ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΙΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΣΤΑΤΙΣΤΙΚΟΥ ΕΛΕΓΧΟΥ ΠΟΙΟΤΗΤΑΣ ΣΕ ΜΙΑ ΠΑΡΑΓΩΓΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Σωκράτης Α. Ζωρόθεος

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

Πειραιάς
Σεπτέμβριος 2014

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΣΤΑΤΙΣΤΙΚΟΥ ΕΛΕΓΧΟΥ ΠΟΙΟΤΗΤΑΣ ΣΕ ΜΙΑ ΠΑΡΑΓΩΓΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Σωκράτης Α. Ζωρόθεος

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

Πειραιάς
Σεπτέμβριος 2014

Η παρούσα διπλωματική εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε 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)
- Ευαγγελάρας Χαράλαμπος
- Αντζουλάκος Δημήτριος

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

APPLICATION OF TECHNIQUES OF STATISTICAL QUALITY CONTROL IN A MANUFACTURING PROCESS

By
Socratis A. Zorotheos

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
September 2014

Στην οικογένειά μου

Απόστολο, Βάγια και Μαρία

Ευχαριστίες

Ευχαριστώ θερμά:

Την οικογένειά μου για την στήριξη που μου παρείχε καθ' όλη τη διάρκεια των σπουδών μου μιας και χωρίς αυτούς δεν θα είχα φτάσει ως εδώ.

Τον καθηγητή κ. Μάρκο Κούτρα για την συνεχή και καθοριστική βοήθειά του στην εκπόνηση της διπλωματικής εργασίας καθώς και για την υπομονή και το ενδιαφέρον που υπέδειξε.

Τους καθηγητές μου στο Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική για τις πολύτιμες γνώσεις που μου παρείχαν αυτά τα δύο χρόνια.

Περίληψη

Η παρούσα διπλωματική εργασία σκοπό έχει την περιγραφή και ανάλυση διάφορων τεχνικών Στατιστικού Ελέγχου Ποιότητας καθώς και την εφαρμογή τους σε μια συγκεκριμένη παραγωγική διαδικασία με πραγματικά δεδομένα. Αυτό θα γίνει μετά από μια αρχική στατιστική ανάλυση που θα μας επιτρέψει να ομαδοποιήσουμε και να ερμηνεύσουμε τους παράγοντες οι οποίοι επηρεάζουν την ποιότητα της βενζίνης που παράγεται στο διυλιστήριο μιας ελληνικής εταιρίας στην Θεσσαλονίκη.

Η βενζίνη είναι ένα ελαφρύ υγρό, πτητικό και εύφλεκτο, καθώς κάι ένα από τα πιο κύρια και χρήσιμα παράγωγα του πετρελαίου και προέρχεται κυρίως από την κλασματική απόσταξή του.

Πολλοί είναι οι παράγοντες που κατά τη διάρκεια της διύλισης επηρεάζουν, άλλοι σε μικρό και άλλοι σε μεγαλύτερο βαθμό, την ποιότητα του εξερχόμενου προϊόντος (βενζίνη). Στην παρούσα διπλωματική εργασία παρουσιάζονται, αναλύονται και ερμηνεύονται οι μεταβλητές εκείνες που συνδέονται άμεσα με παράγοντες που μπορούν να επηρεάσουν την ποιότητα της παραγόμενης βενζίνης κατά τη διαδικασία παραγωγής της. Θα υπολογισθούν επίσης διάφοροι δείκτες επίδοσης και θα κατασκευασθούν κατάλληλα διαγράμματα ελέγχου που θα μας βοηθήσουν να εξάγουμε χρήσιμα συμπεράσματα για τον τρόπο βελτίωσης του τελικού προϊόντος.

Abstract

This thesis aims to describe and analyze various Statistical Quality Control techniques and their application in a manufacturing process with real data. This will be done after an initial statistical analysis that will allow us to group and interpret the factors which influence the quality of gasoline produced at the refinery of a Greek company in Thessaloniki.

Gasoline is a light liquid, volatile and inflammable and also one of the main and most useful petroleum derivatives, mainly extracted from its fractional distillation.

There are several factors which, during the refining, can influence the quality of the outgoing product (gasoline), others to a small and others to a greater extent. In this thesis are presented, analyzed and interpreted the variables that are directly linked to factors that can affect the quality of gasoline produced during the production process. Various performance indices will also be calculated and appropriate control charts will be constructed that will help us come to conclusions about how to improve the final product.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

τ.μ.	τυχαία μεταβλητή
ANOVA	Analysis of Variance
PCA	Principal Component Analysis
SQC	Statistical Quality Control
ΜΠΚ	Μονάδες ΠετροΚάρβουνο
ΙΠ	Ισοδύναμο Πετρελαίου
δ.ε.	διάστημα εμπιστοσύνης

ΜΕΡΟΣ Ι

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Τα δεδομένα της εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η παρουσίαση διαφόρων στατιστικών μεθόδων που χρησιμοποιούνται στο Στατιστικό Έλεγχο Ποιότητας (ΣΕΠ) και η εφαρμογή τους σε ένα συγκεκριμένο σύνολο πραγματικών δεδομένων. Αυτό θα μας οδηγήσει στο συνεχή έλεγχο της ποιότητας της βενζίνης καθώς και στο πως θα μπορέσουμε να βελτιώσουμε την μελλοντική ποιότητα του τελικού προϊόντος. Για να γίνει αυτό όμως θα πρέπει αρχικά να εντοπιστούν και να μελετηθούν οι παράγοντες εκείνοι που θα μπορούσαν να επηρεάσουν την ποιότητα της παραγόμενης βενζίνης σε ένα διυλιστήριο καθώς και να ερμηνευθούν. Σε αυτό το κομμάτι θα μας βοηθήσουν τα εργαλεία της Ανάλυσης Παλινδρόμησης και της Πολυμεταβλητής Ανάλυσης.

Τα δεδομένα που συλλέχθηκαν καλύπτουν ένα χρονικό εύρος ενός έτους και οι μετρήσεις έγιναν από τους μηχανικούς του διυλιστηρίου με ρυθμό μία μέτρηση την ημέρα. Για την ανάλυση που θα γίνει σε επόμενα κεφάλαια χρησιμοποιήθηκε ένα κομμάτι των αρχικών δεδομένων και συγκεκριμένα οι παρατηρήσεις των έξι (6) πρώτων μηνών.

Οι μεταβλητές (χαρακτηριστικά) που καθορίζουν την ποιότητα του εξερχόμενου προϊόντος είναι:

- τα οκτάνια,
- η τάση ατμών της βενζίνης,
- η επί τοις εκατό (%) περιεκτικότητα σε βενζόλιο.

Επίσης, βάσει προγενέστερης γνώσης, πιθανολογείται ότι οι παράγοντες που επηρεάζουν τις παραπάνω τρεις μεταβλητές είναι:

- η σύνθεση του αργού πετρελαίου,
- η ογκομετρική παροχή εισόδου της στήλης και εξόδου του κλάσματος κορυφής,
- το τελικό σημείο ζέσης,
- οι ογκομετρικές παροχές εισόδου του τμήματος ισομερισμού,
- οι ογκομετρικές παροχές εισόδου του τμήματος αναμόρφωσης,

- οι τιμές των οκτανίων τους,
- η θερμοκρασία των αντιδραστήρων αναμόρφωσης της βενζίνης.

Σε πρώτη φάση θα γίνει μια μικρή αναφορά σε έννοιες όπως «πετρέλαιο», «διύλιση» και «ενέργεια» ώστε ο αναγνώστης να εξοικειωθεί με τις παραπάνω έννοιες καθώς και με πιο τεχνικούς όρους που χρησιμοποιούνται κυρίως στη διαδικασία της διύλισης.

1.2 Σκοπός

Σκοπός της εργασίας είναι να αναλύσουμε τα δεδομένα που διαθέτουμε βασιζόμενοι σε μια σειρά από στατιστικά εργαλεία όπως είναι η Περιγραφική Στατιστική Ανάλυση, η Ανάλυση Παλινδρόμησης, η Πολυμεταβλητή Ανάλυση καθώς και ο Στατιστικός Έλεγχος Ποιότητας. Αυτό θα μας δώσει τη δυνατότητα να έχουμε μια συνολική εικόνα για το πώς (ίσως) κατηγοριοποιούνται οι παράγοντες που επηρεάζουν τις μεταβλητές που διαθέτουμε αλλά και σε ποιο βαθμό η επιρροή τους είναι σημαντική. Όπως είναι γνωστό, μέσα στη διαδικασία παραγωγής υπάρχουν διάφοροι σταθεροί και μεταβλητοί παράγοντες που μπορούν να επηρεάσουν την ποιότητα της παραγόμενης βενζίνης. Στόχος μας είναι να απομονώσουμε και να παρακολουθήσουμε αυτούς τους παράγοντες ώστε στο τέλος να εξάγουμε χρήσιμα συμπεράσματα για την ποιότητα του τελικού προϊόντος.

1.3 Διάρθρωση της εργασίας

Η διπλωματική εργασία χωρίζεται σε δύο μέρη.

Το πρώτο μέρος περιλαμβάνει την εισαγωγή, το σκοπό και τη διάρθρωση της εργασίας, μια μικρή αναφορά στο πετρέλαιο, σε ορολογίες και τεχνικούς όρους της διύλισης καθώς και την παράθεση κάποιων αποτελεσμάτων περιγραφικής στατιστικής που θα βοηθήσουν τον αναγνώστη να πάρει μια πρώτη (οπτική κυρίως) γεύση των δεδομένων της έρευνας. Επίσης, θα γίνει περιγραφή των στατιστικών μεθόδων που θα χρησιμοποιηθούν. Τα αντίστοιχα κεφάλαια του πρώτου μέρους έχουν ως εξής:

- Κεφάλαιο 1: Εισαγωγή, σκοπός και διάρθρωση της εργασίας, το πετρέλαιο και η διαδικασία της διύλισης
- Κεφάλαιο 2: Περιγραφή στατιστικών μεθόδων
- Κεφάλαιο 3: Περιγραφική Στατιστική Ανάλυση

Το δεύτερο μέρος περιλαμβάνει τη βασική διερευνητική ανάλυση των δεδομένων με εφαρμογή μεθόδων Ανάλυσης Παλινδρόμησης (εντοπισμός παραγόντων που επηρεάζουν τις

τρεις εξαρτημένες μεταβλητές και σχηματισμός στατιστικού μοντέλου πρόβλεψης) και Πολυμεταβλητής Ανάλυσης (Principal Component Analysis ή Cluster Analysis όπου έχει νόημα). Επίσης θα περιέχει τα αποτελέσματα της εφαρμογής εργαλείων του Στατιστικού Ελέγχου Ποιότητας (εξαγωγή δεικτών επίδοσης και διαγράμματα ελέγχου) καθώς και των συμπερασμάτων που θα μας οδηγήσουν σε τεχνικές μελλοντικής βελτίωσης του προϊόντος. Τα αντίστοιχα κεφάλαια του δεύτερου μέρους έχουν ως εξής:

- Κεφάλαιο 4: Ανάλυση Παλινδρόμησης
- Κεφάλαιο 5: Πολυμεταβλητή Ανάλυση
- Κεφάλαιο 6: Στατιστικός Έλεγχος Ποιότητας
- Κεφάλαιο 7: Συμπεράσματα

1.4 Το πετρέλαιο και η διύλιση

1.4.1 Πηγές ενέργειας και πρώτες ύλες

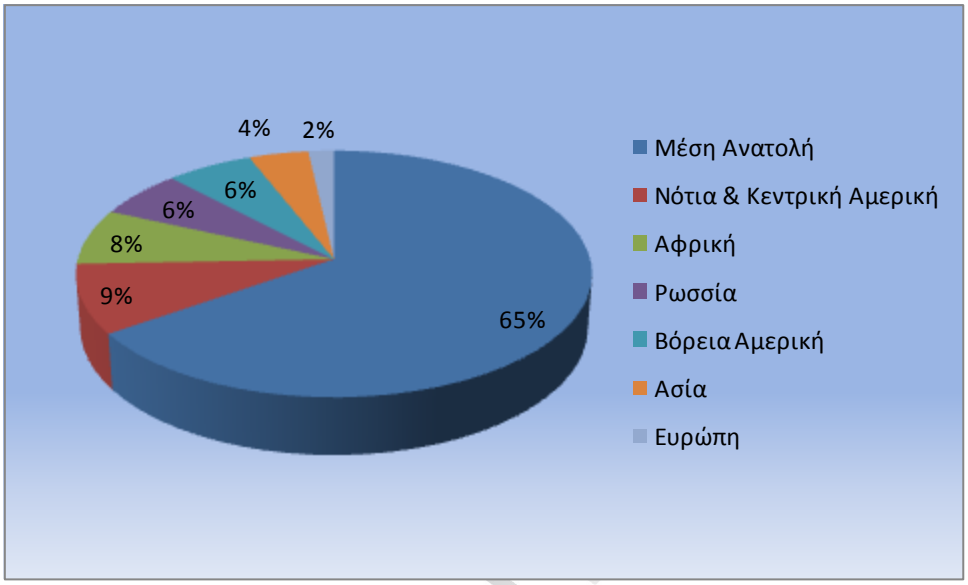
Τα τελευταία 25 χρόνια οι ενεργειακές ανάγκες, σε παγκόσμια κλίμακα, έχουν υπερδιπλασιαστεί και ανήλθαν το 1992 σε $11{,}9 \cdot 10^9\ t$ **ΜΠΚ** (**Μ**ονάδες **Π**ετρο**Κ**άρβουνο), που αναλογούν σε $8{,}34 \cdot 10^9\ t$ **ΙΠ** (**Ι**σοδύναμου **Π**ετρελαίου)[1]. Στα μέσα της δεκαετίας του 70 ο μέσος ετήσιος ρυθμός αύξησης της κατανάλωσης ενέργειας ανέρχονταν σε περίπου 5%, ενώ μειώθηκε αισθητά προς το τέλος της δεκαετίας του 80, λόγω της χρήσης της πυρηνικής ενέργειας από πολλές χώρες.

Ο άνθρακας, το πετρέλαιο και το φυσικό αέριο, τα οποία σχηματίσθηκαν με τη βοήθεια της ηλιακής ενέργειας σε μια χρονική διάρκεια πολλών εκατομμυρίων ετών, πρέπει να καλύψουν σήμερα τις συνεχώς αυξανόμενες ανάγκες. Το πετρέλαιο αποτελεί την ύλη με την περισσότερο αυξανόμενη ζήτηση αλλά παράλληλα και μια συνεχή πηγή πολέμων και προστριβών μεταξύ κρατών μιας και όποιος ελέγχει το πετρέλαιο ελέγχει την παγκόσμια οικονομία. Όμως η υπερβολική του κατανάλωση είναι αιτία μεγάλων περιβαλλοντικών καταστροφών. Τα παγκόσμια αποθεματικά του, όπως αυτά καταγράφηκαν από την υπηρεσία του Ο.Η.Ε. το 2004, φαίνονται στο Σχήμα 1.1.

Σύμφωνα με στοιχεία παγκόσμιων οργανισμών, η παγκόσμια κατανάλωση πρωτογενούς ενέργειας στις αναπτυγμένες χώρες κατανέμεται περίπου κατά το ένα τρίτο στις μεταφορές –

[1] $ΜΠΚ: 1t\ ΜΠΚ\ =\ 7 \cdot 10^6\ kcal\ =\ 29{,}3 \cdot 10^6\ kj\ =\ 0{,}7\ t\ ΙΠ$

συγκοινωνίες, ένα τρίτο στη βιομηχανία και τέλος ένα τρίτο στην οικιακή χρήση, αγροτική οικονομία και σε άλλους τομείς.



Σχήμα 1.1: Παγκόσμια αποθέματα πετρελαίου επί τοις εκατό

Μελέτες της βιομηχανικής ανάπτυξης των κρατών δείχνουν ότι σε μια χρονική περίοδο μισού αιώνα, οι πρωτογενείς πηγές ενέργειας άλλαξαν χρήση. Έτσι στις αρχές της δεκαετίας του 50 η βιομηχανική παραγωγή είχε εξάρτηση 60% από τον άνθρακα και μόνο 25% από το πετρέλαιο ενώ το 2000 οι αριθμοί αντιστράφηκαν. Η σημερινή βιομηχανία χρησιμοποιεί το πετρέλαιο, το φυσικό αέριο αλλά και την πυρηνική ενέργεια σε μερικές εφαρμογές. Οι λόγοι της χρήσης του πετρελαίου και του φυσικού αερίου είναι τα επόμενα πλεονεκτήματα που παρουσιάζουν αυτές οι πηγές πρωτογενούς ενέργειας:

- Η σχετικά εύκολη και οικονομική παραγωγή τους,
- Η πολύπλευρη χρήση τους,
- Η σχετικά εύκολη και οικονομική μεταφορά και διανομή τους.

Σήμερα, σε παγκόσμια κλίμακα, υπάρχει ο προβληματισμός για την εύρεση τρόπων απεξάρτησης από το πετρέλαιο και τον άνθρακα της χρησιμοποιούμενης ενέργειας. Ήδη χρηματοδοτούνται προγράμματα και μελέτες για την παραγωγή ενέργειας από άλλες πηγές, με αποφυγή επιπτώσεων στο περιβάλλον. Τέτοιες πηγές είναι ο ήλιος και ο άνεμος. Πιστεύεται ότι η ηλιακή και η αιολική ενέργεια, σύντομα θα γίνουν οικονομικά ελκυστικές για τους καταναλωτές, αλλά δεν αναμένεται να καλύψουν μεγάλο ποσοστό ενεργειακών αναγκών. Στον Πίνακα 1.1 φαίνεται η συμμετοχή των επί μέρους πηγών ενέργειας στην παγκόσμια κατανάλωση.

Σύμφωνα με στοιχεία του τμήματος Τεκμηρίωσης της Δ/σης Ενεργειακής Πολιτικής του Υπουργείου Ανάπτυξης, στην πατρίδα μας διετέθησαν το έτος 1998 καύσιμα διαφόρων μορφών, τα οποία αναλογούσαν σε $27{,}6 \cdot 10^6\ t\ ΙΠ$. Τα στοιχεία αυτά φαίνονται στον Πίνακα 1.2.

Πίνακας 1.1: Συμμετοχή των επί μέρους πηγών ενέργειας στην παγκόσμια κατανάλωση (σε%)

	1964	**1974**	**1984**	**1994**
Πετρέλαιο	41	48	42	40
Άνθρακας	37	25	27	26
Φυσικό αέριο	15	18	19	21
Πυρηνική ενέργεια	1	3	5	7
Υδροδυναμική, λοιπές μορφές (βιομάζα)	6	6	7	6

Πίνακας 1.2: Κατανάλωση διαφόρων μορφών καυσίμων στην Ελλάδα το έτος 1998

	Στερεά καύσιμα	**Υγρά καύσιμα**	**Φυσικό αέριο**	**Ανανεώσιμες πηγές**	**Σύνολο**
Διάθεση καυσίμων **t, ΙΠ**	$8.9 \cdot 10^6$	$15.5 \cdot 10^6$	$0.8 \cdot 10^6$	$2.4 \cdot 10^6$	$27.6 \cdot 10^6$
Ποσοστό (%)	32.2	56.2	2.9	8.7	100

1.4.2 Πετρέλαιο

Το αργό (ακατέργαστο) πετρέλαιο είναι υγρό πέτρωμα, μίγμα υδρογονανθράκων, δηλαδή ουσιών που περιέχουν άνθρακα και υδρογόνο. Όμως περιέχει και αρκετούς αρωματικούς υδρογονάνθρακες, καθώς και άλλες οργανικές ενώσεις και συναντάται μέσα σε πορώδη πετρώματα στα ανώτερα στρώματα μερικών περιοχών τού φλοιού της Γης.

Για την ερμηνεία της δημιουργίας του πετρελαίου, υπάρχουν πολλές και μάλιστα αλληλοσυγκρουόμενες θεωρίες η πιο σύγχρονη και γενικότερα παραδεκτή από τις οποίες ανάγει την δημιουργία του πετρελαίου σε φυτικές και ζωικές πρώτες ύλες. Ο γεωλόγος Ποτονιέ θεωρεί πως το πετρέλαιο είναι προϊόν αποσύνθεσης ζωικών και φυτικών οργανισμών που εγκλείστηκαν μέσα στα πετρώματα σε μεγάλο βάθος στη Γη. Οπαδοί αυτού δέχονται επίσης πως οι εν λόγω οργανισμοί ήταν κυρίως θαλάσσιοι, ανάλογοι με εκείνους που αποτελούν το πλαγκτόν. Τα λείψανα αυτών των οργανισμών παρασύρθηκαν από θαλάσσια ρεύματα και ανέμους και συγκεντρώθηκαν κατά μεγάλες ποσότητες στους πυθμένες θαλασσίων λεκανών (κόλπων, λιμνοθαλασσών κ.τ.λ.). Στη συνέχεια, οι λεκάνες αυτές αποκλείσθηκαν και καταχώθηκαν από διάφορες αναστατώσεις της επιφάνειας της Γης. Έτσι,

εκ του αποκλεισμένου αυτού οργανικού υλικού προέκυψε με αποσύνθεση, υπό την επίδραση αναερόβιων βακτηρίων, το πετρέλαιο.

Το ορυκτό πετρέλαιο, ή "αργό πετρέλαιο" όπως λέγεται, μπορεί να ποικίλει στην εμφάνιση, τη σύνθεση, και την καθαρότητα. Λαμβάνοντας υπόψη τη σύνθεση των πετρελαίων, αυτά κατατάσσονται σε τρεις βασικές κατηγορίες

- **Παραφινικά πετρέλαια**. Αυτά περιέχουν στερεή παραφίνη και κατά την απόσταξη δίνουν σημαντική αναλογία ελαφρών κλασμάτων που αποτελούνται αποκλειστικά από κορεσμένους υδρογονάνθρακες της αλειφατικής σειράς. Και τα μεν πρώτα της σειράς αυτής μεθάνιο, αιθάνιο, προπάνιο και βουτάνιο παρατηρούνται και στα αέρια που συνοδεύουν το πετρέλαιο στην εξόρυξή του.
- **Ασφαλτικά πετρέλαια**. Αυτά δίνουν περισσότερο βαρέα κλάσματα όπως μαζούτ και ορυκτέλαια. Τα ελαφρά κλάσματα των πετρελαίων αυτών αποτελούνται κυρίως από κεκορεσμένους κυκλικούς υδρογονάνθρακες (ναφθένια) της πολυμεθυλενικής σειράς.
- **Ασφαλτοπαραφινικά πετρέλαια**. Αυτά αποτελούν μίξη των παραπάνω κατηγοριών όπου η μία σειρά δεν υπερτερεί της άλλης.

Το αργό πετρέλαιο χρησιμοποιείται συνήθως για την παραγωγή καυσίμων για μηχανές εσωτερικής καύσης και για το λόγο αυτό είναι μια σημαντική πηγή ενέργειας. Είναι, επίσης, η πρώτη ύλη για πολλά χημικά προϊόντα, συμπεριλαμβανομένων των διαλυτών, των λιπασμάτων, των φυτοφαρμάκων, καθώς και στα συνθετικά προϊόντα όπως των πλαστικών και των απορρυπαντικών ακόμη και ορισμένων εκρηκτικών υλών. Τα προϊόντα που προέρχονται από το πετρέλαιο λέγονται **πετροχημικά** (petrochemicals) και ο κλάδος της Χημείας που ασχολείται με την ανάπτυξή τους Πετροχημεία.

Υπάρχουν διάφορες ταξινομήσεις του αργού πετρελαίου, ανάλογα με την πυκνότητα, την περιεκτικότητα σε θείο, τη βάση αργού, τη μοριακή δομή και ορισμένους δείκτες/συντελεστές. Η ταξινόμηση του αργού πετρελαίου ανάλογα με την πυκνότητά του είναι η πιο χονδρική. Η πυκνότητα ενός πετρελαιοειδούς επηρεάζεται γενικά από τον τύπο των υδρογονανθράκων, και μάλιστα για ίδιες περιοχές βρασμού η πυκνότητα αυξάνει σύμφωνα με τη σειρά:

παραφίνες < ναφθένια < αρωματικά.

Έτσι, αργό χαμηλής πυκνότητας, υποδηλώνει ότι περιέχει παραφίνες σε υψηλότερο ποσοστό, ενώ αντίθετα υψηλής πυκνότητας περιέχει ναφθένια σε υψηλότερο ποσοστό. Ανάλογα με την

πυκνότητα, τα αργά πετρέλαια διακρίνονται σε ελαφρά, μέσα και βαριά. Επίσης, η πυκνότητα των αργών πετρελαίων είναι συνδεμένη με την περιεκτικότητά τους σε θείο και με το ιξώδες τους. Τύποι πετρελαίων χαμηλής πυκνότητας (ελαφριά) είναι κατά κανόνα χαμηλής περιεκτικότητας θείου και έχουν χαμηλότερο ιξώδες (είναι πιο λεπτόρρευστα), ενώ αντίθετα μεγάλης πυκνότητας περιέχουν μεγάλο ποσοστό θείου και έχουν μεγάλο ιξώδες (είναι πιο παχύρρευστα).

Ανάλογα με την περιεκτικότητα σε θείο τα αργά πετρέλαια διακρίνονται σε χαμηλού, μέσου και υψηλού θείου. Ο χαρακτηρισμός αυτός είναι συνδεμένος με την αναμενόμενη περιεκτικότητα σε θείο των βασικών κλασμάτων του πετρελαίου. Όσο υψηλότερο είναι το ποσοστό θείου στο αργό, τόσο περισσότερο θείο αναμένεται να έχουν τα βασικά κλάσματα. Το θείο (ή καλύτερα οι οργανικές θειούχες ενώσεις) είναι ανεπιθύμητο στα προϊόντα του πετρελαίου (βασικά και τελικά), επειδή οι ενώσεις του είναι διαβρωτικές, αλλοιώνουν τις προδιαγραφές των τελικών προϊόντων και τα προϊόντα καύσης του (διοξείδιο και τριοξείδιο του θείου) είναι ισχυροί ρύποι για το περιβάλλον.

Οι μέθοδοι χαρακτηρισμού του αργού πετρελαίου στηρίζονται στο είδος των υδρογονανθράκων (παραφίνες ή ναφθένια) που κυριαρχεί στη μοριακή του δομή. Σύμφωνα με τις μεθόδους αυτές καθορίζεται η λεγόμενη βάση του αργού και γίνεται διαχωρισμός των αργών σε παραφινικής, μεικτής και ναφθενικής βάσης. Ο προσδιορισμός της βάσης του αργού γίνεται με τη μέτρηση της πυκνότητας δύο κλασμάτων, τα οποία χαρακτηρίζονται ως κλάσματα-κλειδιά. Το πρώτο κλάσμα - κλειδί λαμβάνεται με ατμοσφαιρική απόσταξη ($P = 760$ Torr) και σε θερμοκρασίες από 250 °C μέχρι 275 °C και το δεύτερο κλάσμα - κλειδί με απόσταξη με ελαττωμένη πίεση ($P = 40$ Torr) και σε θερμοκρασίες μεταξύ 275 °C ως 300 °C. Στα δύο αυτά κλάσματα-κλειδιά μετριέται η πυκνότητα ($gr \cdot cm^{-3}$) στους 15,6 °C και, σύμφωνα με τους όρους που περιγράφονται στον Πίνακα 1.3, χαρακτηρίζονται σε παραφινικής, μικτής και ναφθενικής βάσης.

Πίνακας 1.3: Χαρακτηρισμός των κλασμάτων – κλειδιών ανάλογα με την πυκνότητά τους

βάση αργού	**$1^{ο}$ κλάσμα κλειδί πυκνότητα ($gr \cdot cm^{-3}$)**	**$2^{ο}$ κλάσμα κλειδί πυκνότητα ($gr \cdot cm^{-3}$)**
παραφίνη	< 0,825	< 0,876
μικτή	0,825 – 0,860	0,876 – 0,934
ναφθενική	> 0,860	> 0,934

Τύποι αργού **παραφινικής** βάσης έχουν χαμηλή πυκνότητα και αποδίδουν μεγάλο ποσοστό βενζινών και λιπαντικών λαδιών. Από τα παραγόμενα κλάσματα το φωτιστικό πετρέλαιο, τα γκαζόιλ κλάσματα και τα βασικά λάδια είναι καλής ποιότητας. Τύποι αργού **ναφθενικής** βάσης έχουν, αντίθετα, υψηλή πυκνότητα και αποδίδουν χαμηλό ποσοστό βενζινών και λιπαντικών λαδιών. Από τα παραγόμενα κλάσματα οι βενζίνες και η άσφαλτος είναι καλής ποιότητας, ενώ τα βασικά λάδια μέτριας ποιότητας.

Για μια ουσιαστικότερη ταξινόμηση των πετρελαίων ή των κλασμάτων τους ως προς τη μοριακή τους δομή, δηλαδή ως προς τη σύσταση τους σε παραφίνες, ναφθένια και αρωματικά, χρησιμοποιείται η ανάλυση δακτυλίων. Η μέθοδος χρησιμοποιεί ως παραμέτρους ανάλυσης την πυκνότητα (d), το δείκτη διάθλασης (n), το σημείο ανιλίνης και τη μέση μοριακή μάζα (M). Από τις τιμές αυτές και με τη βοήθεια νομογραφημάτων μπορούν να βγουν συμπεράσματα γύρω από την περιεκτικότητα σε παραφίνες, ναφθένια και αρωματικά.

Τέλος, μια κάποια αξιολόγηση του αργού πετρελαίου και των κλασμάτων του είναι δυνατή με τη βοήθεια των παρακάτω χαρακτηριστικών αριθμών:

- **Δείκτης συσχετισμού**, ο οποίος υπολογίζεται από σχετικό τύπο στο σημείο βρασμού και με τιμές της σχετικής πυκνότητας. Χαμηλές τιμές του δείκτη δηλώνουν παραφινικό τύπο, ενώ υψηλές τιμές του δείκτη αρωματικό χαρακτήρα.
- **Συντελεστής χαρακτηρισμού**, ο οποίος υπολογίζεται από σχετικό τύπο στο σημείο βρασμού και με τιμές της σχετικής πυκνότητας. Χρησιμοποιείται από την UNIVERSAL OIL PRODUCTS Co. Υψηλές τιμές δείχνουν παραφινικό τύπο.

Μετά τη διαδικασία της ταξινόμησης ακολουθεί η αξιολόγηση του αργού πετρελαίου. Αποφασιστική σημασία στη διαδικασία αυτή παίζουν τόσο οι αποδόσεις όσο και η ποιότητα των επί μέρους βασικών κλασμάτων του (βενζίνη, γκαζόιλ, μαζούτ, λιπαντικά λάδια και άσφαλτος), που λαμβάνονται με απόσταξη. Είναι γεγονός ότι ανεξάρτητα από την ποιότητα του αργού πετρελαίου ή την προέλευσή του, οι προδιαγραφές των προϊόντων καθορίζονται από τις προδιαγραφές που επιβάλει το Χημείο του Κράτους. Οι προδιαγραφές αυτές είναι σταθερές σε μέγιστα και ελάχιστα και αλλάζουν μόνο με απόφαση του Κράτους στην προσπάθεια να βελτιώσει την ποιότητα του προϊόντος.

Σύμφωνα με νεώτερες μελέτες που έγιναν κατά τα τελευταία έτη, τα βεβαιωμένα παγκόσμια αποθέματα στους ταμιευτήρες πετρελαίου αυξήθηκαν με βάση νεώτερες μελέτες. Τα βεβαιωμένα αποθέματα πετρελαίου κατανέμονται κατά περίπου 56% στη Μέση Ανατολή, 16% στη Νότια Αμερική, 9% στην Αφρική, 8% στην Κεντρική και Βόρεια Αμερική, 7% στην

Ευρασία, 3% στην Ασία και στην Ωκεανία και μόνο 1%στην Ευρώπη. Σημειώνεται ότι οι χώρες του OPEC[2] (**O**rganization of **P**etroleum **E**xporting **C**ountries) παρήγαγαν το 2013 το 81% της συνολικής παγκόσμιας παραγωγής.

Πέραν των αποθεμάτων αυτών έχουν βεβαιωθεί και αποθέματα **συνθετικού πετρελαίου** στους **πετρελαϊκούς σχιστόλιθους** και τις **πετρελαϊκές άμμους**, τα οποία υπολογίζονται σε παγκόσμια κλίμακα στο δεκαπλάσιο των αναφερθέντων στην προηγούμενη παράγραφο αποθεμάτων. Τα κοιτάσματα αυτά σήμερα είναι εκμεταλλεύσιμα με την πρόοδο της χημικής βιομηχανίας και τεχνολογίας καθώς και μέσω διαδικασιών εκχύλισης και πυρόλυσης. Χώρες που εκμεταλλεύονται εμπορικά αυτά τα κοιτάσματα είναι η Ρωσία, η Ουκρανία, η Λευκορωσία, ο Καναδάς.

Ο χρονικός ορίζοντας επάρκειας των βεβαιωμένων αποθεμάτων πετρελαίου, υπολογίζεται σε 40 ως 50 χρόνια από σήμερα αν η κατανάλωση παραμείνει σταθερή στις επόμενες 10ετίες. Με την εκμετάλλευση και των αποθεμάτων του συνθετικού πετρελαίου ο χρονικός ορίζοντας ξεπερνά κατά πολύ τα 100 χρόνια.

1.4.3 Η διαδικασία της διύλισης

Το ακατέργαστο πετρέλαιο δε μπορεί να χρησιμοποιηθεί όπως είναι για παραγωγή άλλης μορφής ενέργειας. Θα πρέπει να υποστεί κατεργασία – διύλιση, από την οποία εξάγονται μια σειρά από τελικά προϊόντα, όπως για παράδειγμα βενζίνη super και αμόλυβδη, πετρέλαιο κίνησης και θέρμανσης, κ.τ.λ. Οι προδιαγραφές των προϊόντων της διύλισης είναι καθορισμένες διεθνώς.

Στα διυλιστήρια βασική διεργασία για την παραγωγή προϊόντων από το αργό πετρέλαιο είναι η **κλασματική απόσταξη**. Σκοπός της απόσταξης είναι ο διαχωρισμός δύο ή περισσοτέρων συστατικών μίγματος για παραγωγή προϊόντων με ορισμένες προδιαγραφές. Οι προδιαγραφές προκαθορίζονται με βάση την απαιτούμενη καθαρότητα του προϊόντος ή τα όρια απόσταξης ή τις απαιτήσεις συνέχισης της παραγωγικής διαδικασίας. Η κλασματική απόσταξη είναι η επεξεργασία των προϊόντων βάσει των διαφορών στην πτητικότητα (όπως καθορίζεται από το σημείο βρασμού) και γίνεται μέσα στην στήλη κλασματικής απόσταξης. Η στήλη της κλασματικής απόσταξης είναι ένας κατακόρυφος πύργος με **δίσκους**. Οι δίσκοι διακρίνονται σε **μερικής** ή **ολικής απόληψης** και φέρουν κανάλια ή ποτήρια συλλογής

[2] Ο OPEC περιλαμβάνει τις εξής χώρες: Αλγερία, Αγκόλα, Εκουαδόρ, Ιράν, Ιράκ, Κουβέιτ, Λιβύη, Νιγηρία, Κατάρ, Σαουδική Αραβία, Ηνωμένα Αραβικά Εμιράτα και Βενεζουέλα.

υγρού. Από τους δίσκους ολικής απόληψης λαμβάνεται όλη η ποσότητα του υγρού που συγκεντρώνεται σ' αυτούς. Οι δίσκοι μερικής απόληψης διαθέτουν ανοικτές καπνοδόχους, οι οποίες από τη μια μεριά επιτρέπουν τη διέλευση του ατμού προς τα επάνω και από την άλλη εμποδίζουν με τη βοήθεια ειδικών σκεπών το υγρό να φθάσει στον αμέσως παρακάτω δίσκο. Ο αριθμός των δίσκων ενός πύργου διαφοροποιείται ανάλογα με την εργασία που εκτελεί ο πύργος.

Το αργό πετρέλαιο θερμαίνεται και εισέρχεται σε υγρή και αέρια μορφή θερμοκρασίας 373 °C στη στήλη. Στον πυθμένα του πύργου διοχετεύεται υπέρθερμος ατμός, για να απογυμνώσει το κατερχόμενο υγρό τμήμα της τροφοδοσίας από τυχόν υπολείμματα πτητικότερων συστατικών, τα οποία είναι διαλυμένα σ' αυτό. Με την προσθήκη του υπέρθερμου ατμού πετυχαίνεται μια επιπλέον εξάτμιση των ελαφρύτερων συστατικών του τμήματος απογύμνωσης, λόγω πτώσης της θερμοκρασίας βρασμού τους από την προσθήκη υπέρθερμου ατμού. Για καλύτερο διαχωρισμό των προϊόντων της κλασματικής απόσταξης του αργού ή του ατμοσφαιρικού υπολείμματος είναι απαραίτητη μια ικανοποιητική εσωτερική ροή υγρού **(εσωτερική επαναρροή)** στον πύργο απόσταξης. Κατά τη διάρκεια της ροής αυτής η κατερχόμενη υγρή φάση έρχεται σε στενή επαφή με την ανερχόμενη φάση ατμού και επέρχεται μια θερμοδυναμική ισορροπία μεταξύ των φάσεων αυτών. Αποτέλεσμα της ισορροπίας αυτής είναι τα βαρύτερα συστατικά της ανερχόμενης φάσης ατμού να υγροποιούνται, να μεταφέρονται στην υγρή φάση και να συλλέγονται από τους δίσκους απόληψης. Αντίθετα τα ελαφρύτερα συστατικά της κατερχόμενης υγρής φάσης εξατμίζονται και μεταπηδάνε στη φάση ατμού. Με τον τρόπο αυτό αυξάνεται η διαχωριστική ικανότητα του πύργου.

Τα ελαφρύτερα συστατικά του, κυρίως αέρια, λαμβάνονται ως προϊόντα κορυφής της στήλης απόσταξης. Τα υπόλοιπα κλάσματα λαμβάνονται από τις εξόδους της στήλης σε διαφορετικό ύψος, λόγω διαφορετικής θερμοκρασίας πήξης. Η θερμοκρασία στην ατμοσφαιρική στήλη μεταβάλλεται από πάνω προς τα κάτω, με ψυχρό το πάνω άκρο και θερμό το κάτω. Η σειρά των κλασμάτων είναι: ελαφριά βενζίνη, βαριά βενζίνη (νάφθα), κηροζίνη, ελαφρύ γκαζόιλ, βαρύ γκαζόιλ και το μαζούτ στον πυθμένα της στήλης. Στον Πίνακα 2.4 φαίνονται τα προϊόντα της απόσταξης με βάση τις περιοχές βρασμού. Ανάλογα με την ποιότητα του προϊόντος που θέλουμε να πετύχουμε, βελτιώνουμε τις συνθήκες λειτουργίας της στήλης, προκειμένου να πετύχουμε τις ανάλογες προδιαγραφές του κάθε προϊόντος.

Πίνακας 1.4: Τα προϊόντα της απόσταξης με βάση τις περιοχές βρασμού

προϊόν		**περιοχή βρασμού**	**σειρά απόληψης**
αέρια		< 40 ℃	προϊόν κορυφής
βενζίνη	ελαφριά	40 – 100 ℃	προϊόν κορυφής
	βαριά βενζίνη	100 – 200 ℃	$1^{ο}$ πλευρικό προϊόν
	νάφθα		
κηροζίνη		150 – 250 ℃	$2^{ο}$ πλευρικό προϊόν
γκαζόιλ	ελαφρύ	200 – 300 ℃	$3^{ο}$ πλευρικό προϊόν
	βαρύ	250 – 350 ℃	$4^{ο}$ πλευρικό προϊόν
μαζούτ		> 350 ℃	προϊόν πυθμένα

Οι παράγοντες που επηρεάζουν τα φυσικά και χημικά χαρακτηριστικά των προϊόντων διύλισης διακρίνονται σε παράγοντες πριν τη διύλιση (αργό πετρέλαιο) και σε παράγοντες κατά τη διύλιση (ατμοσφαιρική στήλη, τελικό σημείο ζέσης, κ.τ.λ.).

Η ποιότητα του πετρελαίου καθώς και των παραγώγων του επηρεάζεται τόσο από παράγοντες πριν την διύλιση όσο και από παράγοντες κατά τη διύλιση. Στην πρώτη κατηγορία ανήκουν παράγοντες που σχετίζονται με ανεπιθύμητες προσμίξεις του αργού πετρελαίου με ξένες ουσίες (νερό, ενώσεις αλογόνων, ενώσεις θείου, ενώσεις αζώτου κ.τ.λ.) και με την απόδοση των βασικών κλασμάτων που περιέχονται στους διάφορους τύπους αργού πετρελαίου οι οποίοι μετρούνται με τιμές-κλειδιά κάποιων δεικτών. Στη δεύτερη κατηγορία ανήκουν παράγοντες που σχετίζονται με τη διαδικασία της επεξεργασίας του αργού πετρελαίου, τον εμπλουτισμό του, την αναμόρφωση και τον ισομερισμό του ώστε τα τελικά προϊόντα να έχουν ορισμένες προδιαγραφές, βάσει νομοθεσίας. Αυτές οι προδιαγραφές αφορούν σε φυσικά και χημικά χαρακτηριστικά που είναι η περιοχή βρασμού, το σημείο ανάφλεξης, το ιξώδες, η περιεκτικότητα σε θείο, η περιεκτικότητα σε αρωματικούς υδρογονάνθρακες, κ.τ.λ. και στην ουσία είναι αυτοί που συνδέονται με τα χαρακτηριστικά ποιότητας των τελικών προϊόντων.

Στην παρούσα διπλωματική εργασία θα μελετηθούν οι παρακάτω μεταβλητές-παράγοντες:

X_1 : Περιεκτικότητα σε αργό πετρέλαιο ARABIAN LIGHT,

X_2 : Περιεκτικότητα σε αργό πετρέλαιο ARABIAN EXTRA LIGHT,

X_3 : Περιεκτικότητα σε αργό πετρέλαιο ΡΩΣΙΚΟ,

X_4 : Περιεκτικότητα σε αργό πετρέλαιο IRANIAN LIGHT,

X_5 : Περιεκτικότητα σε αργό πετρέλαιο IRANIAN HEAVY,

X_6 : Κορυφή της αποστακτικής στήλης σε m^3/h,

X_7 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε m^3/h,

X_8 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε HVTO 1075 (βαριά νάφθα) m^3/h,

X_9 : Τελικό σημείο ζέσης (Τ.Σ.Ζ.) τροφοδοσίας αναμόρφωσης σε °C,

X_{10} : Αριθμός οκτανίων προϊόντων αναμόρφωσης (RON),

X_{11} : Η τροφοδοσία σε m^3/h της διαδικασίας του ισομερισμού,

X_{12} : Αριθμός οκτανίων (RON) του προϊόντος του ισομερισμού,

X_{13} : Θερμοκρασία φούρνων (R-400) στους αντιδραστήρες αναμόρφωσης, είσοδος αναμόρφωσης,

Y_1 : Αριθμός οκτανίων τελικού προϊόντος,

Y_2 : Τάση ατμών κατά Reid,

Y_3 : Περιεκτικότητα σε βενζόλιο επί τις εκατό (%).

Οι μεταβλητές Y_1-Y_3 αποτελούν τρία χαρακτηριστικά ποιότητας του τελικού προϊόντος που στην παρούσα διπλωματική εργασία είναι η βενζίνη. Επίσης, σύμφωνα με τη διαδικασία διύλισης που εφαρμόζει το συγκεκριμένο διυλιστήριο, έχουμε και κάποιους παράγοντες που ενδέχεται να επηρεάζουν τα παραπάνω χαρακτηριστικά ποιότητας. Αυτοί οι παράγοντες αποτυπώνονται από τις μεταβλητές X_1-X_{13}, οι οποίες, με τον τρόπο που έχουν συλλεχθεί τα δεδομένα, είναι χωρισμένες σε τέσσερις κατηγορίες. Οι κατηγορίες αυτές είναι το ποσοστό του αργού πετρελαίου στο τελικό προϊόν, η ατμοσφαιρική στήλη, ο εμπλουτισμός και ο ισομερισμός του αργού πετρελαίου. Στην ουσία στις μεταβλητές X_1-X_{13} περιέχονται τα δεδομένα των βασικών σταδίων της διαδικασίας της διύλισης ενώ οι μεταβλητές Y_1-Y_3 περιγράφουν τα χαρακτηριστικά ποιότητας της βενζίνης μετά τη διαδικασία της διύλισης.

ΚΕΦΑΛΑΙΟ 2

Περιγραφή Στατιστικών Μεθόδων

2.1 Εισαγωγή

Στο σημείο αυτό και προτού προχωρήσουμε στην ανάλυση των δεδομένων, καλό θα ήταν να γίνει μια σύντομη παρουσίαση των στατιστικών μεθόδων βάσει των οποίων θα γίνει η ανάλυση. Όπως αναφέρθηκε και στο Κεφάλαιο 1, τα στατιστικά εργαλεία που θα χρησιμοποιηθούν είναι η Περιγραφική Στατιστική Ανάλυση, η Ανάλυση Παλινδρόμησης, η Πολυμεταβλητή Ανάλυση και τέλος ο Στατιστικός Έλεγχος Ποιότητας. Στη συνέχεια περιγράφονται οι παραπάνω μέθοδοι με τη σειρά που αναφέρονται.

2.2 Περιγραφική Στατιστική Ανάλυση

Η Περιγραφική Στατιστική Ανάλυση ασχολείται με την παρουσίαση και την περιγραφή των δεδομένων. Η παρουσίαση πρέπει να γίνει με τέτοιο τρόπο ώστε να μπορεί να γίνει μια πρώτη ερμηνεία των αποτελεσμάτων. Επίσης, είναι σημαντικό να μπορούν να ανιχνευθούν κάποια ιδιαίτερα χαρακτηριστικά των τιμών του δείγματος (ή των δειγμάτων) που πιθανότατα να είναι και χαρακτηριστικά του πληθυσμού (ή των πληθυσμών), τα οποία θα μελετηθούν αναλυτικά σε επόμενα στάδια (Διερευνητική Στατιστική). Στο κομμάτι της παρουσίασης των δεδομένων υπάρχουν δύο κύριοι τρόποι, με πίνακες και με διαγράμματα. Οι πίνακες συνήθως περιέχουν συχνότητες και σχετικές συχνότητες εμφάνισης των πεπερασμένων δυνατών παρατηρήσεων του δείγματος που μελετάται, τα αντίστοιχα επί τοις εκατό (%) ποσοστά, τις κλάσεις που έχουν δημιουργηθεί καθώς και τις κεντρικές τιμές τους. Ένας πίνακας θα πρέπει να παρουσιάζει με απλότητα τα δεδομένα, να είναι ευανάγνωστος και να είναι ξεκάθαρο το τι περιέχει. Όσον αφορά τα διαγράμματα, πριν επιλεγεί το είδος του διαγράμματος, καλό θα είναι να ληφθεί υπόψη το τι ακριβώς θέλουμε να παρουσιάσουμε με το διάγραμμα, τι είδους μεταβλητή εξετάζουμε και τι είδους είναι τα δεδομένα που διαθέτουμε. Ένα διάγραμμα θα πρέπει να είναι παραστατικό ώστε να διευκολύνει την κατανόηση και να παρουσιάζει τα βασικά χαρακτηριστικά της μεταβλητής, να είναι σαφές ώστε να μη δημιουργεί σύγχυση και ακριβές ώστε να μην παραπλανά τον αναγνώστη.

Αντίστοιχα, στο κομμάτι της περιγραφής των δεδομένων έχουμε κάποιους δείκτες – μέτρα που υπολογίζονται από τα δεδομένα και εκφράζουν κάποιες ιδιότητές τους, όπως τη δομή και τη μορφή τους. Αυτά τα μέτρα χωρίζονται σε κατηγορίες ανάλογα με το ποιες από τις ιδιότητες αυτές εκφράζουν. Έτσι, έχουμε τα μέτρα κεντρικής τάσης (αριθμητικός, γεωμετρικός, αρμονικός μέσος) τα μέτρα θέσης (διάμεσος, επικρατούσα τιμή, ποσοστημόρια), τα μέτρα κύμανσης (εύρος, διακύμανση, τυπική απόκλιση), τα μέτρα λοξότητας ή ασυμμετρίας (συντελεστής ασυμμετρίας του Pearson) και τα μέτρα κύρτωσης (συντελεστής κύρτωσης).

Βέβαια, για να καταλήξουμε στον τρόπο παρουσίασης και περιγραφής των δεδομένων, θα πρέπει αρχικά να διαπιστωθεί η φύση των δεδομένων. Τα δεδομένα μπορεί να είναι ποιοτικά (ή αλλιώς ονομαστικά) ή ποσοτικά. Επίσης, προτού κατασκευαστούν κατάλληλα διαγράμματα για τα δεδομένα θα πρέπει να εξετάζεται αν αυτά είναι χρονολογικά, διαστρωμματικά, μικτά κ.τ.λ.

2.2.1 Ποιοτικά δεδομένα

Στην περίπτωση που τα υπό μελέτη δεδομένα είναι ποιοτικά, τα περιγραφικά μέτρα δεν έχουν νόημα και επομένως, η περιγραφή των δεδομένων γίνεται με βάση τις συχνότητές τους. Επίσης, καταρτίζονται πίνακες και κατασκευάζονται κατάλληλα διαγράμματα. Οι πίνακες περιέχουν τις κατηγορίες των δεδομένων καθώς και τις αντίστοιχες συχνότητες και σχετικές συχνότητες εμφάνισής τους. Τα κυριότερα διαγράμματα γι’ αυτό το είδος των δεδομένων είναι τα κυκλικά διαγράμματα και τα ραβδογράμματα. Και τα δύο απεικονίζουν συχνότητες (ή σχετικές συχνότητες) που αντιστοιχούν στις τιμές (κατηγορίες) της μεταβλητής. Στην πρώτη περίπτωση, οι συχνότητες παριστάνονται με κυκλικούς τομείς ενώ στη δεύτερη με κατακόρυφες γραμμές (ή ορθογώνια – ράβδους).

Ένα ραβδόγραμμα πλεονεκτεί έναντι ενός κυκλικού διαγράμματος μιας και κατασκευάζεται πιο εύκολα με το χέρι, διακρίνονται σε αυτό πιο εύκολα μικρές διαφορές μεταξύ συχνοτήτων και μπορεί να χρησιμοποιηθεί όταν έχουμε πολλές κατηγορίες της μεταβλητής. Επίσης, με ένα ραβδόγραμμα μπορούν εύκολα να μελετηθούν περισσότερα από ένα χαρακτηριστικά του πληθυσμού αντίστοιχα και είναι ευκολότερο να γίνουν συγκρίσεις ανάμεσα σε δύο ή περισσότερα διαγράμματα. Από την άλλη, ένα πλεονέκτημα του κυκλικού διαγράμματος είναι ότι παρέχει καλύτερη εικόνα για τη σχέση που έχει μία συχνότητα με το άθροισμα όλων των συχνοτήτων.

2.2.2 Ποσοτικά δεδομένα

Στην περίπτωση που τα υπό μελέτη δεδομένα είναι ποσοτικά, παρουσιάζουμε τα δεδομένα με πίνακες και διαγράμματα και υπολογίζουμε κάποιους από τους περιγραφικούς δείκτες που αναφέρθηκαν προηγουμένως. Σε αυτή την περίπτωση, δεν έχει νόημα να χρησιμοποιήσουμε συχνότητες για τα ακατέργαστα δεδομένα. Αντί γι' αυτό χωρίζουμε τα δεδομένα σε ομάδες – τάξεις και απεικονίζουμε τις συχνότητες αυτών των τάξεων. Το πλάτος τους, δ, δίνεται από τον τύπο:

$$\delta = \frac{R}{k},$$

όπου

R: το εύρος των παρατηρήσεων,

k: το επιθυμητό πλήθος διαστημάτων

Ένας τρόπος επιλογής του k είναι ο εμπειρικός τύπος του Sturges σύμφωνα με τον οποίο ο αριθμός k επιλέγεται να είναι ο μικρότερος ακέραιος που είναι μεγαλύτερος ή ίσος από

$$1 + \frac{\log(n)}{\log(2)},$$

όπου n: το πλήθος των παρατηρήσεων.

Πολλά είναι τα διαγράμματα που μπορούν να κατασκευαστούν για την παρουσίαση ποσοτικών δεδομένων. Τα κυριότερα είναι το ιστόγραμμα συχνοτήτων στο οποίο παριστάνονται οι τάξεις που έχουν δημιουργηθεί, το θηκόγραμμα που παρουσιάζει γραφικά τα μέτρα θέσης μιας μεταβλητής και το κυκλικό διάγραμμα. Επίσης, αν στο ιστόγραμμα συχνοτήτων κατασκευάσουμε και το πολύγωνο συχνοτήτων, μπορούμε να έχουμε μια εικόνα για την κατανομή των δεδομένων που μπορεί να είναι μονοκόρυφη, σχήματος U, σχήματος J, συμμετρική, με θετική ή αρνητική ασυμμετρία κ.τ.λ. Αυτό συμβαίνει κυρίως όταν έχουμε μεγάλο αριθμό διαστημάτων και έτσι το πολύγωνο συχνοτήτων προσεγγίζει μια ομαλή καμπύλη, την καμπύλη συχνοτήτων η οποία με τη σειρά της προσεγγίζει τη θεωρητική κατανομή του πληθυσμού από τον οποίο προήλθε το δείγμα.

Οι κυριότεροι περιγραφικοί δείκτες είναι ο αριθμητικός μέσος, ο γεωμετρικός μέσος, ο αρμονικός μέσος, η διάμεσος, η επικρατούσα τιμή, τα τεταρτημόρια, το εύρος, η διασπορά και η τυπική απόκλιση, τα μέτρα ασυμμετρίας και τα μέτρα κύρτωσης. Όλοι αυτοί οι δείκτες δίνονται από τα στατιστικά πακέτα και είναι σχετικά εύκολοι στον υπολογισμό τους. Επίσης, έχουν κάποια πλεονεκτήματα και μειονεκτήματα. Για παράδειγμα, ένα μειονέκτημα του αριθμητικού μέσου είναι ότι είναι ευαίσθητος σε ακραίες παρατηρήσεις, ενώ η διάμεσος είναι

ανθεκτική σε ακραίες παρατηρήσεις. Δύο πλεονεκτήματα που έχει η επικρατούσα τιμή σε σχέση με άλλα περιγραφικά μέτρα είναι ότι μπορεί να υπολογιστεί και για κατανομές που είναι ανοικτές προς τα πάνω ή προς τα κάτω και μπορεί να χρησιμοποιηθεί και για ονομαστικά δεδομένα. Όμως, ένα μειονέκτημά της είναι ότι μπορεί να υπολογιστεί μόνο για μονοκόρυφες κατανομές. Τέλος, η διαφορά του πρώτου από το τρίτο τεταρτημόριο, δηλαδή η διαφορά $Q_3 - Q_1$ ονομάζεται **ενδοτεταρτημοριακό εύρος**. και όταν η διάμεσος βρίσκεται πιο κοντά στο ένα ή στο άλλο τεταρτημόριο, αυτό μας δείχνει το είδος της ασυμμετρίας.

2.3 Ανάλυση Παλινδρόμησης

Οι τεχνικές της Ανάλυσης Παλινδρόμησης αποτελούν ίσως το πλέον δημοφιλές εργαλείο του εφαρμοσμένου στατιστικού. Τέτοιες τεχνικές χρησιμοποιούνται σχεδόν σε όλες τις περιπτώσεις όπου υπάρχει ανάγκη ταυτόχρονης μελέτης δύο ή περισσοτέρων μεταβλητών (χαρακτηριστικών) με στόχο την πρόβλεψη μιας εξ' αυτών όταν είναι γνωστή η τιμή κάποιας ή κάποιων άλλων μεταβλητών που σχετίζονται με αυτήν. Για παράδειγμα

- ο χρόνος αλλοίωσης ενός γαλακτοκομικού προϊόντος επηρεάζεται αρνητικά από τη θερμοκρασία με την έννοια ότι όσο πιο μεγάλη είναι η θερμοκρασία τόσο μικρότερος θα είναι ο χρόνος που θα υποστεί αλλοίωση το προϊόν.
- η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική εξάρτηση μεταξύ τους με την έννοια ότι όσο πιο μεγάλο είναι το παιδί τόσο μεγαλύτερο βάρος θα έχει.
- το ύψος των αποδοχών των υπαλλήλων μιας εταιρίας εξαρτάται από το χρόνο εργασίας τους.

Σε όλα τα παραπάνω προβλήματα, παρουσιάζει ενδιαφέρον να εξεταστούν οι επιδράσεις που κάποια μεταβλητή ασκεί σε κάποια άλλη μεταβλητή. Αν μάλιστα θα μπορούσε να βρεθεί και ένα απλό μαθηματικό μοντέλο που να εκφράζει τη σχέση αυτή μέσω μιας συνάρτησης, τότε θα μπορούσε να χρησιμοποιηθεί για την πρόβλεψη των τιμών μιας μεταβλητής από τις γνώσεις που διαθέτουμε για τις άλλες μεταβλητές. Ο τομέας της Στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με στόχο την πρόβλεψη μιας από αυτές με χρήση των τιμών μιας ή περισσοτέρων άλλων ονομάζεται **ανάλυση παλινδρόμησης** (regression analysis). Αν μας ενδιαφέρει η ταυτόχρονη μελέτη δύο μεταβλητών με σκοπό την πρόβλεψη μιας εξ αυτών, τότε μιλάμε για **απλή παλινδρόμηση**, ενώ αν έχουμε ταυτόχρονη μελέτη περισσοτέρων από δύο μεταβλητών ώστε να προβλέψουμε μία από αυτές, τότε μιλάμε για **πολλαπλή παλινδρόμηση**.

Στην παρούσα διπλωματική εργασία θα ασχοληθούμε στην πράξη μόνο με μοντέλο πολλαπλής παλινδρόμησης, οπότε στη συνέχεια θα περιγράψουμε όσα σχετίζονται με αυτό, χωρίς αναφορά σε μοντέλα απλής παλινδρόμησης.

2.3.1 Πολλαπλή παλινδρόμηση

Στην απλή παλινδρόμηση γίνεται ανάπτυξη και μελέτη μοντέλων τα οποία περιγράφουν τη σχέση μιας ανεξάρτητης μεταβλητής X και μιας μεταβλητής απόκρισης Y. Συνήθως όμως στην πράξη η συμπεριφορά μιας μεταβλητής απόκρισης Y επηρεάζεται από πολλές μεταβλητές $X_1, X_2, X_3, \dots$ οπότε η χρήση μόνο μιας από αυτές δεν αναμένεται να οδηγήσει σε ισχυρά μοντέλα πρόβλεψης, αφού θα μείνει ανερμήνευτη η μεταβλητότητα της Y που οφείλεται στις άλλες ανεξάρτητες μεταβλητές.

Επομένως, γίνεται επιτακτική η ανάγκη διαμόρφωσης μοντέλων τα οποία να επιτρέπουν την αποτελεσματική πρόβλεψη μιας μεταβλητής απόκρισης Y μέσω πολλών ανεξάρτητων μεταβλητών $X_1, X_2, X_3, \dots$. Στην παρούσα ενότητα θα ασχοληθούμε με τη μελέτη μοντέλων αυτής της μορφής, τα οποία είναι γνωστά ως μοντέλα πολλαπλής παλινδρόμησης.

2.3.1.1 Εκτιμήτριες Ελαχίστων Τετραγώνων σε Μοντέλα με Πολλές Ανεξάρτητες Μεταβλητές

Η πιο απλή περίπτωση μοντέλου πολλαπλής παλινδρόμησης προκύπτει, όταν θέλουμε να εξετάσουμε με ποιο τρόπο δύο ανεξάρτητες μεταβλητές X_1, X_2 επηρεάζουν μια τρίτη μεταβλητή Y, που θα είναι η μεταβλητή απόκρισης ή εξαρτημένη μεταβλητή. Όπως και στο απλό στατιστικό γραμμικό μοντέλο, οι μεταβλητές X_1 και X_2 δε θεωρούνται τυχαίες, ενώ ή μεταβλητή απόκρισης θεωρείται ότι είναι μια τυχαία μεταβλητή.

Όταν έχουμε δύο ανεξάρτητες μεταβλητές X_1, X_2 οι οποίες παίρνουν τιμές x_{i1}, x_{i2} αντίστοιχα, καταλήγουμε σε μια συνάρτηση τριών παραμέτρων β_0, β_1 και β_2 της μορφής

$$g(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right)^2 \qquad (2.3.1)$$

όπου η διαφορά $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}), i = 1,2, \dots, n$ είναι το σφάλμα που αντιστοιχεί στην i μέτρηση. Οι εκτιμήτριες ελαχίστων τετραγώνων που θα προκύψουν, θα συμβολίζονται με $\hat{\beta}_0, \hat{\beta}_1$ και $\hat{\beta}_2$. Επομένως, βάσει αυτών, μπορούμε να γράψουμε το **επίπεδο παλινδρόμησης** της Y πάνω στις X_1 και X_2 το οποίο είναι το εξής:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

Επίσης, έχοντας υπολογίσει τις εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$ και $\hat{\beta}_2$, μπορούμε να πάρουμε τις διαφορές

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}\right), \qquad i = 1,2,\dots,n$$

οι οποίες θα λέγονται και πάλι **εκτιμημένα σφάλματα**, καθώς και την ποσότητα

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}\right)\right)^2,$$

η οποία θα λέγεται **άθροισμα τετραγώνων των** (εκτιμημένων) **σφαλμάτων**.

Όσα παρουσιάστηκαν παραπάνω για τη μελέτη μιας μεταβλητής απόκρισης Y με βάση δύο ανεξάρτητες μεταβλητές X_1 και X_2 μπορούν εύκολα να γενικευτούν για την περίπτωση που διαθέτουμε έναν οποιονδήποτε αριθμό ανεξάρτητων μεταβλητών, έστω $k \geq 1$ το πλήθος. Επομένως, η σχέση (2.3.1) στην περίπτωση που διαθέτουμε τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k, k \geq 1$, γίνεται

$$g(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})\right)^2. \qquad (2.3.2)$$

Ελαχιστοποιούμε και πάλι το άθροισμα αυτό θέτοντας τις μερικές παραγώγους ίσες με μηδέν για να προσδιορίσουμε τις παραμέτρους $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Έτσι οδηγούμαστε σε ένα σύστημα που αποτελείται από $p = k + 1$ γραμμικές εξισώσεις οι οποίες ονομάζονται **κανονικές εξισώσεις**. Το σύστημα αυτό έχει ως αγνώστους τις p παραμέτρους $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Στη συνέχεια, θα σταματήσουμε να χρησιμοποιούμε το σύμβολο k και θα χρησιμοποιούμε μόνο το σύμβολο p, το οποίο εκφράζει το πλήθος των παραμέτρων $\beta_i, i = 0,1,\dots,p-1$ που θέλουμε να εκτιμήσουμε. Έτσι, στη γενική περίπτωση της πολλαπλής παλινδρόμησης θα θεωρούμε απλά ότι διαθέτουμε $p-1$ ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_{p-1}$ όπου $p \geq 2$ για τις οποίες έχουμε συλλέξει παρατηρήσεις

$$x_{i1}, x_{i2}, \dots, x_{i,p-1}, i = 1,2,\dots,n$$

μαζί με τις αντίστοιχες τιμές $y_i, i = 1,2,\dots,n$ της μεταβλητής απόκρισης. Με τις παρατηρήσεις αυτές κατασκευάζουμε το λεγόμενο **πίνακα σχεδιασμού** (design matrix) του μοντέλου ο οποίος συμβολίζεται με X, έχει διάσταση $n \times p$ και έχει ως πρώτη στήλη μια στήλη με μονάδες

Οι ποσότητες $\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}$ που προκύπτουν από τον τύπο $\boldsymbol{\beta} = (X'X)^{-1}X'\mathbf{y}$, ελαχιστοποιούν το άθροισμα τετραγώνων (2.3.2), οπότε θα λέγονται **εκτιμήτριες ελαχίστων τετραγώνων** και θα συμβολίζονται με $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{p-1}$, αντίστοιχα. Επομένως, θα έχουμε

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

Αξίζει να σημειωθεί ότι τόσο ο πίνακας $X'X$ όσο και ο αντίστροφός του $(X'X)^{-1}$, εάν υπάρχει, είναι συμμετρικοί πίνακες.

2.3.1.2 Το Στατιστικό Μοντέλο Πολλαπλής Παλινδρόμησης

Στην ενότητα αυτή θα παρουσιάσουμε ένα στοχαστικό πρότυπο για τη μελέτη μιας μεταβλητής απόκρισης Y, η οποία είναι τυχαία μεταβλητή και οι τιμές της μπορούν να προβλεφθούν με χρήση $k = p - 1$ ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_{p-1}$ οι οποίες δεν είναι τυχαίες μεταβλητές.

Το μοντέλο που θα αναπτυχθεί ονομάζεται **Στατιστικό Μοντέλο Πολλαπλής Παλινδρόμησης** με $p - 1$ ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_{p-1}$ και θα έχει τη μορφή

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,\ i = 1,2, \dots, n \qquad (2.3.3)$$

για το οποίο ισχύουν οι επόμενες υποθέσεις:

Σ1. Οι ποσότητες $\beta_0, \beta_1, \dots, \beta_{p-1}$ είναι άγνωστες παράμετροι.

Σ2. Τα $x_{i1}, x_{i2}, \dots, x_{i,p-1}, i = 1,2, \dots, n$ είναι γνωστοί αριθμοί – πιο συγκεκριμένα είναι οι τιμές των ανεξάρτητων (ή ελεγχόμενων ή προβλεπουσών) μεταβλητών κατά την i επανάληψη του πειράματος. Οι τιμές αυτές καθορίζονται από τον ερευνητή που εκτελεί το πείραμα.

Σ3. Το Y_i είναι η τιμή της εξαρτημένης μεταβλητής (ή μεταβλητής απόκρισης) κατά την i επανάληψη του πειράματος. Το Y_i είναι τυχαία μεταβλητή, και θα συμβολίζουμε με y_i την τιμή που λαμβάνει αυτή, αν εκτελεστεί το πείραμα και καταγραφεί το αποτέλεσμα που παρατηρείται για τη μεταβλητή απόκρισης Y (παρατηρούμενη τιμή της Y, όταν οι ανεξάρτητες μεταβλητές $X_{i1}, X_{i2}, \dots, X_{i,p-1}$ λάβουν τις τιμές $x_{i1}, x_{i2}, \dots, x_{i,p-1}$, αντίστοιχα).

Σ4. Τα $\varepsilon_i, i = 1,2, \dots, n$ είναι τυχαία σφάλματα με μέση τιμή 0 και διακύμανση σ^2, δηλαδή

$$E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2.$$

Σ5. Τα σφάλματα ε_i και ε_j που αντιστοιχούν σε διαφορετικές επαναλήψεις του πειράματος $(i \neq j)$ θεωρούνται ασυσχέτιστα, δηλαδή ισχύει

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ για } i \neq j.$$

Γράφοντας τις n ισότητες της σχέσης (2.3.3) για $i = 1,2,\dots,n$ σε μορφή ισότητας πινάκων, καταλήγουμε στην παρακάτω μορφή του στατιστικού μοντέλου πολλαπλής παλινδρόμησης.

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

όπου X είναι ο πίνακας σχεδιασμού που είδαμε στην ενότητα (2.3.1.1) και **β, Y** και **ε** τα διανύσματα – στήλες

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

διάστασης $p \times 1, n \times 1, n \times 1,$ αντίστοιχα. Σημειώνεται ότι το **β** είναι ένα διάνυσμα παραμέτρων, ενώ τα **Y** και **ε** είναι διανύσματα που περιέχουν τυχαίες μεταβλητές. Αφού το **ε** αποτελεί μια n – διάστατη τυχαία μεταβλητή, το διάνυσμα μέσος της **ε**, (ή απλά ο μέσος της **ε**) θα είναι ίσο με

$$\boldsymbol{\varepsilon} = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ E(\varepsilon_3) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix}$$

και λόγω της υπόθεσης $E(\varepsilon_i) = 0, i = 1,2,\dots,n$ (βλέπε συνθήκη Σ4) βρίσκουμε ότι $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, όπου με $\mathbf{0} = (0,0,\dots,0)'$ έχουμε συμβολίσει το μηδενικό διάνυσμα – στήλη διάστασης $n \times 1$. Επίσης, από την σχέση (2.3.19) προκύπτει ότι

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + E(\varepsilon_i)$$
$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}, \qquad i = 1,2,\dots,n$$

οπότε ο μέσος της n – διάστατης τυχαίας μεταβλητής **Y** θα είναι τελικά ίσος με

$$\mathrm{E}(\mathbf{Y}) = X\boldsymbol{\beta}.$$

Σε κάθε πολυδιάστατη τυχαία μεταβλητή, όπως το **ε**, αντιστοιχεί, πέραν του μέσου, και ένας τετραγωνικός πίνακας ο οποίος περιέχει όλες τις συνδιακυμάνσεις $Cov(\varepsilon_i, \varepsilon_j)$ για $i = 1,2,\dots,n$ και $j = 1,2,\dots,n$. Ο πίνακας αυτός λέγεται **πίνακας διακυμάνσεων – συνδιακυμάνσεων** (variance – covariance matrix) ή απλά πίνακας συνδιακυμάνσεων, συμβολίζεται με $D(\boldsymbol{\varepsilon})$ ή $\sigma^2(\boldsymbol{\varepsilon})$ και ορίζεται ως εξής:

$$\sigma^2(\boldsymbol{\varepsilon}) = D(\boldsymbol{\varepsilon}) = \begin{bmatrix} Cov(\varepsilon_1, \varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_2, \varepsilon_1) & Cov(\varepsilon_2, \varepsilon_2) & \cdots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & Cov(\varepsilon_n, \varepsilon_2) & \cdots & Cov(\varepsilon_n, \varepsilon_n) \end{bmatrix}.$$

Αφού ισχύει ότι $Cov(\varepsilon_i, \varepsilon_i) = V(\varepsilon_i) = \sigma^2, i = 1,2, \dots, n$ (βλέπε συνθήκη Σ4) και $Cov(\varepsilon_i, \varepsilon_j) = 0$ για $i \neq j$, ο πίνακας $\sigma^2(\boldsymbol{\varepsilon})$ θα είναι ίσος με

$$\sigma^2(\boldsymbol{\varepsilon}) = D(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

δηλαδή,

$$\sigma^2(\boldsymbol{\varepsilon}) = D(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

όπου με I_n συμβολίζουμε τον μοναδιαίο πίνακα διάστασης $n \times n$. Λόγω της σχέσης (2.3.19) είναι φανερό ότι

$$V(Y_i) = V(\varepsilon_i) = \sigma^2, i = 1,2, \dots, n$$

και

$$Cov(Y_i, Y_i) = Cov(\varepsilon_i, \varepsilon_i) = 0 \text{ για } i \neq j,$$

οπότε και για τη n – διάστατη τυχαία μεταβλητή $\mathbf{Y}$ του στατιστικού μοντέλου πολλαπλής παλινδρόμησης θα έχουμε

$$\sigma^2(\mathbf{Y}) = D(\mathbf{Y}) = \sigma^2 I_n.$$

Αφού στο στατιστικό μοντέλο πολλαπλής παλινδρόμησης έχουμε μια n – διάστατη τυχαία μεταβλητή $\mathbf{Y}$ της οποίας ο μέσος είναι συνάρτηση των παραμέτρων $\beta_0, \beta_1, \dots, \beta_{p-1}$, και πάλι ενδιαφέρον θα παρουσίαζε η εύρεση εκτιμητριών για αυτές. Όπως και στο απλό γραμμικό μοντέλο, έτσι και στο μοντέλο πολλαπλής παλινδρόμησης, ως κριτήριο καλής εκτίμησης χρησιμοποιούμε την ελαχιστοποίηση του αθροίσματος τετραγώνων

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} \right) \right)^2. \qquad (2.3.4)$$

Μετά λοιπόν τη διαδικασία της ελαχιστοποίησης, καταλήγουμε ότι οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ θα δίνονται από τον τύπο

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1} X'\mathbf{Y}$$

όπου $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})'$. Οι εκτιμήτριες αυτές, ως συναρτήσεις των τυχαίων μεταβλητών $Y_i, i = 1,2, \dots, n$, θα είναι επίσης τυχαίες μεταβλητές.

Επίσης, οι εκτιμήτριες ελαχίστων τετραγώνων του στατιστικού μοντέλου πολλαπλής παλινδρόμησης ικανοποιούν την ιδιότητα που περιγράφει το Θεώρημα των Gauss – Markov, δηλαδή έχουν τη μικρότερη δυνατή διακύμανση ανάμεσα σε όλες τις αμερόληπτες εκτιμήτριες που είναι γραμμικές συναρτήσεις των $Y_1, Y_2, \dots, Y_n$.

Ο πίνακας $X'X$ αναφέρεται συνήθως ως **πίνακας πληροφορίας** (information matrix) του πολλαπλού γραμμικού μοντέλου και είναι πολύ σημαντικός για τη μελέτη των εκτιμητριών ελαχίστων τετραγώνων του μοντέλου αυτού, αφού μας δίνει τη δυνατότητα να λαμβάνουμε πληροφορίες για τη διακύμανση και τις συνδιακυμάνσεις τους.

Ένα μοντέλο με δύο ανεξάρτητες μεταβλητές X_1, X_2, όπως αυτό που εξετάσαμε στην αρχή της ενότητας 2.3.1.1 είναι της μορφής

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

και λέγεται **προσθετικό μοντέλο** (additive model), αφού οι επιδράσεις των ανεξάρτητων μεταβλητών X_1, X_2 δρουν προσθετικά επί της μεταβλητής απόκρισης, χωρίς η μία να επηρεάζεται από την άλλη. Ένα **μη προσθετικό μοντέλο** (non additive model) με δύο ανεξάρτητες μεταβλητές X_1, X_2 είναι το μοντέλο

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

στο οποίο έχει συμπληρωθεί και ο πολλαπλασιαστικός όρος $x_1 x_2$. Σε ένα τέτοιο μοντέλο, αν αυξηθεί κατά μία μονάδα η τιμή της ανεξάρτητης μεταβλητής X_1, δηλαδή από x_1 να γίνει $x_1 + 1$, χωρίς να μεταβληθεί καθόλου η τιμή x_2της X_2, η μέση τιμή της μεταβλητής απόκρισης θα γίνει ίση με $E(Y') = E(Y) + \beta_1 + \beta_3 x_2$.

Από τον τελευταίο τύπο είναι φανερό ότι η αναμενόμενη μεταβολή της μεταβλητής απόκρισης είναι ίση με $\beta_1 + \beta_3 x_2$ και ως εκ τούτου εξαρτάται από την τιμή x_2 που έχει η ανεξάρτητη μεταβλητή X_2. Αντίστοιχο αποτέλεσμα έχουμε όταν αυξηθεί η τιμή της ανεξάρτητης μεταβλητής X_2 κατά μία μονάδα. Όταν λοιπόν εμφανίζεται το παραπάνω φαινόμενο θα λέμε ότι οι ανεξάρτητες μεταβλητές X_1 και X_2 παρουσιάζουν **αλληλεπίδραση** (interaction), ενώ ο όρος $x_1 x_2$ που το προκάλεσε θα λέγεται **όρος αλληλεπίδρασης** (interaction term). Για τη μελέτη μοντέλων με όρους αλληλεπίδρασης γίνεται χρήση των αποτελεσμάτων της Ιδιότητας 1.

2.3.1.3 Στατιστικές Ιδιότητες των Υπολοίπων

Μετά την εύρεση των εκτιμητριών ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ του στατιστικού μοντέλου πολλαπλής παλινδρόμησης, το επόμενο βήμα είναι να μελετήσουμε τις **προσαρμοσμένες τιμές**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}, \qquad i = 1,2,\dots,n \tag{2.3.5}$$

της μεταβλητής απόκρισης Y καθώς επίσης και τα **εκτιμημένα σφάλματα** ή **υπόλοιπα** (residuals)

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}\right), \qquad i = 1,2,\dots,n.$$

Αν ορίσουμε τα διανύσματα – στήλες

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{2.3.6}$$

οι ισότητες (2.3.5) και (2.3.6) μπορούν να γραφτούν σε πινακική μορφή, ως εξής:

$$\widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\beta}} \text{ και } \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$$

όπου X ο γνωστός πίνακας σχεδιασμού του μοντέλου. Αντικαθιστώντας την έκφραση για τις εκτιμήτριες ελαχίστων τετραγώνων, όπως είδαμε στην προηγούμενη ενότητα, στην πρώτη από τις παραπάνω σχέσεις παίρνουμε

$$\widehat{\mathbf{Y}} = X(X'X)^{-1}X'\mathbf{Y} = H\mathbf{Y}$$

όπου θέσαμε

$$H = X(X'X)^{-1}X'.$$

Επομένως, για το διάνυσμα των υπολοίπων $\hat{\boldsymbol{\varepsilon}}$ θα έχουμε ότι

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - H\mathbf{Y} = (I_{\mathrm{n}} - H)\mathbf{Y}.$$

Ο πίνακας H βρίσκει μεγάλη χρησιμότητα στη μελέτη του στατιστικού μοντέλου πολλαπλής παλινδρόμησης και είναι γνωστός στη διεθνή βιβλιογραφία με την ονομασία **hat matrix**.

Στη συνέχεια, έχουμε τις παρακάτω ιδιότητες για το στατιστικό μοντέλο πολλαπλής παλινδρόμησης.

Ιδιότητα 1: Για κάθε $i = 1,2,\dots,n$ η προσαρμοσμένη τιμή (2.3.5) είναι αμερόληπτη εκτιμήτρια της παραμετρικής συνάρτησης

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1},$$

η οποία αποτελεί τη μέση τιμή $E(Y_i)$ της μεταβλητής απόκρισης Y όταν οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_{p-1}$ λάβουν τις τιμές $x_{i1}, x_{i2}, \dots, x_{i,p-1}$ αντίστοιχα.

Ιδιότητα 2: Για τα υπόλοιπα $\hat{\varepsilon}_i = Y_i - \hat{Y}_i, i = 1,2, \dots, n$ ισχύουν οι σχέσεις

$$\text{α.} \sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \qquad \text{β.} \sum_{i=1}^{n} x_{ij}\hat{\varepsilon}_i = 0 \text{ για } j = 1,2, \dots, p-1$$

Ιδιότητα 3: Οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ του στατιστικού μοντέλου πολλαπλής παλινδρόμησης είναι ασυσχέτιστες με καθένα από τα υπόλοιπα $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n$, δηλαδή ισχύει

$$Cov(\hat{\beta}_0, \hat{\varepsilon}_i) = 0, \qquad Cov(\hat{\beta}_1, \hat{\varepsilon}_i) = 0, \qquad Cov(\hat{\beta}_{p-1}, \hat{\varepsilon}_i) = 0$$

για κάθε $i = 1,2, \dots, n$.

Τέλος, έχουμε και τα μέσα αθροίσματα τετραγώνων που δίνονται από τις σχέσεις

$$s^2 = MSE = \frac{SSE}{n-p}, \qquad MSR = \frac{SSR}{p-1}.$$

Το s^2 αποτελεί αμερόληπτη εκτιμήτρια της άγνωστης διακύμανσης σ^2, ενώ η τετραγωνική της ρίζα $s = \sqrt{MSE}$ μπορεί να χρησιμοποιηθεί ως (μη αμερόληπτη) εκτιμήτρια της τυπικής απόκλισης των ε_i (ή των Y_i), $i = 1,2, \dots, n$. Επιπλέον, χρησιμοποιώντας το s^2 στη θέση του σ^2, θα προκύψει ο πίνακας

$$s^2(\hat{\boldsymbol{\beta}}) = s^2(X'X)^{-1}$$

τα στοιχεία του οποίου δίνουν αμερόληπτες εκτιμήτριες των διακυμάνσεων και των συνδιακυμάνσεων των εκτιμητριών ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$.

2.3.1.4 Διαστήματα Εμπιστοσύνης για το Κανονικό Μοντέλο Πολλαπλής Παλινδρόμησης

Προκειμένου να μπορέσουμε να περάσουμε σε στατιστική συμπερασματολογία για τις παραμέτρους του στατιστικού μοντέλου πολλαπλής παλινδρόμησης, είναι απαραίτητο να θεωρήσουμε ότι οι τυχαίες μεταβλητές Y_i (ή ισοδύναμα τα σφάλματα ε_i), $i = 1,2, \dots, n$ ακολουθούν κάποια συγκεκριμένη κατανομή. Ξεκινώντας από ένα στατιστικό μοντέλο πολλαπλής παλινδρόμησης της μορφής (2.3.3), το οποίο ικανοποιεί τις συνθήκες Σ1 – Σ5 που δόθηκαν στην ενότητα 2.3.1.2, και υποθέτοντας ότι τα τυχαία σφάλματα $\varepsilon_i, i = 1,2, \dots, n$

είναι ανεξάρτητα και ακολουθούν την κανονική κατανομή $N(0,\sigma^2)$, φτάνουμε στο **κανονικό μοντέλο πολλαπλής παλινδρόμησης.**

Στη συνέχεια παρουσιάζονται κάποιες πολύ χρήσιμες ιδιότητες για την καλύτερη κατανόηση του κανονικού μοντέλου πολλαπλής παλινδρόμησης.

Ιδιότητα 1

Στο κανονικό μοντέλο πολλαπλής παλινδρόμησης ισχύουν τα ακόλουθα.

α. Οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_i$ των παραμέτρων $\beta_i, i = 1,2,\dots,p-1$ ακολουθούν τη (μονοδιάστατη) κανονική κατανομή $N\big(\beta_i,\sigma^2(\beta_i)\big)$, όπου τα $,\sigma^2(\beta_i)$ μπορούν να προσδιοριστούν από τον τύπο $\sigma^2\big(\hat{\boldsymbol{\beta}}\big) = \sigma^2(X'X)^{-1}$.

β. Ο λόγος

$$\frac{(n-p)s^2}{\sigma^2} = \frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

ακολουθεί την κατανομή χ^2 με $n-p$ βαθμούς ελευθερίας.

γ. Οι τυχαίες μεταβλητές

$$\frac{\hat{\beta}_i - \beta_i}{s\big(\hat{\beta}_i\big)}, i = 1,2,\dots,p-1$$

ακολουθούν την κατανομή t (ή κατανομή του Student) με $n-p$ βαθμούς ελευθερίας.

Το γ σκέλος της Ιδιότητας 2 μας δίνει τη δυνατότητα να δημιουργήσουμε διαστήματα εμπιστοσύνης για τις παραμέτρους του κανονικού μοντέλου πολλαπλής παλινδρόμησης, τα οποία είναι πανομοιότυπα με τα διαστήματα εμπιστοσύνης της απλής παλινδρόμησης. Πιο συγκεκριμένα, έχουμε το επόμενο αποτέλεσμα.

- Διάστημα εμπιστοσύνης για την παράμετρο β_i του μοντέλου

  Το διάστημα

  $$I_i = \big[\hat{\beta}_i - s\big(\hat{\beta}_i\big)t_{n-p}(\alpha/2), \hat{\beta}_i + s\big(\hat{\beta}_i\big)t_{n-p}(\alpha/2)\big]$$

  αποτελεί ένα διάστημα εμπιστοσύνης για την παράμετρο $\beta_i, i = 1,2,\dots,p-1$ του κανονικού μοντέλου πολλαπλής παλινδρόμησης με συντελεστή εμπιστοσύνης $1-\alpha$.

Για τη δημιουργία ταυτόχρονων διαστημάτων εμπιστοσύνης για οποιεσδήποτε r από τις p παραμέτρους $\beta_0, \beta_1, \dots, \beta_{p-1}$ του κανονικού μοντέλου πολλαπλής παλινδρόμησης, έχουμε μια μέθοδο γνωστή ως **μέθοδος Bonferroni** η οποία συνοπτικά δίνεται παρακάτω.

Ας υποθέσουμε λοιπόν ότι θέλουμε να κατασκευάσουμε ένα ταυτόχρονο διάστημα εμπιστοσύνης για τις παραμέτρους $\beta_1, \beta_2, \dots, \beta_r$ με συντελεστή εμπιστοσύνης $1 - a$. Συμβολίζοντας με I_1 το απλό διάστημα εμπιστοσύνης για την παράμετρο β_1 με συντελεστή εμπιστοσύνης $1 - a_1$, με I_2 το απλό διάστημα εμπιστοσύνης για την παράμετρο β_2 με συντελεστή εμπιστοσύνης $1 - a_2$ κ.ο.κ, και τέλος με I_r το απλό διάστημα εμπιστοσύνης για την παράμετρο β_r με συντελεστή εμπιστοσύνης $1 - a_r$, θα έχουμε

$$P(\beta_1 \in I_1) = 1 - \alpha_1, P(\beta_2 \in I_2) = 1 - \alpha_2, \dots, P(\beta_r \in I_r) = 1 - \alpha_r.$$

Μετά από πράξεις καταλήγουμε στην παρακάτω σχέση

$$P(\beta_1 \in I_1, \beta_2 \in I_2, \dots, \beta_r \in I_r) \geq 1 - (\alpha_1 + \alpha_2 + \cdots + \alpha_r)$$

και, αν τα $\alpha_1, \alpha_2, \dots, \alpha_r$ επιλεγούν έτσι ώστε να ισχύει

$$\alpha_1 + \alpha_2 + \cdots + \alpha_r = \alpha$$

συμπεραίνουμε ότι το ταυτόχρονο διάστημα εμπιστοσύνης που ορίζεται από τις ανισότητες

$$\hat{\beta}_i - s(\hat{\beta}_i)t_{n-p}(a_i/2) \leq \beta_i \leq \hat{\beta}_i + s(\hat{\beta}_i)t_{n-p}(a_i/2), i = 1,2, \dots, r$$

αποτελεί ένα διάστημα εμπιστοσύνης για τις παραμέτρους $\beta_1, \beta_2, \dots, \beta_r$ με συντελεστή εμπιστοσύνης **τουλάχιστον** $1 - \alpha$.

Κατά την κατασκευή ταυτόχρονων διαστημάτων εμπιστοσύνης με τη μέθοδο Bonferroni, η πλέον συνήθης επιλογή για τις ποσότητες είναι η συμμετρική

$$\alpha_1 = \alpha_2 = \cdots = \alpha_r = \frac{\alpha}{r}.$$

Καταλήγουμε επομένως στα ακόλουθα συμπεράσματα.

Έστω το κανονικό μοντέλο πολλαπλής παλινδρόμησης και $\widehat{\boldsymbol{\beta}}$ οι εκτιμήτριες ελαχίστων τετραγώνων των παραμέτρων $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})'$. Τότε, η περιοχή που ορίζεται από την ανισότητα

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' X' X (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq p s^2 F_{p,n-p}(a)$$

αποτελεί ένα ταυτόχρονο διάστημα εμπιστοσύνης για τις άγνωστες παραμέτρους $\beta_0, \beta_1, \dots, \beta_{p-1}$ με συντελεστή εμπιστοσύνης $1 - \alpha$.

Επίσης, έστω το κανονικό μοντέλο πολλαπλής παλινδρόμησης και ας συμβολίσουμε με Y_0 την τυχαία μεταβλητή που περιγράφει τη μεταβλητή απόκρισης, όταν οι ανεξάρτητες

μεταβλητές $X_1, X_2, \dots, X_{p-1}$ λάβουν τιμές $x_{01}, x_{02}, \dots, x_{0,p-1}$ αντίστοιχα. Τότε, ένα διάστημα εμπιστοσύνης για τη μέση πρόβλεψη

$$E(Y_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_{p-1} x_{0,p-1}$$

είναι το

$$\left[\hat{Y}_0 - s(\hat{Y}_0) t_{n-p}(\alpha/2), \hat{Y}_0 + s(\hat{Y}_0) t_{n-p}(\alpha/2)\right]$$

όπου

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_{p-1} x_{0,p-1} = \mathbf{x}_0' \boldsymbol{\beta}, \qquad \mathbf{x_0} = [1 \quad x_{01} \quad \cdots \quad x_{0,p-1}]'$$

και $s(\hat{Y}_0)$ η θετική τετραγωνική ρίζα της ποσότητας

$$s^2(\hat{Y}_0) = s^2(\mathbf{x}_0'(X'X)^{-1}\mathbf{x_0}).$$

Τέλος, ένα διάστημα πρόβλεψης με συντελεστή εμπιστοσύνης $1-\alpha$ $(0 < \alpha < 1)$ για την τιμή Y_0 της Y όταν οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_{p-1}$ λάβουν τι τιμές $x_{01}, x_{02}, \dots, x_{0,p-1}$ αντίστοιχα, δίνεται από τον τύπο

$$\left[\hat{Y}_0 - s\left(Y_0^{(n)}\right) t_{n-p}(\alpha/2), \hat{Y}_0 + s\left(Y_0^{(n)}\right) t_{n-p}(\alpha/2)\right]$$

όπου $s\left(Y_0^{(n)}\right)$ είναι η θετική τετραγωνική ρίζα της ποσότητας

$$s^2\left(Y_0^{(n)}\right) = s^2 + s^2(\hat{Y}_0) = s^2\left(1 + (\mathbf{x}_0'(X'X)^{-1}\mathbf{x_0})\right).$$

2.3.1.5 Έλεγχοι Υποθέσεων για το Κανονικό Μοντέλο Πολλαπλής Παλινδρόμησης

Εκμεταλλευόμενοι τη σχέση μεταξύ διαστήματος εμπιστοσύνης και αντίστοιχου αμφίπλευρου ελέγχου, μπορούνε να κατασκευάσουμε την κρίσιμη περιοχή για τους τελευταίους. Πιο συγκεκριμένα έχουμε.

- Έλεγχος αμφίπλευρης υπόθεσης

  Ο κανόνας απόφασης για τον έλεγχο της υπόθεσης

  $$H_0: \beta_i = 0$$
  $$H_1: \beta_i \neq 0$$

  στο κανονικό μοντέλο πολλαπλής παλινδρόμησης $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$ σε επίπεδο σημαντικότητας α, είναι ο εξής:

  - ✓ Αν ισχύει $|T| > t_{n-p}(\alpha/2)$, τότε απορρίπτουμε την H_0.
  - ✓ Αν ισχύει $|T| \leq t_{n-p}(\alpha/2)$, τότε δεν απορρίπτουμε την H_0.

Η συνάρτηση T που χρησιμοποιείται παραπάνω δίνεται από τον τύπο

$$T = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}.$$

Όταν μας ενδιαφέρουν οι μονόπλευροι έλεγχοι, έχουμε τα δύο παρακάτω αποτελέσματα.

- Έλεγχος μονόπλευρης υπόθεσης

  Ο κανόνας απόφασης για τον έλεγχο της υπόθεσης

  $$H_0: \beta_i = 0$$
  $$H_1: \beta_i > 0$$

  στο κανονικό μοντέλο πολλαπλής παλινδρόμησης σε επίπεδο σημαντικότητας α, είναι ο εξής:

  - ✓ Αν ισχύει $T > t_{n-p}(\alpha)$, τότε απορρίπτουμε την H_0.
  - ✓ Αν ισχύει $T \leq t_{n-p}(\alpha)$, τότε δεν απορρίπτουμε την H_0.

- Έλεγχος μονόπλευρης υπόθεσης

  Ο κανόνας απόφασης για τον έλεγχο της υπόθεσης

  $$H_0: \beta_i = 0$$
  $$H_1: \beta_i < 0$$

  στο κανονικό μοντέλο πολλαπλής παλινδρόμησης σε επίπεδο σημαντικότητας α, είναι ο εξής:

  - ✓ Αν ισχύει $T < -t_{n-p}(\alpha)$, τότε απορρίπτουμε την H_0.
  - ✓ Αν ισχύει $T \geq -t_{n-p}(\alpha)$, τότε δεν απορρίπτουμε την H_0.
  - ✓

Μπορούμε επίσης να κατασκευάσουμε και έναν κανόνα απόφασης ο οποίος θα βασίζεται στην λεγόμενη ***p*-value**. Η p-value (ή αλλιώς τιμή p) εκφράζει την πιθανότητα, ενώ ισχύει η $H_0: \beta_1 = 0$, η στατιστική συνάρτηση $T = \hat{\beta}_1 / s(\hat{\beta}_1)$ να λάβει τιμή μεγαλύτερη κατ' απόλυτη τιμή από την τιμή $|t|$, όπου t είναι η τιμή που προκύπτει για την T με βάση τα διαθέσιμα δεδομένα, δηλαδή

$$p = P(|T| > |t||H_0).$$

Επομένως, αν γνωρίζουμε την πιθανότητα αυτή, είμαστε σε θέση να λάβουμε απόφαση για την απόρριψη ή μη της μηδενικής υπόθεσης $H_0: \beta_1 = 0$ ακολουθώντας τον επόμενο κανόνα.

- Κανόνας απόφασης με τη χρήση του p-value

  Όταν γνωρίζουμε την τιμή p του p-value ενός ελέγχου, ο κανόνας απόφασης σε οποιοδήποτε επίπεδο σημαντικότητας α, διαμορφώνεται ως εξής:

  - ✓ Απορρίπτουμε την H_0, αν ισχύει $a > p$.
  - ✓ Δεν απορρίπτουμε την H_0, αν ισχύει $a < p$.

Για να εξετάσουμε το κατά πόσο οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_{p-1}$ είναι χρήσιμες, ως σύνολο, για την πρόβλεψη της Y, θα πρέπει να πραγματοποιήσουμε τον έλεγχο

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$
$$H_1: \beta_1 \neq 0 \text{ ή } \beta_2 \neq 0 \text{ ή } \dots \text{ ή } \beta_{p-1} \neq 0. \qquad (2.3.7)$$

Ο παραπάνω έλεγχος γίνεται με τη χρήση του λόγου F που είναι ίσος με

$$F = \frac{MSR}{MSE}.$$

Επομένως, ο κανόνας απόφασης θα είναι ο εξής:

- Έλεγχος F για το κανονικό μοντέλο πολλαπλής παλινδρόμησης

  Ένας κανόνας απόφασης για τον έλεγχο της υπόθεσης (2.3.7) σε επίπεδο σημαντικότητας α είναι ο εξής:

  - ✓ Αν ισχύει $F > F_{p-1,n-p}(\alpha)$, τότε απορρίπτουμε την H_0.
  - ✓ Αν ισχύει $F \leq F_{p-1,n-p}(\alpha)$, τότε δεν απορρίπτουμε την H_0.

Στη συνέχεια δίνεται ο κανόνας απόφασης στη γενική περίπτωση που θα θέλαμε να ελέγξουμε ταυτοχρόνως οποιεσδήποτε r γραμμικές συνθήκες μεταξύ των παραμέτρων του κανονικού μοντέλου πολλαπλής παλινδρόμησης.

- Έλεγχος της υπόθεσης $H_0: A\boldsymbol{\beta} = \mathbf{c}$

  Στο κανονικό μοντέλο πολλαπλής παλινδρόμησης $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ όπου $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$ θεωρούμε έναν πίνακα A διάστασης $r \times p$ και ένα διάνυσμα – στήλη $\mathbf{c}$ διάστασης $r \times 1$ (των οποίων τα στοιχεία είναι σταθεροί αριθμοί). Τότε ο έλεγχος της υπόθεσης

  $$H_0: A\boldsymbol{\beta} = \mathbf{c}$$
  $$H_1: A\boldsymbol{\beta} \neq \mathbf{c}$$

  σε επίπεδο σημαντικότητας α μπορεί να γίνει με τη βοήθεια της στατιστικής συνάρτησης

$$F = \frac{n-p}{r}\frac{\left(A\hat{\boldsymbol{\beta}}-\mathbf{c}\right)'(A(X'X)^{-1}A')^{-1}\left(A\hat{\boldsymbol{\beta}}-\mathbf{c}\right)}{\left(\mathbf{Y}-X\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y}-X\hat{\boldsymbol{\beta}}\right)} =$$

$$= \frac{\left(A\hat{\boldsymbol{\beta}}-\mathbf{c}\right)'(A(X'X)^{-1}A')^{-1}\left(A\hat{\boldsymbol{\beta}}-\mathbf{c}\right)}{rs^2}$$

χρησιμοποιώντας τον επόμενο κανόνα:

- ✓ Αν ισχύει $F > F_{r,n-p}(\alpha)$, τότε απορρίπτουμε την H_0.
- ✓ Αν ισχύει $F \leq F_{r,n-p}(\alpha)$, τότε δεν απορρίπτουμε την H_0.

2.4 Πολυμεταβλητή Ανάλυση

Όταν ενδιαφερόμαστε μόνο για ένα χαρακτηριστικό των ατόμων ή μελετάμε πολλά χαρακτηριστικά αλλά ανεξάρτητα το ένα από το άλλο, τότε οι κλασικές τεχνικές της περιγραφικής στατιστικής και της στατιστικής συμπερασματολογίας είναι αρκετές για να μας δώσουν μια ικανοποιητική εικόνα του πληθυσμού. Στην πράξη όμως είναι αρκετά συνηθισμένο να ενδιαφερόμαστε να εξετάσουμε περισσότερα από ένα χαρακτηριστικά συγχρόνως. Σε αυτές τις περιπτώσεις είναι αναγκαίο να καταφύγουμε σε τεχνικές της πολυμεταβλητής ανάλυσης (Multivariate Statistical Analysis). Στις επόμενες παραγράφους θα γίνει αναφορά στις τρεις από τις τεχνικές της πολυμεταβλητής ανάλυσης:

α. την ανάλυση κυρίων συνιστωσών (Principal Component Analysis – PCA),

β. την ανάλυση παραγόντων (Factor Analysis),

γ. την ανάλυση κατά συστάδες ή ομάδες (Cluster Analysis).

Στα πλαίσια της παρούσας διπλωματικής εργασίας θα γίνει χρήση και επομένως σύντομη περιγραφή των μεθόδων της Ανάλυσης Κυρίων Συνιστωσών και της Ανάλυσης κατά Συστάδες.

Η Ανάλυση Κυρίων Συνιστωσών είναι μια μέθοδος η οποία έχει ως στόχο να δημιουργήσει ένα μικρό αριθμό από γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους αλλά και να περιέχουν όσο γίνεται μεγαλύτερο μέρος της πληροφορίας που υπάρχει στις αρχικές μεταβλητές. Εφόσον επιτευχθεί αυτό, αντί για τα αρχικά δεδομένα, μπορούμε να αναπαραστήσουμε γραφικά τις πρώτες κύριες συνιστώσες έχοντας έτσι μια αξιόπιστη γραφική παρουσίαση των δεδομένων. Τα οφέλη από την υλοποίηση αυτής της διαδικασίας είναι η οικονομία αποθήκευσης των δεδομένων καθώς και ότι καταλήγουμε σε ένα σύνολο

ασυσχέτιστων μεταβλητών, κάτι το οποίο μπορεί να είναι κρίσιμης σημασίας για τη μελέτη των δεδομένων με εφαρμογή πρόσθετων στατιστικών μεθόδων.

Η Ανάλυση κατά Συστάδες εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών με σκοπό να δημιουργήσει συστάδες (ομάδες) από παρατηρήσεις που μοιάζουν μεταξύ τους. Μια επιτυχημένη εφαρμογή των τεχνικών της θα καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς, ενώ παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο. Με αυτόν τον τρόπο επιτυγχάνουμε την ευκολότερη και αποδοτικότερη επεξεργασία των δεδομένων που διαθέτουμε.

Υπάρχουν πολλά διαφορετικά μέτρα απόστασης που χρησιμοποιούνται στις παραπάνω μεθόδους, τα οποία όμως για λόγους συντομίας δεν θα παρουσιαστούν αναλυτικά παρά μόνο ονομαστικά. Κάποια από αυτά είναι μέτρα που χρησιμοποιούνται για τη μέτρηση της απόστασης ανάμεσα σε συνεχή δεδομένα, όπως η ευκλείδεια απόσταση, η απόσταση του Pearson, η απόσταση Manhattan, η απόσταση Minkowski, η απόσταση Chebyshev και άλλες. Άλλα πάλι είναι μέτρα που χρησιμοποιούνται για τη μέτρηση της απόστασης μεταξύ ατόμων/αντικειμένων στα οποία οι παρατηρήσεις που γίνονται περιγράφονται από δίτιμες μεταβλητές, όπως η απόσταση simple matching, η απόσταση των Rogers και Tanimoto και άλλες.

2.4.1 Ανάλυση Κυρίων Συνιστωσών

2.4.1.1 Εισαγωγή

Ο στόχος της Ανάλυσης Κυρίων Συνιστωσών είναι να αντικαταστήσει ένα σύνολο μεταβλητών $X_1, X_2, \dots, X_p$ με ένα σημαντικά μικρότερο πλήθος μεταβλητών $F_1, F_2, \dots$ το οποίο να αποτελείται από γραμμικούς συνδυασμούς των αρχικών και διατηρεί ένα σημαντικό μέρος των πληροφοριών που περιέχει το αρχικό σύνολο. Η πρώτη μεταβλητή F_1, η οποία θα περιέχει τη μέγιστη δυνατή πληροφορία, λέγεται **πρώτη κύρια συνιστώσα**, η επόμενη στη σειρά F_2 λέγεται **δεύτερη κύρια συνιστώσα** κ.ο.κ. Ως «πληροφορία» μιας ή περισσοτέρων μεταβλητών θεωρούμε στην PCA το ποσοστό της συνολικής διασποράς το οποίο μεταφέρεται σε αυτή.

2.4.1.2 Εύρεση της πρώτης κύριας συνιστώσας

Θεωρώντας τη νέα μεταβλητή

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

η οποία προκύπτει ως γραμμικός συνδυασμός των $X_1, X_2, \dots, X_p$, το ενδιαφέρον μας εστιάζεται στον προσδιορισμό των $a_1, a_2, \dots, a_p \in \mathbb{R}$ έτσι ώστε οι τιμές (scores) των n ατόμων για τη μεταβλητή Y, δηλαδή τα

$$y_i = a_1x_{i1} + a_2x_{i2} + \cdots + a_px_{ip}, \qquad i = 1,2, \dots, n$$

να 'διατηρούν' όσο το δυνατόν περισσότερο τις αποστάσεις που έχουν τα άτομα ως προς όλες τις αρχικές μεταβλητές. Αυτό θα επιτευχθεί αν καταφέρουμε να μεγιστοποιήσουμε την ποσότητα

$$SS_Y = nDis(Y)$$

όπου

$$Dis(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} f_i^2 = \mathbf{f}'\mathbf{f} = (Z\boldsymbol{\alpha})'(Z\boldsymbol{\alpha}) = \boldsymbol{\alpha}'Z'Z\boldsymbol{\alpha}$$

είναι ένα μέτρο μεταβλητότητας για το σύνολο των τιμών $y_1, y_2, \dots, y_n$ και θα λέγεται **διασπορά του συνόλου $\boldsymbol{N = \{x_1, x_2, \dots, x_n\}}$ κατά μήκος του διανύσματος $\boldsymbol{\alpha}$** και συμβολίζεται συνήθως με $Dis_{\boldsymbol{\alpha}}(N)$.

Η ποσότητα

$$Dis(N) = tr(Z'Z)$$

ονομάζεται **διασπορά του συνόλου *N***.

Οι πίνακες που αναφέρονται στους παραπάνω τύπους είναι οι εξής:

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \qquad Z = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix},$$

$$\mathbf{z}_i = \begin{bmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_2 \\ \vdots \\ x_{ip} - \bar{x}_p \end{bmatrix}, \qquad \boldsymbol{\alpha} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \text{ και}$$

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}, \qquad \bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}.$$

Βάσει των παραπάνω, η πρώτη κύρια συνιστώσα θα προκύπτει από το επόμενο αποτέλεσμα.

Πρόταση

Ο πίνακας $Z'Z$ έχει μη αρνητικές ιδιοτιμές, έστω $\lambda_1, \lambda_2, \dots, \lambda_p$. Ας θεωρήσουμε ότι $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ και ας συμβολίσουμε $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ τα αντίστοιχα μοναδιαία ιδιοδιανύσματα. Τότε

i. το διάνυσμα $\boldsymbol{\alpha}$ που μεγιστοποιεί την $\boldsymbol{\alpha}'Z'Z\boldsymbol{\alpha}$ είναι το μοναδιαίο διάνυσμα $\mathbf{u}_1$ που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή λ_1

ii. η μέγιστη τιμή της τετραγωνικής μορφής είναι ίση με λ_1, δηλαδή ισχύει

$$\max_{\|\boldsymbol{\alpha}\|=1} Dis_{\boldsymbol{\alpha}}(N) = Dis_{\mathbf{u}_1}(N) = \lambda_1.$$

Η μεταβλητή Y για την οποία επιτυγχάνεται η προαναφερθείσα μεγιστοποίηση λέγεται **πρώτη κύρια συνιστώσα** (first principal component).

Χρησιμοποιώντας τη γνωστή έκφραση του ίχνους ενός πίνακα ως άθροισμα όλων των ιδιοτιμών του, η διασπορά σου συνόλου σημείων N μπορεί να υπολογιστεί επίσης μέσω της έκφρασης

$$Dis(N) = tr(Z'Z) = \sum_{j=1}^{p} \lambda_j = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

Αξίζει να σημειωθεί ότι, οι βέλτιστες τιμές των συντελεστών $\alpha_1, \alpha_2, \dots, \alpha_p$ εξαρτώνται από τη μονάδα μέτρησης των χαρακτηριστικών $X_1, X_2, \dots, X_p$. Επομένως, αν επιθυμούμε να δημιουργήσουμε ένα γραμμικό συνδυασμό ο οποίος δεν θα επηρεάζεται από τις μονάδες μέτρησης των χαρακτηριστικών αυτών, θα πρέπει να προβούμε σε κανονικοποίηση του κάθε χαρακτηριστικού. Αυτό σημαίνει ότι για κάθε $j = 1,2, \dots, p$ θα πρέπει να γίνει η διαίρεση των αποκλίσεων $x_{ij} - \bar{x}_j, i = 1,2, \dots, n$ με την τετραγωνική ρίζα ενός μέτρου διασποράς των τιμών αυτών π.χ. του d_j.

2.4.1.3 Γεωμετρική ερμηνεία της πρώτης κύριας συνιστώσας

Η πρώτη κύρια συνιστώσα καθορίζει μία κατεύθυνση (ευθεία ε που ορίζεται από το μοναδιαίο διάνυσμα $\boldsymbol{\alpha}$) τέτοια ώστε οι προβολές των διαθέσιμων σημείων (δεδομένων) επάνω σε αυτήν να βρίσκονται όσο πιο κοντά γίνεται στις πραγματικές θέσεις των σημείων. Η κατεύθυνση αυτή ονομάζεται συνήθως **πρώτος κύριος άξονας** (first principal axis).

Προβάλλοντας τις διαθέσιμες παρατηρήσεις επάνω στον πρώτο κύριο άξονα μπορούμε να έχουμε μια αρκετά αξιόπιστη αναπαράσταση των δισδιάστατων παρατηρήσεων σε μια ευθεία (μονοδιάστατη αναπαράσταση). Αν οι αρχικές μας παρατηρήσεις ήταν τρισδιάστατες $(p = 3)$ θα είχαμε αντίστοιχη αναπαράσταση τρισδιάστατων παρατηρήσεων στη μία διάσταση. Όμοιος ισχυρισμός ισχύει και στη γενική περίπτωση $(p > 3)$, τότε όμως δεν υπάρχει αντίστοιχη εποπτική εικόνα. Το γεγονός αυτό, δηλαδή η δυνατότητα αναπαράστασης των παρατηρήσεων σε χώρο μικρότερης διάστασης αναφέρεται στη στατιστική ορολογία ως **dimensionality reduction** των δεδομένων (ελάττωση της διάστασης).

Η ανισότητα $Dis_{\boldsymbol{\alpha}}(N) \leq Dis(N)$ μας δείχνει ότι η πλέον ιδανική περίπτωση όταν αναζητούμε το καλύτερο διάνυσμα $\boldsymbol{\alpha}$ είναι να φτάσουμε στην ισότητα $Dis_{\boldsymbol{\alpha}}(N) = Dis(N)$ (αυτό θα συμβεί αν η ευθεία ε περάσει από όλα τα διαθέσιμα σημεία). Ως ένας δείκτης επίτευξης του στόχου, θα μπορούσε να χρησιμοποιηθεί το πηλίκο

$$\frac{Dis_{\boldsymbol{\alpha}}(N)}{Dis(N)} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \leq 1.$$

Το πηλίκο αυτό, εκφρασμένο ως ποσοστό, αναφέρεται συνήθως ως το **ποσοστό της συνολικής διασποράς των δεδομένων που εξηγείται από τον πρώτο κύριο άξονα** (ή εναλλακτικά από την πρώτη κύρια συνιστώσα).

Τέλος, η ποσότητα

$$CT(\mathbf{x}_i) = \frac{f_i^2}{Dis_{\mathbf{u}_1}(N)} = \frac{f_i^2}{\lambda_1} = \frac{f_i^2}{f_1^2 + f_2^2 + \cdots f_n^2}$$

λέγεται **συμβολή της i-οστής παρατήρησης** στη διαμόρφωση (της κατεύθυνσης) του πρώτου κύριου άξονα.

2.4.1.4 Στατιστική ερμηνεία της πρώτης κύριας συνιστώσας

Σύμφωνα με τα προηγούμενα, το αποτέλεσμα της εφαρμογής της Ανάλυσης Κυρίων Συνιστωσών είναι η δημιουργία μίας νέας μεταβλητής Y της μορφής

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

και η 'αντικατάσταση' των p-διάστατων παρατηρήσεων $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ των n ατόμων με τις (μονοδιάστατες) τιμές $y_i = \mathbf{x}_i'\boldsymbol{\alpha}, i = 1,2, \dots, n.$ Αν λοιπόν χρησιμοποιήσουμε την πρώτη κύρια συνιστώσα, θα πρέπει να πάρουμε $\boldsymbol{\alpha} = \mathbf{u}_1$ οπότε θα προκύψουν οι n τιμές

$$y_1 = \mathbf{x}_1'\mathbf{u}_1, \qquad y_2 = \mathbf{x}_2'\mathbf{u}_1, \dots, \qquad y_n = \mathbf{x}_n'\mathbf{u}_1$$

καθώς και οι κεντρικοποιημένες τιμές

$$f_1 = y_1 - \bar{y}, \qquad f_2 = y_2 - \bar{y}, \dots, \qquad f_n = y_n - \bar{y}.$$

Θέλοντας να διερευνήσουμε τη σχέση που έχει η νέα μεταβλητή Y με τις αρχικές $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ φαίνεται λογικό να καταφύγουμε σε ένα στατιστικό μέτρο συσχέτισης, για παράδειγμα στο δειγματικό συντελεστή συσχέτισης. Χρησιμοποιώντας λοιπόν το δείγμα των n ατόμων και θέλοντας να συγκρίνουμε την Y με την X_j (για κάποιο $j = 1,2, \dots, p$), οι διαθέσιμες πληροφορίες μας είναι

- για την Y οι τιμές $y_1, y_2, \dots, y_n$ (με αντίστοιχο δειγματικό μέσο $\bar{y}$),
- για την X_j οι τιμές $x_{1j}, x_{2j}, \dots, x_{nj}$ (με αντίστοιχο δειγματικό μέσο $\bar{x}_j$).

Επομένως, ο δειγματικός συντελεστής συσχέτισης μεταξύ της πρώτης κύριας συνιστώσας και της j αρχικής μεταβλητής X_j θα δίνεται από το γνωστό τύπο

$$r_{1j} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

ή ισοδύναμα, αφού $f_i = y_i - \bar{y}$, $z_{ij} = x_{ij} - \bar{x}_j$,

$$r_{1j} = \frac{\sum_{i=1}^{n} f_i z_{ij}}{\sqrt{\sum_{i=1}^{n} f_i^2}\sqrt{\sum_{i=1}^{n} z_{ij}^2}}.$$

Σημειώνεται ότι, ο τελευταίος τύπος δίνει ουσιαστικά το συντελεστή συσχέτισης μεταξύ των κεντρικοποιημένων scores $f_1, f_2, \dots, f_n$ και των κεντρικοποιημένων τιμών στην j μεταβλητή

$$z_{1j} = x_{1j} - \bar{x}_j, \qquad z_{2j} = x_{2j} - \bar{x}_j, \dots, \qquad z_{nj} = x_{nj} - \bar{x}_j.$$

Έχοντας βρει τη μέγιστη ιδιοτιμή λ_1 του πίνακα $Z'Z$ και το αντίστοιχο μοναδιαίο ιδιοδιάνυσμα $\mathbf{u}_1$ που αντιστοιχούν στον πρώτο κύριο άξονα, είναι πολύ πιο εύκολο να υπολογίσουμε τους συντελεστές συσχέτισης μέσω της επόμενης πρότασης.

Πρόταση

Ο συντελεστής συσχέτισης μεταξύ του πρώτου κύριου άξονα και της j μεταβλητής X_j δίνεται από τον τύπο

$$r_{1j} = \frac{\sqrt{\lambda_1}u_{1j}}{d_j}, \qquad j = 1,2, \dots, p$$

όπου u_{1j} είναι η j συντεταγμένη του ιδιοδιανύσματος $\mathbf{u}_1$ και

$$d_j^2 = \sum_{i=1}^{n} z_{ij}^2 = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = (n-1)s_j^2.$$

2.4.1.5 Οι υπόλοιπες κύριες συνιστώσες

Σύμφωνα με τη γεωμετρική ερμηνεία της ενότητας 2.4.1.3 η πρώτη κύρια συνιστώσα ορίζει μία ευθεία επάνω στην οποία μπορούμε να προβάλουμε τις p-διάστατες παρατηρήσεις μας έχοντας τη μικρότερη δυνατή απώλεια πληροφορίας. Η ιδέα αυτή θα μπορούσε να επεκταθεί χρησιμοποιώντας αντί ευθείες ένα επίπεδο με $\boldsymbol{\alpha}, \boldsymbol{\beta}$ τα αντίστοιχα μοναδιαία διανύσματα των κάθετων αξόνων που ορίζουν το επίπεδο αυτό.

Η συνολική διασπορά του συνόλου N θα είναι

$$Dis(N) = Dis_{\boldsymbol{\alpha}}(N) + Dis_{\boldsymbol{\beta}}(N) + \sum_{i=1}^{n} K_i M_i^2$$

όπου M_i είναι το σημείο του τρισδιάστατου χώρου στο οποίο αντιστοιχεί η i-οστή παρατήρηση $\mathbf{x}_i$ και K_i η προβολή του σημείου M_i στο επίπεδο. Επίσης, η ποσότητα

$$Dis_{\boldsymbol{\alpha},\boldsymbol{\beta}}(N) = Dis_{\boldsymbol{\alpha}}(N) + Dis_{\boldsymbol{\beta}}(N)$$

θα λέγεται **διασπορά του συνόλου N στο επίπεδο** που καθορίζεται από τα μοναδιαία (και κάθετα μεταξύ τους) διανύσματα $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Θέλοντας λοιπόν να έχουμε όσο το δυνατόν πιστότερη αναπαράσταση των σημείων επάνω στο επίπεδο, θα πρέπει να ελαχιστοποιήσουμε το άθροισμα

$$\sum_{i=1}^{n} K_i M_i^2 = Dis(N) - Dis_{\boldsymbol{\alpha},\boldsymbol{\beta}}(N)$$

ή ισοδύναμα, αφού το $Dis(N)$ δεν επηρεάζεται από την επιλογή του επιπέδου (δηλαδή των $\boldsymbol{\alpha}, \boldsymbol{\beta}$), θα πρέπει να μεγιστοποιήσουμε την ποσότητα $Dis_{\boldsymbol{\alpha},\boldsymbol{\beta}}(N)$. Αυτό μπορεί να γίνει επιλέγοντας αρχικά το $\boldsymbol{\alpha}$ έτσι ώστε να μεγιστοποιείται η $Dis_{\boldsymbol{\alpha}}(N) = \boldsymbol{\alpha}' Z' Z \boldsymbol{\alpha}$ και στη συνέχεια το $\boldsymbol{\beta}$ έτσι ώστε να είναι κάθετο στο $\boldsymbol{\alpha}$ και να μεγιστοποιείται η $Dis_{\boldsymbol{\beta}}(N) = \boldsymbol{\beta}' Z' Z \boldsymbol{\beta}$. Αν πάρουμε το $\boldsymbol{\alpha}$ ίσο με $\mathbf{u}_1$ θα έχουμε $\max_{\|\boldsymbol{\alpha}\|=1} Dis_{\boldsymbol{\alpha}}(N) = Dis_{\mathbf{u}_1}(N) = \lambda_1$ ενώ, το $\boldsymbol{\beta}$ θα πρέπει να επιλεγεί ως το μοναδιαίο ιδιοδιάνυσμα $\mathbf{u}_2$ που αντιστοιχεί στη δεύτερη μεγαλύτερη ιδιοτιμή λ_2 του πίνακα $Z'Z$ και θα ισχύει

$$Dis_{\mathbf{u}_2}(N) = \lambda_2.$$

Σε αντιστοιχία με το διάνυσμα $\mathbf{u}_1 = (\mathbf{u}_{11}, \mathbf{u}_{12}, \dots, \mathbf{u}_{1p})'$, το διάνυσμα $\mathbf{u}_2 = (\mathbf{u}_{21}, \mathbf{u}_{22}, \dots, \mathbf{u}_{2p})'$ θα δημιουργεί έναν δεύτερο γραμμικό συνδυασμό

$$Y_2 = u_{21} X_1 + u_{22} X_2 + \cdots + u_{2p} X_p$$

που θα λέγεται **δεύτερη κύρια συνιστώσα**.

Η διαδικασία κατασκευής κυρίων αξόνων μπορεί να συνεχιστεί με 3ο, 4ο,… άξονα (ο 3ος θα πρέπει να είναι κάθετος στο επίπεδο που ορίζουν οι δύο πρώτοι, από εκεί και πέρα χάνεται η εποπτεία). Χρησιμοποιώντας λοιπόν όλες τις ιδιοτιμές $\lambda_1, \lambda_2, \dots, \lambda_p$ του πίνακα $Z'Z$ και τα αντίστοιχα ιδιοδιανύσματα $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ θα έχουμε

$$Dis_{\mathbf{u}_j}(N) = \mathbf{u}_j' Z'Z \mathbf{u}_j = \lambda_j, j = 1,2, \dots, p$$

και η συνολική διασπορά $Dis(N) = tr(Z'Z)$ θα αναλύεται ως εξής

$$Dis(N) = tr(Z'Z) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Dis_{\mathbf{u}_j}(N) = \sum_{j=1}^{p} \mathbf{u}_j' Z'Z \mathbf{u}_j.$$

Επομένως, χρησιμοποιώντας τις j πρώτες συνιστώσες, το ποσοστό της διασποράς που επιτυγχάνουμε να ερμηνεύσουμε είναι ίσο με

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_j}{Dis(N)} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \quad 1 \le j \le p.$$

2.4.1.6 Επιλογή του πλήθους των κυρίων συνιστωσών που θα διατηρήσουμε

Η επιλογή του πλήθους των κυρίων συνιστωσών που θα διατηρήσουμε ώστε να περιγράψουμε ικανοποιητικά τα διαθέσιμα δεδομένα είναι ίσως το πιο σημαντικό κομμάτι της ανάλυσης, το οποίο δυστυχώς δεν έχει εύκολη και κοινώς αποδεκτή απάντηση. Από τις προηγούμενες παραγράφους είναι φανερό ότι, επιλέγοντας λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, θα προκύψει αναγκαστικά κάποια απώλεια πληροφορίας. Για λόγους οικονομίας και ευκολίας στην ερμηνεία των αποτελεσμάτων ενδιαφερόμαστε για όσο το δυνατόν μικρότερο αριθμό κυρίων συνιστωσών. Όμως, όσο μικρότερος είναι ο αριθμός των κύριων συνιστωσών που διατηρούμε, τόσο μεγαλύτερη είναι η απώλεια της πληροφορίας. Επομένως, γίνεται κατανοητό ότι το πόσες κύριες συνιστώσες θα πρέπει να διατηρήσουμε δεν έχει μόνο μία αποδεκτή λύση. Στη βιβλιογραφία υπάρχουν πολλά κριτήρια, κάποια από τα οποία παρουσιάζονται στη συνέχεια.

α. Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο (π.χ. 75%) και επιλέγουμε τον αριθμό των συνιστωσών έτσι ώστε όλες μαζί να εξηγούν μεγαλύτερο ποσοστό από το όριο που βάλαμε. Ως κριτήριο είναι πολύ απλό και εύκολο αλλά στην πράξη δεν δίνει πάντα καλά αποτελέσματα, ιδίως αν ο στόχος είναι αρκετά υψηλός (οπότε μπορεί να χρειαστεί να διατηρήσουμε ιδιαίτερα μεγάλο πλήθος κυρίων συνιστωσών).

β. Κριτήριο του Kaiser

Ο Kaiser προτείνει να διατηρούμε μόνο τις ιδιοτιμές που είναι μεγαλύτερες από τη μέση τιμή των ιδιοτιμών $\lambda_1, \lambda_2, \dots, \lambda_p$, δηλαδή από την ποσότητα

$$\bar{\lambda} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_p}{p} = \frac{Dis(N)}{p} = \frac{tr(Z'Z)}{p}.$$

Στην περίπτωση που εργαζόμαστε με τον πίνακα συσχετίσεων ισχύει ότι $tr(Z'Z) = p$ οπότε η μέση τιμή των $\lambda_1, \lambda_2, \dots, \lambda_p$ είναι ίση με $\bar{\lambda} = 1$. Επομένως τότε, σύμφωνα με το κριτήριο του Kaiser, διαλέγουμε τόσες συνιστώσες όσες ιδιοτιμές είναι μεγαλύτερες από τη μονάδα. Το κριτήριο αυτό είναι συνήθως η προκαθορισμένη επιλογή σε πολλά στατιστικά πακέτα.

γ. Ποσοστό της διακύμανσης που ερμηνεύεται για κάθε αρχική μεταβλητή

Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο για το ποσοστό της διακύμανσης που ερμηνεύεται για κάθε μία από τις αρχικές μεταβλητές. Επιλέγονται τόσες συνιστώσες ώστε να ερμηνεύεται για κάθε μεταβλητή ποσοστό τουλάχιστον ίσο με το προκαθορισμένο όριο που τέθηκε. Ωστόσο, το όριο αυτό είναι υποκειμενικό. Επίσης, αν κάποια μεταβλητή δεν ερμηνεύεται ικανοποιητικά από τις πρώτες κύριες συνιστώσες, το κριτήριο θα οδηγήσει σε διατήρηση μεγάλου αριθμού συνιστωσών.

δ. Scree plot

Το scree plot είναι ένα γράφημα που απεικονίζει τις ιδιοτιμές με βάση τη σειρά μεγέθους τους. Πιο συγκεκριμένα, στον οριζόντιο άξονα τοποθετείται η σειρά 1,2,3,… και στον κάθετο άξονα η τιμή της κάθε ιδιοτιμής. Εξετάζοντας το scree plot, εντοπίζουμε το σημείο στο οποίο το γράφημα γίνεται περίπου οριζόντιο και διατηρούμε τόσες συνιστώσες όσες υποδεικνύονται από το σημείο αυτό.

2.4.1.7 Το πιθανοθεωρητικό μοντέλο της Ανάλυσης Κυρίων Συνιστωσών

Η μέθοδος της Ανάλυσης Κυρίων Συνιστωσών όπως αυτή αναπτύχθηκε αρχικά από τον Pearson και τον Hotelling το 1933 και ενσωματώθηκε στη συνέχεια στα κλασσικά βιβλία πολυμεταβλητής ανάλυσης, παρουσιάζεται συνήθως μέσω μιας πιθανοθεωρητικής θεμελίωσης και όχι με τον τρόπο που προσεγγίστηκε στις προηγούμενες ενότητες.

Για λόγους συντομίας δεν θα γίνει περεταίρω αναφορά στην αρχική μέθοδο των Pearson και Hotelling. Μπορούμε απλά να αναφέρουμε ότι διαπιστώνεται ότι οι δύο προσεγγίσεις

έχουν μεγάλο κοινό μέρος και φυσικά καταλήγουν σε συγκρίσιμα στατιστικά συμπεράσματα (JOLLIFFE, I.T., 2002. Principal Component Analysis).

2.4.2 Ανάλυση Κατά Συστάδες

2.4.2.1 Εισαγωγή

Η ανάλυση κατά συστάδες εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών με σκοπό να δημιουργήσει συστάδες (ομάδες) από παρατηρήσεις που μοιάζουν μεταξύ τους. Μια επιτυχημένη εφαρμογή των τεχνικών της θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς ενώ παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο. Με αυτόν τον τρόπο επιτυγχάνουμε την ευκολότερη και αποδοτικότερη επεξεργασία των δεδομένων που διαθέτουμε.

2.4.2.2 Μέτρα ομοιότητας

Μια εναλλακτική κατηγορία μέτρων που μπορούν να χρησιμοποιηθούν για να μας δείξουν αν δύο άτομα (παρατηρήσεις) είναι όμοια ή ανόμοια μεταξύ τους είναι τα λεγόμενα μέτρα ομοιότητας (similarity measures ή affinity measures). Αυτά έχουν το χαρακτηριστικό γνώρισμα ότι παρατηρήσεις που μοιάζουν μεταξύ τους, δίνουν πολύ μεγάλη τιμή στο μέτρο της ομοιότητας, ενώ παρατηρήσεις που είναι ανόμοιες του δίνουν πολύ μικρή τιμή. Έχοντας εισάγει κατάλληλα τέτοια μέτρα, θα μπορούμε να τοποθετούμε ζεύγη παρατηρήσεων στην ίδια ή σε διαφορετικές ομάδες ανάλογα με το αν η τιμή του μέτρου είναι μεγάλη ή μικρή αντίστοιχα.

Αν υποθέσουμε ότι για κάθε ζεύγος παρατηρήσεων $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ ορίζεται ένας πραγματικός αριθμός $s_{ij} = s(x_i, x_j)$ έτσι ώστε να ισχύουν οι επόμενες τρεις ιδιότητες:

I1. $s_{ij} \geq 0$ για κάθε i, j και $i = j \Rightarrow s_{ij} = 1$

I2. $s_{ij} \leq 1$

I3. $s_{ij} = s_{ji}$ (συμμετρική ιδιότητα)

τότε θα λέμε ότι η συνάρτηση $s_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$ δίνει ένα μέτρο ομοιότητας.

Το πιο γνωστό μέτρο ομοιότητας για ποσοτικές παρατηρήσεις είναι ο (δειγματικός) συντελεστής συσχέτισης, που δίνεται από τον τύπο

$$s_{ij} = \frac{\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i\cdot})(x_{jr} - \bar{x}_{j\cdot})}{\left(\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i\cdot})^2 \sum_{r=1}^{p}(x_{jr} - \bar{x}_{j\cdot})^2\right)^{1/2}}$$

όπου $\bar{x}_{i\cdot} = \frac{1}{p}\sum_{r=1}^{p} x_{ir}$, $\bar{x}_{j\cdot} = \frac{1}{p}\sum_{r=1}^{p} x_{jr}$.

Τα πιο γνωστά μέτρα ομοιότητας που χρησιμοποιούνται για τις ανάγκες προβλημάτων ομαδοποίησης ατόμων στα οποία παρατηρούνται δίτιμες μεταβλητές δίνονται στον επόμενο πίνακα.

	Ονομασία	Τύπος	Επεξήγηση του μέτρου
1	*Simple matching*	$s_{ij} = \frac{a+d}{a+b+c+d}$	Ίσα βάρη για συμφωνίες «1-1» και «0-0»
2	*Rogers and Tanimoto*	$s_{ij} = \frac{a+d}{(a+d)+2(b+c)}$	Διπλάσιο βάρος για τις ασυμφωνίες
3	*Sokal and Sneath*	$s_{ij} = \frac{2(a+d)}{2(a+d)+(b+c)}$	Διπλάσια βάρη για συμφωνίες «1-1» και «0-0»
4	*Jaccard coefficient*	$s_{ij} = \frac{a}{a+b+c}$	Απουσία συμφωνιών «0-0» από αριθμητή και παρονομαστή
5	*Dice and Sorensen*	$s_{ij} = \frac{2a}{2a+b+c}$	Απουσία συμφωνιών «0-0» από αριθμητή και παρονομαστή. Διπλάσιο βάρος για τις συμφωνίες «1-1»

Παρατηρούμε ότι τα παραπάνω μέτρα ομοιότητας συνδέονται με τις αντίστοιχες αποστάσεις για δίτιμες μεταβλητές που δόθηκαν στην ενότητα 2.4.1, με τη σχέση $s_{ij} = 1 - d_{ij}$.

Σε αντιστοιχία με το μέτρο απόστασης που ορίσαμε για κατηγορικές μεταβλητές, ένα μέτρο ομοιότητας μεταξύ δύο ατόμων/αντικειμένων i και j, για τα οποία τα χαρακτηριστικά που μας ενδιαφέρουν είναι κατηγορικά, είναι το

$$s_{ij} = \frac{u}{p}$$

(simple matching similarity) όπου u είναι ο αριθμός των συμφωνιών, δηλαδή ο αριθμός των μεταβλητών για τις οποίες τα αντικείμενα i και j συμβαίνει να βρίσκονται στην ίδια κατάσταση και p είναι ο συνολικός αριθμός των μεταβλητών.

Αξίζει να σημειωθεί ότι, αν έχουμε ορίσει μία απόσταση d_{ij}, τότε μπορούμε να δημιουργήσουμε ένα αντίστοιχο μέτρο ομοιότητας με τη χρήση του τύπου

$$s_{ij} = \frac{1}{1 + d_{ij}}.$$

Όμοια, αν έχουμε ορίσει ένα μέτρο ομοιότητας s_{ij}, τότε μπορούμε να δημιουργήσουμε ένα αντίστοιχο μέτρο απόστασης με χρήση του τύπου

$$d_{ij} = \sqrt{2(1 - s_{ij})}.$$

Ο τελευταίος τύπος όμως δεν εξασφαλίζει την ισχύ της τριγωνικής ανισότητας. Ο Gower απέδειξε ότι, αν ο πίνακας $[s_{ij}]_{n\times n}$ με στοιχεία τις τιμές του μέτρου ομοιότητας για τα n άτομα, είναι μη αρνητικά ορισμένος, τότε η συνάρτηση d_{ij} που ορίστηκε παραπάνω ικανοποιεί και την τριγωνική ανισότητα.

2.4.2.3 Κατάταξη των μεθόδων ομαδοποίησης

Εκείνο που θα εξετάσουμε στις επόμενες ενότητες είναι πιο συστηματικές μέθοδοι ομαδοποίησης οι οποίες έχουν κάποια μαθηματική/στατιστική βάση. Οι μέθοδοι ομαδοποίησης μπορούν να χωριστούν σε δύο διαφορετικές κατηγορίες ανάλογα με τον τρόπο που προχωρούν στη διαμόρφωση των ομάδων: στις **μη ιεραρχικές** και στις **ιεραρχικές** μεθόδους.

Στις ιεραρχικές μεθόδους οι ομάδες σχηματίζονται σταδιακά είτε με συνένωση μικρότερων ομάδων σχηματίζοντας συνεχώς μεγαλύτερες ομάδες μέχρι να φτάσουμε να έχουμε όλα τα δεδομένα σε μια ομάδα (**συσσωρευτικές μέθοδοι**), είτε με τη διαίρεση ομάδων σε μικρότερες μέχρι να φτάσουμε σε μια κατάσταση όπου κάθε παρατήρηση να είναι από μόνη της μια ομάδα (**διαιρετικές μέθοδοι).**

Στις μη ιεραρχικές μεθόδους θεωρείται ότι ο αριθμός των ομάδων είναι γνωστός από πριν. Σε αυτές χρησιμοποιούμε έναν επαναληπτικό αλγόριθμο για να τοποθετούμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην εκάστοτε παρατήρηση.

Στο ερώτημα «ποια μέθοδο θα πρέπει να χρησιμοποιήσω για να πάρω τα καλύτερα αποτελέσματα» δεν υπάρχει ικανοποιητική γενική απάντηση. Καλό όμως είναι να έχουμε υπόψη τα εξής. Οι ιεραρχικές μέθοδοι καλό είναι να αποφεύγονται να χρησιμοποιούνται για μεγάλο πλήθος δεδομένων αφού απαιτούν πολύ χρόνο, μνήμη και υπολογιστική ισχύ. Επίσης,

υπάρχει η τάση να δημιουργούνται ομάδες με ανομοιογενές μέγεθος. Από την άλλη οι μη ιεραρχικές μέθοδοι ενώ δουλεύουν ικανοποιητικά με μεγάλα δείγματα και δημιουργούν ομάδες παραπλήσιου μεγέθους, επηρεάζονται αρκετά από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

2.4.2.4 Μη ιεραρχικές μέθοδοι ομαδοποίησης

Ο στόχος των μη ιεραρχικών μεθόδων είναι να ομαδοποιήσουν τις n μονάδες των δεδομένων σε k ομάδες, όπου το k είναι καθορισμένο από την αρχή. Αυτό αποτελεί έναν περιορισμό της μεθόδου καθώς, είτε πρέπει να τρέξουμε τον αλγόριθμο με διαφορετικές επιλογές ως προς το πλήθος των ομάδων, είτε πρέπει με κάποιον άλλο τρόπο να έχουμε καταλήξει στον αριθμό των ομάδων.

Ο μηχανισμός λειτουργίας των περισσότερων μη ιεραρχικών μεθόδων είναι

- να θεωρούν k συγκεκριμένα άτομα (μητρικά σημεία – seed points) και γύρω από αυτά να ταξινομούν τα υπόλοιπα στοιχεία έως ότου διαμορφωθούν οι επιθυμητές ομάδες ή
- να ξεκινούν με ένα αρχικό διαμερισμό (initial partition) των ατόμων σε k ομάδες και στη συνέχεια να μετακινούν τα άτομα μεταξύ των ομάδων έως ότου πετύχουν τον καλύτερο διαμερισμό.

Για τον τρόπο δημιουργίας των μητρικών σημείων υπάρχουν πολλές μέθοδοι όπως το να επιλέξουμε τα πρώτα k στη σειρά άτομα από τα δεδομένα, να αριθμήσουμε τα άτομα από το 1 έως το n και να διαλέξουμε αυτά με την αρίθμηση $n/k, 2n/k, \dots, (k-1)n/k$ και n κ.α. Σε κάποιες μη ιεραρχικές μεθόδους ομαδοποίησης είναι προτιμότερος ένας αρχικός διαμερισμός των ατόμων σε k ομάδες από το να βρίσκουμε μητρικά σημεία.

Υπάρχουν διάφοροι αλγόριθμοι υλοποίησης μη ιεραρχικών μεθόδων, οι οποίοι δουλεύουν επαναληπτικά και χρησιμοποιούν την έννοια του κέντρου μιας ομάδας (κέντρου βάρους – centroid) το οποίο δεν είναι τίποτα άλλο από τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Η διαφοροποίηση των μεθόδων έγκειται στο σημείο στο οποίο γίνεται η ανανέωση των κέντρων των ομάδων και η ταξινόμηση των υπολοίπων παρατηρήσεων σε αυτές. Συνήθως η απόσταση που χρησιμοποιείται για την κατάταξη είναι η ευκλείδεια, χωρίς φυσικά να αποκλείεται η χρήση και κάποιας άλλης από τις αποστάσεις που αναφέρθηκαν προηγουμένως.

Πολλές είναι οι μέθοδοι που χρησιμοποιούνται τόσο από τους στατιστικούς όσο και από τα στατιστικά προγράμματα, όπως η μέθοδος που προτάθηκε από τον Forgy, η **μέθοδος**

MacQueen ή ***k*-means method**, η μέθοδος των **αρχικών ομάδων** καθώς και παραλλαγές αυτών. Στην παρούσα διπλωματική εργασία δεν θα αναλυθούν περεταίρω οι μέθοδοι αυτοί.

2.4.2.5 Ιεραρχικές μέθοδοι ομαδοποίησης

Στις μη ιεραρχικές μεθόδους ομαδοποίησης έχουμε ένα διαμερισμό των n ατόμων σε προκαθορισμένο αριθμό k ομάδων. Σε αυτή την ενότητα θα αναφέρουμε μια σειρά από αλγόριθμους που παράγουν μια ιεραρχία ΄δενδροειδούς μορφής' όπου στα διάφορα στάδια το πλήθος k των ομάδων παίρνει όλες τις δυνατές τιμές από το 1 έως το n. Στο ένα άκρο αυτής της ιεραρχίας υπάρχει μία μόνο ομάδα που περιέχει n άτομα (διαιρετικές μέθοδοι) και στο άλλο άκρο υπάρχουν n ομάδες που η καθεμία περιέχει ένα μόνο άτομο (συσσωρευτικές μέθοδοι).

- **Συσσωρευτικές μέθοδοι**

Στην κατηγορία των συσσωρευτικών ιεραρχικών μεθόδων, ανήκουν μέθοδοι όπως η **μέθοδος της απλής συνένωσης ή πλησιέστερου γείτονα** (single linkage method or nearest neighbor method), η **μέθοδος της πλήρους συνένωσης ή του μακρινότερου γείτονα** (complete linkage method or furthest neighbor method), η **μέθοδος των σταθμισμένων μέσων** (weighted average linkage method), η **μέθοδος των κέντρων βάρους** (centroid method), η **μέθοδος του Ward** και άλλες.

Συγκρίνοντας τις διάφορες μεθόδους μεταξύ τους με δεδομένα προσομοίωσης έχει διαπιστωθεί ότι συνήθως, η καλύτερη ομαδοποίηση επιτυγχάνεται με τη μέθοδο του Ward και τη μέθοδο των σταθμισμένων μέσων. Η μέθοδος του κοντινότερου γείτονα είναι αυτή με τη χειρότερη επίδοση. Παρόλα αυτά, σε πολλά προβλήματα δεν είναι ξεκάθαρο ποια μέθοδος είναι προτιμότερη και η καθεμία δουλεύει καλύτερα με συγκεκριμένη μορφή δεδομένων. Φυσικά, αν οι ομάδες είναι αρκετά διαφορετικές μεταξύ τους, όλες οι μέθοδοι θα βρουν τη σωστή ομαδοποίηση.

Τα βασικά μειονεκτήματα των αλγόριθμων ιεραρχικής ομαδοποίησης είναι τα εξής:

- Συνήθως είναι ασύμφορες από άποψη υπολογιστικού φόρτου όταν θέλουμε να αναλύσουμε μεγάλα σύνολα δεδομένων, και ιδιαίτερα όταν το πλήθος των ατόμων που θέλουμε να ομαδοποιήσουμε είναι μεγάλο. Αυτό οφείλεται στο γεγονός ότι πρέπει κανείς να σχηματίσει και να αποθηκεύσει στη μνήμη του υπολογιστή ολόκληρο τον πίνακα αποστάσεων των ατόμων. Έτσι, ακόμα και αν η απόσταση είναι συμμετρική, χρειάζεται να υπολογίσει κανείς και να αποθηκεύσει $n(n-1)/2$

αποστάσεις. Αυτός ο πίνακας πρέπει να ανανεώνεται σε κάθε βήμα και επομένως θα πρέπει συνέχεια να διαβάζουμε και να γράφουμε στη μνήμη του υπολογιστή ένα μεγάλο πλήθος δεδομένων, πράγμα που οδηγεί σε χρονοβόρες υπολογιστικές διαδικασίες.

- Οι ομάδες που φτιάχνονται σε αρχικά βήματα δεν μπορούν να χωρίσουν και επομένως οι παρατηρήσεις που ενώνονται σε αρχικά βήματα μένουν μαζί για πάντα. Γενικά, η πορεία που θα ακολουθήσει ένας ιεραρχικός αλγόριθμος εξαρτάται από τον τρόπο που υπολογίζουμε την απόσταση ανάμεσα σε ομάδες. Πολύ συχνά καταλήγει στη δημιουργία ενός μικρού πλήθους ομάδων με πολλές παρατηρήσεις και αφήνει αρκετές παρατηρήσεις να είναι από μόνες τους ανεξάρτητες ομάδες.

- **Διαιρετικές μέθοδοι**

Παρότι οι ιεραρχικές μέθοδοι είναι σχεδόν συνυφασμένες με τις συσσωρευτικές μεθόδους, υπάρχει και μια άλλη μεγάλη κατηγορία, αυτή των διαιρετικών μεθόδων, οι οποίες εκτελούν την ακριβώς αντίθετη διαδικασία από τις συσσωρευτικές. Όπως αναφέραμε και σε προηγούμενες παραγράφους, οι διαιρετικές μέθοδοι ξεκινούν από μία μόνο ομάδα που περιέχει τα n άτομα για τα οποία έχουν καταμετρηθεί τα χαρακτηριστικά και τη διαιρούν σε όλο και μικρότερες ομάδες. Η λογική στην οποία βασίζονται οι διαιρετικοί αλγόριθμοι είναι, να βρίσκουν υποομάδες των ήδη διαμορφωμένων ομάδων που είναι περισσότερο απομακρυσμένες και να τις διαχωρίζουν. Έτσι, σε κάθε βήμα διαμερίζουν μια ομάδα σε δύο άλλες μικρότερες έως ότου φτάσουν στο σημείο όπου όλες οι ομάδες περιέχουν ένα μόνο στοιχείο.

Ο κύριος λόγος που οι διαιρετικές μέθοδοι δεν είναι αρκετά διαδεδομένες στην πράξη, είναι ότι απαιτούν πολύ περισσότερους υπολογισμούς από ότι οι συσσωρευτικές μέθοδοι. Ένας δημοφιλής αλγόριθμος διαιρετικής ομαδοποίησης είναι αυτός που έχει προταθεί από τους *Edwards & Cavalli-Sforza* (1965). Η διαδικασία που ακολουθεί είναι να επιλέγει σε κάθε βήμα από όλες τις δυνατές διαμερίσεις σε δύο ομάδες εκείνη η οποία ελαχιστοποιεί το άθροισμα των τετραγωνικών αποκλίσεων για τις δύο ομάδες. Η λογική της μεθόδου είναι παρόμοια με αυτήν του αλγόριθμου του Ward.

2.4.2.6 Επιλογή του πλήθους των ομάδων

Στις μεθόδους ομαδοποίησης που εξετάσαμε προκύπτουν δύο βασικά ερωτήματα:

- Σε ποιο σημείο θα πρέπει να σταματήσουμε μια ιεραρχική συσσωρευτική μέθοδο ώστε να έχουμε το βέλτιστο αριθμό ομάδων;
- Σε ποιο αριθμό ομάδων k πρέπει να διαμερίσουμε τα δεδομένα με μια μη ιεραρχική μέθοδο διαμερισμού;

Ένας απλός πρακτικός τρόπος εύρεσης του πλήθους των ομάδων είναι να εξετάσουμε το δενδρόγραμμα που προκύπτει από μία ιεραρχική συσσωρευτική μέθοδο και από αυτό να καθορίσουμε το βέλτιστο πλήθος. Πιο συγκεκριμένα, σε εκείνο το σημείο του δενδρογράμματος που παρατηρείται η μεγαλύτερη μεταβολή της ποσότητας που καταγράφεται στον οριζόντιο άξονα (απόσταση ή μέτρο απόστασης) μπορούμε να φέρουμε μια παράλληλη γραμμή προς τον κατακόρυφο άξονα και να δούμε σε πόσα σημεία τέμνει το δενδρόγραμμα. Το πλήθος k, για το οποίο παρατηρούμε μεγάλες αποστάσεις συνένωσης σε σχέση με το προηγούμενο ($k-1$ ομάδες) αποτελεί μια λογική τιμή για το βέλτιστο πλήθος των ομάδων.

Είναι φανερό ότι η απόφαση με βάση το κριτήριο αυτό είναι σε μεγάλο βαθμό υποκειμενική. Δίνουμε λοιπόν στη συνέχεια κάποιες πιο αντικειμενικές τεχνικές που βασίζονται σε μεθόδους ανάλυσης διακύμανσης και σε αποστάσεις από τα κέντρα των ομάδων.

Α. Αν $\bar{\mathbf{x}}_j$ είναι ο μέσος (κέντρο βάρους) της j ομάδας ($j = 1,2,\dots,k$) και $\bar{\mathbf{x}}$ είναι ο μέσος (κέντρο βάρους) όλων των παρατηρήσεων, τότε αναλύουμε το συνολικό άθροισμα τετραγωνικών αποκλίσεων στα εξής δύο μέρη

$$B = \sum_{j=1}^{k} n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})', \qquad W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{\mathbf{x}}_j)(x_{ij} - \bar{\mathbf{x}}_j)'$$

και επιδιώκουμε το B να είναι όσο το δυνατόν μεγαλύτερο και το W όσο το δυνατόν μικρότερο. Επειδή το κριτήριο αυτό θα οδηγεί πάντοτε σε επιλογές με μεγάλες τιμές του k, συνήθως επιλέγουμε την ομαδοποίηση για την οποία προκύπτει η μεγαλύτερη τιμή του λόγου

$$c = \frac{\frac{tr(B)}{k-1}}{\frac{tr(W)}{n-k}}.$$

Β. Αφού υπολογίσουμε τις ποσότητες

$$W_k = \sum_{i \in G} d(\mathbf{x}_i, \bar{\mathbf{x}}_k)\,, (j = 1,2, \dots, k) \text{ και } P_k = \sum_{j=1}^{k} W_j\,,$$

επιλέγουμε το k για το οποίο το P_k είναι ελάχιστο. Ένα εναλλακτικό κριτήριο προκύπτει αν αθροίσουμε τις αποστάσεις όλων των παρατηρήσεων από τον ολικό μέσο

$$T = \sum_{i=1}^{n} d(\mathbf{x}_i, \bar{\mathbf{x}})$$

και επιλέγουμε το k που μεγιστοποιεί έναν από τους παρακάτω δείκτες

$$R^2 = 1 - \frac{P_k}{T}, \qquad F = \frac{\frac{T - P_k}{k-1}}{\frac{P_k}{n-k}}.$$

Αξίζει να αναφερθεί ότι, η τελευταία στατιστική συνάρτηση μοιάζει με το κριτήριο F της ανάλυσης διακύμανσης.

Τέλος, μια άλλη αποδοτική μέθοδος που έχει προταθεί είναι να γίνεται προσθήκη νέας ομάδας αν η ποσότητα $(n - k + 1)P_k/P_{k+1}$ υπερβαίνει την τιμή 10, αλλιώς να θεωρείται ότι το πλήθος k των ομάδων στο οποίο φτάσαμε είναι το βέλτιστο.

2.5 Στατιστικός Έλεγχος Ποιότητας

2.5.1 Εισαγωγή

Ο Στατιστικός Έλεγχος Ποιότητας αποτελείται από ένα σύνολο μεθόδων ανάλυσης στατιστικών δεδομένων οι οποίες χρησιμοποιούνται για την ανάλυση της διαδικασίας διεκπεραίωσης ενός έργου ή των εκροών του. Το σύνολο αυτό μπορεί να χωριστεί σε τρία βασικά υποσύνολα που το καθένα περιέχει στατιστικές μεθόδους προσανατολισμένες σε διαφορετικές φάσεις της παραγωγικής διαδικασίας. Τα τρία αυτά στάδια είναι τα ακόλουθα

1. **Σχεδιασμός και Ανάλυση Πειραμάτων (Design of Experiments)**
2. **Στατιστικός Έλεγχος Διεργασιών (Statistical Process Control)**
3. **Δειγματοληψία Αποδοχής (Acceptance Sampling)**

Στην παρούσα διπλωματική εργασία θα εστιάσουμε σε θέματα σχετικά με το Στατιστικό Έλεγχο Διεργασιών ο οποίος περιέχει στατιστικές τεχνικές που είναι απαραίτητες για τον έλεγχο της παραγωγικής διεργασίας κατά τη διάρκεια της παραγωγής των προϊόντων.

Για να ικανοποιεί ένα προϊόν το χρήστη, πρέπει να παράγεται σύμφωνα με μία «σταθερή επαναλαμβανόμενη» διεργασία. Η διεργασία πρέπει να είναι ικανή να λειτουργεί με μικρή μεταβλητότητα γύρω από κάποιες τιμές στόχους που έχουν τεθεί στα ποιοτικά χαρακτηριστικά που πρέπει να διακρίνουν το τελικό προϊόν. Ο Στατιστικός Έλεγχος Διεργασιών είναι μια συλλογή εργαλείων που είναι χρήσιμα για την επίβλεψη της σταθερότητας μιας διεργασίας και μπορεί να εφαρμοστεί σε κάθε διαδικασία. Τα επτά κυριότερα εργαλεία που χρησιμοποιεί είναι τα ακόλουθα:

- Το Ιστόγραμμα, Διάγραμμα Μίσχου – Φύλλων (Histogram, Stem-and-Leaf Plot)
- Το Φύλλο Ελέγχου (Check Sheet)
- Το Διάγραμμα Pareto (Pareto Chart)
- Το Διάγραμμα Αιτίας-Αποτελέσματος (Cause-and-Effect Diagram)
- Το Διάγραμμα Συγκέντρωσης Ελαττωμάτων(Defect Concentration Diagram)
- Το Διάγραμμα Διασποράς ή Διακύμανσης (Scatter Plot)
- Το Διάγραμμα Ελέγχου (Control Chart)

Τα παραπάνω εργαλεία αναφέρονται και ως "the magnificent seven". Από αυτά τα εργαλεία, το διάγραμμα ελέγχου είναι αυτό που θα μας απασχολήσει στη συνέχεια.

2.5.2 Στατιστικός Έλεγχος Διεργασιών και Διάγραμμα Ελέγχου

2.5.2.1 Το πρόβλημα του Στατιστικού Ελέγχου Διεργασιών

Το κύριο αντικείμενο του Στατιστικού Ελέγχου Διεργασιών είναι η έγκαιρη ανίχνευση της εμφάνισης ειδικών αιτιών μεταβλητότητας σε μια διεργασία έτσι ώστε να προχωρήσουμε σε έρευνα και να προβούμε στις απαραίτητες διορθωτικές ενέργειες προτού κατασκευαστούν αρκετά προϊόντα **μη συμμορφούμενα** (non conforming) με τις προδιαγραφές. Τα διαγράμματα ελέγχου είναι μια τεχνική που χρησιμοποιείται ευρέως για την ανίχνευση σε πραγματικό χρόνο της εμφάνισης ειδικών αιτιών μεταβλητότητας σε μια διεργασία (on-line process monitoring).

Για να είναι αποτελεσματικός ο Στατιστικός Έλεγχος Διεργασιών θα πρέπει να συνοδεύεται απαραίτητα από ένα εκτός ελέγχου πρόγραμμα δράσης (out-of-control action plan, OCAP) το οποίο θα πρέπει να ενεργοποιείται κάθε φορά που το διάγραμμα ελέγχου παρέχει ενδείξεις εμφάνισης ειδικών αιτιών μεταβλητότητας στη διεργασία. Το *OCAP*

περιγράφει την ακολουθία των ελέγχων που πρέπει να γίνουν με σκοπό τον προσδιορισμό και τελικά την εξάλειψη των ειδικών αιτιών μεταβλητότητας.

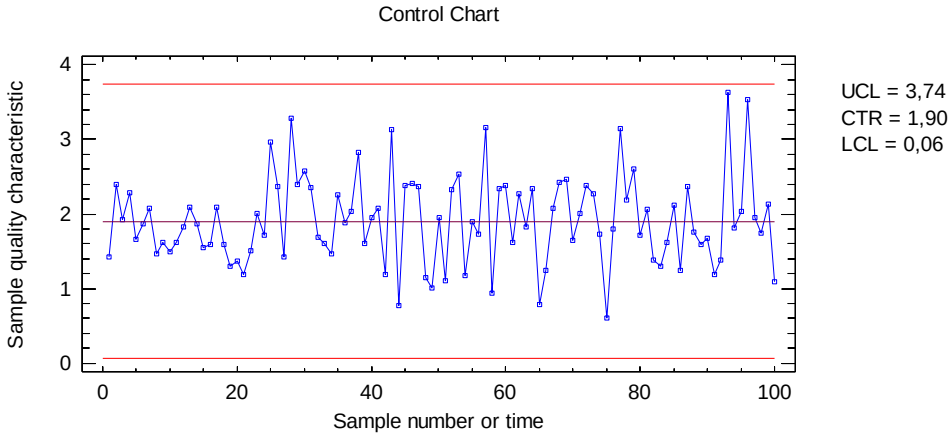
Επίσης, ανάλογα με το αν αναφερόμαστε σε *μη συμμορφούμενο ή ελαττωματικό προϊόν* ή σε *αριθμό ελαττωμάτων ή ατελειών* ενός προϊόντος, έχουμε τις δύο επόμενες κατηγορίες διαγραμμάτων ελέγχου ανάλογα με τον τύπο της μεταβλητής που περιγράφει το ποιοτικό χαρακτηριστικό του προϊόντος.

- Διαγράμματα ελέγχου για συνεχή χαρακτηριστικά – μεταβλητές (control charts for variables)
- Διαγράμματα ελέγχου για διακριτά δεδομένα – ιδιότητες (control charts for attributes)

2.5.2.2 Περιγραφή και Χρήση ενός Διαγράμματος Ελέγχου

Στις παραγωγικές διεργασίες μας ενδιαφέρει η παρακολούθηση της συμπεριφοράς μιας κρίσιμης ποσότητας ενός (μετρήσιμου) χαρακτηριστικού X (τυχαία μεταβλητή) των προϊόντων που παράγονται. Η διαδικασία της παρακολούθησης της κρίσιμης ποσότητας βασίζεται σε μετρήσεις του χαρακτηριστικού X, όπως προκύπτουν από την επιλογή δειγμάτων προϊόντων από την παραγωγή σε διαφορετικές χρονικές στιγμές στα οποία αντιστοιχούν τυχαία δείγματα τιμών του χαρακτηριστικού X, έστω τα $\mathbf{X}_1, \mathbf{X}_2, \dots$. Χρησιμοποιώντας τα τυχαία δείγματα $\mathbf{X}_1, \mathbf{X}_2, \dots$ υπολογίζουμε την τιμή $W_i = g(\mathbf{X}_i), i = 1,2, \dots$,μιας κατάλληλης στατιστικής συνάρτησης που εκτιμά (συνήθως μέσω αμερόληπτης εκτιμήτριας) την κρίσιμη ποσότητα που μας ενδιαφέρει. Έτσι, η παρακολούθηση της συμπεριφοράς της κρίσιμης ποσότητας επιτυγχάνεται με την παρακολούθηση των τιμών που λαμβάνει η στατιστική συνάρτηση W στα διάφορα δείγματα.

Για παράδειγμα, ας υποθέσουμε ότι μας ενδιαφέρει να παρακολουθήσουμε τη συμπεριφορά της μέσης τιμής μ ενός προϊόντος. Για το σκοπό αυτό επιλέγονται τυχαία δείγματα μεγέθους n από την παραγωγή σε διαφορετικά χρονικά διαστήματα και μπορούμε να χρησιμοποιήσουμε τη στατιστική συνάρτηση $W_i = g(\mathbf{X}_i) = (X_{i1} + X_{i2} + \cdots + X_{in})/n$ (η οποία είναι αμερόληπτη εκτιμήτρια του μέσου της κατανομής της X). Ένα τυπικό διάγραμμα ελέγχου είναι μια γραφική παράσταση με την ακόλουθη μορφή.



Όσο οι τιμές της *W* εμφανίζονται εντός των ορίων ελέγχου και η συμπεριφορά τους είναι «τυχαία», μπορούμε να υποθέσουμε ότι η διεργασία παραμένει εντός στατιστικού ελέγχου και δεν χρειάζεται να προβούμε σε κάποια διορθωτική ενέργεια. Αν όμως κάποιο σημείο βρεθεί εκτός των ορίων ελέγχου λέμε ότι υπάρχει **ένδειξη** ότι η διεργασία είναι εκτός στατιστικού ελέγχου, οπότε αντιμετωπίζουμε κατάσταση **συναγερμού** (alarm) και πρέπει να προχωρήσουμε σε έρευνα για να ανακαλύψουμε τις ειδικές αιτίες μεταβλητότητας που είναι υπεύθυνες γι' αυτή τη συμπεριφορά και αν κριθεί απαραίτητο να προβούμε σε διορθωτικές ενέργειες. Βέβαια, θα πρέπει να σημειώσουμε ότι ακόμα κι αν όλα τα σημεία βρίσκονται εντός ορίων ελέγχου, μπορεί να έχουμε ένδειξη για εκτός ελέγχου διεργασία αν τα σημεία αυτά συμπεριφέρονται με ένα **συστηματικό ή μη τυχαίο τρόπο.**

Μια εναλλακτική χρήση των διαγραμμάτων ελέγχου είναι η ανακάλυψη του τύπου της μεταβλητότητας που παρουσιάζεται σε μια διεργασία, η οποία μπορεί να παρουσιάζει στάσιμη συμπεριφορά με ασυσχέτιστα ή αυτοσυσχετιζόμενα δεδομένα ή μη στάσιμη συμπεριφορά σύμφωνα με την οποία τα δεδομένα κινούνται στο χώρο χωρίς λογική.

Σε πολλές περιπτώσεις, για να κάνουμε πιο ευαίσθητο το διάγραμμα ελέγχου ως προς την ικανότητά του να ανιχνεύει πιο γρήγορα εκτός ελέγχου διεργασίες, εκτός από τα όρια ελέγχου, σχεδιάζουμε επίσης και προειδοποιητικά όρια (warning limits) εσωτερικά των ορίων ελέγχου, τα οποία μπορεί να είναι εξωτερικά και εσωτερικά προειδοποιητικά όρια και χρησιμοποιούνται μαζί με κάποιους «κανόνες» που περιγράφουν ενδεχόμενα που σχετίζονται με την εμφάνιση ειδικών ακολουθιών σημείων σε ένα διάγραμμα ελέγχου. Στην περίπτωση που συμβεί το ενδεχόμενο που περιγράφει ο κανόνας, τότε θεωρούμε ότι η διεργασία είναι

εκτός στατιστικού ελέγχου χωρίς απαραίτητα να έχουμε κάποιο σημείο εκτός των ορίων ελέγχου.

2.5.2.3 Κατασκευή ενός Τυπικού Διαγράμματος Ελέγχου Τύπου Shewhart για τη Μέση Τιμή

Έστω ότι υπό συνθήκες φυσικής μεταβλητότητας, κατανομή της X είναι η κανονική με μέση τιμή μ και διακύμανση σ^2, και έστω ότι μας ενδιαφέρει να παρακολουθήσουμε τη συμπεριφορά της μέσης τιμής μ της X. Χρησιμοποιώντας ως εκτίμηση της μέσης τιμής μ της διεργασίας σε κάθε δείγμα το δειγματικό μέσο $W_i = \bar{X}_i$ και επιλέγοντας $z_{a/2} = 3$, δηλαδή $\alpha = 0.0027$ έχουμε ότι ο δειγματικός μέσος $W_i = \bar{X}_i$ θα βρίσκεται εντός του διαστήματος

$$\left[\mu_{\bar{X}_i} - 3\sigma_{\bar{X}_i}, \mu_{\bar{X}_i} + 3\sigma_{\bar{X}_i}\right]$$

με πιθανότητα $100(1-\alpha)\% = 99.73\%$ και εκτός του παραπάνω διαστήματος με πιθανότητα 0.27% υπό την προϋπόθεση ότι η διεργασία βρίσκεται συνεχώς εντός στατιστικού ελέγχου. Το παραπάνω διάστημα αποτελεί την περιοχή μη απόρριψης της μηδενικής υπόθεσης $H_0: \mu = \mu_{\bar{X}_i}$ *έναντι της* $H_1: \mu \neq \mu_{\bar{X}_i}$ σε επίπεδο σημαντικότητας $\alpha = 0.0027$.

Επομένως, θα μπορούσαμε να κατασκευάσουμε ένα διάγραμμα ελέγχου για τον δειγματικό μέσο $W_i = \bar{X}_i$ με $CL = \mu_{\bar{X}_i}, UCL = \mu_{\bar{X}_i} + 3\sigma_{\bar{X}_i}$ και $LCL = \mu_{\bar{X}_i} - 3\sigma_{\bar{X}_i}$. Το διάγραμμα αυτό θα απεικονίζει τα αποτελέσματα των διαδοχικών ελέγχων της υπόθεσης $H_0: \mu = \mu_{\bar{X}_i}$ – $H_1: \mu \neq \mu_{\bar{X}_i}$ με σταθερό σ σε επίπεδο σημαντικότητας $\alpha = 0.0027$. Κάθε σημείο του διαγράμματος που βρίσκεται εντός των ορίων ελέγχου αντιστοιχεί σε μη απόρριψη της μηδενικής υπόθεσης, βάσει του αντίστοιχου τυχαίου δείγματος και συνεπώς μπορούμε να υποθέσουμε ότι το διάγραμμα δείχνει ότι δεν έχει εμφανιστεί ειδική αιτία μεταβλητότητας που έχει ως αποτέλεσμα τη μετατόπιση της μέσης τιμής μ της διεργασίας.

Όπως είναι λογικό γεννάται το ερώτημα «γιατί μας ενδιαφέρει αν μια παραγωγική διαδικασία είναι εκτός ελέγχου στην περίπτωση που οι μετρήσεις βρίσκονται εντός των ορίων προδιαγραφών;». Η απάντηση στο ερώτημα αυτό είναι ότι η εμφάνιση ειδικής μεταβλητότητας μεταφράζεται σε αύξηση των παραγόμενων προϊόντων που έχουν τιμές εκτός των ορίων προδιαγραφών. Επίσης, όσο πιο κοντά στα όρια ελέγχου κινούνται οι τιμές της στατιστικής συνάρτησης W, τόσο περισσότερο φθίνει η ποιότητα του προϊόντος και συνεπώς υπάρχει αυξημένος κίνδυνος να μην ικανοποιεί την ανάγκη για την οποία παράγεται.

2.5.2.4 Διαγράμματα Ελέγχου – Ορολογία – Αρχές

Στην παρούσα ενότητα θα περιγράψουμε ορισμένες βασικές έννοιες και κάποια ιδιαίτερα χαρακτηριστικά που συνοδεύουν την κατασκευή ενός διαγράμματος ελέγχου.

- **Μέγεθος δείγματος και συχνότητα δειγματοληψίας**

Κατά τον σχεδιασμό ενός διαγράμματος ελέγχου πρέπει να καθοριστεί το μέγεθος των δειγμάτων των προϊόντων και η συχνότητα δειγματοληψίας. Γενικά, μεγάλα μεγέθη δειγμάτων κάνουν πιο εύκολη την ανίχνευση μικρών μετατοπίσεων του μέσου επιπέδου της διεργασίας. Γι' αυτό το σκοπό μπορούμε να κατασκευάσουμε μια γραφική παράσταση που να μας δείχνει αυτές τις μεταβολές και ονομάζεται **χαρακτηριστική ή χαρακτηρίζουσα (λειτουργική) καμπύλη** (operating curve). Με τη χρήση αυτής της καμπύλης μπορεί να γίνει κατανοητό ότι το κατάλληλο μέγεθος δείγματος εξαρτάται από το είδος της μετατόπισης (μικρής ή μεγάλης) του μέσου επιπέδου της διεργασίας που θέλουμε να ανιχνεύσουμε.

Οι πολιτικές δειγματοληψίας θα πρέπει να καθορίζονται και με βάση το οικονομικό κόστος. Επομένως, γενικά, η πολιτική που ακολουθείται είναι να γίνεται λήψη μικρών δειγμάτων αρκετά συχνά παρά μεγάλων δειγμάτων λιγότερο συχνά.

- **Μέσο μήκος ροής (Average run length ARL)**

Μια άλλη έννοια που σχετίζεται με τα διαγράμματα ελέγχου είναι το μέσο μήκος ροής του διαγράμματος που ορίζεται από τη σχέση

$$ARL = \frac{1}{p}$$

όπου p συμβολίζει την πιθανότητα να βρεθεί ένα σημείο του διαγράμματος ελέγχου εκτός των ορίων ελέγχου. Είναι προφανές ότι η ποσότητα ARL δηλώνει τον αναμενόμενο αριθμό των σημείων που πρέπει να σχεδιαστούν σε ένα διάγραμμα ελέγχου για να εμφανιστεί ένα σημείο εκτός των ορίων ελέγχου, αφού το μήκος ροής ακολουθεί την Γεωμετρική κατανομή με παράμετρο p, δηλαδή $G(p)$. Για μια διεργασία που βρίσκεται εντός ελέγχου και στην περίπτωση που χρησιμοποιούμε 3σ όρια ελέγχου και η κατανομή της W είναι κανονική, έχουμε ότι το **εντός ελέγχου μέσο μήκος ροής** ARL_0 (in-control average run length) είναι ίσο με

$$ARL_0 = \frac{1}{0.0027} \cong 370.$$

Προφανώς στην πράξη, θέλουμε να έχουμε μεγάλο ARL_0.

Για μια διεργασία που βρίσκεται εκτός ελέγχου λόγω μετατόπισης του μέσου επιπέδου της διεργασίας από μ σε μ^*, το **εκτός ελέγχου μέσο μήκος ροής** ARL_1 (out-of-control average run length) είναι ίσο με

$$ARL_1 = \frac{1}{1-\beta}.$$

Το ARL_1 δηλώνει τον αναμενόμενο αριθμό δειγμάτων που πρέπει να ληφθούν για να εντοπιστεί η αλλαγή στο μέσο επίπεδο της διεργασίας από τη στιγμή που αυτή η μετατόπιση συνέβη. Επομένως, είναι προφανές ότι στην πράξη θέλουμε να έχουμε μικρό ARL_1.

Η χρήση του ARL ως μέτρου για την περιγραφή της απόδοσης μιας διεργασίας έχει υποστεί αρκετή κριτική τα τελευταία χρόνια γιατί το ARL που παρατηρείται στην πράξη διαφέρει συνήθως αρκετά από το «θεωρητικό» ARL, αφού η κατανομή του μήκους ροής είναι πολύ ασύμμετρη και συνεπώς η μέση τιμή δεν μπορεί να θεωρηθεί ως αντιπροσωπευτικό μέτρο κεντρικής τάσης της κατανομής (ιδιαίτερα για μικρές τιμές του p).

Στην πράξη, πέραν του μέσου μήκους ροής, χρησιμοποιείται αρκετά συχνά ο **μέσος όρος μέχρι την εμφάνιση σήματος** ATS (average time to signal) που ορίζεται από τη σχέση

$$ATS = ARL \times h$$

όπου h συμβολίζει το χρόνο που μεσολαβεί για τη λήψη δύο διαδοχικών δειγμάτων (θεωρούμε ότι το h είναι σταθερό). Συνεπώς η ποσότητα ATS δηλώνει το μέσο χρόνο που απαιτείται για να δώσει το διάγραμμα ένδειξη εκτός ελέγχου διεργασίας.

- **Φάση I και Φάση II**

Στη βιβλιογραφία υπάρχουν δύο φάσεις για τον έλεγχο μιας παραγωγικής διεργασίας με τη χρήση διαγραμμάτων ελέγχου, η Φάση I και η Φάση II.

Φάση I (Phase I): Σε αυτή τη φάση τα διαγράμματα ελέγχου χρησιμοποιούνται αναδρομικά για να ελέγξουν αν η διεργασία ήταν εντός ή εκτός ελέγχου εξετάζοντας δείγματα που συλλέχθηκαν σε παρελθόντα χρόνο. Σε αυτή τη φάση τα διαγράμματα ελέγχου βοηθούν τον χειριστή της διαδικασίας να φέρει τη διεργασία εντός στατιστικού ελέγχου. Όταν αυτό επιτευχθεί, τα διαγράμματα ελέγχου που προκύπτουν είναι κατάλληλα για την παρακολούθηση της μελλοντικής συμπεριφοράς της διεργασίας. Αυτή η χρήση των διαγραμμάτων ελέγχου αναφέρεται και ως **αναδρομική** (retrospective).

Φάση II (Phase II): Σε αυτή τη φάση τα διαγράμματα ελέγχου χρησιμοποιούνται προκειμένου να ελέγχουμε συνεχώς αν η διαδικασία παραμένει εντός ελέγχου. Στη φάση αυτή, ο διαχειριστής έχει στα χέρια του ένα πολύτιμο εργαλείο μέσω του οποίου είναι

δυνατόν να παρακολουθεί συνεχώς την παραγωγική διεργασία και να ανιχνεύει εγκαίρως μια πιθανή αλλαγή στο μέσο επίπεδο των χαρακτηριστικών που καθορίζουν την ποιότητα του παραγόμενου προϊόντος. Δηλαδή, σε κάθε χρονική περίοδο που ένα δείγμα λαμβάνεται από τη διεργασία, ο διαχειριστής παίρνει μια απάντηση στο ερώτημα «παραμένει η διεργασία εντός ελέγχου;». Σε αυτή τη φάση ο διαχειριστής αδιαφορεί για τον τρόπο με τον οποίο το μέσο επίπεδο της διεργασίας είχε εκτιμηθεί ή αν αυτό ήταν εκ των προτέρων γνωστό.

Πολλές φορές, η Φάση *II* χαρακτηρίζεται ως **On-Line Control Phase** ή ως **Control to Standard Phase**, ενώ η Φάση *I* χαρακτηρίζεται ως **Off-Line Control Phase** ή ως **Initial Study Phase**.

2.5.2.5 Ταξινόμηση Διαγραμμάτων Ελέγχου

Τα διαγράμματα ελέγχου μπορούν να ταξινομηθούν σε πολλές κατηγορίες ανάλογα με ορισμένα χαρακτηριστικά τους. Έτσι,

1. Ανάλογα με το είδος της μεταβλητής που περιγράφει το ποιοτικό χαρακτηριστικό του προϊόντος, έχουμε **διαγράμματα ελέγχου για μεταβλητές** (control charts for variables) και **διαγράμματα ελέγχου για ιδιότητες** (control charts for attributes).
2. Αν από την παραγωγική διεργασία λαμβάνονται δείγματα μετρήσεων μεγέθους μεγαλύτερου της μονάδας, αναφερόμαστε σε **διαγράμματα ελέγχου για ομάδες** (control charts for rational subgroups), ενώ αν λαμβάνονται δείγματα μετρήσεων μεγέθους ένα, αναφερόμαστε σε **διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις** (control charts for individual observations).
3. Αν οι μετρήσεις που λαμβάνονται σε κάθε χρονική στιγμή t είναι εξαρτημένες από τις μετρήσεις που λαμβάνονται στο χρόνο $t-1$, αναφερόμαστε σε **διαγράμματα ελέγχου για αυτοσυσχετιζόμενες διεργασίες** (control charts for autocorrelated processes), ενώ αν οι μετρήσεις που λαμβάνονται σε κάθε χρονική στιγμή t είναι ανεξάρτητες από τις μετρήσεις που λαμβάνονται στο χρόνο $t-1$, αναφερόμαστε σε **διαγράμματα ελέγχου για ασυσχέτιστες διεργασίες** (control charts for uncorrelated processes).
4. Αν οι μετρήσεις που λαμβάνονται αφορούν ένα ποιοτικό χαρακτηριστικό, αναφερόμαστε σε **μονομεταβλητά διαγράμματα ελέγχου**, ενώ αν οι μετρήσεις αφορούν περισσότερα ποιοτικά χαρακτηριστικά, αναφερόμαστε σε **πολυμεταβλητά διαγράμματα ελέγχου**.

5. Αν οι μετρήσεις που λαμβάνονται ακολουθούν μια γνωστή κατανομή, τότε αναφερόμαστε σε **παραμετρικά διαγράμματα ελέγχου**, ενώ σε αντίθετη περίπτωση, αναφερόμαστε σε **μη παραμετρικά διαγράμματα ελέγχου**.

Επίσης, ανάλογα με τη στατιστική θεωρία που στηρίζει την κατασκευή ενός διαγράμματος ελέγχου, δημιουργούνται κάποιες δευτερεύουσες κατηγορίες διαγραμμάτων οι οποίες είναι οι κάτωθι.

1. **Διαγράμματα ελέγχου τύπου Shewhart** (Shewhart type control chart). Τα διαγράμματα αυτά χρησιμοποιούνται κυρίως στις κάτωθι περιπτώσεις:
   a) Όταν οι αλλαγές του μέσου επιπέδου που θέλουμε να ανιχνεύσουμε δεν είναι μικρές.
   b) Όταν γνωρίζουμε την κατανομή των αρχικών παρατηρήσεων ή του δειγματικού χαρακτηριστικού που απεικονίζεται στο διάγραμμα.
2. **Διαγράμματα ελέγχου τύπου CUSUM** (Cumulative SUM – CUSUM type control charts). Τα διαγράμματα αυτά χρησιμοποιούνται κυρίως στις κάτωθι περιπτώσεις:
   a) Όταν οι αλλαγές του μέσου επιπέδου που θέλουμε να ανιχνεύσουμε είναι μικρές.
   b) Όταν γνωρίζουμε την κατανομή των αρχικών παρατηρήσεων ή του δειγματικού χαρακτηριστικού που απεικονίζεται στο διάγραμμα.
3. **Διαγράμματα ελέγχου τύπου EWMA** (Exponentially weighted moving average – EWMA type control charts). Τα διαγράμματα αυτά χρησιμοποιούνται κυρίως στις κάτωθι περιπτώσεις:
   a) Όταν οι αλλαγές του μέσου επιπέδου που θέλουμε να ανιχνεύσουμε είναι μικρές και σταδιακές (όχι απότομες).
   b) Όταν δε γνωρίζουμε την κατανομή των αρχικών παρατηρήσεων ή του δειγματικού χαρακτηριστικού που απεικονίζεται στο διάγραμμα.

2.5.3 Διαγράμματα Ελέγχου Τύπου Shewhart για Μεταβλητές

Τα διαγράμματα αυτά χρησιμοποιούνται για την παρακολούθηση της μέσης τιμής και της διασποράς των τιμών του χαρακτηριστικού ποιότητας X. Επίσης, χρησιμοποιούνται ώστε να είναι εφικτή η ανίχνευση οποιασδήποτε μεταβολής της διεργασίας. Μια βασική ιδιότητα των διαγραμμάτων αυτών είναι ότι δεν έχουν μνήμη, δηλαδή λαμβάνουν υπόψη τους μόνο τις πρόσφατες παρατηρήσεις. Τα διαγράμματα ελέγχου τύπου Shewhart για μεταβλητές

χωρίζονται σε δύο κατηγορίες, τα διαγράμματα ελέγχου για δείγματα και για μεμονωμένες παρατηρήσεις.

Τα διαγράμματα ελέγχου για δείγματα μπορεί να είναι διαγράμματα ελέγχου για τη μέση τιμή, όπως το $\bar{X}$ διάγραμμα ή διαγράμματα για την παρακολούθηση της διασποράς, όπως το R διάγραμμα, το S διάγραμμα και το S^2 διάγραμμα. Σε κάποιες περιπτώσεις, επειδή δεν συνηθίζεται ή δεν μπορούμε να αναπτύξουμε διαγράμματα ελέγχου ορίων 3σ ή γενικά $L\sigma$, κατασκευάζουμε τα λεγόμενα διαγράμματα ελέγχου με όρια ελέγχου πιθανότητας α, τα οποία βασίζονται στα ποσοστιαία σημεία της κατανομής των δεδομένων. Επίσης, ανάλογα με το αν η μέση τιμή και η διασπορά είναι γνωστές ή όχι (σε πραγματικά προβλήματα δεν είναι), μπορούμε να αναπτύξουμε τα παραπάνω διαγράμματα γι' αυτές τις δύο περιπτώσεις και επομένως θα έχουμε τα διαγράμματα ελέγχου Φάσης I και Φάσης II. Στη Φάση II, οι ποσότητες μ και σ θεωρούνται γνωστές (π.χ. από προηγούμενη εμπειρία) ενώ στη Φάση I δεν θεωρούνται γνωστές και πρέπει να εκτιμηθούν.

Για διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις, δηλαδή στην περίπτωση που το μέγεθος του δείγματος είναι ίσο με 1, ισχύουν αντίστοιχα με τα διαγράμματα ελέγχου για δείγματα με κατάλληλη προσαρμογή των ορίων ελέγχου. Για παράδειγμα, αντί για το $\bar{X}$ διάγραμμα ελέγχου έχουμε πλέον το X διάγραμμα, στο οποίο απεικονίζονται πλέον οι μεμονωμένες παρατηρήσεις X_i. Επίσης, για το διάγραμμα ελέγχου της διασποράς, επειδή για $n = 1$ δεν έχει νόημα να υπολογιστεί το εύρος R μιας μόνο παρατήρησης, χρησιμοποιούμε το κινούμενο εύρος (moving range) των μεμονωμένων παρατηρήσεων.

2.5.4 Διαγράμματα Ελέγχου Τύπου Shewhart για Ιδιότητες

Σε αρκετές περιπτώσεις ταξινομούμε ένα προϊόν σαν **ελαττωματικό ή μη συμμορφούμενο** (defective or nonconforming), αν τουλάχιστον ένα ποιοτικό χαρακτηριστικό του έχει τιμή η οποία βρίσκεται εκτός των ορίων προδιαγραφών. Σε αυτή την περίπτωση λέμε ότι το προϊόν παρουσιάζει τουλάχιστον ένα **ελάττωμα ή ατέλεια** (defect or nonconformity).

Ο αριθμός των ελαττωματικών προϊόντων μιας παραγωγικής διεργασίας, όπως και ο αριθμός των ελαττωμάτων ενός προϊόντος, είναι ποιοτικά χαρακτηριστικά που περιγράφονται από διακριτές τυχαίες μεταβλητές οι οποίες στα πλαίσια του Στατιστικού Ελέγχου Ποιότητας ονομάζονται **ιδιότητες** (attributes).

Στην παρούσα ενότητα θα αναφέρουμε τρία βασικά είδη διαγραμμάτων ελέγχου για ιδιότητες. Το πρώτο διάγραμμα αφορά το ποσοστό p των ελαττωματικών προϊόντων μιας

παραγωγικής διεργασίας, γνωστό ως p διάγραμμα. Μπορούμε επίσης να έχουμε ένα διάγραμμα για τον αριθμό των ελαττωματικών προϊόντων μιας διεργασίας, το οποίο είναι το np διάγραμμα. Το δεύτερο διάγραμμα ελέγχου αφορά το συνολικό αριθμό των ελαττωμάτων σε μια μονάδα επιθεώρησης, γνωστό ως c διάγραμμα και το τρίτο διάγραμμα αφορά το μέσο αριθμό ελαττωμάτων ανά μονάδα επιθεώρησης, γνωστό ως u διάγραμμα. Ο όρος μονάδα επιθεώρησης δε σημαίνει απαραίτητα ένα προϊόν. Μονάδα επιθεώρησης ή απλά μονάδα μπορεί να είναι το ίδιο το προϊόν, ένα μέρος αυτού ή ένα σύνολο προϊόντων.

Όπως στα διαγράμματα ελέγχου τύπου Shewhart για μεταβλητές, έτσι και στα διαγράμματα τύπου Shewhart για ιδιότητες, έχουμε διαφοροποιήσεις στα όρια ελέγχου ανάλογα με τον αν έχουμε Φάση I ή Φάση II. Στην πρώτη περίπτωση, θα πρέπει και πάλι να εκτιμηθούν οι άγνωστες ποσότητες p, c και u. Τέλος, θα πρέπει να λάβουμε υπόψη μας αν το μέγεθος του δείγματος είναι σταθερό ή μεταβλητό καθώς στην περίπτωση του μεταβλητού μεγέθους δείγματος θα πρέπει να γίνουν κάποιες τροποποιήσεις στους τύπους υπολογισμού των ορίων ελέγχου.

2.5.5 Ανάλυση Ικανότητας μιας Διεργασίας

2.5.5.1 Ικανότητα μιας Διεργασίας

Ο όρος ικανότητα μιας διεργασίας (process capability) χρησιμοποιείται για να περιγράψει πόσο αποτελεσματική είναι μια διεργασία ως προς το να παράγει προϊόντα με τιμές X εντός των ορίων προδιαγραφών και όσο το δυνατό πιο κοντά στην τιμή στόχο T. Γνωρίζουμε ότι τα φυσικά όρια ανοχής μιας διεργασίας με μέσο μ και τυπική απόκλιση σ δίνονται από τις σχέσεις $\mu - 3\sigma$ (κάτω όριο) και $\mu + 3\sigma$ (άνω όριο). Αν η κατανομή του υπό μελέτη ποιοτικού χαρακτηριστικού του προϊόντος είναι κανονική, τότε σχεδόν όλα (99.73%) τα προϊόντα που παράγονται έχουν τιμές εντός του διαστήματος $[\mu - 3\sigma, \mu + 3\sigma]$, που αποτελεί το **φυσικό εύρος ανοχής** της κατανομής του ποιοτικού χαρακτηριστικού. Αν δύο διεργασίες Α και Β έχουν ίσο εύρος ανοχής, θα λέμε ότι η Α είναι ικανότερη από τη Β αν η πρώτη παράγει λιγότερα ελαττωματικά προϊόντα από τη δεύτερη. Για να υπολογίσουμε τον αριθμό των ελαττωματικών προϊόντων που παράγει η κάθε διεργασία, θα πρέπει να λάβουμε υπόψη μας το εύρος ανοχής 6σ της διεργασίας αλλά και τη θέση στην οποία βρίσκεται η κατανομή σε σχέση με τα όρια προδιαγραφών της.

Η **ανάλυση της ικανότητας μιας διεργασίας** (process capability analysis) σχετίζει συνήθως το μέσο μ και την τυπική απόκλιση σ της κατανομής ενός ποιοτικού χαρακτηριστικού με τα όρια προδιαγραφών, δίνοντας αριθμητικά μέτρα που ονομάζονται **δείκτες ικανότητας της διεργασίας** (process capability indices) που περιγράφουν την ικανότητα της διεργασίας να παράγει προϊόντα σύμφωνα με τις προδιαγραφές. Άλλες ποσότητες που εμπλέκονται συνήθως σε αυτούς τους δείκτες είναι η τιμή στόχος T και το μέσο του διαστήματος $[LSL, USL]$ που είναι ίσο με $M = \frac{LSL+USL}{2}$. Συνήθως ισχύει $M = T$ (συμμετρικά όρια προδιαγραφών) αλλά υπάρχουν και περιπτώσεις όπου $M \neq T$ (μη συμμετρικά όρια προδιαγραφών). Επίσης, υπάρχουν περιπτώσεις στις οποίες ορίζεται μόνο ένα από τα δύο όρια προδιαγραφών.

Οι δείκτες ικανότητας μιας διεργασίας είναι μη αρνητικές συναρτήσεις των ποσοτήτων μ, σ, LSL, USL και T και δεν επηρεάζονται από τις μονάδες μέτρησης του ποιοτικού χαρακτηριστικού, επιτρέποντας έτσι τη σύγκριση διαφορετικών διεργασιών. Μια μεγάλη τιμή ενός δείκτη ικανότητας διεργασιών δηλώνει συνήθως ένδειξη αυξημένης ικανότητας της διεργασίας.

Για να μπορέσουμε να υπολογίσουμε τους δείκτες ικανότητας μιας διεργασίας θα πρέπει να γνωρίσουμε τις τιμές των μ και σ. Αυτές είναι συνήθως άγνωστες και επομένως θα πρέπει να εκτιμηθούν από ένα ή περισσότερα τυχαία δείγματα, κατά τη διάρκεια συλλογής των οποίων, η διεργασία θα πρέπει να είναι εντός ελέγχου.

Οι δείκτες ικανότητας μιας διεργασίας δίνονται από τα στατιστικά πακέτα και οι κυριότεροι από αυτούς είναι οι κάτωθι:

$$C_p = \frac{USL - LSL}{6\sigma}, C_{pk} = min\{C_{pl}, C_{pu}\}$$

όπου

$$C_{pl} = \frac{\mu - LSL}{3\sigma}, \qquad C_{pu} = \frac{USL - \mu}{3\sigma},$$

$$C_{pm} = \frac{USL - LSL}{6\tau}, C_{pmk} = \frac{C_{pk}}{\sqrt{1 + \left(\frac{\mu - T}{\sigma}\right)^2}} = min\left\{\frac{\mu - LSL}{3\sqrt{\sigma^2 + (\mu - T)^2}}, \frac{USL - \mu}{3\sqrt{\sigma^2 + (\mu - T)^2}}\right\}.$$

Οι δείκτες αυτοί χρησιμοποιούνται για την αξιολόγηση της ικανότητας μιας διεργασίας σε διαφορετικές περιπτώσεις ανάλογα με τη θέση της ποσότητας μ σε σχέση με τα όρια προδιαγραφών αλλά και σε σχέση με άλλες ποσότητες όπως για παράδειγμα την ποσότητα M.

Επίσης, οι δείκτες αυτοί εντάσσονται σε κάποια όρια τιμών, πέρα από τα οποία μια διεργασία μπορεί να χαρακτηριστεί ικανή, μη ικανή ή ενδεχομένως να χρειάζεται παρακολούθηση.

Όταν η διεργασία είναι εκτός ελέγχου, οι δείκτες που υπολογίζονται ονομάζονται **δείκτες απόδοσης** και υπάρχουν πολλοί υποστηρικτές της άποψης ότι δεν πρέπει να χρησιμοποιούνται οι δείκτες αυτοί, επειδή η ικανότητα μιας διεργασίας έχει αξία να υπολογιστεί μόνο για εντός ελέγχου διεργασίες.

Στην περίπτωση που η υπόθεση της κανονικότητας της διεργασίας παραβιάζεται, τότε οι τιμές των δεικτών C_p και C_{pk} που παρουσιάστηκαν μέχρι τώρα θα αλλάξουν, ενώ οι υπόλοιποι δείκτες δεν θα δίνουν την πραγματική ικανότητα της διεργασίας. Όταν παραβιάζεται η υπόθεση της κανονικότητας, έχουμε συνήθως δύο βασικές επιλογές: (α) μετασχηματισμός της κατανομής των δεδομένων έτσι ώστε να προσεγγίζεται ικανοποιητικά από μια κανονική κατανομή και (β) χρήση της πραγματικής κατανομής των δεδομένων και κατασκευή «παρόμοιων» δεικτών ικανότητας. Για τη δεύτερη προσέγγιση χρησιμοποιούνται συνήθως ποσοστιαία σημεία της πραγματικής κατανομής της τυχαία μεταβλητής X που περιγράφει τη διεργασία.

Τέλος, μιας και οι παράμετροι μ και σ μιας διεργασίας συνήθως δεν είναι γνωστές, θα πρέπει να εκτιμηθούν. Έτσι, αν στους δείκτες ικανότητας που αναπτύξαμε παραπάνω αντικαταστήσουμε τις ποσότητες μ και σ με τις αντίστοιχες εκτιμήσεις $\hat{\mu}$ και $\hat{\sigma}$, τότε οι αντίστοιχοι δείκτες ικανότητας αποτελούν εκτίμηση των πραγματικών δεικτών ικανότητας και συμβολίζονται με ανάλογο τρόπο. Με αυτό τον τρόπο, οι δείκτες ικανότητας αποτελούν πλέον σημειακές εκτιμήσεις των πραγματικών δεικτών και επομένως, μια ολοκληρωμένη ανάλυση θα έπρεπε να συνοδεύεται με διαστήματα εμπιστοσύνης και ελέγχους υποθέσεων για τους δείκτες. Τα διαστήματα αυτά, όπως και οι ίδιοι οι δείκτες, δίνονται από τα στατιστικά πακέτα και επομένως για λόγους συντομίας δεν θα γίνει αναλυτική παρουσίασή τους.

2.5.2.2 Διαστήματα Ανοχής

Για την ανάλυση της ικανότητας μιας διεργασίας μέσω των δεικτών ικανότητας ήταν απαραίτητη η γνώση των ορίων προδιαγραφών της διεργασίας. Ωστόσο, η ανάλυση της ικανότητας μιας διεργασίας μπορεί να γίνει και χωρίς τη γνώση των ορίων προδιαγραφών. Σε αυτή την περίπτωση η ικανότητα της διεργασίας περιγράφεται με διαστήματα εντός των οποίων αναμένεται με μεγάλη πιθανότητα να παίρνει τιμές το υπό μελέτη ποιοτικό

χαρακτηριστικό. Οι τεχνικές που χρησιμοποιούμε σε αυτή την περίπτωση είναι το ιστόγραμμα συχνοτήτων, το *normal probability plot* και τα διαστήματα ανοχής.
Η πιο κομψή μέθοδος από τις τρεις που αναφέρθηκαν για την εύρεση διαστημάτων εντός των οποίων βρίσκεται ένα ποσοστό των παρατηρήσεων ενός κανονικού πληθυσμού είναι η μέθοδος κατασκευής των **διαστημάτων ανοχής**, τα οποία διαφέρουν από τα διαστήματα εμπιστοσύνης.

2.5.6 Διαγράμματα Ελέγχου με Μνήμη

Στα διαγράμματα ελέγχου τύπου Shewhart η απόφαση για να χαρακτηρίσουμε μια διεργασία ότι είναι εντός ή εκτός στατιστικού ελέγχου βασίζεται στο αν κάποιο σημείο του διαγράμματος βρεθεί εντός ή εκτός των ορίων ελέγχου. Όμως τα σημεία που σχεδιάζονται στο διάγραμμα αυτό βασίζονται σε πληροφορίες που δίνει ένα μόνο δείγμα, το πιο πρόσφατο, αγνοώντας πληροφορίες που μπορούν να δώσουν προηγούμενα δείγματα. Γι' αυτό το λόγο τα διαγράμματα ελέγχου τύπου Shewhart ονομάζονται διαγράμματα ελέγχου χωρίς μνήμη (control charts without memory). Τα τελευταία 50 χρόνια έχουν αναπτυχθεί διαγράμματα ελέγχου που ο σχεδιασμός ενός σημείου βασίζεται σε πληροφορίες που δίνει όχι μόνο το πιο πρόσφατο δείγμα, αλλά και προγενέστερα δείγματα. Τέτοιου είδους διαγράμματα ελέγχου ονομάζονται διαγράμματα ελέγχου με μνήμη (control charts with memory) και χρησιμοποιούνται κυρίως στη Φάση *II*.

2.5.6.1 Αθροιστικό Διάγραμμα (*CUSUM)*

Ήδη γνωρίζουμε ότι τα διαγράμματα ελέγχου τύπου Shewhart είναι αποτελεσματικά για μετατοπίσεις του μέσου της τάξης του 2σ ή μεγαλύτερες, επομένως, αν μας ενδιαφέρει να ανιχνεύσουμε μικρές μετατοπίσεις του μέσου τότε χρησιμοποιούμε αθροιστικά διαγράμματα ελέγχου γνωστά και ως διαγράμματα ελέγχου τύπου *CUSUM*. Στα διαγράμματα αυτά, έχουμε μια ποσότητα C_t που δηλώνει το συσσωρευτικό άθροισμα των αποκλίσεων των παρατηρήσεων από την τιμή στόχο της διεργασίας. Αν η διεργασία παραμένει εντός ελέγχου, τότε περιμένουμε οι αποκλίσεις των παρατηρήσεων από την τιμή στόχο να είναι μικρές και τα συσσωρευμένα αθροίσματα να κινούνται γύρω από την τιμή 0. Αν όμως ο μέσος μετατοπιστεί στη θέση $\mu_1 > \mu_0$ (ή στη θέση $\mu_1 < \mu_0$), τότε από τη στιγμή που συμβαίνει η μετατόπιση περιμένουμε να υπάρχουν περισσότερες θετικές (αρνητικές) αποκλίσεις και συνεπώς τα συσσωρευμένα αθροίσματα C_t να παρουσιάζουν μια ανοδική (καθοδική) κίνηση.

Το διάγραμμα συσσωρευμένων αθροισμάτων μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε το επίπεδο στο οποίο έχει μετατοπιστεί ο μέσος της διεργασίας. Όσο η διεργασία παραμένει εντός ελέγχου έχουμε ότι $E(C_t) = 0$ και συνεπώς η γραμμή που ενώνει τα σημεία (t, C_t) πρέπει να κινείται γύρω από τον οριζόντιο άξονα που περνά από την αρχή των αξόνων. Όμως, αν τη στιγμή $t_0 + 1$ ο μέσος μετατοπιστεί από τη θέση μ_0 στη θέση μ_1, έχουμε ότι $E(C_t) = (\mu_1 - \mu_0)(t - t_0)$ και συνεπώς από τη στιγμή $t_0 + 1$ και μετά η τεθλασμένη γραμμή που ενώνει τα σημεία (t, C_t) έχει κλίση ίση με $\mu_1 - \mu_0$ που ικανοποιεί τη σχέση

$$\mu_1 - \mu_0 = \frac{C_t - C_{t_0+1}}{t - (t_0 + 1)}.$$

Συνεπώς, η εκτίμηση της κλίσης της παραπάνω ευθείας μπορεί να χρησιμοποιηθεί για την εκτίμηση του μ_1.

Στη βιβλιογραφία υπάρχουν δύο ισοδύναμες μέθοδοι για τη γραφική αναπαράσταση συσσωρευμένων αθροισμάτων με όρια ελέγχου, η **αλγοριθμική μέθοδος** και η μέθοδος της **V μάσκας**.

2.5.6.2 Διαγράμματα Ελέγχου *EWMA* για το Μέσο μιας Διεργασίας

Τα διαγράμματα ελέγχου τύπου *EWMA* έχουν **μη περιορισμένη και μη ομοιόμορφη μνήμη** αφού λαμβάνουν πληροφορίες από όλα τα προηγούμενα δείγματα και το καθένα από αυτά έχει διαφορετική βαρύτητα. Στο διάγραμμα ελέγχου τύπου *EWMA* θα απεικονίζεται η τιμή της στατιστικής συνάρτησης

$$Z_t = (1 - \lambda)Z_{t-1} + \lambda X_t = (1 - \lambda)^t Z_0 + \lambda \sum_{i=1}^{t} (1 - \lambda)^{t-i} X_i \,, 0 < \lambda < 1$$

στην οποία, στην ειδική περίπτωση όπου $\lambda = 1$ προκύπτει διάγραμμα ελέγχου τύπου Shewhart.

Τα όρια ελέγχου και η κεντρική γραμμή δίνονται από τους ακόλουθους τύπους

$$UCL = \mu_{Z_t} + L\sigma_{Z_t} = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2t}]}$$

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

$$LCL = \mu_{Z_t} - L\sigma_{Z_t} = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2t}]}$$

Παρατηρούμε ότι τα όρια ελέγχου είναι μεταβλητά. Ωστόσο, η ποσότητα $(1-\lambda)^{2t}$ τείνει να μηδενιστεί καθώς το t αυξάνει, οπότε σε σχετικά σύντομο χρονικό διάστημα τα όρια ελέγχου σταθεροποιούνται και δίνονται από τις σχέσεις

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

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

Σχετικά με τον υπολογισμό του εντός κι εκτός ελέγχου μέσου μήκους ροής, έχουν προταθεί διάφορες μεθοδολογίες οι οποίες δεν θα αναπτυχθούν στην παρούσα διπλωματική εργασία. Επίσης, οι ποσότητες ARL_0 και ARL_1 μπορούν να βρεθούν μέσω στατιστικών προγραμμάτων για συγκεκριμένες τιμές των ποσοτήτων λ και L.

Για την επιλογή των λ και L καθορίζουμε πρώτα το εντός ελέγχου μέσο μήκος ροής και τη μετατόπιση του μέσου της διεργασίας που θέλουμε να ανιχνεύσουμε και κατόπιν επιλέγουμε τα λ και L από διαθέσιμους πίνακες με κριτήριο το μικρότερο εκτός ελέγχου μέσο μήκος ροής.

Στην πράξη επιλέγουμε το λ έτσι ώστε $0.05 \leq \lambda \leq 0.25$, με πιο δημοφιλείς επιλογές τις $\lambda = 0.05, \lambda = 0.10$ και $\lambda = 0.20$, ενώ για το L έχουμε συνήθως ότι $L = 3$. Βέβαια, όσο μεγαλύτερη είναι η τιμή του λ τόσο μεγαλύτερο είναι το βάρος των πιο πρόσφατων παρατηρήσεων στη διαμόρφωση της τιμής Z_t. Για την ανίχνευση μικρών μετατοπίσεων του μέσου απαιτείται μικρή τιμή του λ. Όπως τα διαγράμματα τύπου *CUSUM*, έτσι και τα διαγράμματα τύπου *EWMA* τα προτιμούμε έναντι των διαγραμμάτων τύπου Shewhart όταν μας ενδιαφέρει να ανιχνεύσουμε μικρές μετατοπίσεις του μέσου, αλλά είναι χειρότερα για την ανίχνευση μεγάλων μετατοπίσεων του μέσου. Ωστόσο, τα διαγράμματα *EWMA* είναι ανώτερα των διαγραμμάτων *CUSUM* για την ανίχνευση μεγάλων μετατοπίσεων του μέσου, ιδιαίτερα για $\lambda > 0.10$.

Επίσης, αξίζει να σημειωθεί ότι η απόδοση ενός διαγράμματος *EWMA* δεν επηρεάζεται σημαντικά από την παραβίαση της υπόθεσης της κανονικότητας των παρατηρήσεων και συνεπώς ένα καλά σχεδιασμένο διάγραμμα *EWMA* είναι (σχεδόν) πάντα η καλύτερη μη παραμετρική λύση για ανίχνευση μικρών μετατοπίσεων του μέσου επιπέδου μιας διεργασίας.

Αν και η μέθοδος των διαγραμμάτων *EWMA* αναπτύχθηκε για μεμονωμένες παρατηρήσεις, μπορεί να τροποποιηθεί έτσι ώστε να καλύψει και την περίπτωση που έχουμε δείγματα μεγέθους $n > 1$. Σε αυτή την περίπτωση η ποσότητα X_t θα πρέπει να αντικατασταθεί με την ποσότητα $\bar{X}_t$ και η ποσότητα σ με την ποσότητα $\sigma/\sqrt{n}$. Συνεπώς, στο διάγραμμα ελέγχου απεικονίζεται η ποσότητα

$$Z_t = (1-\lambda)Z_{t-1} + \lambda\bar{X}_t, 0 < \lambda \leq 1$$

με όρια πιθανότητας και κεντρική γραμμή που δίνονται από τις ακόλουθες σχέσεις

$$UCL = \mu_{Z_t} + L\sigma_{Z_t} = \mu_0 + L\frac{\sigma}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]}$$

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

$$LCL = \mu_{Z_t} - L\sigma_{Z_t} = \mu_0 - L\frac{\sigma}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]}$$

2.5.6.3 Διαγράμματα Ελέγχου Κινούμενου Μέσου

Τα διαγράμματα ελέγχου κινούμενου μέσου έχουν **περιορισμένη και ομοιόμορφη μνήμη** αφού βασίζονται σε πληροφορίες που δίνουν τα πιο πρόσφατα k δείγματα και το καθένα από αυτά έχει την ίδια βαρύτητα $b = 1/k$. Στα διαγράμματα ελέγχου κινούμενου μέσου απεικονίζεται η στατιστική συνάρτηση

$$Y_t = \frac{g(\mathbf{X}_{t-k+1}) + g(\mathbf{X}_{t-k+2}) + \cdots + g(\mathbf{X}_t)}{k}, \qquad t \geq k.$$

Η συνάρτηση $g(\mathbf{X}_t)$ για μεμονωμένες παρατηρήσεις είναι συνήθως η ταυτοτική συνάρτηση, οπότε σε αυτή την περίπτωση απεικονίζεται στο διάγραμμα η ποσότητα

$$Y_t = \frac{\mathbf{X}_{t-k+1} + \mathbf{X}_{t-k+2} + \cdots + \mathbf{X}_t}{k}, \qquad t \geq k.$$

Η κατασκευή ενός διαγράμματος ελέγχου για το μέσο της διεργασίας θα μπορούσε να βασιστεί στην ποσότητα Y_t, $t \geq k$. Τα όρια ελέγχου και η κεντρική γραμμή του διαγράμματος ελέγχου δίνονται από τις ακόλουθες σχέσεις

$$UCL = \mu_{Y_t} + L\sigma_{Y_t} = \mu_0 + L\frac{\sigma}{\sqrt{k}}$$

$$CL = \mu_{Y_t} = \mu_0$$

$$LCL = \mu_{Y_t} - L\sigma_{Y_t} = \mu_0 - L\frac{\sigma}{\sqrt{k}}$$

Για $t < k$ η στατιστική συνάρτηση Y_t ορίζεται συνήθως ως ο μέσος όρος των πρώτων t παρατηρήσεων, δηλαδή

$$Y_t = \frac{X_1 + X_2 + \cdots X_t}{t}, \qquad t < k$$

με

$$E(Y_t) = \mu_0, \qquad Var(Y_t) = \frac{\sigma^2}{t}$$

οπότε τα όρια ελέγχου για τις πρώτες $k - 1$ παρατηρήσεις είναι μεταβλητά.

Γενικά για την ανίχνευση μικρής (μεγάλης) μετατόπισης του μέσου επιπέδου της διεργασίας απαιτείται μεγάλη (μικρή) τιμή της ποσότητας k. Επίσης, αξίζει να σημειωθεί ότι οι ποσότητες Y_n και Y_m για $|n - m| < k$ είναι συσχετισμένες και έτσι έχουμε ακόμη μια επιπρόσθετη δυσκολία να ερμηνεύσουμε πρότυπα στο διάγραμμα ελέγχου κινούμενου μέσου.

Τα όρια ελέγχου στο διάγραμμα ελέγχου κινούμενου μέσου για μεμονωμένες παρατηρήσεις, μπορούν εύκολα να τροποποιηθούν έτσι ώστε να μπορούν να χρησιμοποιηθούν για μεγέθη δειγμάτων $n > 1$. Σε αυτή την περίπτωση χρησιμοποιείται η στατιστική συνάρτηση

$$Y_t = \frac{\bar{X}_{t-k+1} + \bar{X}_{t-k+2} + \cdots + \bar{X}_t}{k}, \qquad t \geq k$$

και τα όρια ελέγχου και η κεντρική γραμμή δίνονται από τις ακόλουθες σχέσεις

$$UCL = \mu_{Y_t} + L\sigma_{Y_t} = \mu_0 + L\frac{\sigma}{\sqrt{nk}}$$

$$CL = \mu_{Y_t} = \mu_0$$

$$LCL = \mu_{Y_t} - L\sigma_{Y_t} = \mu_0 - L\frac{\sigma}{\sqrt{nk}}$$

Γενικά τα διαγράμματα ελέγχου κινούμενου μέσου είναι πιο αποτελεσματικά από τα διαγράμματα ελέγχου τύπου Shewhart για την ανίχνευση μικρών μετατοπίσεων του μέσου της διεργασίας. Ωστόσο, δεν είναι τόσο αποτελεσματικά σε σύγκριση με τα διαγράμματα ελέγχου τύπου *EWMA* και *CUSUM*, και συνεπώς η χρήση τους είναι περιορισμένη.

ΚΕΦΑΛΑΙΟ 3

Περιγραφική Στατιστική Ανάλυση

3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε κάποια περιγραφικά στατιστικά μέτρα και κάποια διαγράμματα προκειμένου να κατανοήσουμε, σε αρχική φάση, τη φύση των δεδομένων και τη συμπεριφορά τους. Τα μέτρα αυτά είναι ο αριθμητικός μέσος, η διάμεσος, η διασπορά κ.α. ενώ μπορούμε να κατασκευάσουμε διαγράμματα ανάλογα με τη φύση των δεδομένων, όπως το ιστόγραμμα συχνοτήτων για να πάρουμε ή όχι ενδείξεις για το αν τα δεδομένα μας είναι κανονικά ή όχι και το κυκλικό διάγραμμα. Επίσης, κατασκευάζοντας κατάλληλα θηκογράμματα, θα είμαστε σε θέση να εντοπίσουμε έκτροπες παρατηρήσεις στις μεταβλητές μας, οι οποίες θα αφαιρεθούν από αυτές, παρέχοντας περισσότερη ομοιομορφία στα δεδομένα μας. Οι έκτροπες αυτές παρατηρήσεις μπορεί να προήλθαν είτε από λάθος καταγραφή ή λόγω κάποιων ιδιαίτερων συνθηκών οι οποίες επικρατούσαν κατά την παρατήρησή τους.

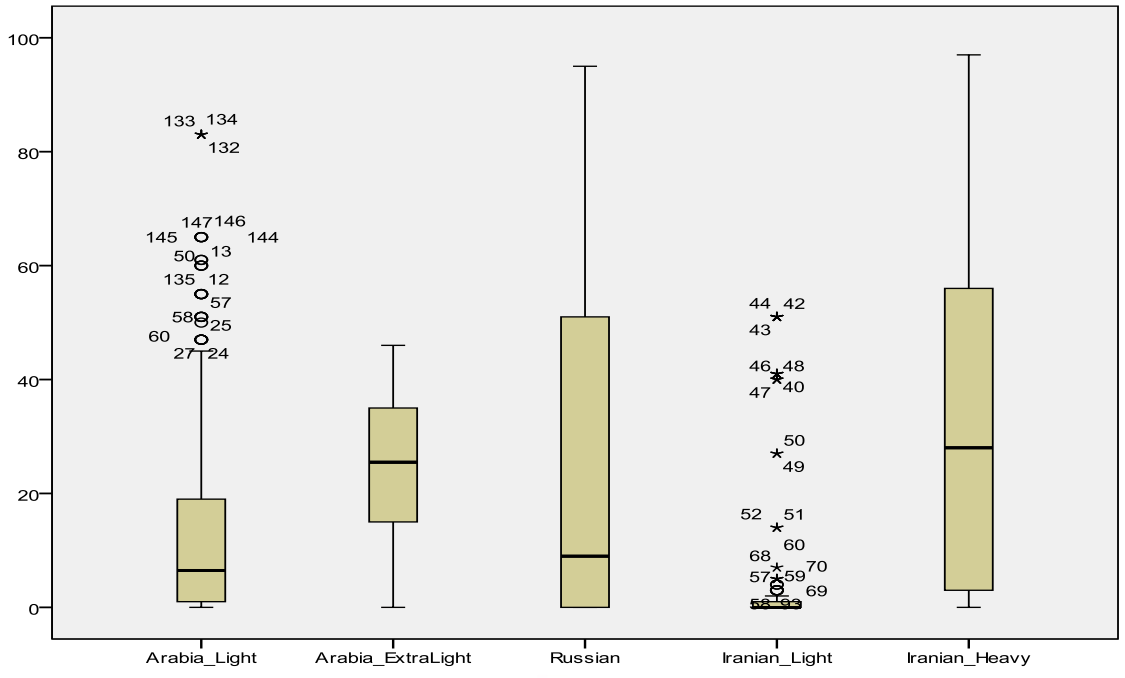
Σε αυτό το σημείο θα πρέπει να επισημάνουμε ότι τα δεδομένα πάνω στα οποία θα εφαρμοστεί η ανάλυση, αφορούν στις (ανεξάρτητες) μεταβλητές $X_1 - X_{13}$ που αναφέρονται σε ποσοστό συμμετοχής διαφόρων τύπων αργού πετρελαίου στη διύλιση καθώς και σε τεχνικά χαρακτηριστικά του αργού πετρελαίου πριν και κατά τη διάρκεια της διύλισης, όπως ο ισομερισμός ή ο αριθμός οκτανίων και στις μεταβλητές (απόκρισης) $Y_1 - Y_3$ οι οποίες αναφέρονται σε χαρακτηριστικά του τελικού προϊόντος, δηλαδή της βενζίνης.

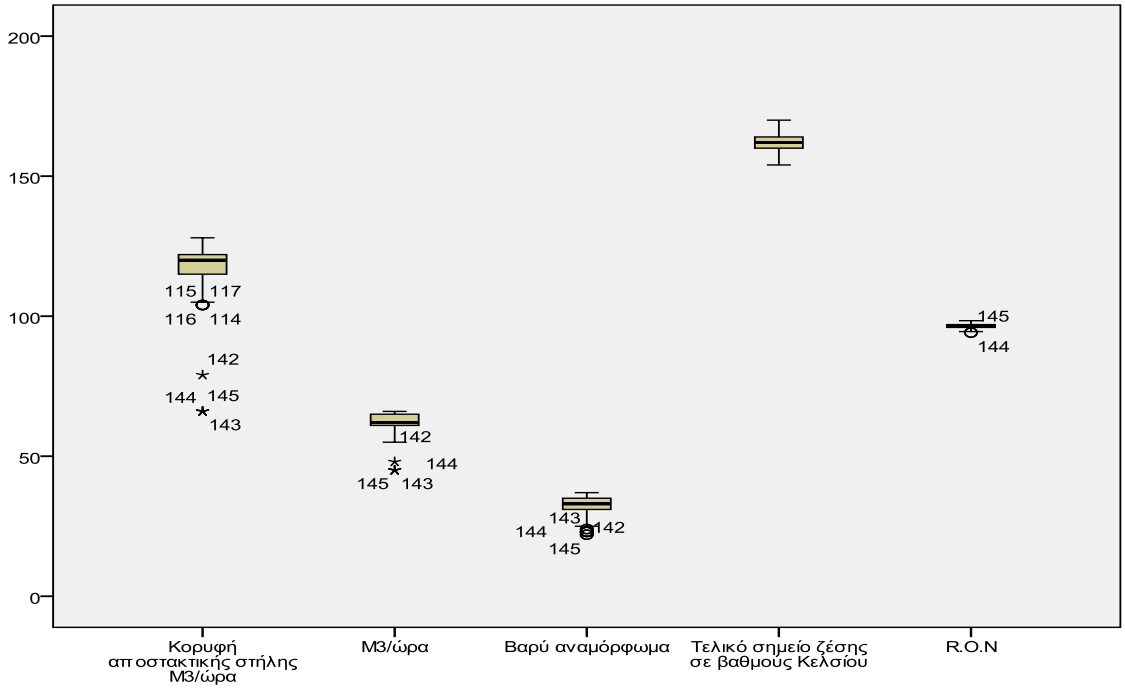
Η ανάλυση του παρόντος κεφαλαίου θα γίνει με τη βοήθεια του στατιστικού προγράμματος SPSS και του προγράμματος Microsoft Excel.

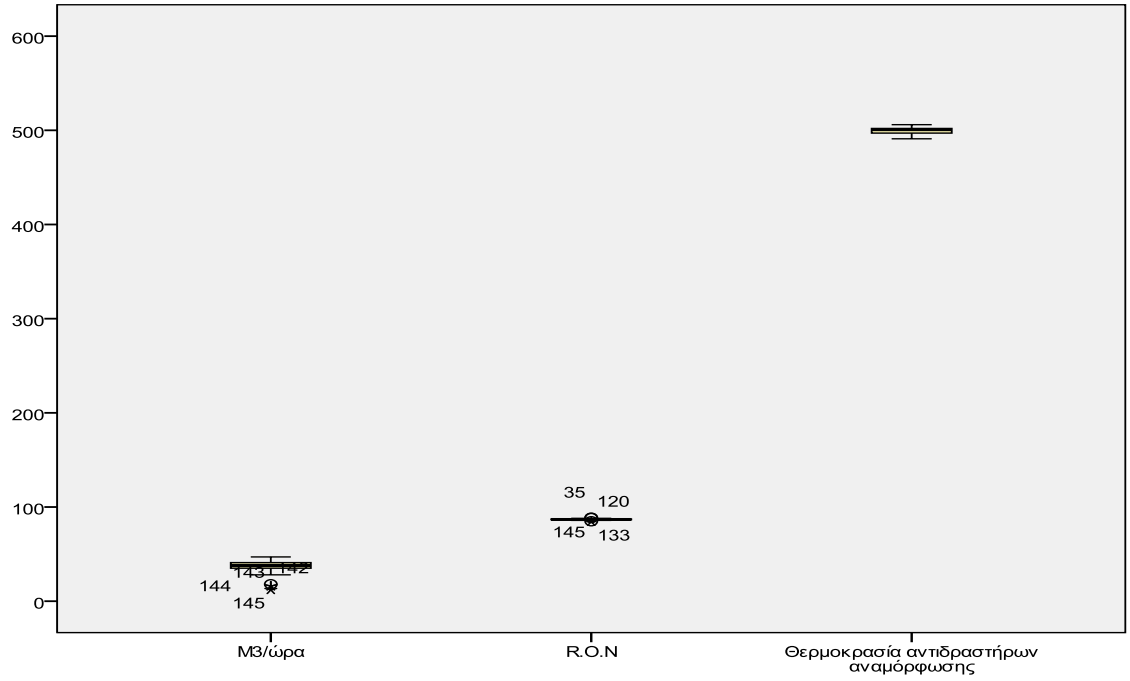
3.2 Διαγράμματα

Η ανάλυσή μας θα ξεκινήσει κατασκευάζοντας κατάλληλα θηκογράμματα για τις μεταβλητές, ώστε μετά την όποια αφαίρεση έκτροπων παρατηρήσεων από αυτές, να έχουμε τα τελικά δεδομένα για περεταίρω ανάλυση.

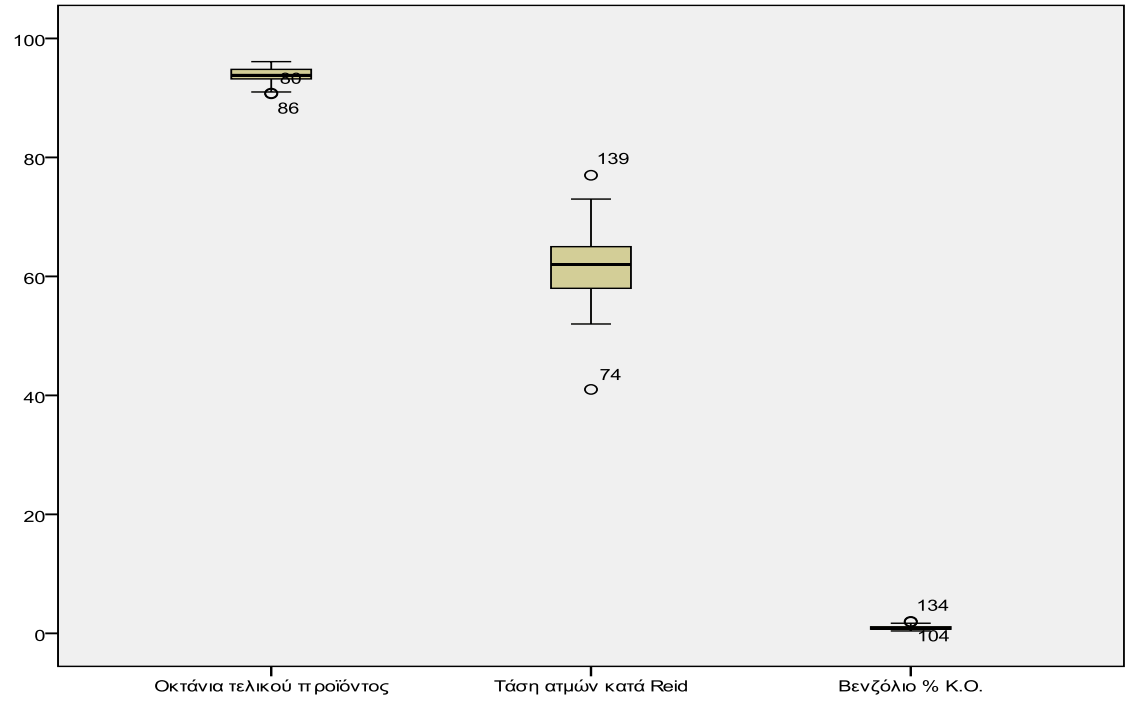
Μέσω του στατιστικού προγράμματος SPSS, τα θηκογράμματα για τις (ανεξάρτητες) μεταβλητές $X_1 - X_{13}$ είναι τα εξής:







Για τις μεταβλητές (απόκρισης) $Y_1 - Y_3$, τα αντίστοιχα θηκογράμματα είναι τα εξής:

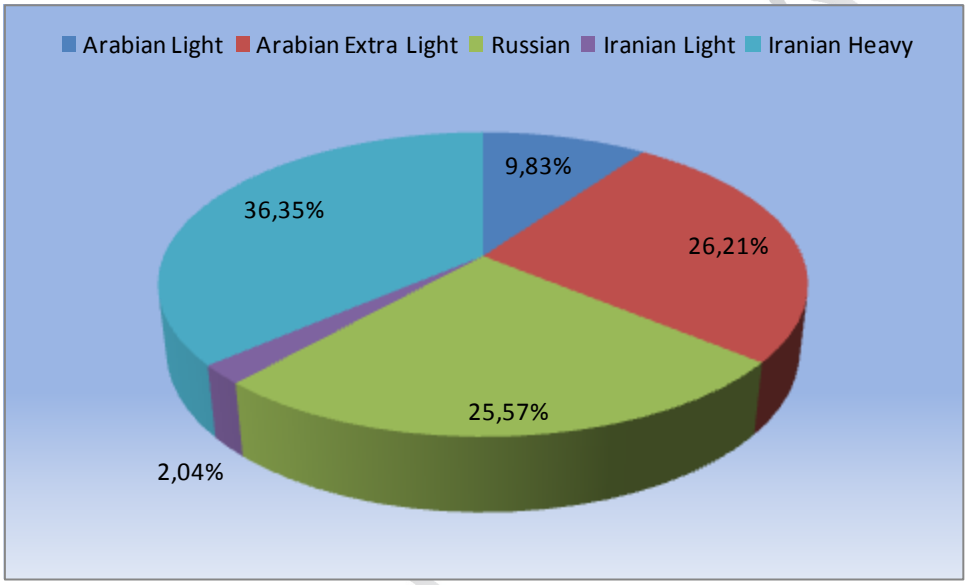


Σχήμα 3.1: Θηκογράμματα για το σύνολο των μεταβλητών

Παρατηρούμε ότι στις μεταβλητές $X_1, X_4, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}, Y_1, Y_2$ και Y_3 υπάρχουν έκτροπες παρατηρήσεις τις οποίες αφού τις αφαιρέσουμε (κοινά για όλες τις μεταβλητές) από τα δεδομένα, ακολουθώντας και πάλι την ίδια διαδικασία, καταλήγουμε σε δεδομένα χωρίς

άλλες έκτροπες παρατηρήσεις. Αυτά θα είναι και τα τελικά μας δεδομένα τα οποία θα χρησιμοποιηθούν και για περεταίρω ανάλυση.

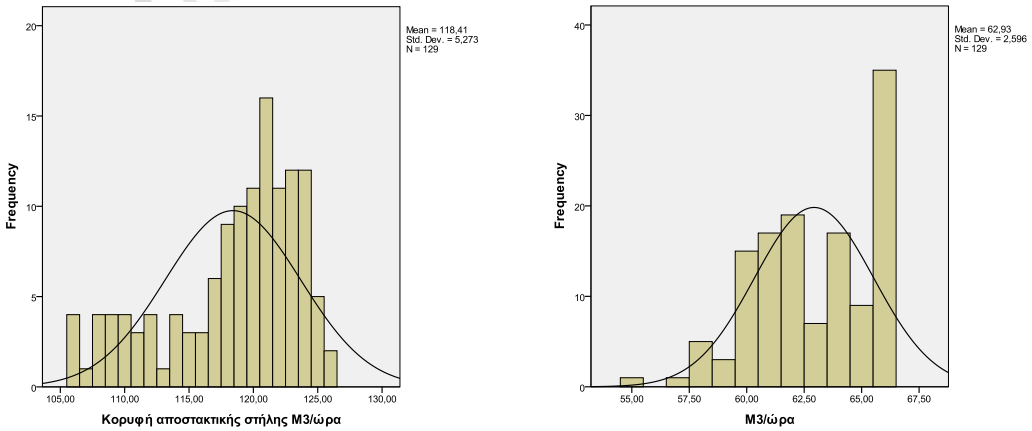
Για τις μεταβλητές $X_1 - X_5$, οι οποίες εκφράζουν ποσοστό συμβολής των διαφόρων τύπων αργού πετρελαίου στο τελικό προϊόν, μπορούμε να κατασκευάσουμε ένα κυκλικό διάγραμμα για καλύτερη οπτική απεικόνιση της συμβολής τους αυτής. Το διάγραμμα είναι το παρακάτω:

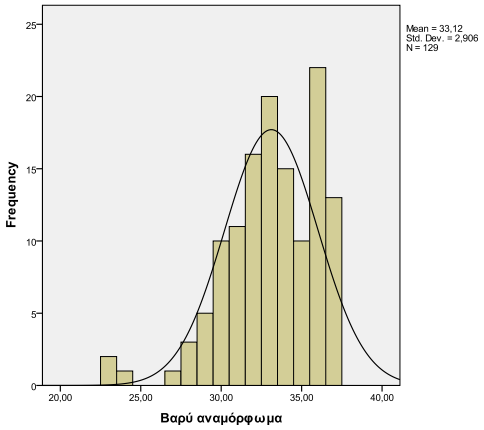


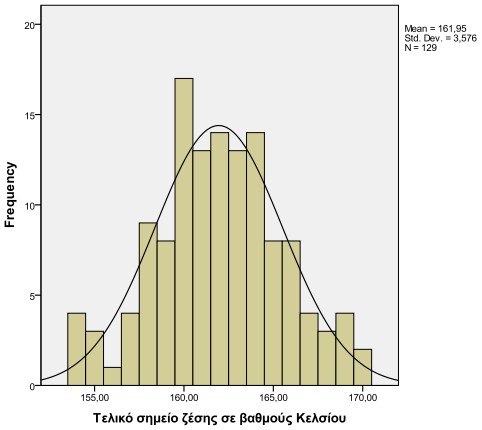
Σχήμα 3.2: Συμμετοχή διαφόρων τύπων αργού πετρελαίου στο τελικό προϊόν

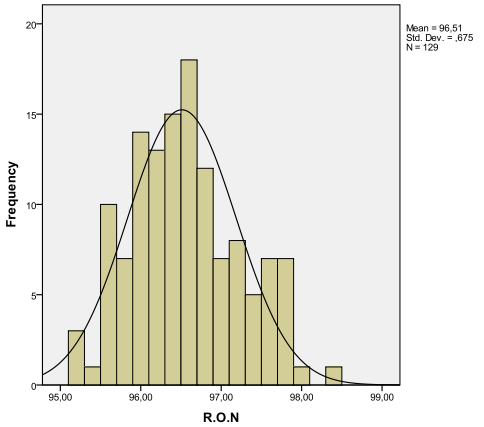
Για τις υπόλοιπες μεταβλητές, τόσο αυτές που εκφράζουν χαρακτηριστικά της διύλισης ($X_6 - X_{13}$) όσο και αυτές που εκφράζουν χαρακτηριστικά του τελικού προϊόντος ($Y_1 - Y_3$), μπορούμε να κατασκευάσουμε ιστογράμματα συχνοτήτων για να έχουμε μια εικόνα για την κατατομή των μεταβλητών αυτών και της συμμετρίας τους ή μη.

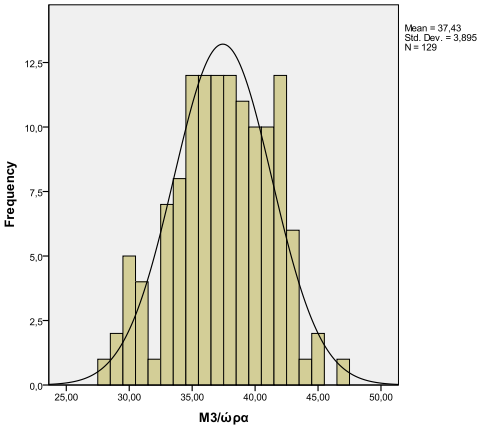
Τα ιστογράμματα συχνοτήτων για τις μεταβλητές $X_6 - X_{13}$ και $Y_1 - Y_3$ είναι τα παρακάτω:

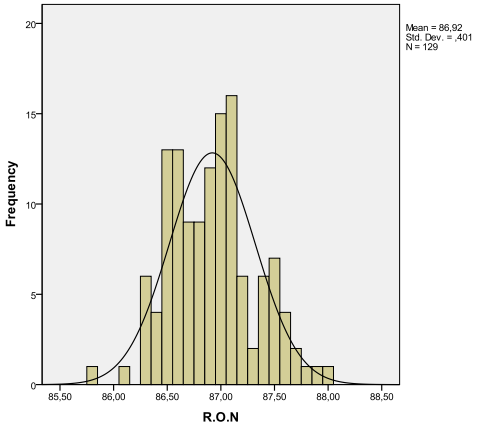


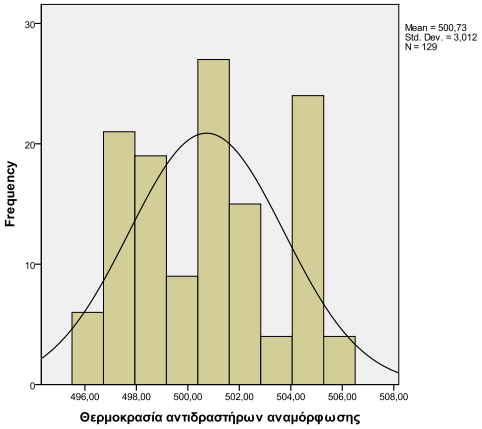




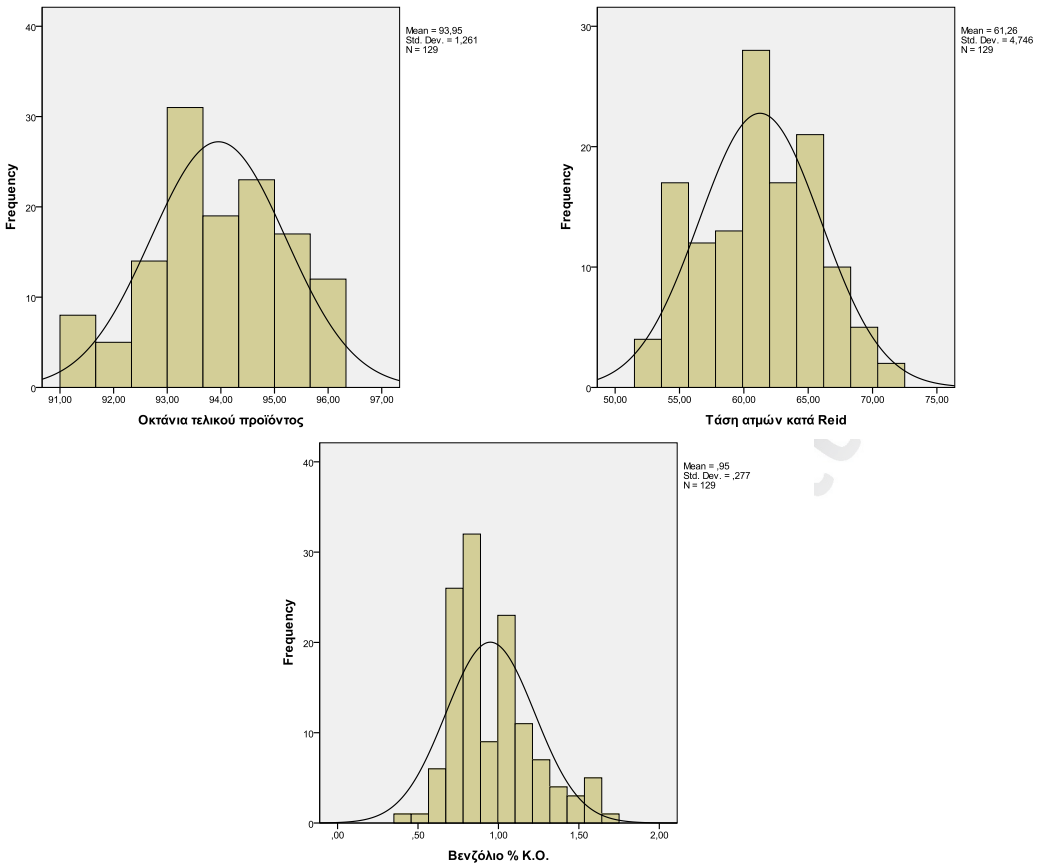








Τα ιστογράμματα συχνοτήτων για τις μεταβλητές $Y_1 - Y_3$ είναι τα παρακάτω:



Σχήμα 3.3: Ιστογράμματα συχνοτήτων για το σύνολο των δεδομένων

3.3 Περιγραφικοί στατιστικοί δείκτες

Στη συνέχεια, μέσω του στατιστικού προγράμματος SPSS, θα υπολογίσουμε κάποια βασικά περιγραφικά μέτρα για τα δεδομένα μας, όπως τον αριθμητικό μέσο, τη διάμεσο, την επικρατούσα τιμή, το εύρος και τη διασπορά που θα μας δώσουν κάποιες πληροφορίες για τις μεταβλητές που χρησιμοποιούμε. Για τις μεταβλητές για τις οποίες κατασκευάσαμε ιστογράμματα συχνοτήτων, ο αριθμητικός μέσος και η τυπική απόκλιση δίνονται άμεσα μαζί με το ιστόγραμμα, όπως μπορούμε να δούμε παραπάνω. Επομένως, ένας συγκεντρωτικός πίνακας με τους δείκτες που αναφέραμε είναι ο κάτωθι.

Πίνακας 3.1: Στατιστικοί περιγραφικοί δείκτες του συνόλου των δεδομένων

Μεταβλητές	Αριθμητικός Μέσος	Διάμεσος	Επικρατούσα τιμή	Τυπική Απόκλιση	Εύρος
Arabian Light (X_1)	9.8295	2	0	15.48695	55
Arabian Extra Light (X_2)	26.2093	28	21	12.44467	46
Russian (X_3)	25.5736	14	0	27.84437	95
Iranian Light (X_4)	2.0388	0	0	8.74857	51
Iranian Heavy (X_5)	36.3488	40	1	31.26916	97
Κορυφή αποστακτικής στήλης M3/ώρα (X_6)	118.4109	120	121	5.27318	20
M3/ώρα (X_7)	62.9302	63	66	2.59563	11
Βαρύ αναμόρφωμα (X_8)	33.1163	33	36	2.90643	14
Τελικό σημείο ζέσης (X_9)	161.9457	162	160	3.57576	16
R.O.N (X_{10})	96.5109	96.5	96.6	0.67537	3.2
M3/ώρα (X_{11})	37.4341	38	35	3.89480	19
R.O.N (X_{12})	86.9202	86.9	87.1	0.40105	2.2
Θερμοκρασία αντιδραστήρων αναμόρφωσης (X_{13})	500.7287	501	501	3.01235	10
Οκτάνια τελικού προϊόντος (Y_1)	93.9543	93.9	93.3	1.26089	5.1
Τάση ατμών κατά Reid (Y_2)	61.2558	62	65	4.74552	20
Βενζόλιο % K.O. (Y_3)	0.9496	0.8	0.8	0.27673	1.3

3.4 Συμπεράσματα

Παρατηρώντας τα όσα αναπτύχθηκαν παραπάνω, μπορούμε να εξάγουμε κάποια συμπεράσματα για τα δεδομένα που διαθέτουμε. Από τα θηκογράμματα, εκτός από τις έκτροπες παρατηρήσεις που ανιχνεύθηκαν και εξαιρέθηκαν, παρατηρούμε ότι η κατανομή των μεταβλητών $X_1 - X_{13}$, εκτός ελαχίστων εξαιρέσεων (π.χ. X_2, X_6, X_9), δε φαίνεται να είναι συμμετρική σε αντίθεση με την κατανομή των μεταβλητών $Y_1 - Y_3$, για την οποία έχουμε ενδείξεις ότι είναι συμμετρική. Αυτό μπορεί να γίνει κατανοητό παρατηρώντας τη θέση της κεντρικής γραμμής του διαγράμματος, που αντιπροσωπεύει τη διάμεσο, σε σχέση με το πάνω και κάτω όριο του ορθογωνίου, που αντιπροσωπεύουν το τρίτο και πρώτο τεταρτημόριο αντίστοιχα.

Από το κυκλικό διάγραμμα, αν επικεντρωθούμε στο ποσοστό αργού πετρελαίου ανά χώρα που περιλαμβάνεται στη διαδικασία της διύλισης στο συγκεκριμένο διυλιστήριο, παρατηρούμε ότι συνολικά το 38.39% προέρχεται από το Ιράν, το 36.04% προέρχεται από την Αραβία και το υπόλοιπο 25.57% προέρχεται από τη Ρωσία. Από την άλλη αν

επικεντρωθούμε στον τύπο αργού πετρελαίου που περιλαμβάνεται στη διαδικασία της διύλισης, μεγαλύτερη συμμετοχή έχει το Ιρανικό βαρύ αργό πετρέλαιο με 36.35%. Από τα ιστογράμματα συχνοτήτων, στα οποία έχουν χαραχθεί και καμπύλες κανονικών κατανομών, λαμβάνουμε ισχυρότερες ενδείξεις ότι οι κατανομές των μεταβλητών $X_6 - X_{13}$ δεν είναι κανονικές και επομένως ούτε συμμετρικές. Τις ίδιες ενδείξεις έχουμε και για τις μεταβλητές $Y_1 - Y_3$, οι κατανομές των οποίων σύμφωνα με τα ιστογράμματα, απέχουν αρκετά από κάποια κανονική κατανομή.

Από τους δείκτες που υπολογίστηκαν, παρατηρούμε ότι η διάμεσος των μεταβλητών X_4 - Iranian Light και Y_3 - Βενζόλιο % Κ.Ο. είναι μηδέν. Αυτό ερμηνεύεται βάσει κυρίως της φύσης των μεταβλητών αυτών μιας και οι τιμές που παίρνουν είναι πολύ κοντά στο μηδέν. Δηλαδή, για τη μεταβλητή X_4 διαπιστώνουμε και μέσω των δεικτών, ότι συμμετέχει σε πολύ μικρό ποσοστό στο τελικό προϊόν ενώ το μέσο κατά όγκο ποσοστό βενζολίου του τελικού προϊόντος, που είναι η βενζίνη, είναι επίσης μικρό, πράγμα που ίσως να σημαίνει μειωμένο αριθμό οκτανίων της βενζίνης που παράγεται. Επίσης, παρατηρούμε ότι ο μέσος αριθμός οκτανίων του τελικού προϊόντος (μεταβλητή Y_1) είναι, όπως αναμένεται αυξημένος σε σχέση με αυτόν του προϊόντος του ισομερισμού (μεταβλητή X_{10}). Αυτό συμβαίνει επειδή το αργό πετρέλαιο εμπλουτίζεται με νάφθες και άλλα συστατικά κατά τη διαδικασία της αναμόρφωσης και του εμπλουτισμού. Τέλος, σημειώνουμε ότι για τις μεταβλητές $X_6 - X_{13}$ και $Y_1 - Y_3$, παρόλο που έχουμε γραφικές ενδείξεις ότι δεν ακολουθούν την κανονική κατανομή, ο αριθμητικός μέσος, η διάμεσος και η επικρατούσα τιμή είναι πολύ κοντά, πράγμα που δηλώνει ότι οι μεταβλητές αυτές ίσως περιγράφονται από μια συμμετρική κατανομή και σε σχέση με αυτά τα τρία μέτρα η τυπική τους απόκλιση είναι πολύ μικρή.

ΜΕΡΟΣ ΙΙ

ΚΕΦΑΛΑΙΟ 4

Ανάλυση Παλινδρόμησης

4.1 Εισαγωγή

Στο κεφάλαιο αυτό θα προσπαθήσουμε να διαπιστώσουμε με ποιο τρόπο οι παράγοντες (ανεξάρτητες μεταβλητές) $X_6 - X_{13}$ επηρεάζουν τα χαρακτηριστικά του τελικού προϊόντος που περιγράφονται από τις (εξαρτημένες) μεταβλητές $Y_1 - Y_3$. Σε αυτή μας την προσπάθεια θα μας βοηθήσουν τα εργαλεία και οι τεχνικές της Ανάλυσης Παλινδρόμησης και πιο συγκεκριμένα, αφού έχουμε περισσότερους από έναν παράγοντες, της ανάλυσης πολλαπλής παλινδρόμησης. Θα κάνουμε την ίδια διαδικασία για κάθε μία από τις τρεις εξαρτημένες μεταβλητές $Y_1 - Y_3$ μεμονωμένα για να διαπιστώσουμε αν μπορούμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης των χαρακτηριστικών της παραγόμενης βενζίνης βασιζόμενοι στα χαρακτηριστικά της διαδικασίας διύλισης.

Οι ανεξάρτητες μεταβλητές που θα χρησιμοποιήσουμε είναι αυτές που έχει νόημα να αναλυθούν, πιο συγκεκριμένα οι $X_6 - X_{13}$. Είναι σημαντικό να αναφέρουμε ότι οι μεταβλητές $Y_1 - Y_3$ που θα χρησιμοποιηθούν στη συνέχεια, θεωρούνται ανεξάρτητες με βάση τον τρόπο συλλογής των δεδομένων καθώς και με βάση τη φύση τους. Επίσης, θα γίνει χρήση του στατιστικού προγράμματος SPSS για την κατασκευή των μοντέλων καθώς και για την αξιολόγησή τους και το επίπεδο σημαντικότητας που θα χρησιμοποιείται για τους ελέγχους υποθέσεων είναι το 5%.

4.2 Ανάλυση παλινδρόμησης για τη μεταβλητή Y_1 – Οκτάνια τελικού προϊόντος

Ξεκινώντας από την εξαρτημένη μεταβλητή Y_1 και χρησιμοποιώντας ως ανεξάρτητες μεταβλητές τις $X_6 - X_{13}$, θα μελετήσουμε αρχικά το μοντέλο πρόβλεψης που δίνεται από την έκφραση

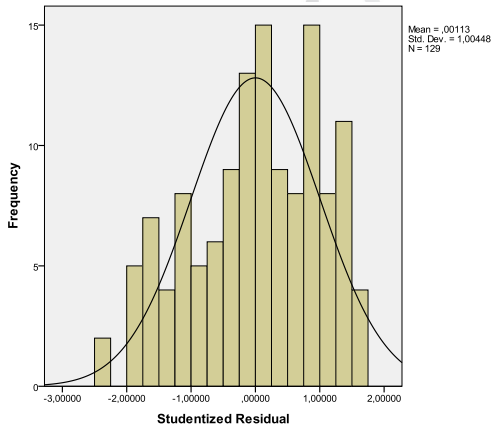
$$Y_1 = \beta_0^{(1)} + \beta_6^{(1)} X_6 + \beta_7^{(1)} X_7 + \beta_8^{(1)} X_8 + \beta_9^{(1)} X_9 + \beta_{10}^{(1)} X_{10} + \beta_{11}^{(1)} X_{11} + \beta_{12}^{(1)} X_{12} + \beta_{13}^{(1)} X_{13} + \varepsilon$$

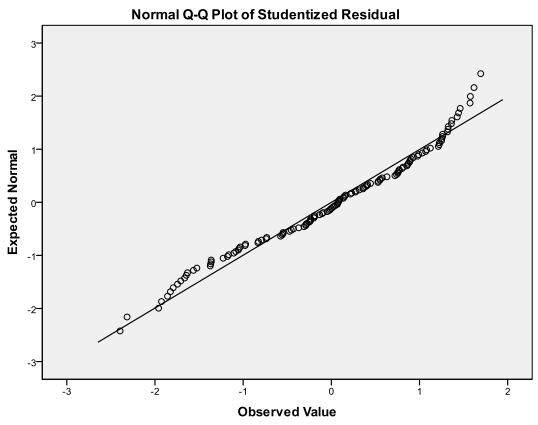
για να διαπιστώσουμε ποιες από τις οκτώ ανεξάρτητες μεταβλητές $X_6 - X_{13}$ είναι στατιστικά σημαντικές και επομένως μπορούν να μας βοηθήσουν στην πρόβλεψη και στην ερμηνεία της Y_1, που εκφράζει τον αριθμό των οκτανίων της βενζίνης.

Μέσω του στατιστικού προγράμματος SPSS, εφαρμόζουμε τη μέθοδο *stepwise* και βρίσκουμε ότι σε επίπεδο σημαντικότητας $\alpha = 5\%$, στατιστικά σημαντικές είναι μόνο οι μεταβλητές $X_7, X_8, \text{και } X_{11}$ μιας και έχουν $p - value < \alpha$. Επίσης, ο σταθερός όρος του μοντέλου είναι στατιστικά σημαντικός. Το μοντέλο παλινδρόμησης που διαμορφώνεται είναι το εξής:

$$Y_1 = 93.691 + 0.231 \cdot X_7 - 0.24 \cdot X_8 - 0.169 \cdot X_{11} \qquad (4.2.1)$$

Κάνοντας χρήση του ελέγχου F, διαπιστώνουμε ότι το μοντέλο (4.2.1) είναι στατιστικά σημαντικό μιας και από τον πίνακα ANOVA που μας δίνεται από το στατιστικό πρόγραμμα, έχουμε ότι οι παράμετροι $\beta_7^{(1)}, \beta_8^{(1)}, \text{και } \beta_{11}^{(1)}$ δεν μπορούν να είναι ταυτόχρονα ίσοι με το μηδέν $(p - value = 0 < 0.05 = \alpha)$. Επίσης, το μοντέλο (4.2.1) έχει συντελεστή προσδιορισμού ίσο με 64.3% και εκτίμηση της διασποράς ίση με 0.568. Για να είμαστε όμως σε θέση να το χρησιμοποιήσουμε για μελλοντική πρόβλεψη του αριθμού των οκτανίων της βενζίνης, θα πρέπει να προβούμε σε περεταίρω ανάλυση. Αυτό θα γίνει μελετώντας τα τυποποιημένα υπόλοιπα (Studentized residuals) και κατασκευάζοντας αρχικά το ιστόγραμμα συχνοτήτων τους καθώς και το κανονικό Q-Q διάγραμμα. Αυτά όπως δίνονται από το πρόγραμμα είναι:



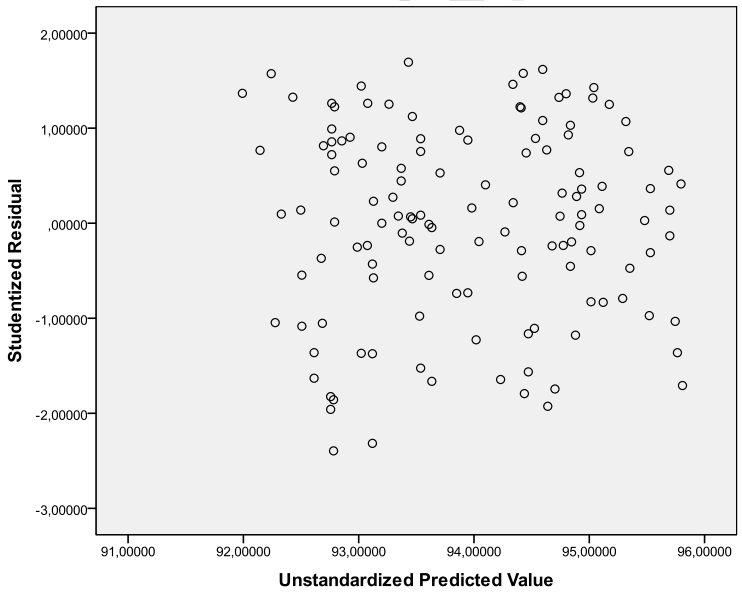


Το ιστόγραμμα συχνοτήτων των υπολοίπων απέχει αρκετά από την κανονική καμπύλη σε αντίθεση με το κανονικό Q-Q διάγραμμα στο οποίο βλέπουμε ότι οι αναμενόμενες τιμές των υπολοίπων δεν απέχουν πολύ από τις παρατηρούμενες. Οι ενδείξεις που έχουμε από τα δύο

αυτά διαγράμματα είναι αντικρουόμενες και όχι απόλυτα ξεκάθαρες, οπότε δεν μας οδηγούν με σιγουριά στο να ισχυριστούμε ότι τα υπόλοιπα του μοντέλου είναι κανονικά.

Γι' αυτό υπάρχει η ανάγκη, οι γραφικές ενδείξεις να επιβεβαιωθούν και από στατιστικές αποδείξεις. Εφαρμόζοντας τον έλεγχο *Kolmogorov – Smirnov* με τη διόρθωση συνέχειας του *Lilliefors*, διαπιστώνουμε ότι πράγματι τα υπόλοιπα της παλινδρόμησης είναι κανονικά ($p - value = 0.086 > 0.05 = a$). Αυτό είναι ιδιαίτερα σημαντικό αφού μας δίνει τη δυνατότητα να προχωρήσουμε σε στατιστική συμπερασματολογία για το μοντέλο (4.2.1).

Κάτι τελευταίο που θα πρέπει να επιβεβαιώσουμε είναι η ομοσκεδαστικότητα των υπολοίπων. Αυτό μπορεί να ελεγχθεί γραφικά, απεικονίζοντας στο ίδιο διάγραμμα τη συμπεριφορά των εκτιμώμενων τιμών της Y_1 σε σχέση με τα Studentized residuals. Το διάγραμμα διασποράς που απεικονίζει τη σχέση αυτή είναι το εξής.



Από το διάγραμμα αυτό γίνεται φανερό ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από την οριζόντια γραμμή που περνάει από το μηδέν. Επομένως, συμπεραίνουμε ότι τα υπόλοιπα έχουν σταθερή διακύμανση.

Κάτι άλλο που μας δίνει το στατιστικό πρόγραμμα είναι διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου. Τα διαστήματα αυτά είναι 95% διαστήματα εμπιστοσύνης και δηλώνουν ότι με βεβαιότητα 95% οι παράμετροι θα βρίσκονται μέσα σε αυτά. Αναλυτικά τα διαστήματα αυτά για τις παραμέτρους του (4.2.1) καθώς και για τον σταθερό όρο δίνονται στον παρακάτω πίνακα.

Πίνακας 4.1: 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου (4.2.1)

Παράμετροι	Κάτω όριο διαστήματος	Πάνω όριο διαστήματος
Σταθερός όρος - $\beta_0^{(1)}$	90.475	96.908
$\beta_7^{(1)}$	0.169	0.293
$\beta_8^{(1)}$	-0.303	-0.178
$\beta_{11}^{(1)}$	-0.212	-0.126

Τέλος, θα μπορούσαμε να δώσουμε και το 95% διάστημα πρόβλεψης της Y_1 για δεδομένες τιμές των μεταβλητών X_7, X_8 και X_{11} όπως για παράδειγμα τις τιμές $\bar{X}_7, \bar{X}_8$ και $\bar{X}_{11}$. Με βάση τη σχέση (4.2.1) για τις τιμές $\bar{X}_7 = 62.9302, \bar{X}_8 = 33.1163$ και $\bar{X}_{11} = 37.4341$ βρίσκουμε $\hat{Y}_1 = 93.9536$. Επομένως σύμφωνα με τη σχέση που έχει δοθεί στην Ενότητα 2.3.1.4, το 95% διάστημα πρόβλεψης της Y_1 είναι το ακόλουθο

$$\left[\hat{Y}_1 - s(\hat{Y}_1)t_{n-p}(\alpha/2), \hat{Y}_1 + s(\hat{Y}_1)t_{n-p}(\alpha/2)\right] =$$
$$[93.9536 - 1.01764 \cdot 1.96\,, 93.9536 + 1.01764 \cdot 1.96] =$$
$$[91.959\,, 95.948]$$

4.3 Ανάλυση παλινδρόμησης για τη μεταβλητή Y_2 – Τάση ατμών κατά Reid

Την ίδια διαδικασία θα ακολουθήσουμε και για τη μεταβλητή απόκρισης Y_2 μελετώντας αρχικά το μοντέλο πρόβλεψης που δίνεται από την έκφραση

$$Y_2 = \beta_0^{(2)} + \beta_6^{(2)}X_6 + \beta_7^{(2)}X_7 + \beta_8^{(2)}X_8 + \beta_9^{(2)}X_9 + \beta_{10}^{(2)}X_{10} + \beta_{11}^{(2)}X_{11} + \beta_{12}^{(2)}X_{12} + \beta_{13}^{(2)}X_{13} + \varepsilon$$

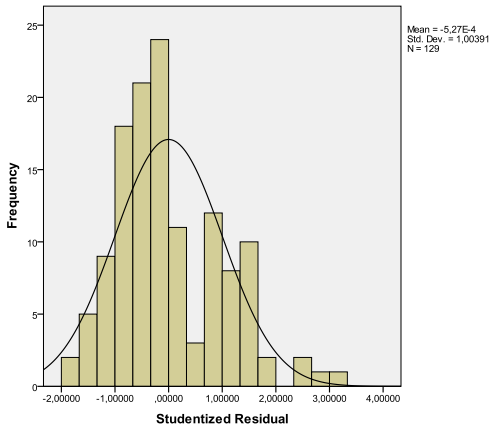
για να διαπιστώσουμε ποιες από τις οκτώ ανεξάρτητες μεταβλητές $X_6 - X_{13}$ είναι στατιστικά σημαντικές και επομένως μπορούν να μας βοηθήσουν στην πρόβλεψη και στην ερμηνεία της Y_2, που εκφράζει την τάση ατμών κατά Reid της βενζίνης.

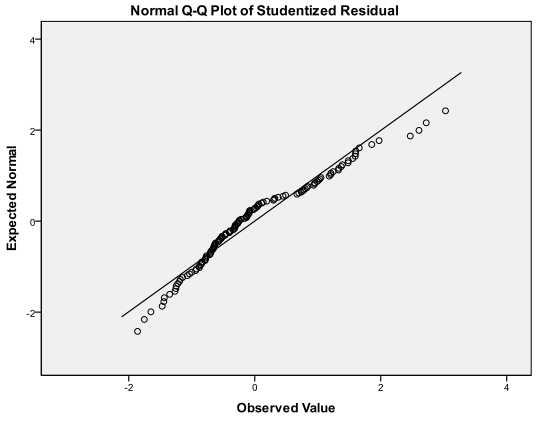
Εφαρμόζουμε και πάλι τη μέθοδο *stepwise* και βρίσκουμε ότι σε επίπεδο σημαντικότητας $\alpha = 5\%$, στατιστικά σημαντικές είναι μόνο οι μεταβλητές X_6, X_9 και X_{13} καθώς και ο σταθερός όρος της παλινδρόμησης μιας και έχουν $p - value < \alpha$. Το μοντέλο παλινδρόμησης που διαμορφώνεται είναι το εξής:

$$Y_2 = 281.556 + 0.251 \cdot X_6 - 0.234 \cdot X_9 - 0.44 \cdot X_{13} \qquad (4.2.2)$$

Κάνοντας και πάλι χρήση του ελέγχου F, μπορούμε να πούμε ότι το μοντέλο (4.2.2) είναι στατιστικά σημαντικό ή γραμμικό μιας και από τον πίνακα ANOVA που μας δίνεται από το στατιστικό πρόγραμμα, έχουμε ότι οι παράμετροι $\beta_6^{(2)}, \beta_9^{(2)}$ και $\beta_{13}^{(2)}$ δεν μπορούν να είναι

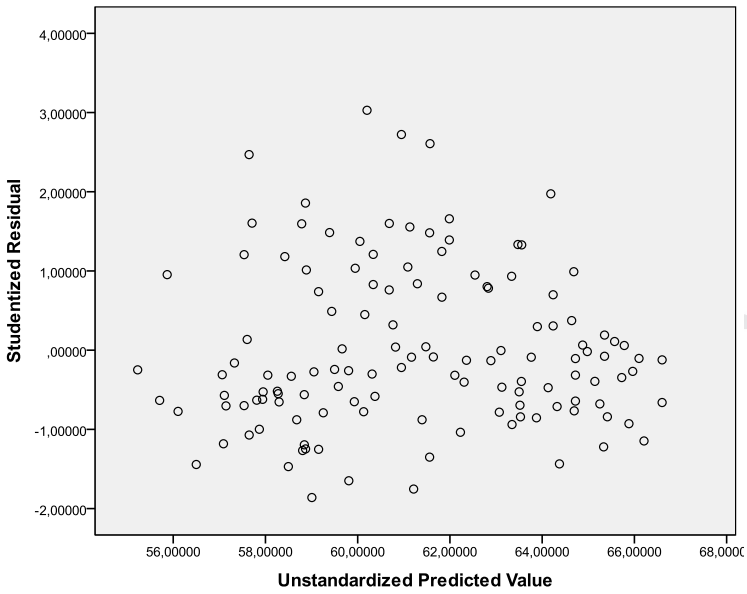
ταυτόχρονα ίσοι με το μηδέν ($p - value = 0 < 0.05 = \alpha$). Επίσης, το μοντέλο (4.2.2) έχει συντελεστή προσδιορισμού ίσο με 36.2% και εκτίμηση της διασποράς ίση με 14.37. Για να είμαστε όμως σε θέση να το χρησιμοποιήσουμε για μελλοντική πρόβλεψη της τάσης ατμών κατά Reid, θα πρέπει να προβούμε σε περεταίρω ανάλυση. Αυτό θα γίνει και πάλι μελετώντας τα τυποποιημένα υπόλοιπα (Studentized residuals) και κατασκευάζοντας αρχικά το ιστόγραμμα συχνοτήτων τους καθώς και το κανονικό Q-Q διάγραμμα. Αυτά όπως δίνονται από το πρόγραμμα είναι:





Παρόμοια εικόνα έχουμε και σε αυτή την περίπτωση αφού το ιστόγραμμα συχνοτήτων των υπολοίπων δεν είναι πολύ ικανοποιητικό, ενώ από το κανονικό Q-Q διάγραμμα βλέπουμε ότι οι αναμενόμενες τιμές των υπολοίπων δεν απέχουν πολύ από τις παρατηρούμενες. Παρόλα αυτά, θα πρέπει και πάλι τις γραφικές ενδείξεις για κανονικότητα των δεδομένων να τις επιβεβαιώσουμε και με στατιστικά κριτήρια. Εφαρμόζοντας τον έλεγχο *Kolmogorov – Smirnov* με τη διόρθωση συνέχειας του *Lilliefors*, διαπιστώνουμε ότι πράγματι τα υπόλοιπα της παλινδρόμησης είναι κανονικά ($p - value = 0.067 > 0.05 = a$).

Ελέγχοντας και πάλι για ομοσκεδαστικότητα των υπολοίπων του μοντέλου (4.2.2), έχουμε το ακόλουθο διάγραμμα διασποράς.



Από το διάγραμμα δεν έχουμε ενδείξεις για μη τυχαία συμπεριφορά των υπολοίπων, μιας και φαίνεται να κινούνται ακανόνιστα γύρω από την οριζόντια γραμμή που περνάει από το μηδέν. Επομένως, και σε αυτή την περίπτωση μπορούμε να πούμε ότι τα υπόλοιπα του μοντέλου (4.2.2) είναι ομοσκεδαστικά.

Τα 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου (4.2.2) καθώς και για τον σταθερό όρο δίνονται στον παρακάτω πίνακα.

Πίνακας 4.2: 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου (4.2.2)

Παράμετροι	**Κάτω όριο διαστήματος**	**Πάνω όριο διαστήματος**
Σταθερός όρος - $\beta_0^{(2)}$	127.084	436.028
$\beta_6^{(2)}$	0.072	0.431
$\beta_9^{(2)}$	-0.447	-0.021
$\beta_{13}^{(2)}$	-0.711	-0.170

Τέλος, θα μπορούσαμε να δώσουμε και το 95% διάστημα πρόβλεψης της Y_2 για δεδομένες τιμές των μεταβλητών X_2, X_9 και X_{13} όπως για παράδειγμα τις τιμές $\bar{X}_6, \bar{X}_9$ και $\bar{X}_{13}$. Με βάση τη σχέση (4.2.1) για τις τιμές $\bar{X}_6 = 118.4109, \bar{X}_9 = 161.9457$ και $\bar{X}_{13} = 500.7287$ βρίσκουμε $\hat{Y}_2 = 53.0612$. Επομένως σύμφωνα με τη σχέση που έχει δοθεί στην Ενότητα 2.3.1.4, το 95% διάστημα πρόβλεψης της Y_2 είναι το ακόλουθο

$$\left[\hat{Y}_2 - s(\hat{Y}_2)t_{n-p}(\alpha/2), \hat{Y}_2 + s(\hat{Y}_2)t_{n-p}(\alpha/2)\right] =$$

$$[53.0612 - 2.87872 \cdot 1.96\,, 53.0612 + 2.87872 \cdot 1.96] =$$

$$[47.4189\,, 58.7035]$$

4.4 Ανάλυση παλινδρόμησης για τη μεταβλητή Y_3 – Βενζόλιο % κατ' όγκο

Ακολουθώντας για τρίτη και τελευταία φορά την ίδια διαδικασία για τη μεταβλητή Y_3, το αρχικό μοντέλο πρόβλεψης που θα μελετήσουμε θα δίνεται από την έκφραση

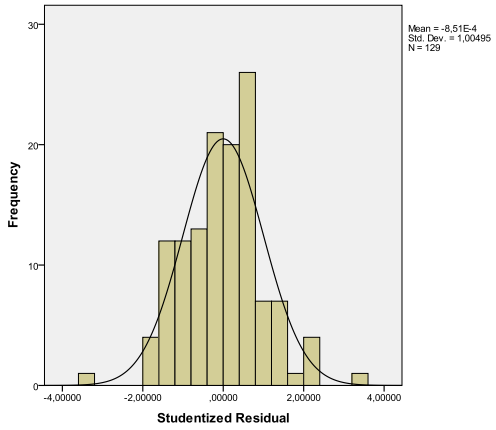
$$Y_3 = \beta_0^{(3)} + \beta_6^{(3)}X_6 + \beta_7^{(3)}X_7 + \beta_8^{(3)}X_8 + \beta_9^{(3)}X_9 + \beta_{10}^{(3)}X_{10} + \beta_{11}^{(3)}X_{11} + \beta_{12}^{(3)}X_{12} + \beta_{13}^{(3)}X_{13} + \varepsilon$$

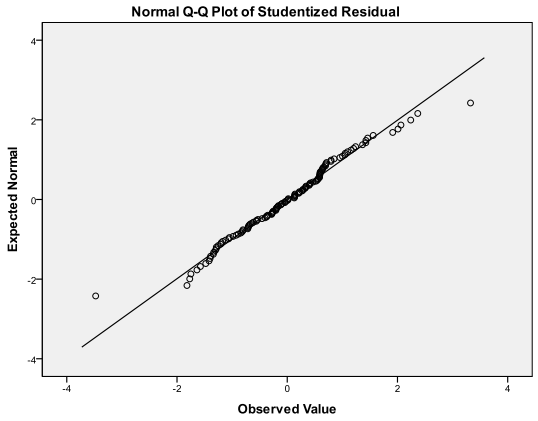
και μέσω του στατιστικού προγράμματος SPSS θα καταλήξουμε στο ποιες από τις οκτώ ανεξάρτητες μεταβλητές $X_6 - X_{13}$ είναι στατιστικά σημαντικές και επομένως μπορούν να μας βοηθήσουν στην πρόβλεψη και στην ερμηνεία της Y_3, που εκφράζει το επί τοις εκατό ποσοστό βενζολίου κατ' όγκο στην παραγόμενη βενζίνη.

Εφαρμόζουμε και πάλι τη μέθοδο *stepwise* και βρίσκουμε ότι σε επίπεδο σημαντικότητας $\alpha = 5\%$, στατιστικά σημαντικές είναι μόνο οι μεταβλητές $X_6\ X_7, X_8, X_{12}$και X_{13} καθώς και ο σταθερός όρος της παλινδρόμησης μιας και έχουν $p - value < \alpha$. Το μοντέλο παλινδρόμησης που διαμορφώνεται είναι το εξής:

$$Y_3 = -31.064 - 0.022 \cdot X_6 + 0.037 \cdot X_7 - 0.018 \cdot X_8 + 0.127 \cdot X_{12} + 0.044 \cdot X_{13} \quad (4.2.3)$$

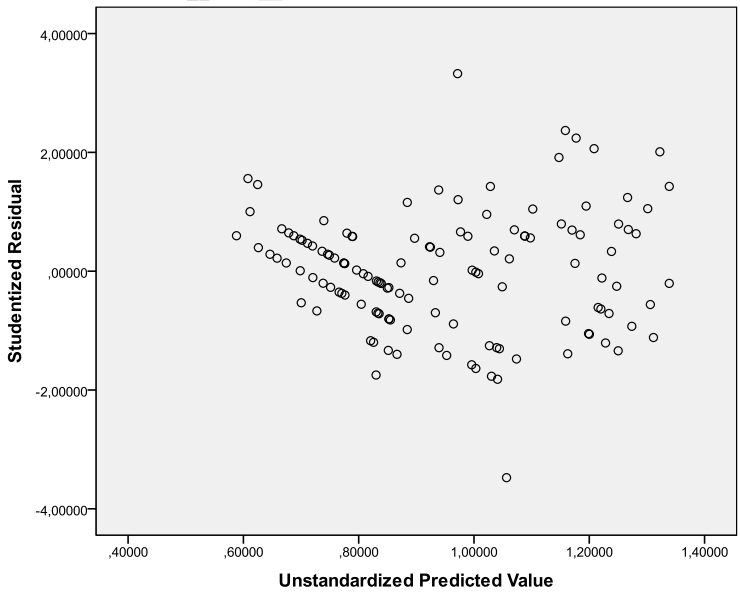
Κάνοντας χρήση του ελέγχου F, μπορούμε να πούμε ότι το μοντέλο (4.2.3) είναι στατιστικά σημαντικό ή γραμμικό μιας και από τον πίνακα ANOVA που μας δίνεται από το στατιστικό πρόγραμμα, έχουμε ότι οι παράμετροι $\beta_6^{(3)}, \beta_7^{(3)}, \beta_8^{(3)}, \beta_{12}^{(3)}$ και $\beta_{13}^{(3)}$ δεν μπορούν να είναι ταυτόχρονα ίσοι με το μηδέν $(p - value = 0 < 0.05 = \alpha)$. Επίσης, το μοντέλο (4.2.3) έχει συντελεστή προσδιορισμού ίσο με 51.6% και εκτίμηση της διασποράς ίση με 0.037. Για να είμαστε όμως σε θέση να χρησιμοποιήσουμε το μοντέλο (4.2.3) για μελλοντική πρόβλεψη του επί τοις εκατό ποσοστού βενζολίου κατ' όγκο στην παραγόμενη βενζίνη, θα πρέπει να προβούμε σε περεταίρω ανάλυση. Αυτό θα γίνει και πάλι μελετώντας τα τυποποιημένα υπόλοιπα (Studentized residuals) και κατασκευάζοντας αρχικά το ιστόγραμμα συχνοτήτων τους καθώς και το κανονικό Q-Q διάγραμμα. Αυτά όπως δίνονται από το πρόγραμμα είναι:





Στην περίπτωση αυτή, η εικόνα είναι σαφώς βελτιωμένη σε σχέση με τις δύο προηγούμενες μιας και τόσο από το ιστόγραμμα συχνοτήτων, όσο και από το κανονικό Q-Q διάγραμμα έχουμε πιο ισχυρές ενδείξεις ότι τα υπόλοιπα του μοντέλου (4.2.3) είναι κανονικά. Ακολουθώντας το ίδιο σκεπτικό με τα προηγούμενα και εφαρμόζοντας τον έλεγχο *Kolmogorov – Smirnov* με τη διόρθωση συνέχειας του *Lilliefors*, διαπιστώνουμε όντως ότι τα υπόλοιπα είναι κανονικά ($p - value = 0.2 > 0.05 = a$).

Ακολουθούμε την ίδια διαδικασία με πριν για τον έλεγχο της ομοσκεδαστικότητας των υπολοίπων και έχουμε το παρακάτω διάγραμμα διασποράς.



Από το διάγραμμα δεν έχουμε ενδείξεις για μη τυχαία συμπεριφορά των υπολοίπων, μιας και φαίνεται να κινούνται ακανόνιστα γύρω από την οριζόντια γραμμή που περνάει από το

μηδέν. Επομένως, και σε αυτή την περίπτωση μπορούμε να πούμε ότι τα υπόλοιπα του μοντέλου (4.2.3) είναι ομοσκεδαστικά.

Τα 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου (4.2.3) καθώς και για τον σταθερό όρο δίνονται στον παρακάτω πίνακα.

Πίνακας 4.3: 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου (4.2.3)

Παράμετροι	Κάτω όριο διαστήματος	Πάνω όριο διαστήματος
Σταθερός όρος - $\beta_0^{(3)}$	-41.915	-20.213
$\beta_6^{(3)}$	-0.029	-0.014
$\beta_7^{(3)}$	0.018	0.056
$\beta_8^{(3)}$	-0.036	-0.001
$\beta_{12}^{(3)}$	0.041	0.212
$\beta_{13}^{(3)}$	0.029	0.058

Τέλος, θα μπορούσαμε να δώσουμε και το 95% διάστημα πρόβλεψης της Y_3 για δεδομένες τιμές των μεταβλητών X_6, X_7, X_8, X_{12} και X_{13} όπως για παράδειγμα τις τιμές $\bar{X}_6, \bar{X}_7, \bar{X}_8, \bar{X}_{12}$ και $\bar{X}_{13}$. Με βάση τη σχέση (4.2.1) για τις τιμές $\bar{X}_6 = 118.4109, \bar{X}_7 = 62.9302, \bar{X}_8 = 33.1163, \bar{X}_{12} = 86.9202$ και $\bar{X}_{13} = 500.7287$ βρίσκουμε $\hat{Y}_3 = 1.1342$. Επομένως σύμφωνα με τη σχέση που έχει δοθεί στην Ενότητα 2.3.1.4, το 95% διάστημα πρόβλεψης της Y_3 είναι το ακόλουθο

$$\left[\hat{Y}_3 - s(\hat{Y}_3)t_{n-p}(\alpha/2), \hat{Y}_3 + s(\hat{Y}_3)t_{n-p}(\alpha/2)\right] =$$
$$[1.1342 - 0.20242 \cdot 1.96\,, 1.1342 + 0.20242 \cdot 1.96] =$$
$$[0.7375\,, 1.5309]$$

4.5 Συμπεράσματα

- **Για το μοντέλο παλινδρόμησης (4.2.1)**

Εφαρμόζοντας τεχνικές πολλαπλής παλινδρόμησης στο σύνολο των δεδομένων που είχαμε στη διάθεσή μας, κατασκευάσαμε ένα μοντέλο το οποίο χρησιμοποιεί ένα υποσύνολο των δεδομένων αυτών. Οι μεταβλητές που εμπεριέχονται στο μοντέλο (4.2.1) είναι οι:

X_7 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε m^3/h,

X_8 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε HVTO 1075 (βαριά νάφθα) m^3/h,

X_{11} : Η τροφοδοσία σε m^3/h της διαδικασίας του ισομερισμού.

Σύμφωνα με το μοντέλο που διαμορφώθηκε, μόνο η X_7 συμβάλει θετικά στη διαμόρφωση του αριθμού των οκτανίων της παραγόμενης βενζίνης, ενώ οι άλλες δύο συμβάλουν αρνητικά.

Το μοντέλο (4.2.1) ελέγχθηκε ως προς τις βασικές υποθέσεις που πρέπει να πληροί ένα μοντέλο πολλαπλής παλινδρόμησης. Αυτές είναι η γραμμικότητα του μοντέλου καθώς και η κανονικότητα και η ομοσκεδαστικότητα των υπολοίπων του μοντέλου. Με κατάλληλους ελέγχους, αυτές οι υποθέσεις επιβεβαιώθηκαν και σε συνδυασμό με την υπόθεση της ανεξαρτησίας των δεδομένων που κάναμε στην αρχή, καταλήγουμε στο γενικό συμπέρασμα ότι το μοντέλο είναι κατάλληλο για την περιγραφή των δεδομένων καθώς και για μελλοντική πρόβλεψη του αριθμού των οκτανίων της παραγόμενης βενζίνης.

Επίσης, το μοντέλο καταφέρνει να εξηγήσει το 64.3% της συνολικής μεταβλητότητας που υπάρχει στα δεδομένα, το οποίο είναι ένα ικανοποιητικό ποσοστό αν αναλογιστούμε την διαφορετικότητα της φύσης των ανεξάρτητων μεταβλητών του μοντέλου.

- **Για το μοντέλο παλινδρόμησης (4.2.2)**

Εφαρμόζοντας τεχνικές πολλαπλής παλινδρόμησης στο σύνολο των δεδομένων που είχαμε στη διάθεσή μας, κατασκευάσαμε ένα μοντέλο το οποίο χρησιμοποιεί ένα υποσύνολο των δεδομένων αυτών. Οι μεταβλητές που εμπεριέχονται στο μοντέλο (4.2.2) είναι οι:

X_6 : Κορυφή της αποστακτικής στήλης σε m^3/h,

X_9 : Τελικό σημείο ζέσης (Τ.Σ.Ζ.) τροφοδοσίας αναμόρφωσης σε °C,

X_{13} : Θερμοκρασία φούρνων (R-400) στους αντιδραστήρες αναμόρφωσης, είσοδος αναμόρφωσης.

Σύμφωνα με το μοντέλο που διαμορφώθηκε, μόνο η X_6 συμβάλει θετικά στη διαμόρφωση της τάσης των ατμών κατά Reid της παραγόμενης βενζίνης, ενώ οι άλλες δύο συμβάλουν αρνητικά.

Το μοντέλο (4.2.2) ελέγχθηκε ως προς τις βασικές υποθέσεις που πρέπει να πληροί ένα μοντέλο πολλαπλής παλινδρόμησης. Αυτές είναι η γραμμικότητα του μοντέλου καθώς και η κανονικότητα και η ομοσκεδαστικότητα των υπολοίπων του μοντέλου. Με κατάλληλους ελέγχους, αυτές οι υποθέσεις επιβεβαιώθηκαν και σε συνδυασμό με την υπόθεση της ανεξαρτησίας των δεδομένων που κάναμε στην αρχή, καταλήγουμε στο γενικό συμπέρασμα ότι το μοντέλο είναι κατάλληλο για την περιγραφή των δεδομένων καθώς και για μελλοντική πρόβλεψη της τάσης των ατμών κατά Reid της παραγόμενης βενζίνης.

Επίσης, το μοντέλο καταφέρνει να εξηγήσει μόλις το 36.2% της συνολικής μεταβλητότητας που υπάρχει στα δεδομένα. Το ποσοστό αυτό δεν είναι ικανοποιητικό και επομένως θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην χρήση του μοντέλου αυτού, παρόλο που πληροί τις υποθέσεις που αναφέραμε.

- **Για το μοντέλο παλινδρόμησης (4.2.3)**

Εφαρμόζοντας τεχνικές πολλαπλής παλινδρόμησης στο σύνολο των δεδομένων που είχαμε στη διάθεσή μας, κατασκευάσαμε ένα μοντέλο το οποίο χρησιμοποιεί ένα υποσύνολο των δεδομένων αυτών. Οι μεταβλητές που εμπεριέχονται στο μοντέλο (4.2.3) είναι οι:

X_6 : Κορυφή της αποστακτικής στήλης σε m^3/h,

X_7 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε m^3/h,

X_8 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε HVTO 1075 (βαριά νάφθα) m^3/h,

X_{12} : Αριθμός οκτανίων (RON) του προϊόντος του ισομερισμού,

X_{13} : Θερμοκρασία φούρνων (R-400) στους αντιδραστήρες αναμόρφωσης, είσοδος αναμόρφωσης.

Σύμφωνα με το μοντέλο που διαμορφώθηκε, μόνο οι X_7, X_{12} και X_{13} συμβάλουν θετικά στη διαμόρφωση της επί τοις εκατό περιεκτικότητας σε βενζόλιο της παραγόμενης βενζίνης, ενώ οι άλλες δύο συμβάλουν αρνητικά.

Το μοντέλο (4.2.3) ελέγχθηκε ως προς τις βασικές υποθέσεις που πρέπει να πληροί ένα μοντέλο πολλαπλής παλινδρόμησης. Αυτές είναι η γραμμικότητα του μοντέλου καθώς και η κανονικότητα και η ομοσκεδαστικότητα των υπολοίπων του μοντέλου. Με κατάλληλους ελέγχους, αυτές οι υποθέσεις επιβεβαιώθηκαν και σε συνδυασμό με την υπόθεση της ανεξαρτησίας των δεδομένων που κάναμε στην αρχή, καταλήγουμε στο γενικό συμπέρασμα ότι το μοντέλο είναι κατάλληλο για την περιγραφή των δεδομένων καθώς και για μελλοντική πρόβλεψη της επί τοις εκατό περιεκτικότητας σε βενζόλιο της παραγόμενης βενζίνης.

Επίσης, το μοντέλο καταφέρνει να εξηγήσει το 51.6% της συνολικής μεταβλητότητας που υπάρχει στα δεδομένα. Το ποσοστό αυτό μπορεί να θεωρηθεί ικανοποιητικό δεδομένης της διαφορετικότητας της φύσης των ανεξάρτητων μεταβλητών του μοντέλου.

ΚΕΦΑΛΑΙΟ 5

Πολυμεταβλητή Ανάλυση

5.1 Εισαγωγή

Στο κεφάλαιο αυτό θα επικεντρωθούμε στις μεταβλητές $Y_1 - Y_3$ οι οποίες περιγράφουν τα χαρακτηριστικά του τελικού προϊόντος που είναι η βενζίνη. Θα εφαρμόσουμε δύο μεθόδους πολυμεταβλητής ανάλυσης, την Ανάλυση Κυρίων Συνιστωσών ή PCA και την Ανάλυση κατά Συστάδες ή Cluster Analysis. Η χρήση των δύο αυτών μεθόδων θα γίνει με σκοπό αφενός να μειώσουμε τη διάσταση των δεδομένων ώστε να μπορούν να περιγραφούν από τον πρώτο κύριο άξονα και αφετέρου για να δούμε αν και με ποιο τρόπο ομαδοποιούνται οι τρεις αυτές μεταβλητές. Αυτά είναι ιδιαίτερα σημαντικά αφού θα μας βοηθήσουν να ερμηνεύσουμε καλύτερα τα χαρακτηριστικά που εκφράζουν και να κατανοήσουμε με τον τρόπο αυτό τη σχέση που έχουν μεταξύ τους.

Μέσω της Ανάλυσης Κυρίων Συνιστωσών θα προσπαθήσουμε να δημιουργήσουμε γραμμικούς συνδυασμούς των αρχικών δεδομένων οι οποίοι να είναι ασυσχέτιστοι μεταξύ τους και να περιέχουν όσο γίνεται μεγαλύτερο μέρος της πληροφορίας που υπάρχει στις αρχικές μεταβλητές, επιτυγχάνοντας με αυτό τον τρόπο μια αξιόπιστη οπτική παρουσίαση των δεδομένων. Μέσω της Ανάλυσης κατά Συστάδες θα προσπαθήσουμε να ομαδοποιήσουμε τις μεταβλητές δημιουργώντας ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους, επιτυγχάνοντας με αυτό τον τρόπο ευκολότερη και αποδοτικότερη επεξεργασία των δεδομένων.

Θα γίνει και πάλι χρήση των στατιστικών προγραμμάτων SPSS και Minitab για την εφαρμογή των μεθόδων και το επίπεδο σημαντικότητας που θα χρησιμοποιείται για τους ελέγχους υποθέσεων είναι το 5%.

5.2 Ανάλυση Κυρίων Συνιστωσών

Προτού ξεκινήσουμε την ανάλυσή μας θα πρέπει να ελέγξουμε την καταλληλότητα των δεδομένων για εφαρμογή της μεθόδου της Ανάλυσης Κυρίων Συνιστωσών, υπολογίζοντας τις

συσχετίσεις ανάμεσα στις μεταβλητές καθώς και τον πίνακα διακυμάνσεων – συνδιακυμάνσεων. Αντίστοιχα, τα αποτελέσματα είναι τα παρακάτω.

Πίνακας 5.1: Συσχετίσεις μεταξύ των μεταβλητών Y_1, Y_2, Y_3

	Y_1	Y_2
Y_2	-0.654 0.000 (*p*-value)	
Y_3	0.663 0.000 (*p*-value)	-0.622 0.000 (*p*-value)

Πίνακας 5.2: Διακυμάνσεις – συνδιακυμάνσεις μεταξύ των μεταβλητών Y_1, Y_2, Y_3

	Y_1	Y_2	Y_3
Y_1	1.589845		
Y_2	-3.915552	22.519985	
Y_3	0.231427	-0.817478	0.076582

Από τον πίνακα συσχετίσεων παρατηρούμε ότι όλες οι μεταβλητές είναι συσχετισμένες μεταξύ τους ($p - value = 0 < 0.05 = a$). Επομένως, από την ανάλυσή μας δεν μπορεί να εξαιρεθεί κάποια από τις τρεις μεταβλητές. Από τον πίνακα διακυμάνσεων – συνδιακυμάνσεων παρατηρούμε ότι η μεταβλητή Y_2 – τάση ατμών κατά Reid έχει μεγάλη διακύμανση σε σχέση με τις άλλες μεταβλητές πράγμα που σημαίνει ότι θα είναι αυτή που θα συμμετέχει περισσότερο στο σχηματισμό των κυρίων συνιστωσών. Αυτό όμως δεν είναι επιθυμητό γιατί μπορεί να οφείλεται στις μονάδες μέτρησης του μεγέθους που εκφράζει η μεταβλητή. Απλά, για να 'ομαλοποιήσουμε' περισσότερο τα δεδομένα μας θα χρειαστεί να τυποποιήσουμε και να χρησιμοποιήσουμε τον πίνακα συσχετίσεων των αρχικών δεδομένων που θα είναι ο πίνακας διακυμάνσεων – συνδιακυμάνσεων των τυποποιημένων. Από όσα αναφέρθηκαν προηγουμένως, καταλήγουμε στο ότι τα δεδομένα μας είναι κατάλληλα για εφαρμογή των μεθόδων της Ανάλυσης Κυρίων Συνιστωσών.

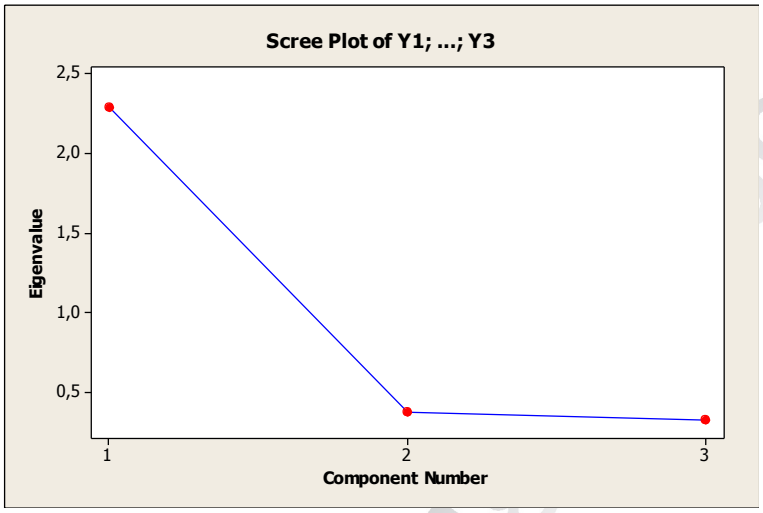
Ξεκινώντας την ανάλυσή μας θα υπολογίσουμε αρχικά τις ιδιοτιμές του πίνακα συσχετίσεων των αρχικών δεδομένων ο οποίος είναι ο κάτωθι.

Πίνακας 5.3: Ιδιοτιμές του πίνακα συσχετίσεων των δεδομένων

Eigenvalue	2.2936	0.3780	0.3284
Proportion	0.765	0.126	0.109
Cumulative	0.765	0.891	1

Για να καθορίσουμε το βέλτιστο πλήθος κυρίων συνιστωσών, θα κάνουμε χρήση του κριτηρίου Kaiser που περιγράψαμε στην Ενότητα 2.4.1.6. Σύμφωνα με αυτό, το βέλτιστο πλήθος κυρίων συνιστωσών είναι μία μιας και μόνο η πρώτη ιδιοτιμή είναι μεγαλύτερη της

μονάδας. Ένας γραφικός τρόπος για την επιλογή του πλήθους των κυρίων συνιστωσών είναι το Scree Plot που περιγράψαμε επίσης στην Ενότητα 2.4.1.6 . Για τα δεδομένα μας προκύπτει το παρακάτω Scree Plot.



Παρατηρούμε ότι το διάγραμμα συμφωνεί με το κριτήριο του Kaiser ως προς την απόφαση μιας και η μεταβολή από την δεύτερη ιδιοτιμή και μετά δεν είναι σημαντική. Επομένως, έχοντας υπολογίσει το βέλτιστο πλήθος, η πρώτη κύρια συνιστώσα, όπως μας δίνεται από το πρόγραμμα είναι η παρακάτω.

Πίνακας 5.4: Πρώτη κύρια συνιστώσα

Μεταβλητή	PC1
Y_1	0.584
Y_2	-0.572
Y_3	0.575

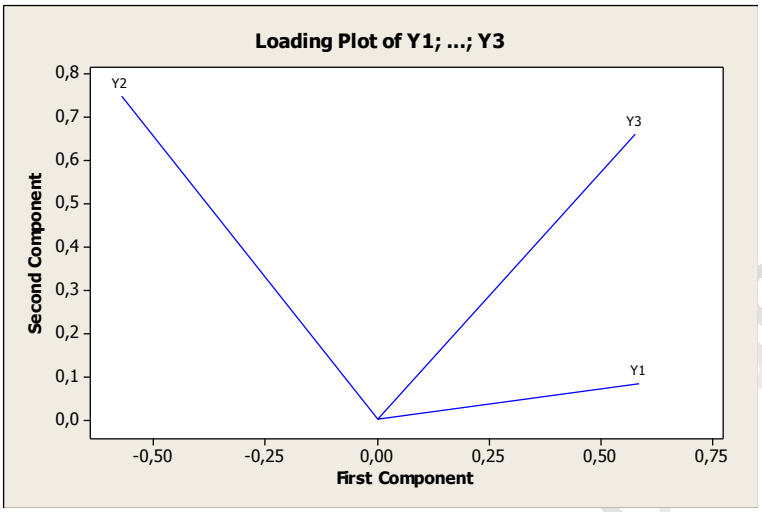
Σε αυτό το σημείο θα πρέπει να αναφερθεί ότι οι μεταβλητές Z_1, Y_1, Y_2 και Y_3 είναι τυποποιημένες και ως εκ τούτου ο τύπος της πρώτης κύριας συνιστώσας δίνεται από την ακόλουθη σχέση:

$$\frac{Z_1 - \bar{Z}_1}{s_{\bar{Z}_1}} = 0.584 \cdot \frac{Y_1 - \bar{Y}_1}{s_{\bar{Y}_1}} - 0.572 \cdot \frac{Y_2 - \bar{Y}_2}{s_{\bar{Y}_2}} + 0.575 \cdot \frac{Y_1 - \bar{Y}_1}{s_{\bar{Y}_3}} \Rightarrow$$

$$Z_1 = 0.584 \cdot \frac{Y_1 - \bar{Y}_1}{s_{\bar{Y}_1}} \cdot s_{\bar{Z}_1} - 0.572 \cdot \frac{Y_2 - \bar{Y}_2}{s_{\bar{Y}_2}} \cdot s_{\bar{Z}_1} + 0.575 \cdot \frac{Y_1 - \bar{Y}_1}{s_{\bar{Y}_3}} \cdot s_{\bar{Z}_1} + \bar{Z}_1$$

Όπως βλέπουμε από τον Πίνακα 5.3, το ποσοστό της συνολικής μεταβλητότητας που ερμηνεύει ο πρώτος κύριος άξονας είναι το 76.5%, ποσοστό πολύ ικανοποιητικό δεδομένου ότι χρησιμοποιούμε μόνο έναν κύριο άξονα.

Για να δούμε καλύτερα τη συμβολή της κάθε μεταβλητής στη διαμόρφωση των κυρίων αξόνων, έχουμε το παρακάτω διάγραμμα το οποίο ονομάζεται διάγραμμα φορτίων ή Loading Plot.



Από αυτό το διάγραμμα μπορούμε να πούμε ότι η μεταβλητή Y_1 συμβάλει στη διαμόρφωση μόνο της πρώτης συνιστώσας ενώ οι άλλες δύο συμβάλουν περίπου στο ίδιο μέγεθος στη διαμόρφωση τόσο του πρώτου όσο και του δεύτερου κύριου άξονα, η μία αρνητικά και η άλλη θετικά. Επίσης αν θέλουμε να δώσουμε ερμηνεία στον πρώτο κύριο άξονα, θα πρέπει πρώτα να υπολογίσουμε τις τιμές scores για την πρώτη κύρια συνιστώσα και στη συνέχεια να δούμε με ποιες μεταβλητές έχει ισχυρή συσχέτιση. Από το στατιστικό πρόγραμμα Minitab υπολογίζουμε τις τιμές αυτές, έστω η μεταβλητή Score1, επομένως οι συσχετίσεις του Pearson μεταξύ των Y_1, Y_2, Y_3 και της Score1 είναι οι εξής:

Πίνακας 5.5: Συσχετίσεις των μεταβλητών Y_1, Y_2, Y_3 με την μεταβλητή Score1

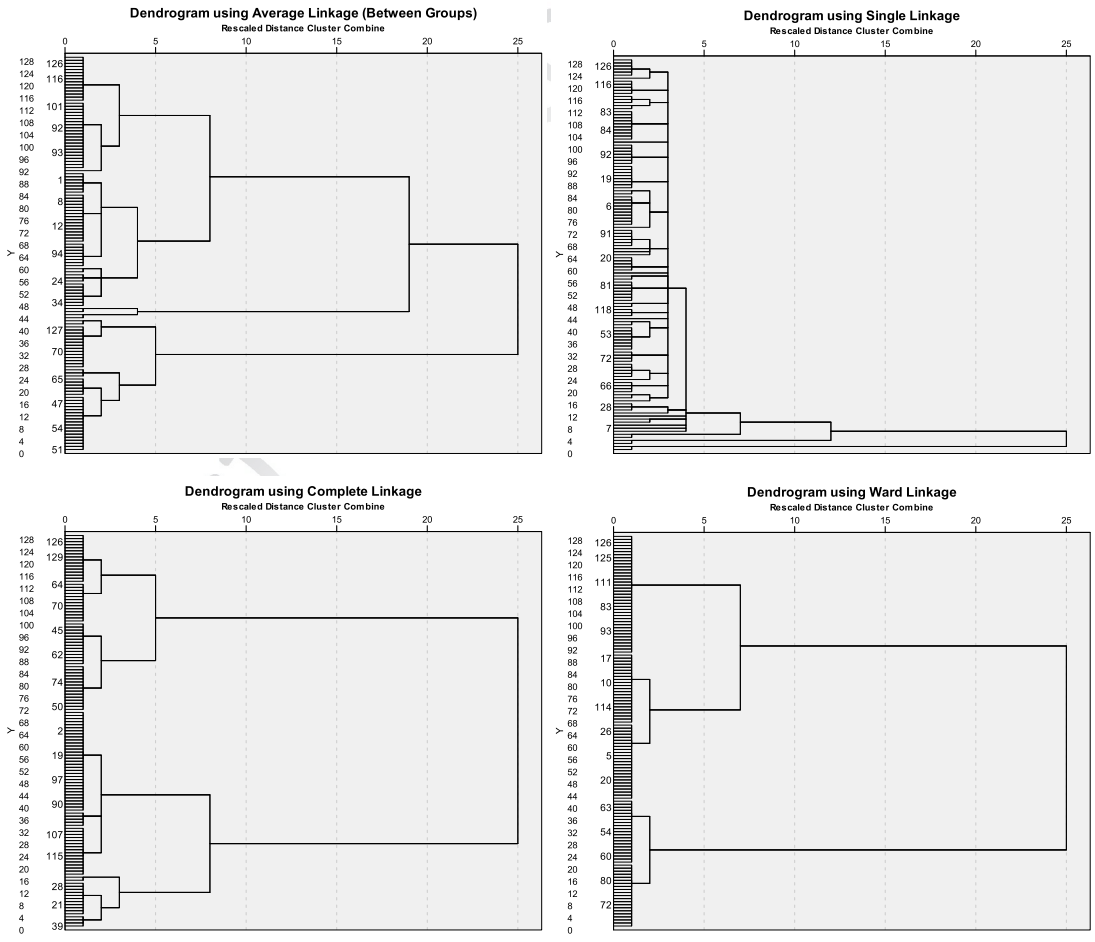
	Y_1	Y_2	Y_3
Y_2	-0.654 0.000 (*p*-value)		
Y_3	0.663 0.000 (*p*-value)	-0.622 0.000 (*p*-value)	
Score1	0.885 0.000 (*p*-value)	-0.867 0.000 (*p*-value)	0.871 0.000 (*p*-value)

Για την πρώτη κύρια συνιστώσα μπορούμε να πούμε ότι σχετίζεται περισσότερο με τις μεταβλητές Y_1 και Y_3. Αν δούμε τη φύση των μεταβλητών αυτών μπορούμε να πούμε ότι ίσως η πρώτη κύρια συνιστώσα σχετίζεται με τεχνικά ποιοτικά χαρακτηριστικά του τελικού προϊόντος όπως αυτά καθορίζονται κατά τη διαδικασία της αναμόρφωσης, του εμπλουτισμού και γενικά μέσα στη διαδικασία της διύλισης. Επομένως, μπορούμε να πούμε ότι η πρώτη κύρια συνιστώσα θα μπορούσε να χρησιμοποιηθεί ως δείκτης για την περιγραφή των ποιοτικών χαρακτηριστικών της βενζίνης.

5.3 Ανάλυση Κατά Συστάδες

Στην ενότητα αυτή θα προσπαθήσουμε να δημιουργήσουμε ομάδες οι οποίες θα περιέχουν όμοιες παρατηρήσεις και θα ερμηνεύσουμε (υποκειμενικά) τα αποτελέσματα της ομαδοποίησης. Για να γίνει αυτό θα πρέπει πρώτα να καθοριστεί ο κατάλληλος αριθμός ομάδων που θα δημιουργηθεί. Γι' αυτό το λόγο θα συνδυάσουμε τα αποτελέσματα κάποιων ιεραρχικών μεθόδων, όπως τη μέθοδο της συνένωσης μεταξύ των ομάδων (between-groups linkage), του πλησιέστερου και μακρινότερου γείτονα καθώς και τη μέθοδο του Ward. Το αποτέλεσμα που θα πάρουμε ως προς το πλήθος των ομάδων που πρέπει να δημιουργήσουμε, θα το χρησιμοποιήσουμε για την εφαρμογή μιας μη-ιεραρχικής μεθόδου, για παράδειγμα της μεθόδου των k-μέσων (k-means method).

Ξεκινώντας λοιπόν με τις τέσσερις συσσωρευτικές ιεραρχικές μεθόδους και χρησιμοποιώντας την τετραγωνική Ευκλείδεια απόσταση (squared Euclidean distance), έχουμε τα ακόλουθα δενδρογράμματα.



Τα δενδρογράμματα, με τη σειρά που εμφανίζονται είναι με τη μέθοδο between-groups linkage, με τη μέθοδο του κοντινότερου γείτονα, με τη μέθοδο του μακρινότερου γείτονα και με τη μέθοδο του Ward. Αν και με την πρώτη μέθοδο δεν γίνεται ξεκάθαρος ο αριθμός των ομάδων που πρέπει να δημιουργηθούν, με τις άλλες τρεις μεθόδους μπορούμε να ισχυριστούμε ότι ο κατάλληλος αριθμός ομάδων θα πρέπει να είναι δύο. Επίσης, κάποιος θα μπορούσε να ισχυριστεί ότι μπορούν να δημιουργηθούν 3 ομάδες με τη μέθοδο του Ward, επιβεβαιώνοντας το πόσο υποκειμενικές μπορεί να είναι οι απόψεις ως προς την επιλογή του κατάλληλου πλήθους ομάδων μέσω των δενδρογραμμάτων.

Επομένως, χρησιμοποιώντας τον αριθμό αυτό, στη συνέχεια εφαρμόζουμε την μη-ιεραρχική μέθοδο των k-μέσων και έχουμε τους ακόλουθους πίνακες.

Πίνακας 5.6: Αρχικά και τελικά κέντρα των ομάδων

	Αρχικά Κέντρα		Τελικά Κέντρα	
	Cluster		Cluster	
	1	2	1	2
Οκτάνια τελικού προϊόντος	95.40	91.40	94.96	93.26
Τάση ατμών κατά Reid	52	72	56.47	64.59
Βενζόλιο % κ.ο.	1.10	0.70	1.16	0.81

Πίνακας 5.7: Αριθμός παρατηρήσεων σε κάθε ομάδα

Cluster	1	53
	2	76
Valid		129
Missing		0

Από τον Πίνακα 5.6 βλέπουμε τη μεταβολή των κέντρων των ομάδων μετά την τελική κατάταξη των παρατηρήσεων σε αυτές. Μπορούμε να πούμε ότι εκτός από τη μεταβλητή ‘τάση ατμών κατά Reid’, τα κέντρα των ομάδων σύμφωνα με τις άλλες δύο μεταβλητές δεν μεταβλήθηκαν πολύ. Αυτό είναι ίσως μια ένδειξη ότι από μόνες τους οι αρχικές παρατηρήσεις είχαν την τάση να ομαδοποιούνται όπως και έγινε τελικά μετά από επτά επαναλήψεις του αλγόριθμου. Επίσης, η απόσταση μεταξύ των τελικών κέντρων των ομάδων είναι 8.304 όπως δίνεται από το στατιστικό πρόγραμμα.

Όπως γνωρίζουμε, οι ιεραρχικές μέθοδοι δουλεύουν ικανοποιητικά για μεγάλα μεγέθη δειγμάτων και δημιουργούν ομάδες με παραπλήσιο αριθμό παρατηρήσεων. Βλέπουμε λοιπόν από τον Πίνακα 5.7 ότι κάτι τέτοιο ισχύει μιας και η πρώτη ομάδα περιέχει 53 παρατηρήσεις και η δεύτερη 76.

Αν θέλουμε να δώσουμε ερμηνεία στον τρόπο με τον οποίο χωρίστηκαν οι παρατηρήσεις σε αυτές τις ομάδες, μπορούμε να δούμε ποιες παρατηρήσεις ανήκουν στην κάθε ομάδα και να ελέγξουμε αν υπάρχει κάποια σύνδεση, για παράδειγμα, με τη χρονική στιγμή που λήφθηκε κάθε μία από αυτές. Με τη βοήθεια του προγράμματος, ταξινομούμε τα δεδομένα σύμφωνα με την ομάδα που ανήκουν (πρώτα οι παρατηρήσεις της πρώτης και μετά της δεύτερης ομάδας) και παρατηρούμε ότι οι παρατηρήσεις της πρώτης ομάδας είναι στο μεγαλύτερο μέρος τους μετρήσεις που λήφθηκαν κατά το πρώτο τρίμηνο του έτους, ενώ της δεύτερης ομάδας μετρήσεις που λήφθηκαν κατά το δεύτερο τρίμηνο του έτους. Επομένως, μπορούμε να ισχυριστούμε ότι στην πρώτη ομάδα έχουν την τάση να κατατάσσονται παρατηρήσεις οι οποίες λήφθηκαν υπό διαφορετικές, πιθανώς, περιβαλλοντικές συνθήκες από αυτές που λήφθηκαν οι παρατηρήσεις της δεύτερης ομάδας.

5.4 Συμπεράσματα

- **Ανάλυση Κυρίων Συνιστωσών**

Μέσω της Ανάλυσης Κυρίων Συνιστωσών καταφέραμε να δημιουργήσουμε μια νέα μεταβλητή με βάση τις τρεις που είχαμε αρχικά, η οποία εμπεριέχει ένα μεγάλο μέρος της πληροφορίας (76.5%) αυτών. Δηλαδή, μειώσαμε τη διάσταση των δεδομένων από τις τρεις διαστάσεις στη μία. Αυτό είναι ιδιαίτερα σημαντικό όταν φτάνουμε στο κομμάτι της ερμηνείας του πρώτου κύριου άξονα που εκφράζεται από τη νέα μας μεταβλητή. Από την ανάλυση που έγινε, καταλήξαμε στο ότι η πρώτη κύρια συνιστώσα σχετίζεται με χαρακτηριστικά της βενζίνης τα οποία καθορίζονται κατά τη διαδικασία της διύλισης. Μεγαλύτερο ρόλο στη διαμόρφωση της πρώτης κύριας συνιστώσας παίζουν οι μεταβλητές Y_1 και Y_3 που συμβάλουν θετικά στη διαμόρφωσή της, ενώ η Y_2 συμβάλει αρνητικά, σχεδόν στον ίδιο βαθμό.

- **Ανάλυση Κατά Συστάδες**

Χρησιμοποιώντας έναν συνδυασμό ιεραρχικών και μη-ιεραρχικών μεθόδων, καταφέραμε να ομαδοποιήσουμε τις παρατηρήσεις που είχαμε στη διάθεσή μας σε δύο ομάδες με άνισο μεν, κοντινό δε, αριθμό παρατηρήσεων η καθεμία. Τα κέντρα των ομάδων δεν μεταβλήθηκαν σημαντικά μετά την τελική ομαδοποίηση, πράγμα που πιθανώς σημαίνει ότι τα δεδομένα από μόνα τους λόγω της φύσης τους, είχαν την τάση να ταξινομούνται στις δύο ομάδες με παρόμοιο τρόπο. Αυτό μας οδήγησε στο να συμπεράνουμε ότι οι παρατηρήσεις που

ταξινομούνται στις δύο ομάδες, είναι πιθανό να επηρεάζονται σε κάποιο βαθμό από τη χρονική στιγμή που λαμβάνονται (πρώτο ή δεύτερο τρίμηνο του έτους).

ΚΕΦΑΛΑΙΟ 6

Στατιστικός Έλεγχος Ποιότητας

6.1 Εισαγωγή

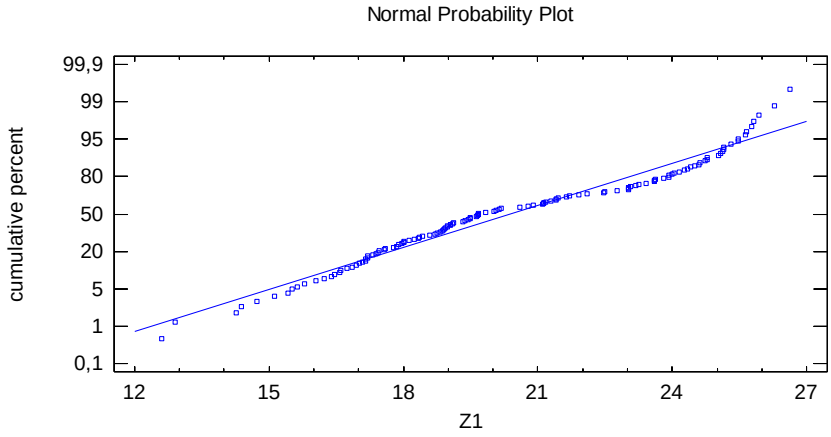
Στο προηγούμενο κεφάλαιο, με τη βοήθεια της Ανάλυσης Κυρίων Συνιστωσών, καταφέραμε να μειώσουμε τη διάσταση των δεδομένων των τελικών ποιοτικών χαρακτηριστικών της βενζίνης από τρισδιάστατα σε μονοδιάστατα. Στη μία διάσταση τα αναπαραστήσαμε με την πρώτη κύρια συνιστώσα Z_1, η οποία καταφέρνει να ερμηνεύσει το 76.5% της συνολικής μεταβλητότητας των δεδομένων.

Στο κεφάλαιο αυτό, θα χρησιμοποιήσουμε τη Z_1 και με τη βοήθεια των εργαλείων του Στατιστικού Ελέγχου Ποιότητας θα προσπαθήσουμε να ανιχνεύσουμε έγκαιρα την εμφάνιση ειδικών αιτιών μεταβλητότητας στην διεργασία παραγωγής της βενζίνης, έτσι ώστε να προχωρήσουμε σε έρευνα και να προβούμε στις απαραίτητες διορθωτικές ενέργειες προτού κατασκευαστούν αρκετά προϊόντα μη συμμορφούμενα με τις προδιαγραφές. Αυτό θα γίνει με την κατασκευή κατάλληλων διαγραμμάτων ελέγχου όπως το διάγραμμα ελέγχου τύπου *Shewhart*, το διάγραμμα *EWMA* και το διάγραμμα *CUSUM* για μεμονωμένες παρατηρήσεις.

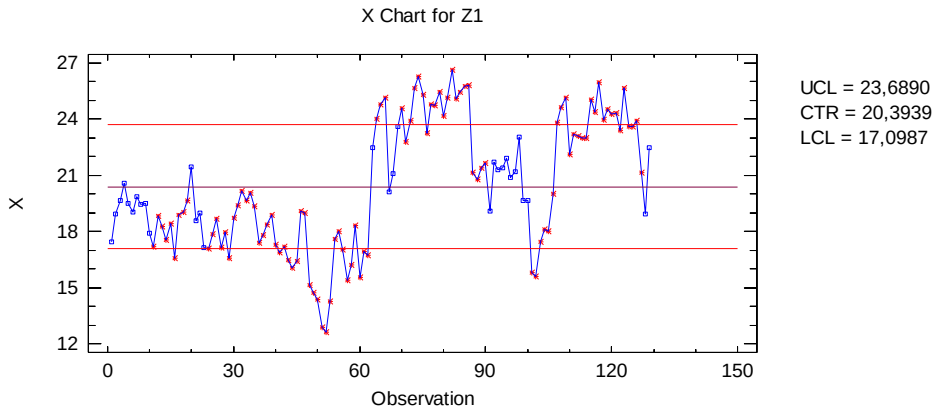
Επίσης, θα γίνει χρήση του στατιστικού προγράμματος Statgraphics για την εφαρμογή των μεθόδων και το επίπεδο σημαντικότητας που θα χρησιμοποιείται για τους ελέγχους υποθέσεων είναι το 5%.

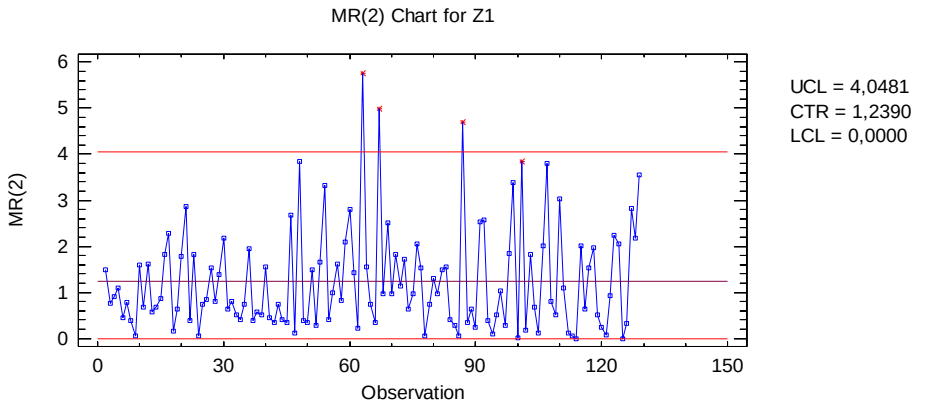
6.2 Αρχική μελέτη των δεδομένων και κατασκευή διαγράμματος ελέγχου τύπου *Shewhart*

Πριν ξεκινήσουμε την ανάλυσή μας και κατασκευάσουμε κατάλληλα διαγράμματα για τα δεδομένα, θα πρέπει να εξετάσουμε αν αυτά ακολουθούν την κανονική κατανομή. Από το Statgraphics, τόσο με τον γραφικό τρόπο του *Normal Probability Plot*, όσο και με τον έλεγχο κανονικότητας των *Kolmogorov – Smirnov*, καταλήγουμε ότι σε επίπεδο σημαντικότητας $\alpha = 5\%$, η πρώτη κύρια συνιστώσα Z_1 είναι κανονική ($p - value = 0{,}181336 > 0.05 = \alpha$). Το *Normal Probability Plot* όπως δίνεται από το πρόγραμμα φαίνονται παρακάτω.



Για να ελέγξουμε αν η διεργασία είναι εντός στατιστικού ελέγχου θα κατασκευάσουμε ένα διάγραμμα *X* για μεμονωμένες παρατηρήσεις και ένα MR διάγραμμα με όρια ελέγχου 3σ. Τα διαγράμματα που παίρνουμε είναι τα ακόλουθα.

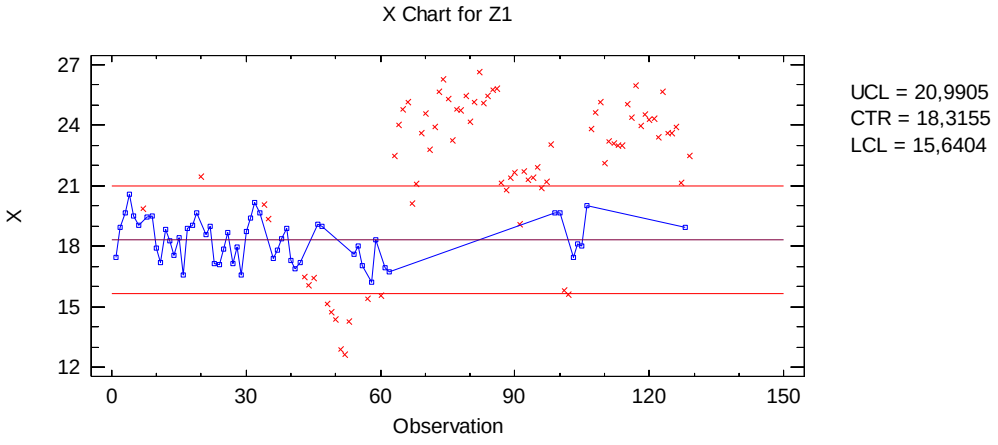


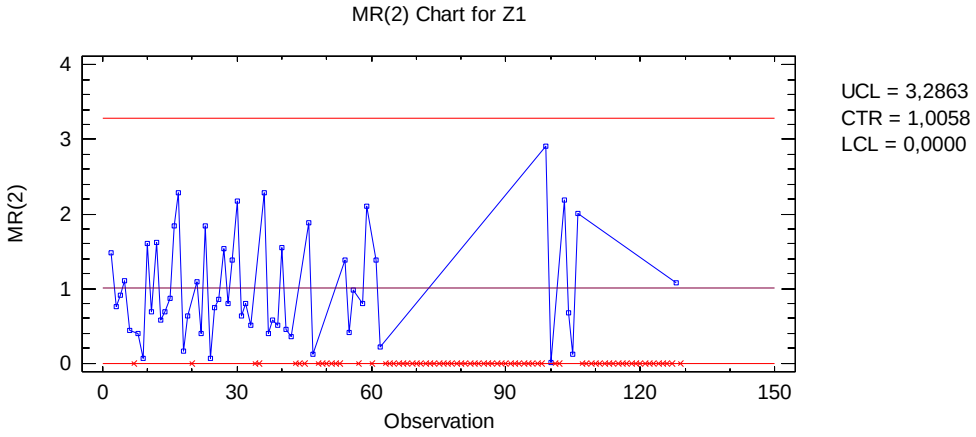


Από την αρχική εικόνα που παίρνουμε μπορούμε να πούμε ότι η διεργασία είναι εκτός στατιστικού ελέγχου. Παρατηρούμε ότι, από την παρατήρηση 63 μέχρι και την παρατήρηση 86 καθώς και από την παρατήρηση 107 μέχρι και την 126 υπάρχει μια μετατόπιση του μέσου της διεργασίας προς τα πάνω, ενώ από την παρατήρηση 41 μέχρι και την 62 υπάρχει μια μετατόπιση του μέσου της διεργασίας προς τα κάτω. Το διάγραμμα ελέγχου τύπου *Shewhart* που κατασκευάσαμε, παρατηρούμε ότι ανιχνεύει την πρώτη μεγάλη μεταβολή του μέσου της διεργασίας περίπου από το σημείο 41 ενώ μια μικρή μεταβολή στην αρχή της διεργασίας, περίπου στο σημείο 16.

Οι μεταβολές στο μέσο μπορούν ως ένα βαθμό να εξηγηθούν αν αναλογιστούμε το γεγονός ότι στη διαμόρφωση της πρώτης κύριας συνιστώσας που αναπαραστήσαμε στα παραπάνω διαγράμματα ελέγχου, οι μεταβλητές Y_1 και Y_3 συμβάλουν θετικά ενώ η Y_2 αρνητικά. Επομένως, αν δούμε τις τιμές των παρατηρήσεων που αναφέραμε, θα δούμε ότι όταν παρατηρείται αύξηση του μέσου της διεργασίας, οι τιμές των Y_1 και Y_3 είναι υψηλές ενώ της Y_2 χαμηλές, ενώ το αντίστροφο συμβαίνει όταν παρατηρείται μείωση του μέσου της διεργασίας.

Αφαιρώντας τα σημεία που είναι εκτός των ορίων ελέγχου καθώς και τα σημεία που παρουσιάζουν μη τυχαία συμπεριφορά, βάσει των κανόνων που υπάρχουν, από τα μακρινότερα προς τα κοντινότερα στα όρια ελέγχου, έχουμε τα ακόλουθα.





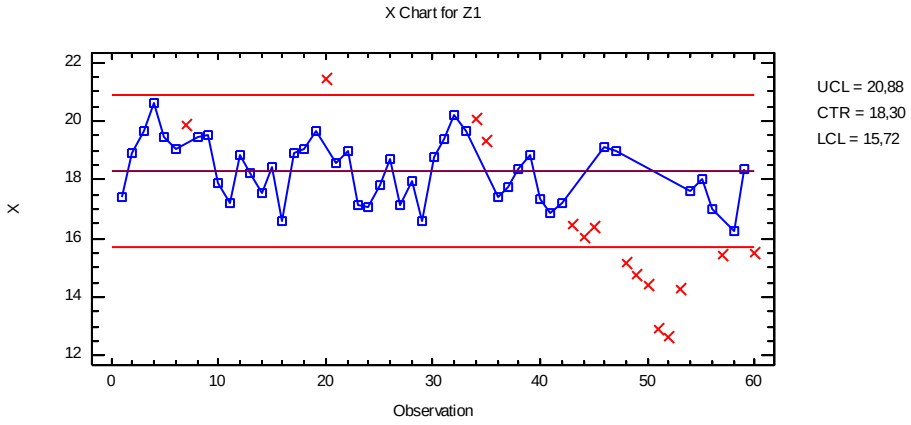
Τα όρια ελέγχου που μπορούν να χρησιμοποιηθούν για μελλοντική παρακολούθηση σε πραγματικό χρόνο είναι:

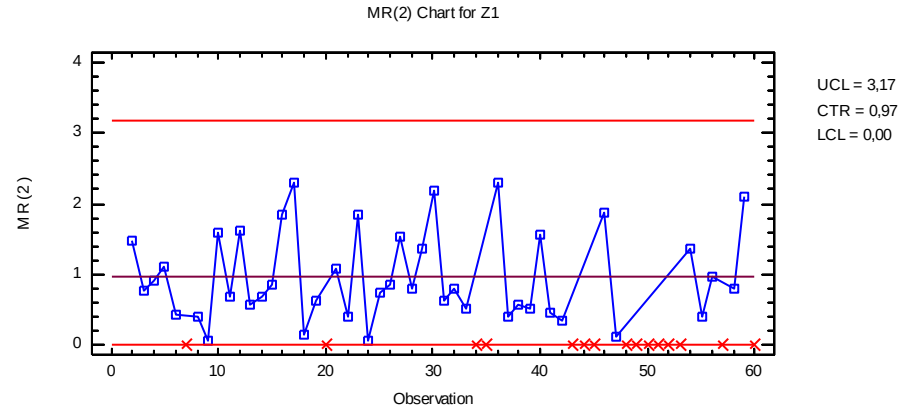
- Για τον μέσο $LCL = 15.6404$ και $UCL = 20.9905$
- Για το κινούμενο εύρος $LCL = 0$ και $UCL = 3.2863$

Οι εντός ελέγχου εκτιμήσεις της μέσης τιμής και της τυπικής απόκλισης της διαδικασίας είναι αντίστοιχα 18.3155 και 0.891673 .

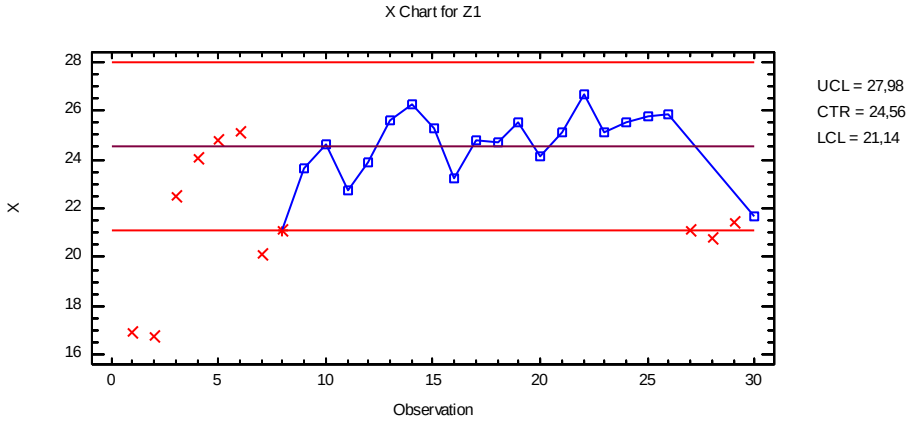
Παρατηρώντας τον τρόπο που κινούνται τα σημεία του αρχικού διαγράμματος X, μπορούμε να ισχυριστούμε ότι τα σημεία συγκεντρώνονται σε δύο ή τρεις ομάδες, ισχυροποιώντας έτσι τα συμπεράσματα που εξήχθησαν από την Ανάλυση κατά Συστάδες. Αυτό πολύ πιθανόν να οφείλεται στην ύπαρξη εποχικότητας στα δεδομένα. Για το λόγο αυτό, θα χωρίσουμε τα δεδομένα σε τρεις ομάδες, παρατηρήσεις $1-60$, $61-90$ και $91-129$, και θα κατασκευάσουμε αντίστοιχα διαγράμματα ελέγχου. Τα διαγράμματα X για μεμονωμένες παρατηρήσεις καθώς και τα MR διαγράμματα με όρια ελέγχου 3σ για τις τρεις παραπάνω ομάδες, μετά την αφαίρεση των σημείων που είναι εκτός ορίων ελέγχου, είναι τα ακόλουθα.

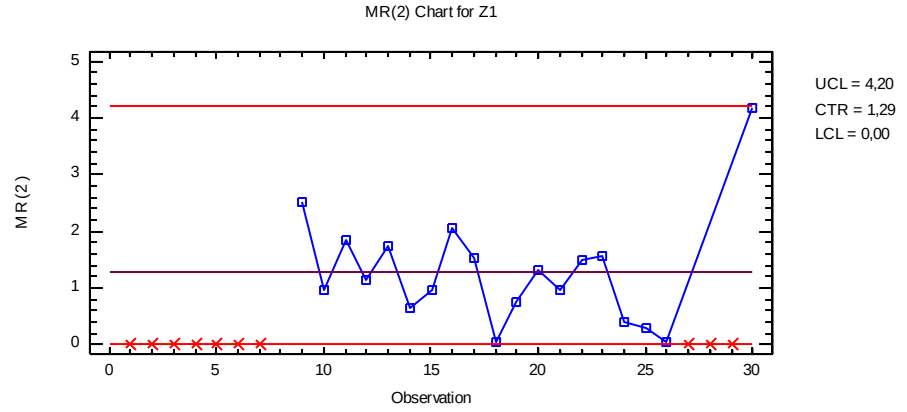
- Για τις παρατηρήσεις 1 – 60



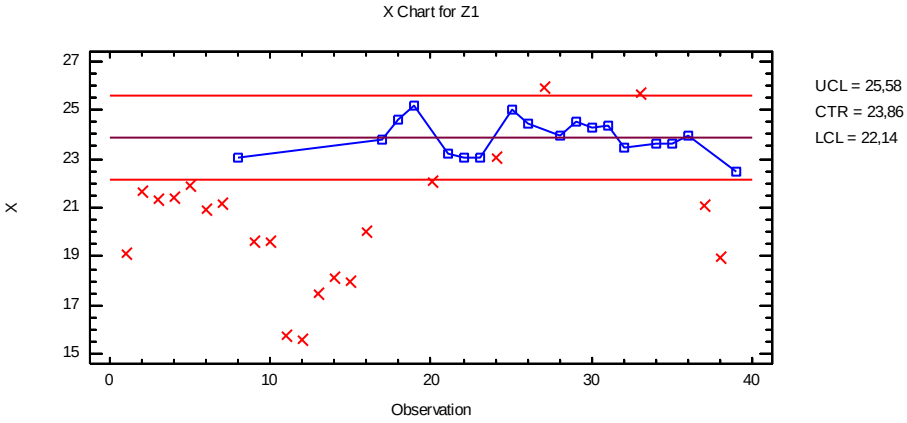


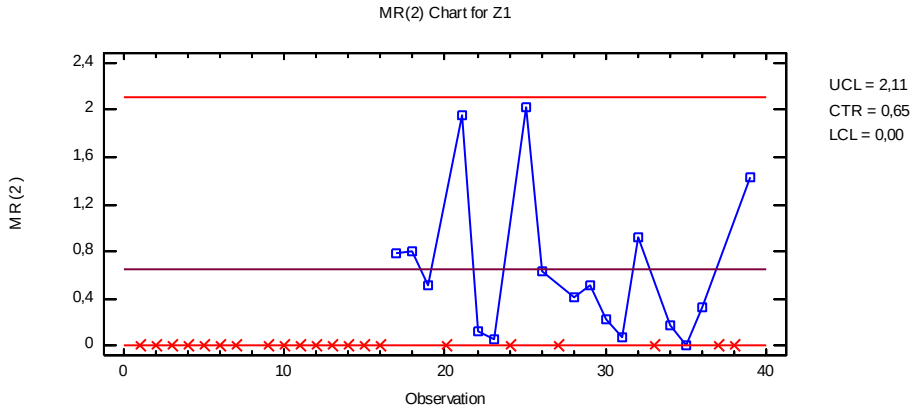
- Για τις παρατηρήσεις 61 – 90





- Για τις παρατηρήσεις 91 – 129





Τα όρια ελέγχου που μπορούν να χρησιμοποιηθούν για μελλοντική παρακολούθηση σε πραγματικό χρόνο είναι:

- Για τις παρατηρήσεις $1 - 60$

Για τον μέσο $LCL = 15.7176$ και $UCL = 20.8828$

Για το κινούμενο εύρος $LCL = 0$ και $UCL = 3.1727$

Οι εντός ελέγχου εκτιμήσεις της μέσης τιμής και της τυπικής απόκλισης της διαδικασίας είναι αντίστοιχα 18.3002 και 0.860857 .

- Για τις παρατηρήσεις $61 - 90$

Για τον μέσο $LCL = 21.136$ και $UCL = 27.98$

Για το κινούμενο εύρος $LCL = 0$ και $UCL = 4.2039$

Οι εντός ελέγχου εκτιμήσεις της μέσης τιμής και της τυπικής απόκλισης της διαδικασίας είναι αντίστοιχα 24.558 και 1.14066 .

- Για τις παρατηρήσεις $91 - 129$

Για τον μέσο $LCL = 22.1398$ και $UCL = 25.5784$

Για το κινούμενο εύρος $LCL = 0$ και $UCL = 2.1121$

Οι εντός ελέγχου εκτιμήσεις της μέσης τιμής και της τυπικής απόκλισης της διαδικασίας είναι αντίστοιχα 23.8591 και 0.573094 .

Παρατηρούμε ότι οι δύο τελευταίες ομάδες παρατηρήσεων έχουν πολύ κοντινούς μέσους σε σχέση με αυτόν της πρώτης ομάδας που απέχει σημαντικά από τους άλλους δύο. Ενδεχομένως αυτό να οφείλεται στον τρόπο με τον οποίο συμμετέχουν οι διάφοροι τύποι αργού πετρελαίου στη διαμόρφωση του τελικού προϊόντος, οι συνθήκες που επικρατούν στο διυλιστήριο τη συγκεκριμένη περίοδο ακόμα και οι κλιματολογικές συνθήκες την περίοδο αυτή.

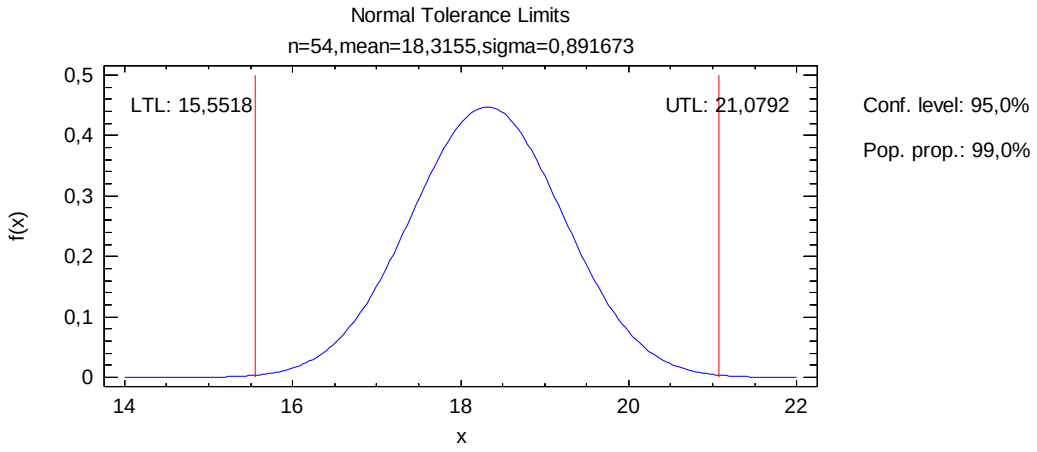
Στη συνέχεια θα εξετάσουμε την ικανότητα της διεργασίας να παράγει όσο το δυνατό μικρότερο αριθμό ελαττωματικών προϊόντων, που στην περίπτωσή μας είναι οι παρατηρήσεις του πρώτου κύριου άξονα. Μέσω της Z_1, υπολογίζουμε την εκτίμηση του μέσου και της τυπικής απόκλισης της διεργασίας (προτού δηλαδή αφαιρέσουμε τα σημεία από το διάγραμμα ελέγχου για το μέσο) έτσι ώστε να υπολογίσουμε τα όρια προδιαγραφών. Τα τελευταία έχουν τιμές $USL = 23{,}689$ και $LSL = 17{,}0988$. Επομένως μέσω του Statgraphics έχουμε τους παρακάτω δείκτες ικανότητας για την διεργασία.

Πίνακας 6.1: Δείκτες ικανότητας της διεργασίας

	Short-Term Capability
Sigma	1,09838
Cp/Pp	0,999999
Cpk/Ppk	0,999993

Παρατηρούμε ότι οι δείκτες ικανότητας $\hat{C}_p$ και $\hat{C}_{pk}$ είναι και οι δύο περίπου ίσοι με τη μονάδα. Αυτό σημαίνει ότι η διεργασία δεν είναι πολύ ικανή και χρειάζεται παρακολούθηση μιας και παράγει αρκετά ελαττωματικά προϊόντα. Θα χρειαστεί να γίνει αναθεώρηση του σχεδιασμού και βελτιωτικές ενέργειες ώστε να αυξηθεί η απόδοση της διεργασίας.

Μπορούμε επίσης να κατασκευάσουμε και το κανονικό διάστημα ανοχής, το οποίο να περιέχει το 99% της πρώτης κύριας συνιστώσας με βεβαιότητα 95%, όταν η διεργασία είναι εντός στατιστικού ελέγχου. Με τη βοήθεια του προγράμματος, υπολογίζουμε αυτό το διάστημα το οποίο είναι το $(15.5518\,, 21.0792)$. Το αντίστοιχο διάγραμμα γι' αυτό το διάστημα ανοχής είναι το ακόλουθο.



Επίσης μπορούμε να υπολογίσουμε μέσω του προγράμματος το μέγεθος του δείγματος ώστε το μη-παραμετρικό διάστημα ανοχής $[X(1)\,, X(n)]$ για τη μεταβλητή Z_1 να περιέχει το 99% των διαθέσιμων τιμών της με πιθανότητα 0.92, όταν η διεργασία βρίσκεται εντός στατιστικού ελέγχου. Το μέγεθος αυτό είναι $n = 473$.

Κλείνοντας την ενότητα αυτή, μπορούμε να πούμε ότι το διάγραμμα ελέγχου τύπου *Shewhart* για μεμονωμένες παρατηρήσεις με 3σ όρια ελέγχου που κατασκευάστηκε για την παρακολούθηση του μέσου της πρώτης κύριας συνιστώσας Z_1, είναι σε θέση να ανιχνεύσει μόνο τις μεγάλες μετατοπίσεις του μέσου. Τέτοιες μετατοπίσεις φαίνονται για παράδειγμα από τις μεγάλες τιμές των σημείων 63 με 86 ή 107 με 126. Το διάγραμμα αυτό μας έδωσε

πρώτη φορά ένδειξη για εκτός ελέγχου διεργασία στο σημείο 11. Το εντός ελέγχου μέσο μήκος ροής, ARL_0 στο διάγραμμα αυτό είναι 370, πράγμα που σημαίνει ότι αναμένεται να σχεδιαστούν 370 σημεία στο διάγραμμα ελέγχου για να εμφανιστεί ένα σημείο εκτός των ορίων ελέγχου.

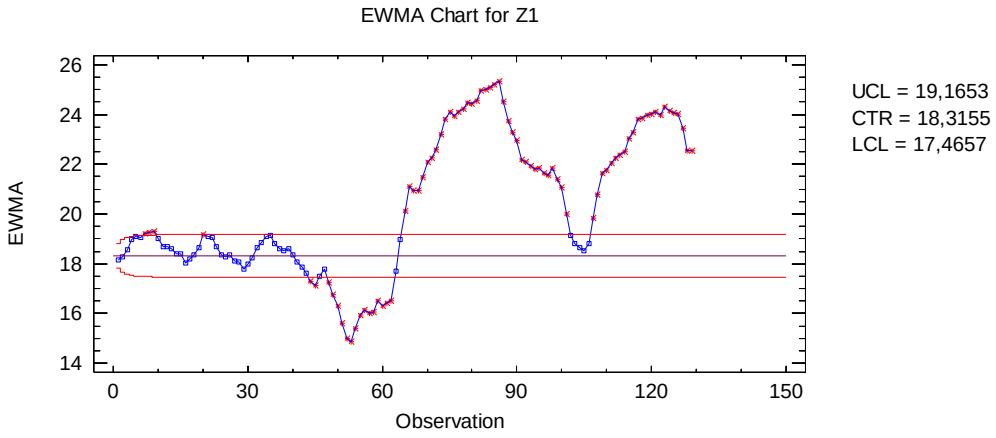
Αν θέλουμε να ανιχνεύσουμε πολύ νωρίς μικρές μετατοπίσεις του μέσου, μπορούμε να κατασκευάσουμε ένα διάγραμμα ελέγχου με μνήμη, όπως είναι το *EWMA* που θα δούμε στη συνέχεια.

6.3 Κατασκευή διαγράμματος ελέγχου με μνήμη – *EWMA*

Κατασκευάζοντας το διάγραμμα ελέγχου τύπου *Shewhart* για την παρακολούθηση του μέσου, εφαρμόσαμε ανάλυση Φάσης *I*, αφού δεν γνωρίζαμε τις τιμές της μέσης τιμής και της τυπικής απόκλισης της διεργασίας και επομένως έπρεπε να τις εκτιμήσουμε. Για την κατασκευή ενός διαγράμματος ελέγχου *EWMA*, θα εφαρμόσουμε ανάλυση Φάσης *II*, δηλαδή θα χρησιμοποιήσουμε τις εκτιμήσεις του μέσου και της τυπικής απόκλισης για τον υπολογισμό των ορίων ελέγχου.

Επίσης, θα πρέπει να δώσουμε κατάλληλες τιμές στις παραμέτρους λ και L ώστε να επιτύχουμε εντός ελέγχου μέσο μήκος ροής για το διάγραμμα ίσο περίπου με 370. Αυτό είναι ιδιαίτερα σημαντικό γιατί μας δίνει τη δυνατότητα να κάνουμε συγκρίσεις ανάμεσα στα δύο διαγράμματα.

Ένα διάγραμμα *EWMA* που πληροί τα παραπάνω είναι αυτό με $L = 2.859\sigma$ όρια ελέγχου και παράμετρο $\lambda = 0.2$ και είναι το ακόλουθο.



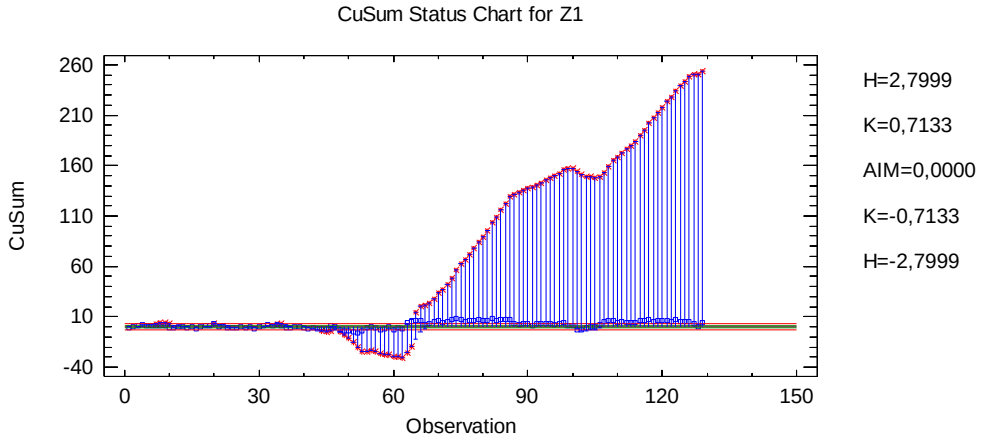
Το διάγραμμα αυτό έχει εντός ελέγχου μέσο μήκος ροής ίσο με 370 και επομένως είναι συγκρίσιμο με το αντίστοιχο διάγραμμα ελέγχου *Shewhart* που κατασκευάσαμε στην Ενότητα 6.2 . Κατά συνέπεια, μπορούμε να πούμε ότι για μια μικρή μετατόπιση του μέσου, το διάγραμμα *EWMA* έδωσε για πρώτη φορά σήμα για εκτός ελέγχου διεργασία μόλις στο σημείο 7 σε αντίθεση με το αντίστοιχο διάγραμμα *Shewhart* που έδωσε σήμα λίγο αργότερα, στο σημείο 16. Επίσης, για την πρώτη μεγάλη μετατόπιση του μέσου, το διάγραμμα *EWMA* έδωσε για πρώτη φορά σήμα για εκτός ελέγχου διεργασία στο σημείο 44 σε αντίθεση με το αντίστοιχο διάγραμμα *Shewhart* που έδωσε σήμα λίγο νωρίτερα, στο σημείο 41.

Τα στοιχεία αυτά επιβεβαιώνουν όσα ήδη γνωρίζουμε για τα δύο αυτά διαγράμματα, ότι δηλαδή ένα καλά σχεδιασμένο *EWMA* διάγραμμα ελέγχου είναι, σχεδόν πάντα, η καλύτερη επιλογή αν θέλουμε να εντοπίσουμε γρήγορα μικρές μετατοπίσεις του μέσου σε αντίθεση με το διάγραμμα ελέγχου τύπου *Shewhart* που είναι ικανό στο να εντοπίζει πιο γρήγορα μεγάλες μετατοπίσεις του μέσου της διεργασίας.

6.4 Κατασκευή διαγράμματος ελέγχου με μνήμη – *CUSUM*

Για την κατασκευή ενός διαγράμματος ελέγχου *CUCUM*, θα εφαρμόσουμε και πάλι ανάλυση Φάσης *II*, δηλαδή θα χρησιμοποιήσουμε τις εκτιμήσεις του μέσου και της τυπικής απόκλισης για τον υπολογισμό των ορίων ελέγχου.

Το διάγραμμα ελέγχου *CUSUM* με παραμέτρους $k = 0.8$ και $h = 3.14$ είναι το ακόλουθο.



Από το διάγραμμα αυτό παρατηρούμε ότι για τη μικρή μετατόπιση του μέσου της διεργασίας έχουμε ένδειξη για εκτός ελέγχου διεργασία πολύ νωρίς, στο σημείο 7, όπως είχαμε και στο *EWMA*. Για την πρώτη μεγάλη μετατόπιση του μέσου τα αποτελέσματα και

πάλι συμπίπτουν με αυτά του *EWMA* διαγράμματος, δηλαδή έχουμε ένδειξη για εκτός ελέγχου διεργασία στο σημείο 44.

Αυτό που παρατηρούμε δηλαδή είναι ότι τα διαγράμματα που κατασκευάστηκαν έως τώρα παρουσιάζουν την ίδια συμπεριφορά, δηλαδή είτε νωρίτερα είτε αργότερα, ανιχνεύουν τις ίδιες μεταβολές στο μέσο της διεργασίας.

Μπορούμε επίσης να κατασκευάσουμε κι άλλα διαγράμματα ελέγχου *EWMA* και *CUSUM*, μεταβάλλοντας τις παραμέτρους λ, L και h, k αντίστοιχα έτσι ώστε τα διαγράμματα να έχουν σχεδόν το ίδιο εντός ελέγχου μέσο μήκος ροής, κοντά στο 370. Αυτό θα μας επιτρέψει να επιλέξουμε το καλύτερο για την ανίχνευση μικρών μεταβολών του μέσου, το οποίο θα είναι εκείνο που έχει το μικτότερο εκτός ελέγχου μέσο μήκος ροής, ARL_1 Επομένως, για διάφορες τιμές των παραμέτρων αυτών καθώς και για τα ήδη κατασκευασμένα διαγράμματα έχουμε τον παρακάτω συγκεντρωτικό πίνακα.

Πίνακας 6.2: Εντός κι εκτός ελέγχου μέσα μήκη ροών

Διάγραμμα	$\boldsymbol{ARL_0}$	$\boldsymbol{ARL_1, \delta = 0.40}$	$\boldsymbol{ARL_1, \delta = 0.1.6}$
Shewhart με 3σ όρια ελέγχου	370.4	200.1	12.4
EWMA με $\lambda = 0.2$ και $L = 2.859$	370	55.4	4.8
EWMA με $\lambda = 0.1$ και $L = 2.701$	370	41.2	5.4
CUSUM με $k = 0.8$ και $h = 3.14$	370	82.9	4.7
CUSUM με $k = 0.8$ και $h = 3.14$	370.5	39.8	7.3

Αυτό που προκύπτει από τον Πίνακα 6.2 είναι ότι για μικρή μετατόπιση του μέσου της τάξης του 0.4σ, ικανότερο διάγραμμα φαίνεται να είναι το διάγραμμα *CUSUM* με $k = 0.8$ και $h = 3.14$ με το διάγραμμα *EWMA* με $\lambda = 0.1$ και $L = 2.701$ να είναι επίσης ικανοποιητικό με την έννοια ότι τα δύο διαγράμματα έχουν παραπλήσιο ARL_1. Επίσης, για μεγαλύτερη μετατόπιση της τάξης του 1.6σ, καταλληλότερα φαίνονται να είναι τα διαγράμματα *EWMA* με $\lambda = 0.2$ και $L = 2.859$ και *CUSUM* με $k = 0.8$ και $h = 3.14$, τα οποία έχουν σχεδόν ίσο ARL_1. Πρέπει πάντως να αναφέρουμε ότι σε καμία μετατόπιση του μέσου, μικρή ή μεγάλη, το διάγραμμα ελέγχου τύπου *Shewhart* με 3σ όρια ελέγχου δεν ήταν σε θέση να την εντοπίσει έγκαιρα.

6.4 Συμπεράσματα

Ξεκινώντας την ανάλυσή μας σε αυτό το κεφάλαιο, ελέγξαμε την καταλληλότητα των δεδομένων για την εφαρμογή μεθόδων Στατιστικού Ελέγχου Ποιότητας. Τα δεδομένα ήταν κανονικά κι έτσι προχωρήσαμε σε πρώτο στάδιο στην κατασκευή του κλασικού

διαγράμματος ελέγχου τύπου *Shewhart* με 3σ όρια ελέγχου. Από αυτό συμπεράναμε αμέσως ότι η διεργασία ήταν εκτός στατιστικού ελέγχου και έπρεπε να κάνουμε ενέργειες για να την φέρουμε εντός. Επίσης, έγινε εκτίμηση κάποιων δεικτών ικανότητας της διεργασίας από τους οποίους, όπως ήταν αναμενόμενο, καταλήξαμε στο ότι η διεργασία δεν είναι ικανή και χρειάζεται παρακολούθηση.

Στη συνέχεια για να διαπιστώσουμε αν μπορούμε να υπολογίσουμε τυχόν μικρές μετατοπίσεις του μέσου της διεργασίας, πέραν των μεγάλων που ήταν προφανείς, κατασκευάσαμε κατάλληλα διαγράμματα ελέγχου με μνήμη όπως είναι τα διαγράμματα *EWMA* και το *CUSUM* για μεμονωμένες παρατηρήσεις. Το αποτέλεσμα ήταν να ανιχνεύεται μια μικρή μετατόπιση του μέσου πολύ νωρίς και από τα δύο διαγράμματα.

Τέλος, έγινε μια σύγκριση διαφόρων διαγραμμάτων με βάση το εκτός ελέγχου μέσο μήκος ροής του καθενός, έχοντας αρχικά εξασφαλίσει ίσο σχεδόν εντός ελέγχου μέσο μήκος ροής για όλα. Ανάλογα με το μέγεθος της μετατόπισης που θέλαμε να ανιχνεύσουμε, η επιλογή μας ήταν ανάμεσα στα διαγράμματα με μνήμη, ενώ αξιοσημείωτο είναι ότι το διάγραμμα ελέγχου τύπου *Shewhart* που κατασκευάσαμε δεν ήταν ικανό να ανιχνεύσει ούτε την μεγάλη μετατόπιση του μέσου νωρίτερα από τα άλλα είδη διαγραμμάτων.

ΚΕΦΑΛΑΙΟ 7

Συμπεράσματα

7.1 Περιγραφική Στατιστική

Ξεκινώντας την ανάλυσή μας, χρειάστηκε να γίνει μια σύντομη περιγραφική ανάλυση, κατασκευάζοντας κάποια γραφήματα και υπολογίζοντας κάποιους δείκτες. Απώτερος σκοπός αυτών ήταν να πάρουμε μια πρώτη εικόνα της φύσης και της συμπεριφοράς των μεταβλητών που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία.

Πιο συγκεκριμένα μετά από κάποια διαγράμματα, όπως για παράδειγμα τα ιστογράμματα συχνοτήτων, οι γραφικές (και επομένως υποκειμενικές) ενδείξεις που λάβαμε ήταν ότι κάποιες από τις μεταβλητές $X_6 - X_{13}$ ίσως προσεγγίζονται από την κανονική κατανομή, ενώ άλλες όχι. Οι μεταβλητές $Y_1 - Y_3$ φάνηκε να πλησιάζουν σε κάποια κανονική κατανομή μιας και το ιστόγραμμά τους δεν απείχε αρκετά από την σχεδιασμένη κανονική καμπύλη.

Οι περιγραφικοί δείκτες που υπολογίστηκαν, μας κατέδειξαν κάποια βαθύτερα χαρακτηριστικά των μεταβλητών. Για παράδειγμα, το μέσο κατ' όγκο ποσοστό βενζολίου του τελικού προϊόντος, που είναι η βενζίνη, είναι κοντά στο 95%, πράγμα που ίσως να σημαίνει σημαντική συμμετοχή της μεταβλητής Y_3 στη διαμόρφωση της ποιότητας του τελικού προϊόντος. Επίσης, παρατηρήσαμε ότι ο μέσος αριθμός οκτανίων του τελικού προϊόντος (μεταβλητή Y_1) είναι, όπως αναμένεται αυξημένος σε σχέση με αυτόν του προϊόντος του ισομερισμού (μεταβλητή X_{10}). Τέλος, σημειώνουμε ότι για τις μεταβλητές $X_6 - X_{13}$ και $Y_1 - Y_3$, παρόλο που έχουμε γραφικές ενδείξεις ότι δεν ακολουθούν την κανονική κατανομή, ο αριθμητικός μέσος, η διάμεσος και η επικρατούσα τιμή είναι πολύ κοντά, πράγμα που δηλώνει ότι οι μεταβλητές αυτές ίσως περιγράφονται από μια συμμετρική κατανομή και σε σχέση με αυτά τα τρία μέτρα η τυπική τους απόκλιση είναι πολύ μικρή.

7.2 Ανάλυση Παλινδρόμησης

Με την Ανάλυση Παλινδρόμησης πήγαμε ένα βήμα παρακάτω, προσπαθώντας να διαπιστώσουμε αν και με ποιο τρόπο οι μεταβλητές που περιγράφουν τα διάφορα χαρακτηριστικά της διαδικασίας της διύλισης επηρεάζουν τις μεταβλητές που περιγράφουν

τα χαρακτηριστικά του τελικού προϊόντος. Η ανάλυση έγινε για κάθε μία ξεχωριστά μιας και σε αυτό το σημείο θέλαμε απλά να ελέγξουμε αν μπορούμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης γι' αυτές τις μεταβλητές.

Ακολουθώντας την ίδια διαδικασία για κάθε μία από τις $Y_1 - Y_3$, διαμορφώσαμε μέσω της μεθόδου *stepwise* ένα μοντέλο για κάθε μεταβλητή και στη συνέχεια τα εξετάσαμε ως προς τη στατιστική τους σημαντικότητα. Γι' αυτό χρησιμοποιήσαμε τα *Studentized residuals* και καταλήξαμε ότι και τα τρία μοντέλα που διαμορφώθηκαν είναι στατιστικά σημαντικά. Πληρούν δηλαδή τα κριτήρια της κανονικότητας και της ομοσκεδαστικότητας. Η ανεξαρτησία των μεταβλητών εξασφαλίστηκε από την αρχή βάσει του τρόπου συλλογής των δεδομένων καθώς και βάσει της φύσης τους. Επίσης, κατασκευάστηκαν τα αντίστοιχα ιστογράμματα συχνοτήτων και τα κανονικά Q-Q διαγράμματα για μια πρώτη εικόνα των υπολοίπων κάθε μοντέλου.

Πιο αναλυτικά διαπιστώσαμε τα κάτωθι.

- Η μεταβλητή Y_1 – αριθμός οκτανίων του τελικού προϊόντος επηρεάζεται από τις μεταβλητές

  X_7 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε m^3/h,

  X_8 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε HVTO 1075 (βαριά νάφθα) m^3/h,

  X_{11} : Η τροφοδοσία σε m^3/h της διαδικασίας του ισομερισμού.

  Σύμφωνα με το μοντέλο που διαμορφώθηκε, μόνο η X_7 συμβάλει θετικά στη διαμόρφωση του αριθμού των οκτανίων της παραγόμενης βενζίνης, ενώ οι άλλες δύο συμβάλουν αρνητικά. Επίσης, το μοντέλο καταφέρνει να εξηγήσει το 64.3% της συνολικής μεταβλητότητας που υπάρχει στα δεδομένα, το οποίο είναι ένα ικανοποιητικό ποσοστό αν αναλογιστούμε την διαφορετικότητα της φύσης των ανεξάρτητων μεταβλητών του μοντέλου.

- Η μεταβλητή Y_2 - τάση ατμών κατά Reid του τελικού προϊόντος επηρεάζεται από τις μεταβλητές

  X_6 : Κορυφή της αποστακτικής στήλης σε m^3/h,

  X_9 : Τελικό σημείο ζέσης (Τ.Σ.Ζ.) τροφοδοσίας αναμόρφωσης σε °C,

  X_{13} : Θερμοκρασία φούρνων (R-400) στους αντιδραστήρες αναμόρφωσης, είσοδος αναμόρφωσης.

  Σύμφωνα με το μοντέλο που διαμορφώθηκε, μόνο η X_6 συμβάλει θετικά στη διαμόρφωση της τάσης των ατμών κατά Reid της παραγόμενης βενζίνης, ενώ οι άλλες

δύο συμβάλουν αρνητικά. Το μοντέλο καταφέρνει να εξηγήσει μόλις το 36.2% της συνολικής μεταβλητότητας που υπάρχει στα δεδομένα. Το ποσοστό αυτό δεν είναι ικανοποιητικό και επομένως θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην χρήση του μοντέλου αυτού, παρόλο που πληροί τις υποθέσεις που αναφέραμε.

- Η μεταβλητή Y_3 – επί τοις εκατό ποσοστό βενζολίου στο τελικό προϊόν επηρεάζεται από τις μεταβλητές

  X_6 : Κορυφή της αποστακτικής στήλης σε m^3/h,

  X_7 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε m^3/h,

  X_8 : Τροφοδοσία διαδικασίας Αναμόρφωσης σε HVTO 1075 (βαριά νάφθα) m^3/h,

  X_{12} : Αριθμός οκτανίων (RON) του προϊόντος του ισομερισμού,

  X_{13} : Θερμοκρασία φούρνων (R-400) στους αντιδραστήρες αναμόρφωσης, είσοδος αναμόρφωσης.

  Σύμφωνα με το μοντέλο που διαμορφώθηκε, μόνο οι X_7, X_{12} και X_{13} συμβάλουν θετικά στη διαμόρφωση της επί τοις εκατό περιεκτικότητας σε βενζόλιο της παραγόμενης βενζίνης, ενώ οι άλλες δύο συμβάλουν αρνητικά. Επίσης, το μοντέλο καταφέρνει να εξηγήσει το 51.6% της συνολικής μεταβλητότητας που υπάρχει στα δεδομένα. Το ποσοστό αυτό μπορεί να θεωρηθεί ικανοποιητικό δεδομένης της διαφορετικότητας της φύσης των ανεξάρτητων μεταβλητών του μοντέλου.

7.3 Πολυμεταβλητή Ανάλυση

Έχοντας εντοπίσει τα χαρακτηριστικά της διαδικασίας της διύλισης που επηρεάζουν τα χαρακτηριστικά του τελικού προϊόντος ξεχωριστά, θέλαμε να δούμε αν οι μεταβλητές $Y_1 - Y_3$ σχετίζονται μεταξύ τους και επομένως αν μπορούν να χρησιμοποιηθούν μέθοδοι Πολυμεταβλητής Ανάλυσης για την ανάλυσή τους.

Είδαμε ότι οι τρεις αυτές μεταβλητές παρουσιάζουν σημαντική συσχέτιση μεταξύ τους, οπότε έχει νόημα κι ενδιαφέρον να μελετηθούν ταυτόχρονα. Αρχικά εφαρμόσαμε Ανάλυση Κυρίων Συνιστωσών για να διαπιστώσουμε αν τα δεδομένα μας μπορούν να εκφραστούν μέσω ενός άξονα και επομένως να μειωθεί η διάστασή τους. Κατασκευάστηκε έτσι ένας κύριος άξονας ο οποίος ερμηνεύει το 76.5% της συνολικής μεταβλητότητας των αρχικών δεδομένων και γι' αυτόν μπορούμε να πούμε ότι σχετίζεται σχεδόν το ίδιο και με τις τρεις μεταβλητές Y_1, Y_2 και Y_3. Αν δούμε τη φύση των μεταβλητών αυτών μπορούμε να πούμε ότι ίσως η πρώτη κύρια συνιστώσα σχετίζεται με τεχνικά χαρακτηριστικά του τελικού προϊόντος

όπως αυτά καθορίζονται κατά τη διαδικασία της αναμόρφωσης, του εμπλουτισμού και γενικά μέσα στη διαδικασία της διύλισης.

Στη συνέχεια, χρησιμοποιήσαμε εργαλεία από την Ανάλυση κατά Συστάδες για να δημιουργήσουμε ομάδες οι οποίες θα περιέχουν όμοιες παρατηρήσεις. Έγινε συνδυασμός κάποιων ιεραρχικών μεθόδων, όπως τη μέθοδο της συνένωσης μεταξύ των ομάδων (between-groups linkage), του πλησιέστερου και μακρινότερου γείτονα καθώς και τη μέθοδο του Ward για να καθοριστεί το πλήθος των ομάδων που θα δημιουργηθούν. Το αποτέλεσμα που πήραμε ως προς το πλήθος των ομάδων που πρέπει να δημιουργήσουμε, το χρησιμοποιήσαμε για την εφαρμογή μιας μη-ιεραρχικής μεθόδου, για παράδειγμα της μεθόδου των k-μέσων (k-means method).

Το αποτέλεσμα της τελευταίας μεθόδου ήταν να δημιουργηθούν δύο ομάδες με αριθμό παρατηρήσεων 53 και 76 αντίστοιχα. Επίσης, τα κέντρα των ομάδων δεν μεταβλήθηκαν σημαντικά μετά την τελική ομαδοποίηση, πράγμα που πιθανώς σημαίνει ότι τα δεδομένα από μόνα τους λόγω της φύσης τους, είχαν την τάση να ταξινομούνται στις δύο ομάδες με παρόμοιο τρόπο. Αυτό μας οδήγησε στο να συμπεράνουμε ότι οι παρατηρήσεις που ταξινομούνται στις δύο ομάδες, είναι πιθανό να επηρεάζονται σε κάποιο βαθμό από τη χρονική στιγμή που λαμβάνονται (πρώτο ή δεύτερο τρίμηνο του έτους).

7.4 Στατιστικός Έλεγχος Ποιότητας

Έχοντας δημιουργήσει έναν κύριο άξονα που εκφράζει τα ποιοτικά χαρακτηριστικά της βενζίνης, θέλαμε να δούμε αν η διαδικασία με την οποία αυτή παράγεται είναι εντός ελέγχου ανιχνεύοντας έγκαιρα την εμφάνιση ειδικών αιτιών μεταβλητότητας, έτσι ώστε να προχωρήσουμε σε έρευνα και να προβούμε στις απαραίτητες διορθωτικές ενέργειες προτού κατασκευαστούν αρκετά προϊόντα μη συμμορφούμενα με τις προδιαγραφές. Αυτό έγινε με την κατασκευή κατάλληλων διαγραμμάτων ελέγχου όπως το διάγραμμα ελέγχου τύπου *Shewhart*, το διάγραμμα *EWMA* και το διάγραμμα *CUSUM* για μεμονωμένες παρατηρήσεις.

Αρχικά ελέγχθηκε η πρώτη κύρια συνιστώσα ως προς την καταλληλότητά της, προκειμένου να μπορούν να εφαρμοστούν τα παραπάνω εργαλεία Ελέγχου Ποιότητας. Με τη βοήθεια του κανονικού P-P διαγράμματος καθώς και του αντίστοιχου ελέγχου *Kolmogorov – Smirnov* διαπιστώθηκε ότι τα δεδομένα μας ήταν κανονικά. Αρχικά κατασκευάστηκε ένα διάγραμμα ελέγχου τύπου *Shewhart* με 3σ όρια ελέγχου το οποίο κατέδειξε ότι η διεργασία ήταν εκτός ελέγχου. Φέρνοντάς την εντός ελέγχου, υπολογίσαμε και την εντός ελέγχου μέση

τιμή και τυπική απόκλιση, οι οποίες χρησιμοποιήθηκαν με τη σειρά τους για τον υπολογισμό κάποιων δεικτών ικανότητας της διεργασίας. Οι δείκτες αυτοί έδειξαν ότι η διεργασία δεν είναι ικανή και χρειάζεται παρακολούθηση.

Στη συνέχεια, κατασκευάστηκαν κάποια διαγράμματα ελέγχου με μνήμη με σκοπό να ανιχνευθούν μικρές μετατοπίσεις στο μέσο της διεργασίας και έγιναν συγκρίσεις με βάση το εκτός ελέγχου μέσο μήκος ροής, ARL_1. Από τις συγκρίσεις αυτές καταλήξαμε στο ότι ανάλογα με το μέγεθος της μετατόπισης που θέλαμε να ανιχνεύσουμε, η επιλογή μας ήταν ανάμεσα στα διαγράμματα με μνήμη, ενώ αξιοσημείωτο είναι ότι το διάγραμμα ελέγχου τύπου *Shewhart* που κατασκευάσαμε δεν ήταν ικανό να ανιχνεύσει ούτε την μεγάλη μετατόπιση του μέσου νωρίτερα από τα άλλα είδη διαγραμμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Wickipedia (el.wickipedia.org/wiki/Πετρέλαιο).

Εφραιμίδης Νικόλαος, Λαφτσής Ιγνάτιος και Τζίκας Χρήστος (2007). Μελέτη της διεργασίας παραγωγής βενζίνης με χρήση στατιστικών τεχνικών διασφάλισης ποιότητας.

Πολίτης Κωνσταντίνος (2012-2013). Σημειώσεις περιγραφικής στατιστικής, Πειραιάς.

Κούτρας Μάρκος και Ευαγγελάρας Χαράλαμπος (2010). Ανάλυση παλινδρόμησης, θεωρία και εφαρμογές, Εκδόσεις Σταμούλη Α.Ε. , Αθήνα.

Κούτρας Μάρκος (2012). Εφαρμοσμένη πολυμεταβλητή ανάλυση, Πειραιάς.

Αντζουλάκος Δημήτριος (2010). Στατιστικός έλεγχος ποιότητας, Πειραιάς.

Ταγαράς Γιώργος (2001). Στατιστικός έλεγχος ποιότητας, Εκδόσεις Ζήτη, Θεσσαλονίκη.

Κούτρας Μάρκος (2008). Προηγμένα εργαλεία και μέθοδοι για τον έλεγχο της ποιότητας, Ελληνικό Ανοικτό Πανεπιστήμιο, Σχολή Θετικών Επιστημών και Τεχνολογίας, Πάτρα.