης, σύστημα εξισώσεων διαφορών [βλ. *Mohtashemi et al.*(2006) ]:

$$\begin{aligned} S_{n+1} &= S_n - \beta S_n I_n \\ I_{n+1} &= I_n + \beta S_n I_n - \delta I_n \\ R_{n+1} &= R_n + \delta I_n \end{aligned} \quad (4.1)$$

Όπου τα  $S_n, I_n, R_n$  είναι τα πλήθη των ατόμων με ευαισθησία, των ατόμων που έχουν μολυνθεί και των ατόμων που έχουν αναρρώσει αντίστοιχα, την  $n$ -οστή ημέρα της επιτήρησης,  $\beta$  είναι το ποσοστό μετάδοσης της νόσου και  $1/\delta$  είναι ο μέσος χρόνος (σε ημέρες) που κάποιος ασθενεί ή αλλιώς  $\delta$  είναι το μέσο ποσοστό ανάρρωσης από την νόσο.

Πρόκειται για ένα δυναμικό μοντέλο, με την έννοια ότι οι τιμές σε κάθε διαμέριση αλλάζουν με το χρόνο. Η σημασία του δυναμικού μοντέλου φαίνεται σε ενδημικές ασθένειες με μικρό χρόνο μέχρι την ανάρρωση, όπως η ιλαρά στην Αγγλία πριν ανακαλυφθεί το εμβόλιο το 1968. Αυτές οι ασθένειες επαναλαμβάνονται περιοδικά στο χρόνο λόγω της διακύμανσης του αριθμού των ατόμων με ευαισθησία. Κατά τη διάρκεια μίας επιδημίας ο αριθμός των ατόμων με ευαισθησία μειώνεται δραστικά καθώς όλο και περισσότερα από αυτά τα άτομα εκδηλώνουν την ασθένεια και άρα εισέρχονται, διαδοχικά, στους υποπληθυσμούς των ασθενών και των ατόμων που ανάρρωσαν. Η επιδημία δεν μπορεί να επανεμφανιστεί μέχρι ο αριθμός των ατόμων με ευαισθησία να επανέλθει σε φυσιολογικά επίπεδα. Φυσικά δεν πρέπει να αγνοήσουμε ότι ένα άτομο, παρόλο που έχει ευαισθησία, μπορεί να μην νοσήσει ποτέ, ούτε να αγνοήσουμε εκροές από το σύστημα (ασθενείς που

αποβιώνουν) ή εισροές (κάποιος που ασθενεί χωρίς να ανήκε προηγουμένα στην ομάδα υψηλού κινδύνου). Συνήθως θεωρούμε ότι ο χρόνος παραμονής ενός ατόμου στο σύστημα κατανέμεται εκθετικά. Μπορούν όμως να χρησιμοποιηθούν και πιο σύνθετες κατανομές, όπως η κατανομή *Erlang*.

Συχνά τα  $S_n$  και  $R_n$  δεν μπορούν να καταγραφούν συστηματικά. Μόνο το  $I_n$  μπορεί να δοθεί έμμεσα, μέσω κάποιων υπολογισμών, χρησιμοποιώντας τον αριθμό των ασθενών  $v_j$  που επισκέπτονται το τμήμα επειγόντων περιστατικών των νοσοκομείων στην ημέρα  $j$ . Για το λόγο αυτό οι *Mohtashemi et al.* (2006) πρότειναν το παρακάτω προσεγγιστικό μοντέλο για το  $I_n$  :

$$I_n = \sum_{j=n-d+1}^n v_j$$

Όπου  $d$  είναι ο μέσος αριθμός ημερών από την διάγνωση μέχρι την ανάρρωση ανά ασθενή. Σκοπός μας είναι να οριστεί το  $I_n$  από μία πιο αυστηρή εξίσωση ώστε να μην εξαρτάται από μεταβλητές που δεν καταγράφονται με ακρίβεια. Από τις δύο πρώτες εξισώσεις του συστήματος (4.1) μπορούμε να καταλήξουμε στην επόμενη εξίσωση όπου το  $I_n$  δεν θα εξαρτάται από την μεταβλητή  $S_n$ . Είναι:

$$I_{n+2} = \frac{I_{n+1}^2}{I_n} - \beta I_{n+1} [I_{n+1} - (1 - \delta) I_n] \quad (4.2)$$

Η εξίσωση (4.2) είναι μια μη γραμμική διαφορική εξίσωση δεύτερης τάξης, η οποία δεν είναι εύκολο να ερμηνευθεί. Επίσης έχουμε να εκτιμήσουμε μόνο μια παράμετρο το  $\beta$ . Οι *Mohtashemi et al.* (2006) χρησιμοποιούν την εξίσωση αυτή για τον υπολογισμό του  $\beta$  εφαρμόζοντας παλινδρόμηση με την μέθοδο ελαχίστων τετραγώνων για κάθε χρονολογικό παράθυρο προκαθορισμένου μήκους. Θεωρούν επίσης ότι το  $\beta$  μεταβάλλεται σχεδόν πάντα και μπορεί να θεωρηθεί περίπου σταθερό μόνο κατά τη διάρκεια μικρών χρονικών διαστημάτων. Αν μπορέσουμε να μοντελοποιήσουμε τη συμπεριφορά του  $\beta$ , υπό κανονικές συνθήκες, τότε αυτό μπορεί να χρησιμοποιηθεί για την ανίχνευση αποκλίσεων από τη συνηθισμένη συμπεριφορά της υποκείμενης διεργασίας. Η πιο σημαντική ιδιότητα, της παραμέτρου αυτής, είναι ότι αντιδρά πολύ γρήγορα σε στατιστικά σημαντικές αλλαγές.

Για κάθε ημέρα, επιλέγεται ένα χρονικό παράθυρο του μήκους  $T$  (από το παρόν προς το παρελθόν), σχηματίζονται  $T - 2$  εξισώσεις και εκτιμάται το  $\beta$  με παλινδρόμηση ελαχίστων τετραγώνων. Στην εργασία τους επιλέχθηκε ένα χρονολογικό παράθυρο με  $T = 7$ .

Αυτή η επιλογή είναι λογική καθώς αντισταθμίζει την επίδραση της ημέρας της εβδομάδας (*day-of-the-week effect*). Επίσης μπορούμε να θεωρήσουμε παράθυρα διάρκειας 14 ή 28 ημερών που σε αρκετές περιπτώσεις μπορούν να δώσουν έγκαιρα συναγερμό. Δεν αποκλείονται βέβαια και παράθυρα με μήκος μικρότερο του 7.

Μετά την εκτίμηση της παραμέτρου, κάθε ημέρα, θα πρέπει να τη συγκρίνουμε με το μέσο του ποσοστού μετάδοσης για την ίδια ημέρα του έτους από τα ιστορικά δεδομένα, εφόσον υπάρχουν. Αν η εκτιμήτρια της τρέχουσας ημέρας είναι μεγαλύτερη από το ανώτατο όριο με βάση τα ιστορικά δεδομένα τότε η μέθοδος ενεργοποιεί το συναγερμό. Οι *Mohtashemi et al.* (2006) χρησιμοποίησαν δεδομένα 7 ετών από περιστατικά συνδρόμου του αναπνευστικού σε παιδιατρικό νοσοκομείο αστικής περιοχής. Σύμφωνα με τη μέθοδο που χρησιμοποίησαν, αν τη  $n$ -οστή ημέρα του έτους ( $n = 365$ ) το  $\beta_n$  ξεπεράσει μια προκαθορισμένη τιμή τότε ενεργοποιείται ο συναγερμός και η ημέρα αυτή θεωρούμε ότι είναι η πρώτη ημέρα της επιδημίας. Συγκεκριμένα αν:

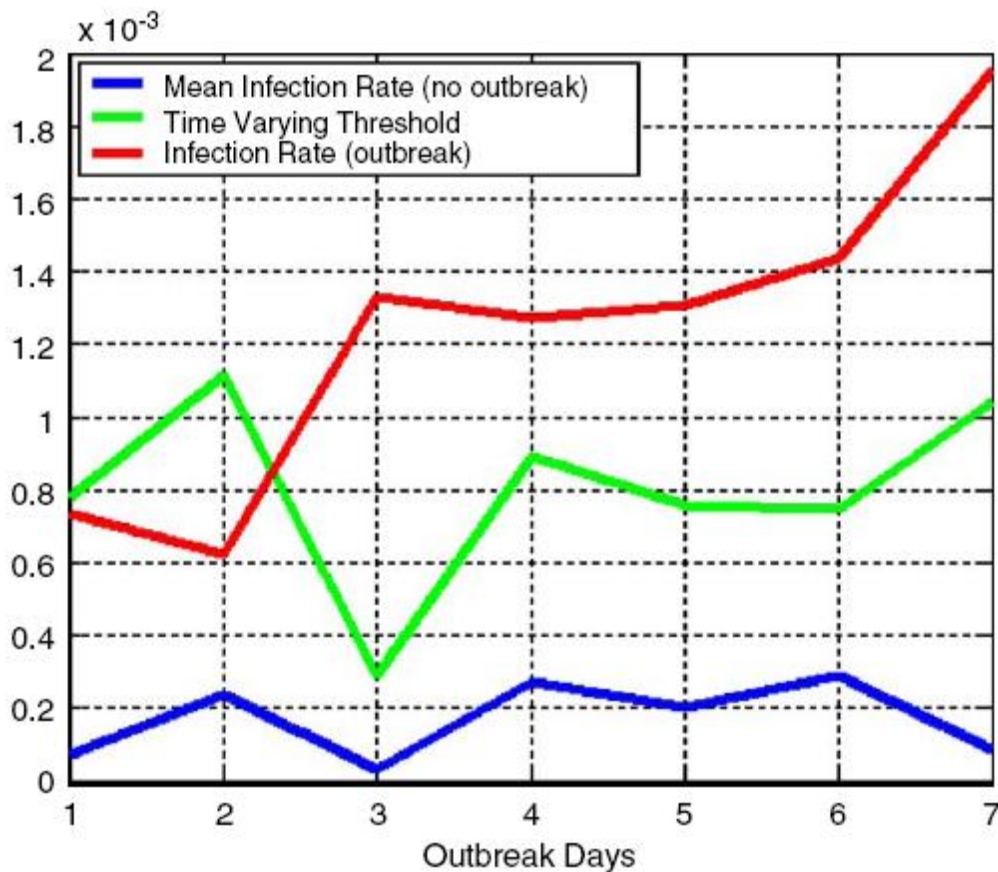
$$\beta_n \geq \bar{\beta}_n + 2\sigma_n$$

και

$$\beta_n > T$$

όπου  $\bar{\beta}_n$  είναι η δειγματική μέση τιμή του ποσοστού μετάδοσης την  $n$ -οστή ημέρα του έτους υπολογισμένη από τα δεδομένα των προηγούμενων ετών,  $\sigma_n$  η τυπική απόκλιση του ποσοστού μετάδοσης την ημέρα αυτή αντίστοιχα και  $T = 2 \times 10^{-4}$  είναι ένα αριθμητικό κατώτατο όριο. Με προσομοίωση εκτιμήθηκαν δεδομένα για ένα ξέσπασμα της νόσου διάρκειας 7 ημερών. Στο Διάγραμμα 4.1 παρατηρούμε ότι την τρίτη ημέρα, από την έναρξη της επιδημίας, έχουμε ένδειξη ότι η διεργασία είναι σε κατάσταση εκτός στατιστικού ελέγχου καθώς τότε το ποσοστό μετάδοσης (κόκκινη γραμμή) ξεπερνάει το άνω όριο (πράσινη γραμμή). Προφανώς το όριο του ελέγχου μεταβάλλεται με τον χρόνο.

Οι *Mohtashemi et al.* (2006) έλεγξαν την απόδοση της μεθόδου χρησιμοποιώντας προσομοίωση. Το ερώτημα που εύλογα γεννάται είναι τι θα πρέπει να κάνουμε όταν δεν υπάρχουν ιστορικά δεδομένα ή αν υπάρχουν αλλά δεν μπορούν να αξιοποιηθούν άμεσα λόγω αλλαγών στις θεραπευτικές μεθόδους ή στον τρόπο καταγραφής κλπ. Τη λύση στο πρόβλημα αυτό μπορούν να δώσουν τα αυτοπαλίνδρομα μοντέλα και τα μοντέλα κινητού μέσου.



**Διάγραμμα 4.1** Ανίχνευση έξαρσης της μεταδοτικότητας αναπνευστικού συνδρόμου [Πηγή Mohtashemy et al. (2006)]

### 4.3 Εφαρμογή του Αυτοπαλίνδρομου Μοντέλου (Autoregressive, AR) στη Βιοεπιτήρηση

Προκειμένου να κατασκευάσουμε ένα αυτοπαλίνδρομο μοντέλο για την έγκαιρη ανίχνευση του ξεσπάσματος μιας επιδημίας θα χρησιμοποιήσουμε το SIR μοντέλο όπως ορίστηκε με το σύστημα εξισώσεων (4.1) και την εξίσωση (4.2).

$$I_{n+2} = a_1 I_{n+1} + a_2 I_{n+2} \quad (4.3)$$

όπου  $\alpha_1 = I_{n+1}/I_n - \beta(I_{n+1} - I_n)$  (4.4)

και  $\alpha_2 = -\beta \delta I_{n+1}$  (4.5)

Η εξίσωση (4.3) είναι μία μη εξίσωση με συντελεστές που εξαρτώνται από το χρόνο και μεταβάλλονται αργά. Μπορεί να αποδειχτεί ότι όταν το γινόμενο  $\beta(I_{n+1} - I_n)$  της εξίσωσης

(4.4) είναι μικρότερο από 0.004 μπορεί να θεωρηθεί αμελητέο [βλ. *Shtatland & Shtatland* (2008)]. Συνεπώς η παράμετρος  $\alpha_1$  μπορεί να οριστεί μόνο από τον λόγο  $I_{n+1}/I_n$  και άρα να αποτελέσει μία καλή εκτιμήτρια της παραμέτρου [βλ. *Hurwicz* (1950), *Zielenski* (1999) και *Luger* (2005)]. Επίσης πολλές φορές στην αρχή μιας επιδημίας η παράμετρος  $\alpha_2$  μπορεί να λάβει αμελητέες τιμές αφού το ποσοστό μετάδοσης της ασθένειας θα είναι ακόμη μικρό, δηλ. και το  $\beta$  και η διαφορά  $\delta = I_{n+1} - I_n$  θα είναι μικρά. Έτσι αν εφαρμόσουμε την εξίσωση (4.3) με  $\alpha_2 = 0$  και αλλαγή του δείκτη, προκύπτει η γραμμική αναδρομική εξίσωση πρώτης τάξεως:

$$I_{n+1} = a_1 I_n$$

Αυτή η εξίσωση μπορεί να θεωρηθεί ως ένδειξη εκθετικής αύξησης. Θα πρέπει, όμως, να ήμασταν ιδιαίτερα προσεκτικοί καθώς οι απλουστεύσεις που κάναμε μπορεί να ισχύουν στα πρώτα στάδια μιας επιδημίας αλλά αν η παρακολούθηση διαρκέσει μεγάλο χρονικό διάστημα τότε οι όροι που απαλείψαμε μπορεί να έχουν σημαντική επίδραση στα δεδομένα. Μία ακόμη υπόθεση που μπορούμε να κάνουμε στην αρχή της επιδημίας είναι, ότι το πλήθος των ατόμων με ευαισθησία την  $n$ -οστή ημέρα της παρακολούθησης θα είναι το σχεδόν ίδιο με το πλήθος των ατόμων με ευαισθησία την ημέρα που ξέσπασε η επιδημία και ότι άτομα με ευαισθησία είναι όλος ο πληθυσμός εκτός αυτών που ήδη ασθενούν ή αναρρώνουν. Αναφερόμαστε πάντα σε ένα κλειστό σύστημα, χωρίς γεννήσεις ή θανάτους, σε ένα μεγάλο πληθυσμό με  $N = S + I + R$ .

Εάν θέσουμε όλες τις παραπάνω υποθέσεις στην εξίσωση (4.3) καταλήγουμε ότι:

$$I_{n+1} \approx (1 + \beta N - \delta) I_n \quad (4.6)$$

Θεωρούμε τώρα την ποσότητα  $\beta^* = \beta N$ , έτσι ποσοτικοποιούμε την μεταδοτικότητα της νόσου σε διαφορετική κλίμακα. Τόσο το  $\beta$  όσο και το  $\beta^*$  είναι πολύ δημοφιλή στην επιδημιολογική κοινότητα. Αντικαθιστώντας στην (4.6) προκύπτει η εξίσωση:

$$I_{n+1} \approx (1 + \beta^* - \delta) I_n \quad (4.7)$$

Τέλος, έχουμε κάνει ένα πλήθος υποθέσεων προκειμένου να καταλήξουμε στην γραμμική εξίσωση (4.7) από την ισχυρά μη γραμμική εξίσωση (4.2). Κάθε μία από αυτές τις υποθέσεις προκαλεί ένα μικρό σφάλμα το οποίο δεν είναι εύκολο να το συνυπολογίσουμε ξεχωριστά.

Επιπλέον τα  $\beta$  και  $\delta$  μπορεί να έχουν κάποια διακύμανση. Αν τώρα συνδυάσουμε όλη αυτή την αβεβαιότητα σε μία εξίσωση λαμβάνουμε το τελικό μοντέλο. Πρόκειται για το αυτοπαλίνδρομο με την πρώτη μόνο υστέρηση (*lag*) ή υστέρηση πρώτης τάξης, το  $AR(1)$  :

$$I_{n+1} = (1 + \beta^* - \delta)I_n + \varepsilon_n$$

όπου  $\varepsilon_n$  είναι το σφάλμα και συχνά αποκαλείται λευκός θόρυβος (*white noise*). Τα  $\varepsilon_n$  είναι ανεξάρτητες, ισόνομες, τυχαίες μεταβλητές από την κανονική κατανομή.

Αξίζει να σημειωθεί ότι όλα τα μοντέλα  $ARMA$ ,  $ARIMA$  στην βιοεπιτήρηση για παθήσεις του αναπνευστικού (γρίπη, *SARS* κλπ) χρησιμοποιούν αυτοπαλίνδρομο πρώτης τάξης [βλ. Shtatland (2007)]. Για αυτό και στην παρακολούθηση πριν το ξέσπασμα κάποιας επιδημίας φαίνεται λογική η χρήση του μοντέλου  $AR(1)$ . Βέβαια, σκοπός μας είναι η ανίχνευση μιας σημαντικής αλλαγής στην διαδικασία που σημαίνει την έναρξη μιας επιδημίας. Έτσι αν  $\beta^* - \delta < 0$ , δηλ αν ο ρυθμός ανάρρωσης είναι μεγαλύτερος από το ρυθμό μετάδοσης, η διαδικασία είναι εντός ελέγχου, στάσιμη. Όταν  $\beta^* - \delta > 0$ , υπάρχει εκθετική αύξηση και ξεκινάει η επιδημία. Τέλος αν  $\beta^* - \delta = 0$  τότε έχουμε την περίπτωση της μοναδιαίας ρίζας, μη στάσιμη διαδικασία.

Επίσης, για λόγους απλότητας θεωρούμε  $\varphi = 1 + \beta^* - \delta$  και για να συμβαδίζουμε με την ευρύτερη βιβλιογραφία στο αντικείμενο των χρονολογικών σειρών αντικαθιστούμε τον συμβολισμό  $I_n$  με το  $y_n$  οπότε γράφουμε:

$$y_{n+1} = \varphi y_n + \varepsilon_n$$

ή ισοδύναμα

$$y_n = \varphi y_{n-1} + \varepsilon_n \quad (4.8)$$

Η θεωρία γύρω από την εκτίμηση της παραμέτρου του μοντέλου  $AR(1)$  είναι ιδιαίτερα ανεπτυγμένη. Πιο συχνά χρησιμοποιείται η μέθοδος ελαχίστων τετραγώνων, οι εκτιμήσεις με τις μεθόδους *Yule-Walker*, *Burg* και τροποποιημένες εκτιμήτριες ελαχίστων τετραγώνων για παράδειγμα αυτές που πρότειναν οι *Provost & Sanjel* (2005).

#### 4.4 Εφαρμογή του Μοντέλου ARMA στην Βιοεπιτήρηση: Πλεονεκτήματα και Μειονεκτήματα

Στην προηγούμενη παράγραφο είδαμε πως ορίζεται το αυτοπαλίνδρομο μοντέλο. Σ' αυτή την παράγραφο θα εξετάσουμε το γενικό μοντέλο ARMA. Για να το κάνουμε αυτό θα ορίσουμε πρώτα το μοντέλο του κινητού μέσου (*moving average, MA*). Η διαδικασία κινητού μέσου τάξης  $q$ , ή πιο σύντομα η διαδικασία  $MA(q)$  δίνεται από την σχέση:

$$y_t = \varepsilon_t + a_1\varepsilon_{t-1} + \dots + a_q\varepsilon_{t-q}$$

όπου  $\varepsilon_t$  είναι ο λευκός θόρυβος και  $y_t = Y_t - \mu$ . Δηλαδή η ακολουθία των  $y_t$  είναι το άθροισμα  $q + 1$  σταθμισμένων όρων από μία διαδικασία λευκού θορύβου.

Οι εξισώσεις (4.3) και (4.8) δίνουν το αυτοπαλίνδρομο μοντέλο δεύτερης και πρώτης τάξης αντίστοιχα. Γενικά, μία αυτοπαλίνδρομη διαδικασία τάξης  $p$ , ή χάρη συντομίας  $AR(p)$  δίνεται από την σχέση:

$$y_t = \varphi_1 y_{t-1} + \dots + \varphi_p y_{t-p} + \varepsilon_t$$

Προφανώς μπορούμε να συνδυάσουμε αυτά τα δύο μοντέλα σε ένα μοντέλο αυτοπαλίνδρομου κινητού μέσου τάξης  $p$  για το αυτοπαλίνδρομο μέρος και τάξης  $q$  για το κομμάτι που αφορά τον κινούμενο μέσο. Για λόγους συντομίας στο εξής θα συμβολίζουμε το μοντέλο αυτό με  $ARMA(p, q)$ . Είναι :

$$y_t = \varphi_1 y_{t-1} + \dots + \varphi_p y_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \dots + a_q \varepsilon_{t-q}$$

Στην πιο συχνή περίπτωση που τα δεδομένα μας αφορούν τις επισκέψεις ασθενών στο τμήμα επειγόντων περιστατικών των νοσοκομείων επιλέγεται το μοντέλο ARMA που προσαρμόζεται καλύτερα στα δεδομένα και εφαρμόζεται στα κατάλοιπα (*residuals*) αφού προσαρμόσουμε την τάση και την εποχικότητα. Σε εργασίες των *Reis, Pagano & Mandl* (2003) και των *Reis & Mandl* (2003) δεδομένα που προέρχονται από τα επείγοντα περιστατικά αναλύονται με μοντέλα ARMA χαμηλής τάξης όπως  $ARMA(1,2)$ ,  $ARMA(2,1)$  και

$ARMA(1,1)$ . Οι Earnest et al. (2005) χρησιμοποίησαν μοντέλα  $AR(1)$ ,  $ARMA(1,1)$ ,  $ARMA(1,2)$ ,  $ARMA(1,3)$ ,  $ARMA(1,4)$ ,  $MA(1)$ ,  $MA(2)$ ,  $MA(3)$ .

Είναι γεγονός ότι δεν υπάρχει ουσιαστική διαφορά ανάμεσα στο αυτοπαλίνδρομο και στο μοντέλο κινούμενου μέσου. Υπό κατάλληλες συνθήκες [βλ. Box & Jenkins (1976)] ένα  $AR$  μπορεί να γραφτεί ως  $MA$  και αντίστροφα:

- Οι συνθήκες για τη στασιμότητα του αυτοπαλίνδρομου μοντέλου είναι οι συνθήκες για την αντιστρεψιμότητα μιας διαδικασίας κινούμενου μέσου
- Μια  $AR$  διαδικασία μπορεί να εκφραστεί ως μια μη πεπερασμένη  $MA$  διαδικασία και αντίστροφα.
- Μια διαδικασία  $AR$  τάξης  $p$  έχει συνάρτηση μερικής αυτοσυσχέτισης (*partial autocorrelation function, PACF*) με σημείο αποκοπής την  $p$ -οστή υστέρηση (*lag*) και μια διαδικασία  $MA$  τάξης  $q$  έχει συνάρτηση αυτοσυσχέτισης (*autocorrelation function, ACF*) με σημείο αποκοπής την  $q$ -οστή υστέρηση.

Στην παρούσα εργασία θα ασχοληθούμε με τα απλούστερα και πιο συχνά χρησιμοποιούμενα μοντέλα  $AR(1)$ ,  $MA(1)$  και  $ARMA(1,1)$ . Έχουν χρησιμοποιηθεί για την ανίχνευση της νόσου *SARS* στην Κίνα [βλ. Lai (2005)] και την Σιγκαπούρη [βλ. Earnest et al (2005)], την ανίχνευση νόσων του αναπνευστικού [βλ. Reis & Mendl (2003)], κλπ.

Για την επιλογή του κατάλληλου μοντέλου χρησιμοποιούνται μέτρα σύγκρισης όπως ο συντελεστής μεταβλητότητας  $R^2$ , ο προσαρμοσμένος συντελεστής μεταβλητότητας (*adjusted R<sup>2</sup>*) και το κριτήριο πληροφορίας του Akaike (*Akaike Information Criterion, AIC*). Αξίζει να σημειωθεί ότι αυτά τα μέτρα μπορούν περισσότερο να χρησιμοποιηθούν ως εμπειρικοί τρόποι λήψης αποφάσεων παρά ως αυστηροί κανόνες στατιστικής συμπερασματολογίας.

Ένα από τα βασικά πλεονεκτήματα των μοντέλων  $ARMA$  είναι η ικανότητά τους να διορθώνουν τάσεις μικρής διάρκειας στα δεδομένα. Δηλ. ότι έγινε χτες συμπεριλαμβάνεται στην πρόβλεψη για το τι θα γίνει σήμερα. Αυτό, για παράδειγμα, είναι ιδιαίτερα χρήσιμο όταν η νόσος που μελετάμε είναι η κοινή γρίπη. Με την χρήση του μοντέλου αυτού, τα υψηλά ποσοστά επισκέψεων στα νοσοκομεία για μεγάλη χρονική περίοδο προσαρμόζονται το μοντέλο. Διαφορετικά θα είχαμε ένδειξη συναγερμού σε καθημερινή βάση για όλη την διάρκεια της γρίπης.

Από την άλλη, αν το μοντέλο  $ARIMA$  «προσαρμόζεται» σε ένα ξέσπασμα της επιδημίας αντί να την ανιχνεύει τότε μία επιδημία που απλώνεται με αργούς ρυθμούς μπορεί να μην



ανιχνευτεί εξ' αιτίας αυτής της διόρθωσης. Επομένως, είναι σημαντικό να χρησιμοποιούμε παράλληλα και άλλα μοντέλα για την ανίχνευση μιας επιδημίας.

Τα μοντέλα *ARMA* που περιγράφονται από τους *Reis & Mandl* (2003) και *Reis, Pagano & Mandl* (2003) απαιτούν μεγάλο όγκο από ιστορικά δεδομένα προκειμένου να αρχίσει η διαδικασία της βιοεπιτήρησης. Αυτό είναι ένα ουσιώδες μειονέκτημα. Σε κάποιες περιπτώσεις, μάλιστα, αυτό δεν είναι διαθέσιμο αλλά δεν είναι και αναγκαίο [βλ. *Moore et al.* (2002)]. Επίσης ο συνδυασμός ιστορικών δεδομένων και πρόσφατων τάσεων είναι αρκετά εφικτός. Ένα ακόμη μειονέκτημα των *ARMA* μοντέλων είναι ότι η διαδικασία της ανίχνευσης δεν είναι ευαίσθητη σε αργές αυξήσεις.

Σύμφωνα με τους *Rizzo et al.* (2005), ξεσπάσματα που εκτυλίσσονται μέσα σε λίγες ημέρες μπορούν να ανιχνευτούν από μοντέλα *ARMA* που δίνουν πρόβλεψη μίας μέρας χρησιμοποιώντας ιστορικά δεδομένα. Όταν η εξάπλωση της νόσου γίνεται σταδιακά είναι πιο εύκολο να ανιχνευτεί με μεθόδους όπως η *CUSUM* [βλ. *Hawkins & Olwell* (1998)].

#### 4.5 Συνδυασμός των μοντέλων αυτοπαλίνδρομου κινούμενου μέσου και των μεθόδων παλινδρόμησης

Οι *Shtatland et al.* (2006) πρότεινα τα παρακάτω βήματα για την ανάλυση των δεδομένων της βιοεπιτήρησης συνδυάζοντας τα μοντέλα *ARMA* με τις μεθόδους παλινδρόμησης: Λογιστική (*Logistic Regression*), παλινδρόμηση Poisson (*Poisson Regression*), γενικευμένα γραμμικά μεικτά μοντέλα (*General Linear Mixed Models, GLMM*) κλπ.

- i. Για ανάλυση χωρικών – χρονολογικών δεδομένων εργαζόμαστε πρώτα με τα χρονολογικά και μετά με τα χωρικά δεδομένα.
- ii. Με βάση τον αριθμό επισκέψεων στα επείγοντα κατασκευάζουμε ένα μοντέλο *ARMA* για κάθε ταχυδρομικό κωδικό αντί για ένα ενιαίο μοντέλο.
- iii. Επιλέγουμε τα μοντέλα *ARMA* με κριτήριο την μεγαλύτερη τιμή του συντελεστή μεταβλητότητας ή του κριτηρίου *AIC*. Αυτή η προσέγγιση είναι λιγότερο αυστηρή αλλά με αυτό τον τρόπο αποφεύγουμε τους πολλαπλούς ελέγχους υποθέσεων.
- iv. Με βάση τους ταχυδρομικούς κωδικούς των περιστατικών που διακομίστηκαν στα επείγοντα με ασθενοφόρο επιλέγουμε αυτά που αντιστοιχούν στα μοντέλα *ARMA* που έχουμε ήδη θεωρήσει.

- v. Με αυτό τον τρόπο περιορίζουμε τον εξαιρετικά μεγάλο όγκο δεδομένων της χωρικής – χρονολογικής βιοεπιτήρησης που συνήθως οδηγεί σε πιο σύνθετα μοντέλα.
- vi. Οι *Kleinman et al.* (2004) πρότεινα την χρήση συμμεταβλητών μεμονωμένων επιπέδων (*individual-level covariates*) όποτε αυτό είναι δυνατό, λόγω της προβλεπτικής τους αξίας (*predictive value*). Από την άλλη κάτι τέτοιο κάνει τους υπολογισμούς ιδιαίτερα δύσκολους. Οι ίδιοι το 2005 πρότειναν ότι μεταβλητές όπως η ηλικία ή το φύλλο δεν θα πρέπει να υπαισέρχονται στο μοντέλο.
- vii. Αφού μειώσουμε τα δεδομένα της βιοεπιτήρησης μπορούμε να χρησιμοποιήσουμε μεμονωμένες συμμεταβλητές με αυτόματη μέθοδο επιλογής. Έτσι έχουμε ένα δισταδιακό σύστημα παρακολούθησης που συνδυάζει τα πλεονεκτήματα των μοντέλων *ARMA* και των μοντέλων λογιστικής παλινδρόμησης ή *Poisson* παλινδρόμησης ή γενικευμένα γραμμικά μεικτά μοντέλα με βάση τα περιστατικά που διακομίστηκαν με ασθενοφόρο.

## ΒΙΒΛΙΟΓΡΑΦΙΑ

### Ελληνική

Αντζουλάκος, Δημήτριος. (2007). Στατιστικός Έλεγχος Ποιότητας, Πανεπιστημιακές Σημειώσεις, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, ΠΜΣ «Εφαρμοσμένη Στατιστική».

Ταγαράς, Γεώργιος. (2001). Στατιστικός Έλεγχος Ποιότητας, Εκδόσεις Ζήτη.

### Ξένη

Aleni, F., Rom, W. and Eisenstein, E. (1996). Risk-adjusted control charts for health care assessment. *Annals of Operations Research*, **67**, 45-60.

Alm, S.E. (1983). On the distribution of the scan statistic of a Poisson process. *In Probability and Mathematical Statistics*, Gut A, Heist L (eds). Upsalla University Press, Upsalla.

Benneyan, J.C. and Kaminsky, F.C. (1995). Another view on how to measure health care quality. *Quality Progress*, **28**, 120-124.

Benneyan, J.C. and Kaminsky, F.C. (1994). Modeling discrete data in SPC: the g and h control charts. *ASQC Annual Quality Congress Transactions*, 32-42.

Beran, J. (1994) *Statistics for Long Memory Processes*. New York: Chapman & Hall.

Chang, M., Glynn, M.K. and Groseclose, S.L. (2003). Endemic, notifiable bioterrorism-related diseases, United States, 1992–1999. *Emerging Infectious Diseases*, **9**, 556–564.

deLaval, M.R., Francois, K., Bull, C., Brown W. and Spiegelhalter D. (1994). Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *Journal of Thoracic Surgery*, **107**, 914–924.

Duncan, A.J. (1986). *Quality Control and Industrial Statistics*, 5<sup>th</sup> ed. Homewood, IL, Richard D. Irwin, Inc.

Dwas, M. (1957). Modified randomization tests for nonparametric hypotheses. *Ann Math. Statist.*, **28**, 181-187.

Earnest, A., Chen M.I. and Sin, L.Y. (2005). Using ARIMA models to predict and monitor the numbers of beds occupied during a SARS outbreak in tertiary hospital in Singapore. *BMC Health Services Research*, **5**, 36.

Finison, L.J., Finison, K.S. and Bliersback, C.M. (1993). The use of control charts to improve health care quality. *Journal of Healthcare Quality*, **15**, 9-23.

Fisz, M. (1963). *Probability Theory and Mathematical Statistics*, Wiley, New York.

- Jensen, W.A., Jones-Farmer, L.A., Champ, C.W. and Woodall, W.H. (2006). Effects of parameter estimation on control chart performance: a literature review. *J. Qual. Technol.*, **38**, 340–364.
- Joner, Jr, M.D., Woodall, W.H. and Reynolds, Jr, M.R. (2008). On detecting a rate increase using a Bernoulli based scan statistic. *Statistics in Medicine*, *27*, **14**, 2555-2575.
- Knaus, W.A., Draper, E.A., Wagner, D.P. and Zimmerman, J.E. (1985). APACHE II: A severity of disease classification system, *Critical Care Medicine*, *13*, **10**.
- Kulldorf, M., Rand, K., Gherman, G., Williams, G. and DeFrancesco, D. (1998b). *SatScan v 2.1: Software for the Spatial and Space-Time Scan Statistics*. Bethesda: National Cancer Institute.
- Kulldorf, M. (1997). A spatial scan statistic. *Communication in Statistics—Theory and Methods*, **26**, 1481–1496.
- Kulldorf, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society*, *164*, **1**, 61-72.
- Lai, D. (2005). Monitoring the SARS epidemic in China: time series analysis. *Journal of Data Science*, **3**, 279- 293.
- Lucas, J.M. (1985). Counted Data CUSUM's. *Technometrics*, **2** , 129-144.
- Mohhamed, M.A., Worthington, P. and Woodall, W.H. (2008). Plotting basic control charts: tutorial notes for healthcare practitioners. *Qual Saf Health Care*, **17**, 137-145.
- Mohtashemi, M. and Levins, R. (2001). Transient dynamics and early diagnostics in infectious disease. *J. Math. Biol.* **43**, 446–470.
- Mohtashemi, M., Szolovits, P., Duniak, J., Kenneth, D. and Mandl, K. (2006). A susceptible-infected model of early detection of respiratory infection outbreaks on a background of influenza. *Journal of Theoretical Biology*, **241**, 954–963.
- Montgomery, D.C. and Mastrangelo, C.M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, **23**, 179–193.
- Montgomery, D.C. (1991). *Introduction to Statistical Quality Control*, 2<sup>nd</sup> ed. New York, NY, John Wiley and Sons, Inc.
- Naus, J. and Wallenstein, S. (2006). Temporal surveillance using scan statistics. *Statistics in Medicine*, **25**, 311–324.
- Naus, J. (1966). A power comparison of two tests of non-random clustering. *Technometrics*, **8**, 493 –517.
- Page, E.S. (1954). Continuous inspection schemes. *Biometrika* ,**41**, 100 –115.

- Ramjee, R. (2000). *Quality control charts and persistent processes*. Ph.D. thesis, Department of Mathematical Sciences, Stevens Institute of Technology.
- Reis, B.Y. and Mandl K.D. (2003). Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, **3**, **2**.
- Reis, B.Y., Pagano, M. and Mandl, K.D. (2003). Using temporal context to improve biosurveillance. *Proceedings National Academy of Sciences, USA*, **100**, **4**, 1961-1965.
- Ries, L.A., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L. and Edwards, B.K. (1999). *SEER Cancer Statistics Review, 1973 – 1996*. Bethesda: National Cancer Institute.
- Rogerson, P.A. (1997). Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine*, **16**, 2081-2093.
- Shmueli, G. (2006). *Fundamentals of Statistical Monitoring: The Good, Bad & Ugly in Biosurveillance*. Presentation Slides, DIMACS Working Group on BioSurveillance Data Monitoring and Information Exchange, February 22 – 24. <http://dimacs.rutgers.edu/Workshops/Surveillance/slides/shmueli.pdf>
- Shmueli, G. (2007). Statistical challenges in modern biosurveillance. *Technometrics. Special issue on Anomaly Detection*.
- Shtatland, E.S., Kleinman, K. and Cain, E.M. (2006). Biosurveillance and outbreak detection using the ARIMA and LOGISTIC procedures. *SUGI'31 Proceedings*, **197**.
- Shtatland, E.S. and 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. *SGF Proceedings*, **363**.
- Shtatland, E.S. (2007). Low-order autoregressive models in early detection of epidemic outbreaks and explosive behaviors in economic and financial time series. *NESUG'20 Proceedings*, Paper PO19.
- Vander Wiel, S.A. (1996). Monitoring processes that wander using integrated moving average models. *Technometrics*, **38**, 139–151.
- Wallenstein, S. and Naus, J. (2004). Statistics for temporal surveillance of bioterrorism. In *Syndrome Surveillance: Reports from a National Conference, 2003. Morbidity and Mortality Weekly Report*, **53**, 74–78.
- Wallenstein, S. and Neff, N. (1987). An approximation for the distribution of the scan statistic. *Statistics in Medicine*, **6**, 197–207.
- Wardell, D.G., Moskowitz, H. and Plante, R.D. (1992). Control charts in the presence of data correlation. *Management Science*, **38**, 1084–1105.
- Weinstock, M. (1981). A generalized scan statistic for the detection of clusters. *International Journal of Epidemiology*, **10**, 289–293.

Young, W.W. (1985). *Measuring the cost of care using patient management categories: Final report of the Health Care Finance Administration*, Baltimore, MD, Grant No. 18P97063/3.

MEMORANDUM FOR THE SECRETARY OF HEALTH AND HUMAN SERVICES